NEXT GENERATION SEQUENCING TO IDENTIFY MULTIPLE CLINICALLY RELEVANT GENETIC LESIONS FOR THE DIAGNOSIS OF ACUTE MYELOID LEUKAEMIA

NICHOLAS TELFORD

A thesis submitted in partial fulfilment of the requirements of
Manchester Metropolitan University for the degree of
Professional Doctorate

School of Healthcare Science
Manchester Metropolitan University
in collaboration with
The Christie NHS Foundation Trust

2018

Abstract

Routine cytogenetic and molecular genetic investigations aid the diagnosis of acute myeloid leukaemia (AML) and are critical for prognostic stratification to optimise therapy and enhance survival of patients. Advances in the understanding of the genomics of AML by next generation sequencing (NGS) technology have identified a mutational landscape, which has the potential to improve risk assessment and identify new targets for therapy. This project developed a novel, custom-designed NGS panel for the resequencing of genomic DNA (gDNA), to identify multiple types of genetic lesion in AML. Regions of recurrent genetic abnormalities were targeted, to reproduce the output of conventional testing in a single assay. Solution hybrid capture and Illumina-based NGS were used to analyse 36 AML samples, without use of normal control to represent the typical diagnostic workflow. A panel of 42 genes, including those most frequently mutated, was used to test for clinically relevant abnormalities, including common duplications and gene fusions. Sequence data was analysed with a pipeline of relevant bioinformatic tools and the output was compared to standard results and sequencing from an alternative NGS platform.

Following variant annotation, a total of 143 likely oncogenic variants were detected across all samples. This included all 13 *NPM1* insertions, 10 *FLT3*-ITD, and the 7 fusion genes found by routine tests. There was strong concordance between NGS platforms for mutation detection. Multiple new findings included two *KMT2A*-PTD, a *TP53* mutation in a patient with a complex karyotype, and a rare *NUP98-DDX10* gene fusion. Patients were regrouped by a new prognostic scheme based on genomic features. Eight patients were reclassified; seven changed from the Intermediate group, three to Favourable and four to Adverse. The successful detection of genomic lesions demonstrated the principle that the new NGS assay could reliably detect a variety of genomic abnormalities and that it could be refined for use in the diagnostic laboratory, with the potential to rationalise multidisciplinary workflows. The feasibility of implementation is discussed. A potential clinical utility was inferred and suggests that benefit could be derived for its validation for mainstream diagnosis for the clinical management of AML.

Acknowledgements

I would like to thank my supervisors, Dr Stefan Meyer and Dr Chris Murgatroyd, my Director of Studies, Dr Ciaren Graham and former Director of Studies, Professor Bill Gilmore. My thanks also go to Dr Carol Taylor, Programme Leader for the Professional Doctorate and to all who gave their time to teaching on the modular components of the course.

I am indebted to The Christie NHS Foundation Trust for allowing me to study for the Professional Doctorate and for their financial support. I am extremely grateful to the Alan Gowenlock bursary, which partly supported tuition fees.

I would like to thank all staff of the Oncology Cytogenetics for their hard work and dedication providing an excellent service and for performing the routine investigations that were used for the project. In particular, I would like to thank Pragnya Patel who coordinated the collection of surplus samples and helped with DNA extraction.

Technical advice in the new technology was gratefully received from many sources and the following deserve particular mention; Lee Silcock and Ivan Lesende (Agilent), Freddie Sharkey and Kicki Bergefall (Thermo Fisher Scientific), Matthew Bashton and Sarra Ryan (Newcastle University). In particular, Matthew provided bioinformatics advice and support by allowing me use of Newcastle University Pindel and Lumpy analytical software. Erica Ashton, kindly accessed NHS Summary Care Record for missing overall survival data.

The main next generation sequencing was performed in the Core Facility of the CRUK Manchester Institute. The Ion Torrent sequencing was performed in the Thermo Fisher Scientific laboratory in Darmstadt, Germany. All sample preparation and post sequencing data processing and analysis was performed by me.

And last but by no means least, to Karen, Cerys, Elise and George. The project would not have been possible without their love and support.

Contents

1.0 Introduction

1.1 Genomic Science and the Human Genome Project

Modern biomedical and clinical practice is in the midst of a revolution in genomic science which has the potential to transform healthcare. Personal genomics is set to transform health decision making for individuals by customising disease assessment and by designing specific therapeutic plans accordingly. Treatment can be optimised by providing personalised risk information about cancer and other diseases such as diabetes, heart disease and obesity. Pharmacogenomics will inform how individuals metabolise conventional drugs and a better understanding of the molecular basis of their disease will provide specific targets for therapy. A different landscape for healthcare is envisaged, in which treatment will be tailored to a patient's particular constitutional and disease profile and they would actively participate in directing their care. This is set to challenge our understanding about what constitutes health and well-being and represents a paradigm shift in the way disease is defined and managed for the 21st Century.

We have come a very long way with our knowledge of genomics since DNA was first isolated in 1869 and Watson and Crick described the structure of the DNA double helix in 1953 (Watson & Crick, 1953). Since the Human Genome Project (HGP) delivered the first draft sequences of the human genome, a period of intense genomic research activity and discovery ensued; the start of a "Genomic Era" (Guttmacher & Collins, 2003). HGP was hailed as one of the greatest, collaborative scientific achievements ever; Biology's first 'Big Science' project (Hood & Rowen, 2013; Green *et al.*, 2015; Vermeulen, 2016). Amidst the highest profile publicity, it was promised that the new understanding of genetics would reveal the causes of common diseases, including cancer, and transform therapeutic medicine, enabling cures previously impossible (Collins, 2010b; Hood & Rowen, 2013). There is no doubt that the biosciences have been "profoundly and irreversibly affected by access to the complete DNA sequence of the human genome" (Collins, 2010b; Green *et al.*, 2015). An immense amount of information has been accrued by elucidating the structure of the human genome and the function encoded in it. Furthermore, HGP led to the development of an infrastructure for a new technoscience and facilitated a new way of performing scientific investigations, which represented a departure from traditional scientific method (Barnes & Dupre, 2008; Hood & Rowen, 2013; Green *et al.*, 2015). One of the key achievements of HGP has been the generation of large, publicly available, comprehensive sets of reference data providing the scientific resources that comprise a toolkit for genomics-based research, the most obvious examples being genomic maps and databases of DNA sequence and variation (Butler, 2010; Hood & Rowen, 2013). This has necessitated a parallel drive to upgrade computational functionality and Information Technology infrastructure to provide the analytical resource required for genomic study.

The rhetoric that accompanied the release of the draft human genome sequence was accompanied by scepticism about the scientific validity of HGP, which was reinforced in subsequent years by the apparent anti-climax that the promised health benefits did not move into general public awareness or contribute to health and social change. The revolution in personalised medicine that has been anticipated for almost two decades has been slow to deliver. Unexpected findings and the extreme complexity of the genome has confounded the understanding of the data, even challenging the concept of the gene itself (Barnes & Dupre, 2008; Butler, 2010; Griffiths & Stotz, 2013; Arney, 2016). This has resulted in a re-evaluation of the original estimates of the timelines for the delivery and general adoption of personalised medicine (Collins, 2010a; Hamburg & Collins, 2010). The investigation of common inherited conditions requires further intensive study and it may take many more years to transform medical practice in the way it was first conceived (Manolio *et al.*, 2009; Collins, 2010a). It has been argued that there have been few intellectual or economic benefits, relative to the huge outlay of HGP (US$ 3 billion over 13 years) (Vermeulen, 2009; Wade, 2010). However, the limited success needs to be weighed against the technological, financial, ethical and cultural challenges encountered and the scientific advances that are emerging and still have the potential to be revealed. Perhaps HGP promised too much, too soon. It is now recognised that HGP was only the starting point on the path to genomic medicine (Butler, 2010; Eric Green director of NHGRI quoted in Wadhwa, 2014). Very gradually, new genetic techniques and modernised information technology are being used to interpret the information encoded in the genome, and are being translated into new diagnostic procedures, which will lead to the optimisation of treatment and improvements in clinical practice.

1.1.1 Genomics and Cancer

Cancer is understood to be a genomic disease, involving multiple genes and genomic regions and their interrelationships, with a combined influence on the growth and development of a tumour, rather than single genes having an isolated effect (World Health Organization, 2017). New treatments will benefit from increasing knowledge of the molecular mechanisms of the control of cell growth in healthy individuals and in disease processes, placing oncological clinical practice at the vanguard of personalisation of medicine. For many years, some cancers have been categorised not solely by traditional methods of microscopic examination of tissue morphology and techniques to determine cell type, but by their genetic characteristics (Roug *et al.*, 2014; Arber *et al.*, 2016). This has intensified with the accumulation of evidence that certain genetic defects can define a patient's specific cancer type and therefore provide the most accurate diagnosis (Harris & McCormick, 2010). This has been particularly relevant for the haematological malignancies; for example, the chromosomal translocation t(15;17) in acute promyelocytic leukaemia (APML) (Zelent *et al.*, 2001) is characteristic of the disease entity and whilst not specific, t(9;22) in chronic myeloid leukaemia and the V600E mutation in hairy cell leukaemia are found almost exclusively as the disease-driving genetic changes associated with a morphological subtype (Arber *et al.*, 2016; Swerdlow *et al.*, 2016). Conversely, the emergence of genetic driver mutations, as some of the most powerful diagnostic characteristics, is leading to disease classifications being reframed, making genetic abnormalities, rather than morphological appearance, the main diagnostic determinants of a subtype (Arber *et al.*, 2016; Swerdlow *et al.*, 2016).

Importantly, genetic prognostic biomarkers offer information about the likelihood of survival when patients are given the standard of care for their disease, influencing the choice of conventional treatment if options exist. Increasingly, genetic lesions identified in tumour tissue have predictive value for response to a specific type of therapy. Improved cancer outcomes will result not just from better treatment stratification but also from recognition of clinically actionable targets and implementation of highly specific therapies that are tailored to an individual's cancer mutational profile. This 'precision medicine' could improve cancer outcomes by neutralising the core molecular defects that cause the disease, offering the opportunity of more effective therapy by specifically retarding the growth of cancer cells from disruption of their metabolism and inducing cell death. This has the added benefit of reducing the undesirable side-effects of damage to a patient's normal cells.

Genome analysis is now starting to guide more treatment decisions in cancer and the revolution in disease diagnosis and therapy is gradually revealing itself to improve patients' treatment outcomes (Tian *et al.*, 2012). Development pipelines for drug and biomarker discovery are becoming aligned, offering an increasing armoury of drug candidates and the opportunity to salvage previously failed products by presenting new targets to examine (Miles *et al.*, 2015). We are moving from the one-size-fits-all approach of cytotoxic chemotherapy to strategies based upon molecularly targeted drugs that exploit the particular genetic addictions, dependencies and vulnerabilities of cancer cells (Hoelder *et al.*, 2012; Hollingsworth & Biankin, 2015). Despite remarkable progress in the identification and characterisation of novel oncogenic mechanisms, conventional randomised clinical trials take time to demonstrate the efficacy of a particular approach and the conversion of genomic information to clinical validated therapies is slow (Rubin & Gilliland, 2012). Furthermore, individual molecular markers can be uncommon which would be unworkable for accrual of patients and the cost of sequencing multiple patients to find those eligible would be unacceptable (Hollingsworth, 2015). Cancer genomics is therefore driving innovation in study design whereby potential participants are sequenced in 'umbrella' or 'basket' trials in which multiple treatment arms are possible, depending on the genetic composition of a tumour (Hollingsworth & Biankin, 2015). The uniqueness of each patient's tumour and intense customisation of therapy essentially results in a single person (*N*-of-1) approach but with standardisation of design the data can be cumulative across different trials (Schork, 2015).

Tumour clonal heterogeneity and the evolution of resistance from selective pressure to drive clonal development along new pathways with secondary mutations are recurrent themes, confounding targeted drug development and thereby the precision medicine philosophy (Fisher *et al.*, 2013). Nevertheless, a large and increasing number of targeted or biological agents have been approved by the US Food and Drug Administration (FDA), many for the same and most common tumour types (National Cancer Institute, 2014). However, less than 20 specific molecular genetic targets that also require a companion diagnostic test have been demonstrated to be reliable biomarkers for selecting patients (see Table 1.1, adapted from Dietel *et al.*, 2015; Mertens *et al.*, 2015; Abramson, 2016).

Table 1.1. Targeted drugs in clinical use requiring companion diagnostic genetic tests (Compiled from Dietel *et al.*, 2015; Mertens *et al.*, 2015; Abramson, 2016; McDermott, 2017)

Tumour type	Affected gene(s)	Type of alteration	Method for detection	Related treatment
Acute lymphoblastic leukaemia	*BCR-ABL1*	Gene fusion	Cytogenetics, FISH, PCR	Imatinib, Dasatinib, Bosutinib, Ponatinib
Breast cancer	*HER2*	Amplification	IHC, FISH	Trastuzumab, Ado-trastuzumab, Lapatinib, Pertuzumab
	PIKCA	SNV	Sequencing	Reduced response to anti-HER2 treatment
Chronic lymphocytic leukaemia	*CD52*	n/a	*TP53* deletion (FISH)	Alemtuzumab, Venetoclax
Chronic myeloid leukaemia	*BCR-ABL1*	Gene fusion	Cytogenetics, FISH,	Imatinib, Dasatinib, Bosutinib, Ponatinib, Nilotinib
Colorectal cancer	*RAS (KRAS, NRAS)*	SNV	Sequencing	Cetuximab, Panitumumab
Dermatofibrosarcoma protuberans	*COL1A1–PDGFRB*	Gene fusion	FISH, PCR	Imatinib
Gastric adenocarcinoma	*HER2*	Amplification	IHC/FISH	Trastuzumab
Gastro-Intestinal Stromal Tumour (GIST)	*KIT*	SNV, indel	Sequencing	Imatinib, Sunitinib
	PDGFRA	SNV	Sequencing	Imatinib, Sunitinib
Hairy cell leukaemia	*BRAF*	V600E mutation	Sequencing	Vemurafenib
Chronic eosinophilic leukaemia	*FIP1L1–PDGFRA*	Gene fusion	FISH, PCR	Imatinib
Lymphoma	*MYC*	Gene fusion, amplification	IHC, FISH	BET & Protein translation inhibitors
	BCL2	Gene fusion, amplification	IHC, FISH	BH3 mimetics
Malignant melanoma	*BRAF*	SNV	Sequencing	Vemurafenib, Dabrafenib, Trametinib
	KIT	SNV, indel	Sequencing	Sunitinib, Dasatinib, Imatinib
	MEK			Cobimetinib, Trametinib
MDS/MPN	*PDGFR* fusions			Imatinib
NSCLC	*EGFR*	SNV, MNV, indel	Sequencing	Gefitinib, Erlotinib, Afatinib, Dacomitinib
	ALK	Gene fusion	FISH, IHC, Sequencing	Crizotinib, Ceritinib, Alectinib
	ROS1	Gene fusion	FISH, Sequencing	Crizotinib
	MET	Amplification	FISH, Sequencing	Resistance to EGFR TKIs
Ovarian cancer	*BRCA1/BRCA2*	SNV, MNV, indel	Sequencing	Olaparib

1.1.2 Technological advances in genomic testing

The sequencing of the human genome was not only the start of an era of great understanding of genomic science but it also stimulated a significant development in technology. Completion of the first human whole genome sequences were performed by Sanger DNA sequencing, which had dominated the science for almost two decades and resulted in many key advances (Sanger *et al.*, 1977). During the HGP, this procedure was scaled up and made more efficient by evolutionary technical advances, such as automation and miniaturisation and using new methods, such as capillary-based sequencing and single nucleotide polymorphism (SNP) genotyping (Metzker, 2010). Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies to increase the capacity for genomic analyses; to sequence larger regions and greater numbers of human genomes, in reasonable timescales, at lower cost.

A significant breakthrough in genomic investigation has been 'Next Generation Sequencing' (NGS) of nucleic acids (also known as *second generation*, *massively parallel* or *clonal sequencing*), referring to a technology developed concurrently by several commercial companies (Metzker, 2010; Goodwin *et al.*, 2016). Similar to traditional sequencing methods, NGS uses DNA polymerase to catalyse the incorporation of deoxyribonucleotide triphosphates (dNTPs) using a DNA template strand during sequential cycles of DNA synthesis. Although the precise methodology varies, NGS instruments share the common principle of miniaturisation of sequencing reactions so that clusters of millions of DNA fragments, fixed on a stable substrate, can be scanned and sequenced at the same time in the flow cell of a sequencing machine (see Figure 1.2) (Bentley *et al.*, 2008). NGS requires the preparation of amplified libraries of template DNA prior to the sequencing of these DNA clones; however, the need for laborious fragment-cloning methods that were used with Sanger sequencing is obviated. Thus, the capability of NGS to produce multiple parallel reads facilitates high-throughput sequencing that can reproduce the output of capillary electrophoresis sequencing many times over, accurately determining the sequence of short lengths of DNA in much faster timescales. NGS also offers the opportunity for achieving ‘deep’ DNA sequencing, a large number of sequence reads covering each base or region of a genome, and sequencing of multiple samples at much reduced cost (Metzker, 2010).

The increase in scale not only lends itself to more efficient sequencing, but also to new applications not possible by the Sanger method. This offers enormous opportunities and will be the main transformative benefit from NGS. The sequencing of multiple genomic targets in a single assay enables characterisation of large regions or whole genomes in practical timescales, and also the detection of large scale structural variation. The magnitude of sequencing reactions permits evaluation of the relative frequency of reads to detect copy number variation and improves depth of sequencing, enabling the detection of low frequency gene aberrations such as minor cancer clones. This facilitates novel applications, such as the detection of low level circulating tumour DNA in ‘liquid biopsies’ for cancer diagnosis where a solid tumour cannot be identified or biopsied (Siravegna *et al.*, 2017; Wan *et al.*, 2017). Variation of NGS techniques can be used for wide scale RNA sequencing (e.g. whole transcriptome shotgun sequencing or RNA-seq) (Wang *et al.*, 2009; Ozsolak & Milos, 2011) to characterise coding and noncoding transcriptional activity in a specimen or to target a specific subset of transcripts. Furthermore, analysis of target DNA by methylation-sensitive sequencing can study the epigenetic regulation of gene expression and its emerging significance in neoplastic disease (Hirst, 2013).

Therefore, NGS is not simply the next step in the evolution to more efficiently derive short read sequence information, but has the potential to revolutionise how all genetic information is derived. NGS will transform the provision of laboratory diagnostic services and make possible novel applications in healthcare. NGS can improve on existing laboratory techniques by replacing single gene testing where, increasingly, panels of multiple genetic targets are necessary for comprehensive diagnosis of disorders. Additional information can be simply obtained, facilitating the use of genomic profiling rather than single mutations to more accurately diagnose conditions and stratify treatment. It can be envisaged how cytogenetic investigations could be improved by increasing the resolution of testing to the ultimate DNA sequence level and making the detection of low level variants feasible, improving on the accuracy and sensitivity of traditional analysis. It is possible to identify different types of genomic variation by a single methodology. In time, multiple technologies will be replaced, with laboratory diagnostics converging for most applications in a single technology, challenging how laboratory services are organised and delivered, thereby reducing the requirement for multidisciplinary workflows.

NGS is in a developmental phase whereby it is being refined into a robust and reproducible technology for clinical applications. Assays are being validated for use in the accredited diagnostic laboratory environment. NGS has been used extensively in research, generating vast quantities of genomic information in health and disease. During this time, the technology has been under continual development and commercial companies are developing instruments and analytical pipelines that are applicable to routine laboratory practice; such as Illumina® (e.g. NextSeq™ and MiniSeq™), Thermo Fisher Scientific (e.g. Ion Torrent PGM™) and QIAGEN (e.g. GeneReader®) (Goodwin *et al.*, 2016). There is a definite change of focus amongst equipment suppliers towards NGS being applied to routine clinical practice to improve genetic diagnosis (Metzker, 2010; ten Bosch & Grody, 2008; Tucker *et al.*, 2009; Voelkerding *et al.*, 2009; Pareek *et al.*, 2011; Kohlmann *et al.*, 2012) and competition amongst suppliers is driving down costs (Mardis, 2011). Cost effective bench-top analysers are commercially available that are practical for use in routine laboratory diagnosis (Loman *et al.*, 2012), meaning that NGS technology is indeed egalitarian, by allowing both small and large medical laboratories to use the technology in diagnostic investigations, if they can demonstrate the cost efficiency benefits to justify the change.

Generating the sequence data is no longer a limitation. The cost of routine sequencing is still challenging for regular applications, but genomics has recently celebrated achieving the much-coveted “$1000 genome” milestone (Vertitas Genetics, 2015), illustrating that DNA sequencing is becoming more affordable (see Figure 1.1). Furthermore, it is possible to be selective about the amount of sequencing performed on each sample, by using a variety of targeted or candidate gene sequencing systems as opposed to sequencing whole genomes (Mamanova *et al.*, 2010; Samorodnitsky *et al.*, 2015). Whole genome sequencing (WGS) provides more sequence information, including from non-coding regions (Lam *et al.*, 2011). WGS is not subject to biases intrinsic to the target selection process, which leads to better uniformity of read coverage (Meynert *et al.*, 2014). However, target enrichment prior to DNA sequencing makes optimal use of the capacity of clinical grade sequencers. This can dramatically reduce costs and the turnaround time for diagnostic investigations, making routine sequencing realistic. Target selection also permits greater depth of sequencing of selected regions, which is necessary for the detection of low level allele frequencies in cancer samples. In particular, whole exome sequencing (WES) strategies are available to provide a global screen of the protein-coding DNA. By targeting the regions of the genome, which contain the majority of the known clinically significant mutations, WES provides a cost-effective alternative to WGS (Rabbani *et al.*, 2014).

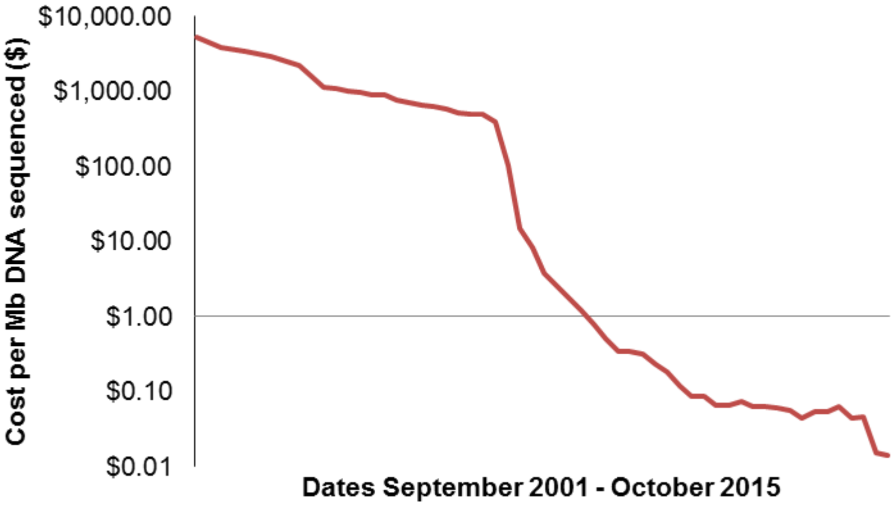


Figure 1.1. The decreasing cost ($ logarithmic scale) of DNA sequencing from September 2001 – October 2015 (data from National Human Genome Research Institute (National Human Genome Research Institute, 2016)

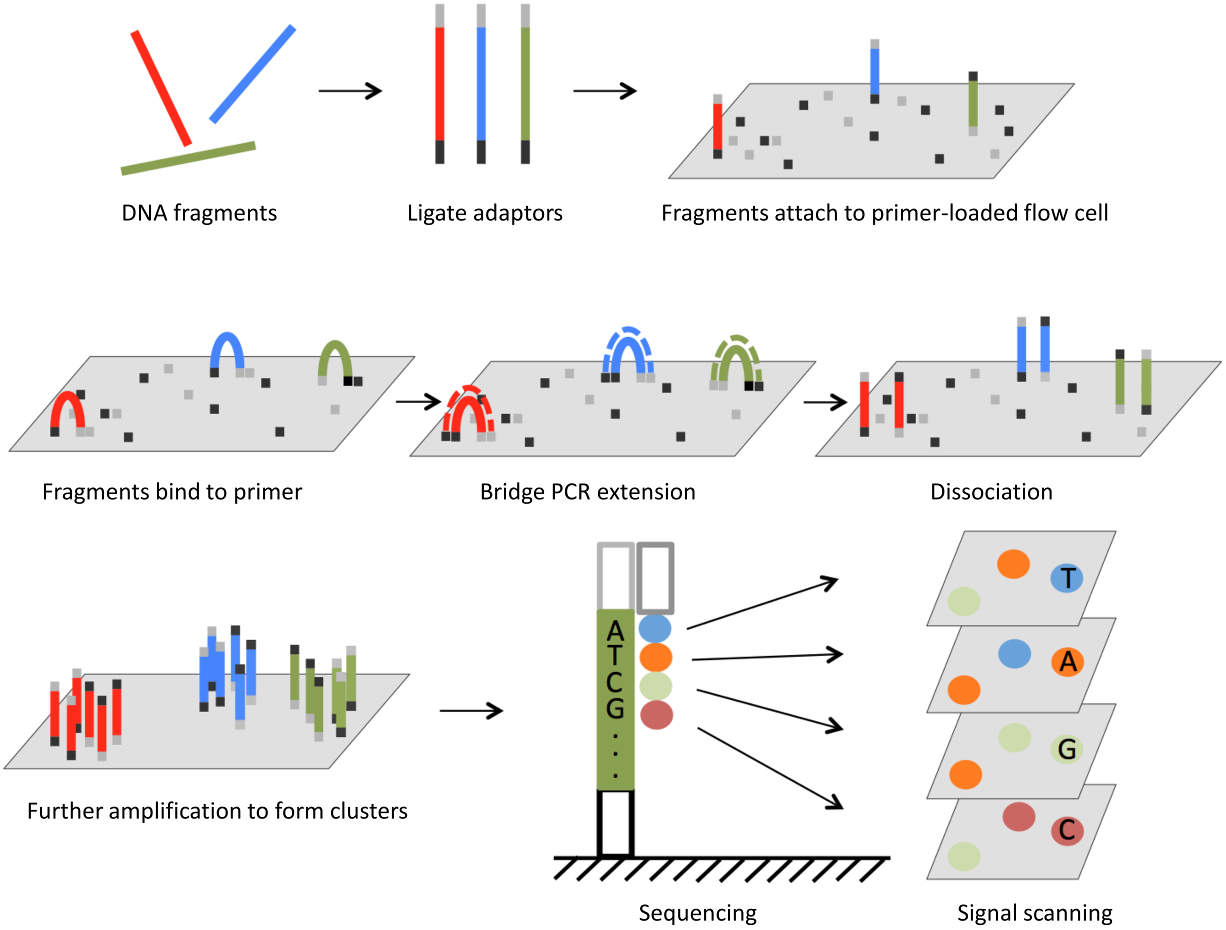


Figure 1.2. Outline of Illumina sequencing process. Adaptors are ligated to the ends of DNA fragments. Fragments bind to primer-loaded flow cell and bridge PCR reactions amplify each bound fragment to produce clusters of fragments. During each sequencing cycle, one nucleotide with an attached base-specific fluorophore is added to the elongating complementary DNA strand. The fluorphores in all the fragments that are being sequenced in the flow cell are simultaneously excited by a laser and an optical scanner records the signals from each fragment cluster. The sequencing terminator is then removed and another sequencing cycle begins. Multiple rounds of additional nucleotides are incorporated until the programmed cycles of sequencing is complete (adapted from Lu *et al.*, 2016).

The simultaneous development of the bioinformatics tools for interpretation of genomic data has been necessary to permit the confident detection of mutations for clinical grade assays (Delon & Scott, 2016). This includes the development of robust algorithms for the accurate alignment, mapping and filtering sequencing data and the calling and annotation of variants. Bioinformatics for genomics has needed to advance rapidly, to parallel the dramatic increase in NGS activity. For research purposes, bioinformatics has become highly sophisticated and capable of producing high-quality, reliable data. Research institutes have made custom pipelines which are made publicly available. Commercial manufacturers of sequencing equipment and reagents provide access to pre-made bioinformatics tools for the analysis of sequence data. These computational scripts can be tailored to specific analytical applications but require bioinformatics expertise to operate. Significant further development is required to make these tools convenient and practical to use for genomic scientists and universally applicable for use in routine diagnostics.

The interpretation of the clinical significance of sequence data is also a developing area. The systematic collection and publication of genomic variation in the normal population and in specific disease cohorts is necessary to continue to investigate and understand its occurrence, and its functional and clinical impact. The further clarification of uncertain and incidental sequence findings, to reduce ambiguity in clinical interpretation, is also necessary. The current challenge, therefore, is not lack of availability of a technology primed to improve genetic diagnosis or its affordability. The principle of precision medicine is sound but supporting practices need to be developed further, to fulfil the unmet clinical need in many diseases where conventional approaches have reached their potential. We are now truly entering a period where NGS technology is ready for the transition to diagnostic laboratories. In this environment, NGS needs to meet clinical standards for speed and accuracy and be validated in robust diagnostic protocols to rigorous accreditation standards. This development will parallel the greater understanding of disease and move to provide universal applications for precision medicine, making genomics accessible for more clinical applications and a greater number of deprived patients.

1.2 Acute Myeloid Leukaemia; the potential clinical utility of genomic investigations

Acute myeloid leukaemia (AML) is a heterogeneous group of haematopoietic neoplastic diseases, resulting from the clonal expansion of immature myeloid cells in the bone marrow and blood. There is a worldwide incidence of approximately 3 in 100,000 people, with a slight male predominance (Swerdlow *et al.*, 2008). AML occurs at any age but is generally considered a disease of the elderly; the incidence rises steeply after 55 years with a median age of ~70 years. However, AML is also found in children with a peak incidence of 3~4 years (Heim & Mitelman, 2009). The overgrowth by malignant blood precursor cells suppresses normal haematopoiesis and leads to a reduction in numbers of functional mature cells in the peripheral blood, which is responsible for the main symptoms of the disease. Very high numbers of leukaemia cells in the blood can cause problems with circulation from leucostasis.

One or more myeloid cell lineages can be involved and, by definition, a minimum of 20% myeloid blast cells in the bone marrow are required for the diagnosis of AML, although this is no longer critical if a recurrent genetic abnormality is present (see Table 1.2 below). Morphological appearance varies depending on the predominant cell type and the degree of cellular maturation, as well as the proportion of abnormal cells and infiltration of different tissues. Therefore, traditionally, the diseases were characterised by cell morphology supplemented by cytochemical staining; a number of different morphological subgroups were defined by cell specificity and stage of maturation. This formed the basis of the original French-American-British (FAB) classification of AML which, over a number of iterations and addition of immunophenotyping, became the prominent functional classification of the disease for more than two decades (Bennett *et al.*, 1976; Bennett *et al.*, 1985). The recognition that there are genetic hallmarks of certain morphological subtypes led to the inclusion of genetic factors and resulted in a major revision of leukaemia classification in the current *World Health Organisation Classification of Tumours of Haematopoietic and Lymphoid Tissues* (WHO Classification) (Arber *et al.*, 2016; Rose *et al.*, 2017). A post-modern approach utilising a full range of recurrent genetic abnormalities as the main determinants of disease subtype, including those at single-nucleotide resolution, is predicted to be the next phase of disease classification (Roug *et al.*, 2014).

AML is now understood to be a number of different diseases, with varying underlying characteristics (Arber *et al.*, 2016). The somatic cell theory of cancer (Hanahan & Weinberg, 2011) dictates that AML must originate from a mutation in a growth regulating gene in a single progenitor cell, resulting in the clonal expansion of the cell. More genetic abnormalities accumulate, with further dysregulation of growth and cell maturation, until full-blown neoplastic disease results. Genetic alterations are therefore the molecular drivers and, to differing degrees, influence the appearance and clinical characteristics of the diseases. This underpins the current disease classification (Arber *et al.*, 2016) (see also Section 1.3 below).

1.2.1 Current standard of AML care

Conventional treatment of AML typically comprises an induction phase of high-dose chemotherapy, usually consisting of two cytotoxic drugs, cytarabine (Ara-C) and an anthracycline (daunorubicin or idarubicin), in attempt to achieve remission; the morphological absence of leukaemic cells in the bone marrow and elimination of symptoms of leukaemia. Remission status can also be assessed more specifically and sensitively by flow cytometry, molecular genetics or cytogenetics, if a suitable biomarker of the disease is available (Ommen, 2016). Induction is followed by a consolidation phase of additional chemotherapy, allogeneic stem cell transplantation (SCT), or occasionally autologous SCT, with the aim of prolonging remission and ultimately curing the disease (Tutt, 2012; Dombret & Gardin, 2016; Cornelissen & Blaise, 2016; Döhner *et al.*, 2016). Not all patients are eligible for SCT and remission needs to be achieved before transplantation can be performed. Regimens comprising these combinations of high dose chemotherapy and SCT have been shown to be the most effective treatment in AML, compared to other cytotoxic drugs in multiple clinical trials over many years.

The current AML19 clinical trial (‘Adults with Acute Myeloid Leukaemia or High-Risk Myelodysplastic Syndrome’ trial) represents the standard of care in the UK for younger adults (aged 18 to 60) and is examining different risk-adapted consolidation therapies. AML19 randomises to DA (daunorubicin and cytarabine) or FLAG-IDA (idarubicin and cytarabine, with fludarabine, a purine analog interfering with DNA synthesis and granulocyte colony stimulating factor, to support neutrophil recovery), and tests the addition of gemtuzumab ozogamacin (GO, also known as Myelotarg®, an anti-CD33 monoclonal antibody with cytotoxic ligand) (Cardiff University, 2015). However, the clinical biology of AML is highly heterogeneous and whilst the disease can be cured in some patients, response to therapy varies as a result of patient- and disease-related prognostic factors. In particular, more than 60 per cent of newly diagnosed patients are older than 60 years, a proportion which is increasing. This group is enriched for patients with a higher frequency of high risk genetics, increased multidrug-resistance and prior haematological neoplasia and to an extent is considered biologically and clinically distinct (Erba, 2015; Döhner *et al.*, 2016). Combined with the lower performance status in many elderly patients, the higher prevalence of comorbid conditions, and impaired bone marrow stem cell reserve, AML in the over 60s is typically difficult to treat. AML treatment is associated with a high treatment-related mortality and the common perception is that elderly patients cannot tolerate intensive therapy, meaning that they will not have access to the most appropriate treatment for their disease and will be incurable. As a result, only 50% of patients older than 60 years achieve complete remission after induction treatment compared to 75% of adults less than 60 years' of age (Oran & Weisdorf, 2012; Ossenkoppele & Löwenberg, 2015). In particular, patients with high risk cytogenetics have a dismal prospect and conventional treatment is largely ineffective in achieving prolonged remission.

The basic paradigm of therapy for AML has not substantially changed in the past three decades (Burnett, 2012), with the exception of anthracycline dose intensification during induction (Luskin *et al.*, 2016), modification of schedules and improved SCT techniques (Dombret & Gardin, 2016; Cornelissen & Blaise, 2016). As a result outcomes have not significantly improved. Conventional chemotherapeutic agents have been successful in treating AML but now appear to have reached their maximum potential and it is necessary to find more effective and safer treatments for AML.

1.3 Genetic, cytogenetic and genomic influences in AML

AML has been the subject of intense genomic research and more patients have had their genomic profiles studied than any other disease. As a result, the genetic landscape of AML is now very well-characterised and knowledge is probably more advanced than in any other neoplastic disease of this complexity (Figure 1.3). Through genomic research the common pathogenic abnormalities of AML have been characterised and now help to define the diseases and directly influence the outcome of treatment (Heim & Mitelman, 2009). An enormous increase of genomic information has been provided by large scale sequencing studies in recent years. Many excellent reviews have described the genetics and genomics of AML and the existing clinical utility of genetic profiling (Marcucci *et al.*, 2011; Abdel-Wahab *et al.*, 2011; Abdel-Wahab, 2012; Meyer & Levine, 2014; Naoe & Kiyoi, 2014; Ohgami & Arber, 2015). This section will focus on the current knowledge of chromosomal abnormalities and molecular genetics; their clinical utility and for the potential for integrating molecular profiling into clinically relevant, functional diagnostic and prognostic groups. This background was used to design the key features of the new diagnostic assay for this research project.

1.3.1 Cytogenetics of AML

The earliest genetic abnormalities to be described in cancer were by the appearance of chromosomes under the microscope; conventional cytogenetic analysis to demonstrate an abnormal karyotype. The seminal report of the first description of a non-random, acquired cancer-associated genetic abnormality was the 'Philadelphia chromosome' in chronic myeloid leukaemia (CML) (Nowell & Hungerford, 1960). This was later found to be a t(9;22) chromosomal translocation (Rowley, 1973a) and was followed by the identification of translocations t(8;21) and t(15;17) in AML (Rowley, 1973b; Rowley *et al.*, 1977). Several hundred recurrent chromosome abnormalities have been identified that can aid the diagnosis of AML, which are collected in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. They can show evidence of clonality and cell lineage but also, often a specific disease association (Chapter 5; Acute Myeloid Leukaemia in Heim & Mitelman, 2009; Mitelman *et al.*, 2017).

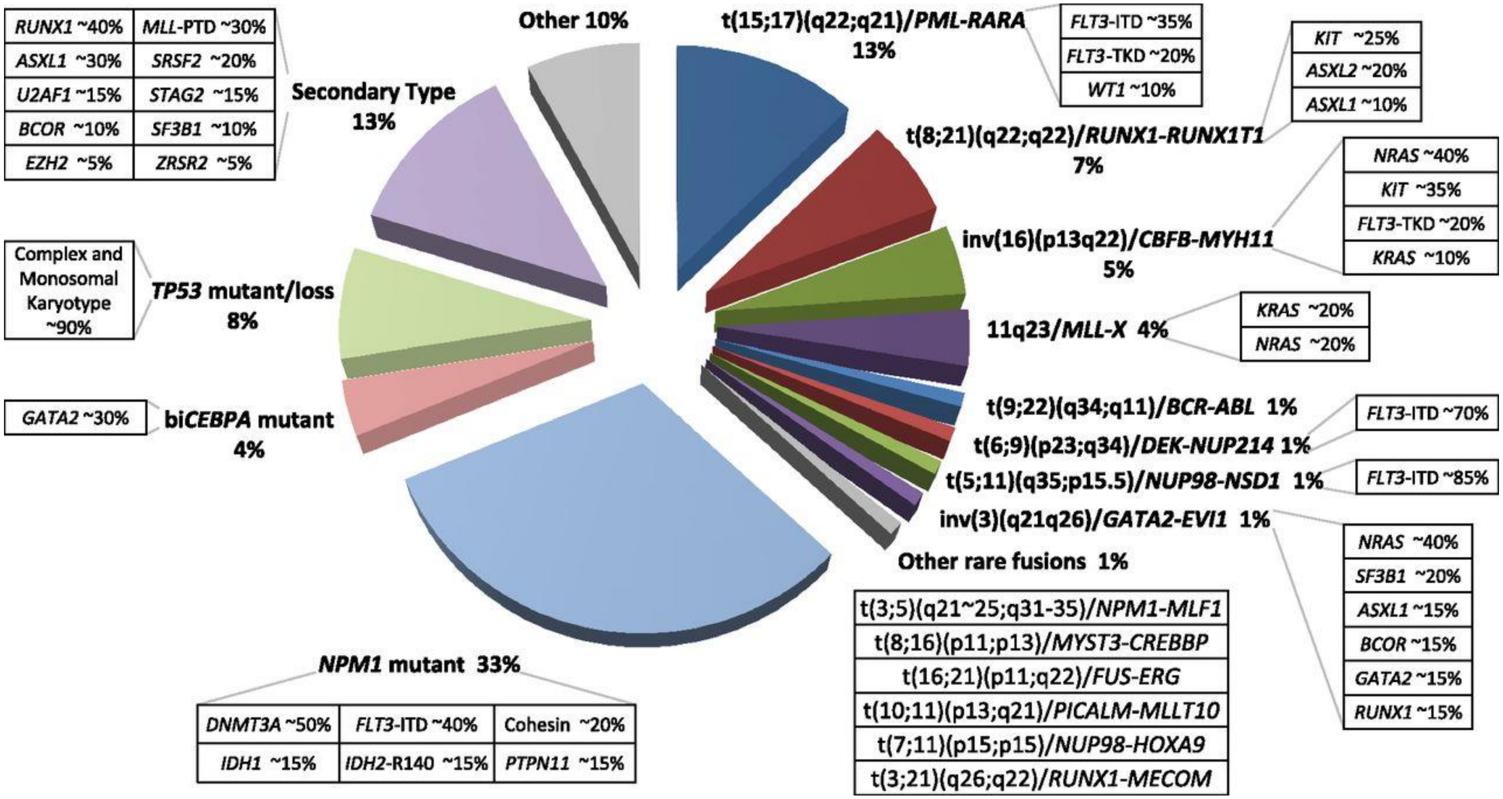


Figure 1.3. Distribution of cytogenetically and molecularly defined subsets of AML presenting in younger adults (reproduced from Grimwade *et al.*, 2016)

1.3.2 Limitations of Cytogenetic analysis

Cytogenetics provides a global genomic screen, albeit at low resolution by NGS standards, and is still routinely used in the diagnosis of AML (Roug *et al.*, 2014; Döhner *et al.*, 2016). Despite the pre-eminence of cytogenetics and its powerful diagnostic and prognostic significance, the analysis of structural variation by conventional karyotype analysis does not reveal the full range of significant genetic abnormalities. A fresh sample and successful growth of leukaemic cells in culture is necessary. Overgrowth by non-cancerous cells and variable chromosome morphology are occasional problems. At best, conventional cytogenetic analysis will reveal structural genomic rearrangements at a resolution of 5Mb of DNA (Heim & Mitelman, 2009) and it requires a trained scientist to provide a subjective analysis. Unfortunately many abnormalities are cryptic to conventional analysis. Approximately 45% of confirmed AML patients have a normal karyotype ('cytogenetically normal-' or CN-AML), meaning the driving genetic alterations are submicroscopic. The large numbers of cases lacking visible chromosome abnormalities are assigned an intermediate risk score but really represent a heterogeneous group in which the genetic lesions have not yet been identified. Furthermore, 20~30% have uncommon or apparently non-specific abnormalities, the prognostic significance of which are uncertain due to insufficient evidence or limited statistical power due to sample size.

1.3.3 The WHO Classification of AML

Certain genetic abnormalities now define specific disease subtypes and the compendium for tumour diagnosis, the WHO Classification, recognises the category 'AML with recurrent genetic abnormalities' which includes class-defining balanced translocations and gene mutations (Arber *et al.*, 2016) (Table 1.2). In particular, t(8;21) and inv(16) involve *RUNX1* and *CBFB* genes respectively, which are heterodimeric components of Core Binding Factor (CBF) and are commonly known as CBF-AML with a combined frequency of 7%. The t(15;17), present in ~5% of AML, is the most specific genetic abnormality in the disease, associated with the distinct subgroup APML. Variant translocations involving *RARA* with different partner genes occur in 1~2% of APML and are important to identify as they may not respond to standard APML treatment (Adams & Nassiri, 2015). t(9;11) and multiple variant translocations of *KMT2A* at 11q23, involving more than 80 partner genes, have variable significance and the classification makes a distinction between these subtypes. The ubiquitous t(9;22) with *BCR-ABL1* gene fusion is included as a provisional entity. The current classification also includes 'AML with mutated *NPM1*' and 'AML with biallelic *CEBPA* mutations' (Arber *et al.*, 2016). The classification has also newly included *de novo* AML with *RUNX1* molecular mutations as a provisional entity (Arber *et al.*, 2016); this disease appears to be biologically distinct with a worse prognosis than other AML types (Tang *et al.*, 2009; Gaidzik *et al.*, 2011; Schnittger *et al.*, 2011b; Mendler *et al.*, 2012; Gaidzik *et al.*, 2016).

Although not specific to these subtypes, cytogenetic profiles also help to define other disease subclasses; 'AML with myelodysplasia-related changes' and 'Therapy-related myeloid neoplasms' (Arber *et al.*, 2008; Vardiman *et al.*, 2008). Whilst most cases of Myelodysplastic Syndrome (MDS) and AML are sporadic, there is increasing recognition of familial disease which is associated with predisposing germline mutations (West *et al.*, 2014a). The category 'Myeloid neoplasms with germline predisposition' is now defined (Arber *et al.*, 2016). Different disease types are involved including MDS, MDS/MPN, as well as AML. The range includes AML with germline *CEBPA* and myeloid malignancies with *DDX41*, *RUNX1*, *ANKRD26*, *ETV6* and *GATA2* mutations, as well as myeloid neoplasms associated with BM failure syndromes. With an underlying genetic defect and often with a leukaemia-predisposition syndrome, the diseases have quite different dynamics from sporadic disease and screening of asymptomatic family members for mutations is indicated (Arber *et al.*, 2016). *GATA2* can have initiating mutations and is mainly associated with familial MDS and AML (Hou *et al.*, 2015). A draft report of a family with three siblings with an inherited *GATA2* germline mutation is presented in Appendix 7.4.2.

Table 1.2. WHO classification of acute myeloid leukaemia (AML) and related neoplasms and acute leukaemias of ambiguous lineage (Arber *et al.*, 2016)

Acute myeloid leukaemia (AML) and related neoplasms	
1. AML with recurrent genetic abnormalities	**4. AML, NOS**
AML with t(8;21)(q22;q22.1); *RUNX1-RUNX1T1*	AML with minimal differentiation
AML with inv(16)(p13.1q22); *CBFB-MYH11*	AML without maturation
APL with *PML-RARA*	AML with maturation
AML with t(9;11)(p21.3;q23.3); *MLLT3-KMT2A*	Acute myelomonocytic leukaemia
AML with t(6;9)(p23;q34.1); *DEK-NUP214*	Acute monoblastic/monocytic leukaemia
AML with inv(3)(q21.3q26.2); *GATA2, MECOM*	Pure erythroid leukaemia
AML with t(1;22)(p13.3;q13.3); *RBM15-MKL1*	Acute megakaryoblastic leukaemia
Provisional entity: AML with BCR-ABL1	Acute basophilic leukaemia
AML with mutated *NPM1*	Acute panmyelosis with myelofibrosis
AML with biallelic mutations of *CEBPA*	**5. Myeloid sarcoma**
Provisional entity: AML with mutated RUNX1	**6. Myeloid proliferations related to Down syndrome**
2. AML with myelodysplasia-related changes	Transient abnormal myelopoiesis (TAM)
3. Therapy-related myeloid neoplasms	Myeloid leukaemia associated with Down syndrome

Acute leukaemias of ambiguous lineage
Acute undifferentiated leukaemia
Mixed phenotype acute leukaemia (MPAL) with t(9;22)(q34.1;q11.2); *BCR-ABL1*
MPAL with t(v;11q23.3); *KMT2A* rearranged
MPAL, B/myeloid, NOS
MPAL, T/myeloid, NOS

1.3.4 Cytogenetic and molecular genetic risk factors and treatment outcomes

Genetic profiling already has a significant impact on the treatment of AML. Chromosomal alterations are the most powerful baseline prognostic factors for response to induction therapy and for survival in AML; cytogenetic analysis of the leukemic cells is used extensively in risk-stratified treatment regimens to guide standard and investigational therapy (Grimwade *et al.*, 2010; Döhner *et al.*, 2016) (see Table 1.3). Younger adult patients (aged ≤60) are commonly categorized into 3 risk groups, favourable, intermediate, or adverse. The subset with favourable outcomes, including CBF-AML, typically achieve long-term survival with conventional chemotherapy, dose intensified in some studies. A survival benefit using GO in CBF-AML, demonstrated in the MRC AML15 study, requires confirmation (Döhner & Gaidzik, 2011). Even within this favourable group, there is some variation in disease behaviour. By contrast, high risk patients, such as those with monosomy 7 or complex karyotype, typically suffer dismal outcomes with conventional chemotherapy and may derive benefit from intensive consolidation therapy in first remission, such as allogeneic SCT (Döhner *et al.*, 2016).

The majority of patients, however, fall within an 'intermediate' risk group in which the genetic abnormality is classified as neither favourable nor adverse. It is possible that the driving genetic alteration truly identifies a disease of moderate severity, as measured by relapse free survival and overall survival. In many cases, however, it means that the underlying, driving stemline mutation has insufficient statistical evidence for it to be classified or that it may not be detected by conventional testing. CN-AML falls within the 'intermediate' group. Substantial outcome variability remains for this group and for many patients, treatment cannot optimally be directed. Confirmation of new prognostic markers is likely to lead to improved genetic stratification of AML and eventually to better prognostication and best treatment of the disease for more patients.

Table 1.3. 2017 European LeukemiaNet risk stratification by genetics (Table 5 from Döhner *et al.*, 2016)

Genetic group	Subsets
Favourable	• t(8;21)(q22;q22); *RUNX1-RUNX1T1* • inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* • Mutated *NPM1* without *FLT3*-ITD or with *FLT3*-ITD^{low} • Biallelic mutated *CEBPA*
Intermediate	• Mutated *NPM1* and *FLT3*-ITD^{high} • Wild-type *NPM1* without *FLT3*-ITD or with *FLT3*-ITD^{low} (w/o other adverse risk genetics) • t(9;11)(p22;q23); *MLLT3-KMT2A* • Cytogenetic abnormalities not classified as favourable or adverse
Adverse	• inv(3)(q21q26.2) or t(3;3)(q21;q26.2); *GATA2, MECOM (EVI1)* • t(6;9)(p23;q34); *DEK-NUP214* • t(v;11)(v;q23.3); *KMT2A* rearranged • t(9;22)(q34.1:q11.2); *BCR-ABL1* • -5 or del(5q); -7; -17/abn(17p) • Complex karyotype (≥3 abnormalities)/ monosomal karyotype • Mutated *RUNX1* • Mutated *ASXL1* • Mutated *TP53*

1.3.5 Refinement of cytogenetic profiling from molecular genetics evidence

In recent years, molecular genetics studies have investigated the involvement of mutations in known oncogenes in AML and a number that are recurrent and common have been studied extensively. Several genes that contribute to the pathogenesis of AML and are becoming important diagnostic factors are now well-characterised (Marcucci *et al.*, 2011; Abdel-Wahab *et al.*, 2011; Abdel-Wahab, 2012; Meyer & Levine, 2014; Naoe & Kiyoi, 2014; Ohgami & Arber, 2015). The number of reported mutations has increased dramatically in recent years; it is likely that many more biomarkers will be available for assessment for risk stratification (see Table 1.4). This will provide better information to redefine the conventional 'intermediate' risk group, particularly CN-AML. The evaluation in cohort studies of multiple, commonly mutated genes in AML, such as *IDH1* and *IDH2*, *DNMT3A*, *RAS*, *PHF6*, *KMT2A*-PTD, *WT1*, *RUNX1*, *KIT* and *TP53*, gives strong indication that they should be considered for more extensive study.

However, to date, only a few molecular abnormalities have been shown to have prognostic significance and are used in laboratory diagnosis (Meyer & Levine, 2014; Döhner *et al.*, 2016). In particular, genetic predictors of favourable outcome, especially in patients without evidence of cytogenetic markers, include mutated *NPM1* without *FLT3*-ITD (Döhner *et al.*, 2005; Suzuki *et al.*, 2005; Thiede *et al.*, 2006) and *CEBPA* double mutation (Fröhling *et al.*, 2004; Wouters *et al.*, 2009; Schlenk *et al.*, 2008). These are incorporated into the European LeukemiaNet (ELN) prognostic system (Table 1.3) (Döhner *et al.*, 2016). *FLT3*-ITD (Thiede *et al.*, 2002; Schlenk *et al.*, 2008; Gale *et al.*, 2008; Kottaridis *et al.*, 2001; Whitman *et al.*, 2001; Fröhling *et al.*, 2002) but not *FLT3*-TKD (Mead *et al.*, 2007) is recognised as a marker of adverse prognosis. Many other gene mutations have been identified that are recurrently involved in AML development and have been strongly suggested to be of prognostic significance, requiring further evidence and confirmation of clinical utility (Table 1.4).

Several groups of related genes in novel pathways have been identified as mutated in AML (Figures 1.4 and 1.5). Individually, genes within their group might be rare but evidence is emerging that they are generally mutually exclusive and are thought to be interchangeable with genes of similar function within their group. Mutations in the group of epigenetic modifying genes are common and include; *DNMT3A* (DNA methylation), *TET2*, *IDH1* and *IDH2* mutations (DNA hydroxymethylation) and *EZH2*, *KMT2A* and *ASXL1* mutations (histone modification) (Abdel-Wahab *et al.*, 2011). Spliceosome complex genes encode proteins which catalyse splicing of precursor mRNA which is an essential step in the control of gene expression (Will & Lührmann, 2011). Mutations are thought to lead to transcriptional dysfunction (Inoue *et al.*, 2016; Joshi *et al.*, 2017). Spliceosome mutations are particularly associated with MDS and AML that has evolved from MDS (Cho *et al.*, 2015).

Recent genome-wide sequencing studies have identified frequent mutations in members of the cohesin complex gene group, in AML and other myeloid malignancies. These may act by increasing chromatin accessibility to transcription factors, altering the growth characteristics of haematopoietic progenitor cells (Fisher *et al.*, 2017; Mazumdar & Majeti, 2017). Several genes are involved in the four core subunits of the cohesion complex, including *STAG1*, *STAG2*, *SMC1A*, *SMC3*, and *RAD21*. Each gene mutation is mutually exclusive, suggesting that any one mutation in the group is sufficient to disrupt the cohesion complex and promote AML.

The conventional approach to genetic diagnosis of AML, therefore, relies on a combination of different techniques for classification and prognostication; cytogenetic analysis, supplemented by FISH and molecular genetics techniques (PCR) for individual gene rearrangements. An argument against generalised mutation profiling of AML as part of the standard diagnostic work-up, is the complexity of the information derived; by breaking genomic groups down into increasingly small and statistically indistinguishable sub-fractions, any diagnostic and prognostic significance will be lost (Fig. 1.3). However, a number of powerful studies have combined NGS of AML genomes with patients' outcomes in large clinical trials which have revealed important patterns of gene interactions and the recognition of biologically and clinically distinct subgroups (Grossmann *et al.*, 2012; Patel *et al.*, 2012; The Cancer Genome Atlas Research Network, 2013; Papaemmanuil *et al.*, 2016; Metzeler *et al.*, 2016). The findings of The Cancer Genome Atlas of co-occurring and mutually exclusive mutations in genes and gene groups is shown in Figure 1.5. In a recent study of

1,540 patients with AML who were analysed by targeted NGS, where data was combined with cytogenetic results, it was possible to segregate patients with AML into 11 distinct phenotypic, clinical and prognostic groups (see Figure 1.4) (Papaemmanuil *et al.*, 2016). In addition to the standard groups, three more functional groups emerged; Chromatin/spliceosome, TP53 mutation/chromosomal aneuploidy and a provisional group with *IDH2*^{R172} mutation, as the sole class-defining lesion. This new classification scheme was able to unambiguously categorise at least 80% of AML into single groups based upon the underlying genetic abnormalities. This classification is used to redefine the diagnostic profiles for the patients in this research (see Results, Table 3.13).

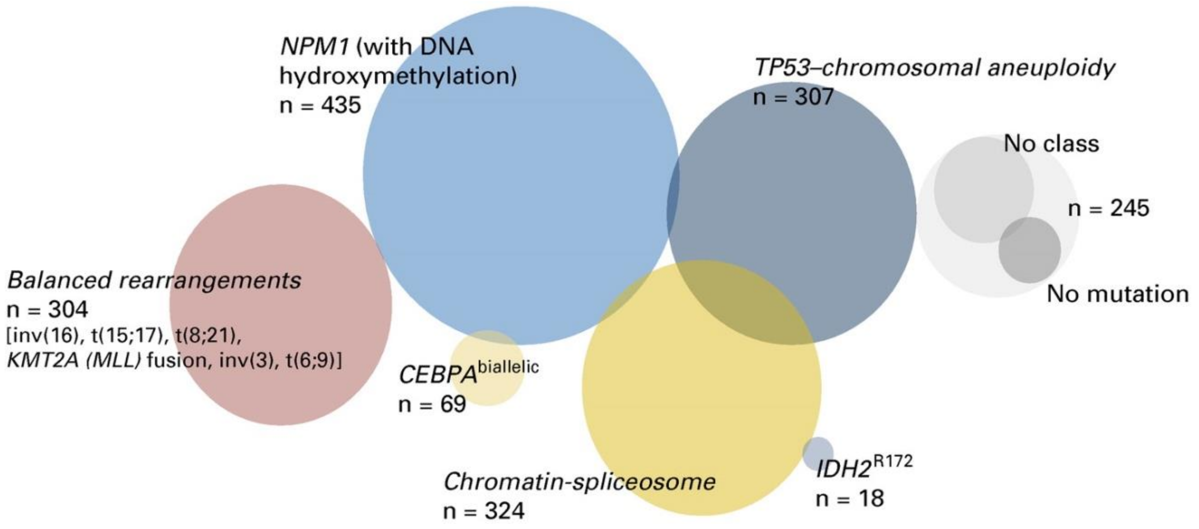


Fig 1.4. Genomic classification of 1,540 AML patients from Papaemmanuil et al (2016). 86% were classified with minimal overlap into 11 clinically relevant groups (including six genomic groups characterised by translocations and/or inversions which are displayed as one group) (Bullinger *et al.*, 2017)

Table 1.4. Frequently mutated genes in the genomic landscape of AML

Gene	Molecular background	Other clinical and pathology features
NPM1	Nucleophosmin (NPM1) protein, mutations altered nuclear signal transduction ((Döhner *et al.*, 2005; Schnittger *et al.*, 2005; Suzuki *et al.*, 2005). Found in 33% of AML and ~85% of cases have a normal karyotype. Heterozygous mutations in exon 12 and rarely exons 9 or 11. 70~80% are a specific TCTG tetranucleotide duplication at codons 956 to 959 ("mutation A") but about 40 mutant variants are described (Falini *et al.*, 2007; Albiero *et al.*, 2007). ~40% have *FLT3*-ITD	Distinctive monocytoid phenotype. Favourable prognosis in the absence of cytogenetic prognostic indicators and *FLT3*-ITD (Haferlach *et al.*, 2009; Döhner & Gaidzik, 2011). Prognosis may be age-dependant and lost in the elderly (Ostronoff *et al.*, 2015). Indicates against HSCT in first CR (Döhner *et al.*, 2016) and in elderly patients who may benefit from intensive chemotherapy (Falini *et al.*, 2005; Falini *et al.*, 2007). Variants may be shown to have differing diagnostic significance (Alpermann *et al.*, 2016).
FLT3	*FLT3* encodes a receptor transmembrane tyrosine kinase. Occurring in 20~30% of cases of AML, most commonly Internal Tandem Duplications (ITD) of 15~180 bp (multiples of 3 base pairs) in the juxtamembrane domain (exons 14 and 15) *FLT3*-ITD result in auto-dimerization and autophosphorylation and constitutive activation of the receptor. *FLT3*-ITD is associated with t(6;9)/*DEK-NUP214* and *NUP98-NSD1*. Missense mutations in the tyrosine kinase activation domain (TKD) at codon 835 and 836 (*FLT3*-TKD) in 5~10% (Lagunas-Rangel & Chavez-Valencia, 2017).	*FLT3*-ITD is the most significant prognostic marker in CN-AML, predicting poor outcomes (Nakao *et al.*, 1996; Gilliland & Griffin, 2002). The highest risk may be conferred by high allelic dosage (Schnittger *et al.*, 2011a), or in patients with high aberrant *FLT3* expression (Fitzgibbon *et al.*, 2005). Allogeneic HSCT is recommended (Pfeiffer *et al.*, 2013; Döhner *et al.*, 2016). *FLT3* is potentially an actionable therapeutic target (reviewed in Meyer & Levine, 2014) and inhibitors are being evaluated as supplements to conventional chemotherapy in randomised phase III trials such as MRC AML19 (Cardiff University, 2015). The prognostic significance of *FLT3*-TKD is uncertain (Mead *et al.*, 2007; Whitman *et al.*, 2008; Marcucci *et al.*, 2011).
CEBPA	CCAAT/enhancer binding protein alpha (*CEBPA*) is a transcription factor involved in myeloid differentiation. Mutations are specific to AML and occur in 6~18% of *de novo* AML, but with a higher frequency in CN-AML. 5~10% of cases are germline mutations. CEBPA mutations occur as single or double mutations and at multiple positions in the coding sequence of the gene (Pabst *et al.*, 2001; reviewed in Bienz *et al.*, 2005). Most frequent mutations are (1) N-terminal frame-shift (nonsense/out of frame indels) dominant-negative protein or (2) C-terminal in-frame indel affecting DNA-binding function. 40–50% of *CEBPA* mutated cases are bi-allelic. Single allelic mutation frequently are associated with *FLT3* (30–35% of cases) or *NPM1* mutations (30–35% of cases).	Double mutations are associated with more specific features and are class-defining. Bi-allelic mutations are associated with improved prognosis; single allelic mutation shows no prognostic significance (Wouters *et al.*, 2009; Taskesen *et al.*, 2011; Fasan *et al.*, 2014). HSCT is not recommended in CN-AML with double CEBPA mutations.
KIT	Receptor tyrosine kinase; acts as a receptor for stem cell factor. Mutations found in 20~30% of CBF-AML (Lavallée *et al.*, 2015). Frequently in the kinase domain at D816, primarily in exon 17 (31%) and exon 8 (6%).	Do not respond to tyrosine kinase inhibitors. Exon 17 mutations have a strong adverse effect on the relapse and survival of adult t(8;21) AML patients (Park *et al.*, 2011; Qin *et al.*, 2014) and inv(16) patients (Paschka *et al.*, 2006).
WT1	Wilms Tumour protein (WT1) is a zinc finger transcription factor (Krauth *et al.*, 2015).	Clinical significance is uncertain. WT1mut had an independent adverse impact on EFS (Krauth *et al.*, 2015).

JAK2	Janus-kinase-2 (JAK2) is a non-receptor tyrosine kinase frequently mutated in myeloproliferative neoplasms; mutations result in activation which is independent of upstream cytokine signalling.	Rare mutations in *de novo* AML. Associated with CBF-AML with t(8;21)(q22;q22); *RUNX1-RUNXT1* or AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11.*
KRAS & NRAS	Membrane associated signal transduction GTPase; activating mutations increase cellular proliferation and decrease apoptosis. NRAS and KRAS mutations, typically in codons 12, 13 and 61, are present in about 25% and 15% of AML. Common findings (30%) in paediatric CBF-AML (Goemans *et al.*, 2005).	Mutations may not have an impact on clinical prognosis.
TP53	Tumour suppressor protein, inactivating point mutations and indels associated with complex karyotype and therapy-related AML.	Mutations associated with poor prognosis (Rucker (Rücker *et al.*, 2012).
RUNX1	Transcription factor. Intragenic mutations lead to a pre-leukaemic state that predisposes to AML. *RUNX1* mutation occurs in approximately 10% of AML, with an incidence increasing with age, peaking in the elderly (Gaidzik *et al.*, 2016). Associated with trisomy 13 and mutations in *ASXL1*, *IDH2*, and *EZH2*, the spliceosome genes (SRSF2 and SF3B1), the cohesin complex gene *STAG2* (Thota *et al.*, 2014) and *BCOR6* and *PHF6* (Gaidzik *et al.*, 2016).	*RUNX1* mutations are class-defining recurrent genetic abnormalities associated with specific presenting clinical and pathologic features, including immature disease phenotype, male predominance and an inferior prognosis.
Mutations of epigenetic modifiers (Abdel-Wahab *et al.*, 2011)		
DNMT3A	**DNA methylation**. DNA methyltransferase 3A (*DNMT3A*), inactivating point mutations most commonly at amino acid D882 occur in up to 25% of AML, particularly monocytic subtypes.	Mutations associated with inferior survival in CN-AML, independent of *FLT3* mutation status (Ley *et al.*, 2010; Thol *et al.*, 2011), particularly without *NPM1* or *FLT3* mutations (Ribeiro *et al.*, 2012) and in monocytic subtypes (Yan *et al.*, 2011).
KMT2A	**Histone modification**. Histone methyltransferase that also interacts with RUNX1 to effect cellular functions. Fusion genes from chromosomal translocations and partial tandem duplications (PTD) between exons 5 and 11 or 12 is inserted, in frame, into intron 4. Found in 8~10% of CN-AML (Basecke *et al.*, 2006).	Prognostic impact still under study but may confer poorer prognosis. Reported to confer an adverse prognosis in CN-AML (Schnittger *et al.*, 2000; Döhner *et al.*, 2002) but not in elderly patients (Whitman *et al.*, 2012). Better outcomes are reported following early allogeneic and autologous HSCT transplant (Whitman *et al.*, 2007).
TET2	**DNA hydroxymethylation** to reverse methylation effects on DNA. Mutually exclusive to *IDH1* and *IDH2* mutations and can be accompanied by deletion of the second allele (Bacher *et al.*, 2010; Bacher *et al.*, 2012; Weissmann *et al.*, 2012).	Some studies point to poorer prognosis. An adverse prognosis has been demonstrated in CN-AML patients without *NPM1* or *FLT3* mutations (Chou *et al.*, 2011) or favourable risk CN-AML NPM1+ or CEBPA+/FLT- AML (Bacher *et al.*, 2010; Metzeler *et al.*, 2011b; Weissmann *et al.*, 2012) but some studies are equivocal (Damm *et al.*, 2014).
IDH1 & IDH2	**DNA hydroxymethylation**. Isocitrate dehydrogenase -1 and -2 (*IDH1, IDH2*) are enzymes that convert isocitrate to alpha-ketoglutarate in the Krebs cycle; mutations result in production of 2-hydroxyglutarate which inhibits hydroxymethylation of DNA. Three mutually exclusive recurrent mutations in *IDH1* and *IDH2*	Mutations are likely to have different prognostic implications; IDH1 R132 adverse prognosis with wild-type FLT3 (Abbas *et al.*, 2010; Green *et al.*, 2010; Paschka *et al.*, 2010), IDH2 R140 favourable (Green *et al.*, 2011) and the IDH1 R172 mutation is neutral. Clinical trials with IDH targeted inhibitors are ongoing.

ASXL1	**Histone modification**. Additional sex combs-like gene (ASXL1) is a chromatin-binding protein likely involved in methylation of histone proteins. Mutations of exon 12 are mutually exclusive of *NPM1* mutations and absence of *FLT3*-ITD mutated *CEBPA*. *ASXL1* have been found most commonly in patients >60 years and AML with myelodysplasia related changes (Devillier *et al.*, 2012).	*ASXL1* mutations are markers of poor prognosis in some studies, in favourable and intermediate risk AML and in AML overall (Metzeler *et al.*, 2011a; Pratcorona *et al.*, 2012).
EZH2	**Histone modification**. Enhancer of zeste homologue 2 (*EZH2*) is the catalytic component of the PRC2 complex and functions to trimethylate histone tails	Rarely mutated in AML, mutations associated with poor prognosis in some studies.
Cohesin Complex		
Cohesin Complex Genes	*STAG1, STAG2, SMC1A, SMC3,* and *RAD21* genes involved in the control of sister chromatid separation. Individually mutations uncommon but overall the group is mutated in 5~10% of AML and >50% of Down Syndrome-AMKL. Mutations are mutually exclusive with others within Cohesin group. Loss of function / null mutations are common, complete protein deficiency in the X-linked genes (*STAG2* and *SMC1A*) in males. Co-occur with NPM1, *DNMT3A*, *TET2* or *RUNX1* mutations (Fisher *et al.*, 2017; Mazumdar & Majeti, 2017).	Included in a high risk group in recent genomic risk stratification (Papaemmanuil *et al.*, 2016).
Spliceosome Complex		
Spliceosome Complex Genes	The genes of the ribonucleoproteins of the spliceosome complex involved in the control of gene expression (Will & Lührmann, 2011). *SF3B1*, *U2AF1*, *SRSF2*, *ZRSR2*, *SF3A1*, *PRPF40B*, *U2AF2*, and *SF1* mutated in AML and associated with evolution from MDS (Cho *et al.*, 2015).	*U2AF1* mutation results in abnormal splicing of genes involved in myeloid differentiation and proliferation and confer a poor prognosis and may be associated with multilineage dysplasia. *SF3B1* mutations are common (20%) in MDS with ring sideroblasts, and associated with a favourable prognosis (Papaemmanuil *et al.*, 2011).

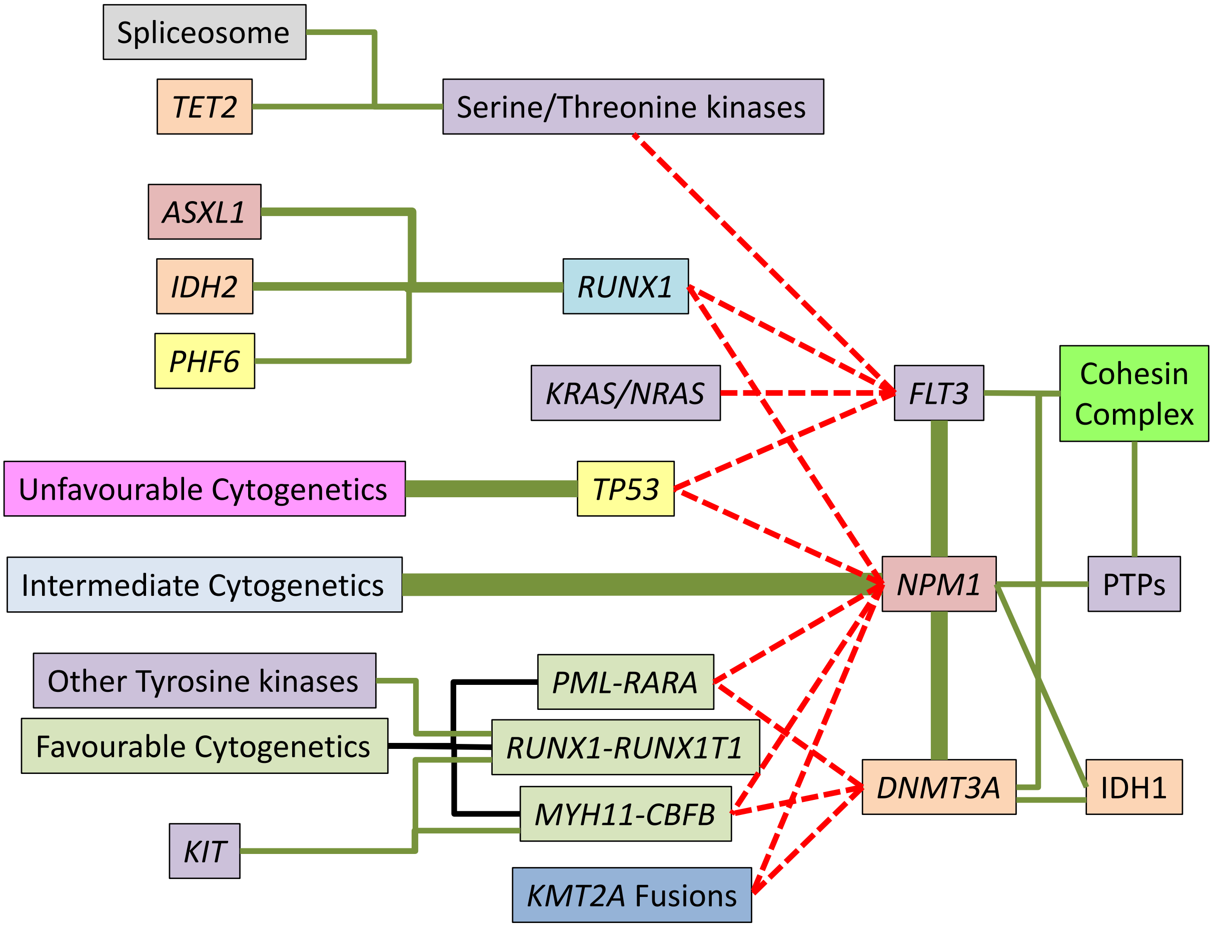


Figure 1.5. Co-occurring and mutually exclusive mutations in genes/groups. Boxes represent genes, gene groups, or cytogenetic risk groups. Green lines edges connect genes that co-occur in a significant number of samples. Red dashed lines connect nodes that are mutually exclusive. Black lines indicate gene fusions that define favourable cytogenetics. The thickness of each green line approximately corresponds to the strength of the association as determined from the frequency from TCGA dataset (adapted from The Cancer Genome Atlas Research Network, 2013).

1.3.6 Complex karyotype and Monosomal karyotype

Complex karyotype (CK) is an important category in AML because it identifies a group of usually elderly AML patients consistently associated with very poor treatment outcomes after chemotherapy, achieving highest risk in prognostic scoring systems. This occurs in up to 12% of AML and is defined by the number of visible karyotypic abnormalities in the cells, variably defined by different studies as having between 3 and 5 cytogenetic abnormalities regardless of type, in the absence of favourable cytogenetics. However, the pattern of abnormalities within complex karyotypes is non-random and is frequently characterised by chromosomal imbalances, such as deletions of 5q, 7q and 17p, and gains of 8q, 11q, and 21q (Mrózek, 2008). The descriptive numerical definition of complex karyotype is non-specific and there is no single cytogenetic hallmark of the subgroup. Recently, an entity 'Monosomal karyotype' (MK) has been defined as multiple chromosome loss (or any single chromosome loss with structural abnormalities) (Anelli *et al.*, 2017). Many of these karyotypes qualify as both CK and MK but they achieve prognostic distinction in some scoring systems (Breems & Löwenberg, 2011; Kayser *et al.*, 2012; Grove & Vassiliou, 2014) but not in the largest UK study (Grimwade *et al.*, 2010). MK may best be used to identify patients that have an extremely poor prognosis (Haferlach *et al.*, 2012).

CK and MK are necessarily defined by conventional cytogenetic studies and any attempt to progress to a molecular definition of this highest risk group of patients requires identification of a reliable molecular marker of karyotype instability. Recently, it was reported that approximately 70% of patients with CK have missense mutations or biallelic deletions of the *TP53* gene, and the absence of regulatory properties of functional p53 would explain genomic instability (Schoch *et al.*, 2005; Haferlach *et al.*, 2008; Seifert *et al.*, 2009; Rücker *et al.*, 2012). *TP53* mutations are uncommon in other cytogenetic subgroups (2%). *TP53* alterations define a subgroup of AML with the worst overall survival (OS at 3 years of 0%). *TP53* mutation was therefore used as a surrogate marker for these patients in a molecular genetic classification of AML (Grossmann *et al.*, 2012). However, ~30% of AML patients with CK lack *TP53* mutation and it is unclear if other genes in the p53 signaling pathway are also involved and have the same effect (Medeiros, 2012) or if patients with CK but without *TP53* have the same predictive value (Lazarus & Litzow, 2012). Inevitably, the significance of other mutations will be identified and a more specific and relevant classification will emerge.

1.3.7 Different considerations in Childhood AML

Paediatric AML has a different profile of cytogenetic abnormalities to the disease of young adults and the elderly, including a number of abnormalities that are not found outside of this group (see Table 1.6) (Harrison *et al.*, 2010; Creutzig *et al.*, 2016). Outcome data from children treated in paediatric AML trials confirmed similar prognosis as adults, e.g. CBF-AML and monosomy 7 (Harrison *et al.*, 2010). t(9;11) and other *KMT2A* translocations are the most frequent abnormalities (16%), particularly in infants (50%) (Harrison *et al.*, 2010). An independent international study reported an adverse prognosis in children with specific *KMT2A* rearrangements (Balgobind *et al.*, 2009). The presence of 12p abnormalities predicted a poor outcome (von Neuhoff *et al.*, 2010; Harrison *et al.*, 2010). Recently, rare cryptic chromosomal abnormalities have been described, specific to paediatric AML, which confer a poor outcome; t(5;11)(q35;p15.5)/*NUP98-NSD1* is associated with *FLT3*-ITD (Akiki *et al.*, 2013; Hollink *et al.*, 2011; Shiba *et al.*, 2013), t(7;12)(q36;p13)/*MNX1-ETV6* occurs mainly in infants and is often accompanied by a deletion of the long arm of chromosome 7 (von Bergh *et al.*, 2006; Tosi *et al.*, 2000) and inv(16)(p13.3q24.3)/*CBFA2T3-GLIS2* (Gruber *et al.*, 2012; Masetti *et al.*, 2013). The current MyeChild 01 trial combines prognostic information from these key studies for cytogenetic risk group stratification (Table 1.6) (University of Birmingham, 2015).

Table 1.6. The cytogenetic and molecular risk group assignment for MyeChild 01; expected incidence and estimated number of cases Paediatric AML cytogenetic risk groups (MyeChild 01 Protocol v1.0, University of Birmingham, 2015)

Good Risk	Incidence	Poor Risk	Incidence
t(8;21)(q22;q22)/ *RUNX1-RUNX1T1*	12%	-7	4%
inv(16)(p13q22)/ t(16;16)(p13;q22)/*CBFB-MYH11*	6%	-5/del(5q)	~1%
Double mutation of *CEBPA* without *FLT3*-ITD	5%	inv(3)(q21q26)/ t(3;3)(q21;q26)/abn(3q26)	~1%
Mutation of *NPM1* without *FLT3*-ITD	5%	t(6;9)(p23;q34)/*DEK-NUP214*	~1%
		t(9;22)(q34;q11)/*BCR-ABL1*	~1%
Intermediate risk		12p abnormalities	2~4%
t(9;11)(p21;q23)/*KMT2A-MLLT3* t(11;19)(q23;p13.3)/*KMT2A-MLLT1* Other *KMT2A* rearrangements not classified as poor risk	11%	*KMT2A* rearrangements classified as poor risk; t(6;11)(q27;q23)/ *KMT2A-MLLT4* t(4;11)(q21;q23)/ *KMT2A-AFF1* t(10;11)(p11-p14;q23)/ *KMT2A-MLLT10*	5%
All other abnormalities which are neither good nor poor risk	25%	t(5;11)(q35;p15.5)/ *NUP98-NSD1*	<5%
		t(7;12)(q36;p13)/ *MNX1-ETV6*	<1%
		inv(16)(p13.3q24.3)/ *CBFA2T3-GLIS2*	<2%
		FLT3-ITD without *NPM1* or CBF	10%

1.3.8 Genetics factors predictive of therapeutic response

Significant advances in the management of AML are possible by evaluating genomic profiles for their efficacy as predictive markers of outcome with specific treatments or targeted therapies. The major success is exemplified in APML, defined by *PML-RARA* gene fusion, which indicates a specific treatment regimen of ATRA combined with idarubicin, daunorubicin or arsenic trioxide. Excellent responses are achieved with long term, complete remission rates in 90% and cure rates of at least 80% (Coombs *et al.*, 2015).

Using standard chemotherapy without the need for stem cell transplantation, CBF-AML patients typically achieve high CR rates in the region of 88% and 42% cure rate (relapse free survival (RFS) at 10 years) (Solh *et al.*, 2014; Sinha *et al.*, 2015). However, a third of patients have *KIT* mutations, worsening prognosis in some studies (Paschka *et al.*, 2006). Clinical trials are examining the use of the tyrosine kinase inhibitor, dasatinib as supplement to conventional treatment (Döhner & Gaidzik, 2011). The frequency and high risk associated with *FLT3*-ITD makes this mutation a compelling drug target, and whilst a durable treatment has proved elusive to date, several second generation inhibitors are showing more promise in clinical trials (Stein & Tallman, 2016; Grunwald & Levis, 2015). *IDH1* and *IDH2* mutations are in use as a target in early phase trials which are showing significant promise of improved outcomes (Stein & Tallman, 2016; Dombret & Gardin, 2016). There are currently no effective therapies to target the common *NPM1* mutations but GO and ATRA have been used to treat these leukaemias with mixed results so far (Grunwald & Levis, 2015). Although epigenetic modulators have been used in the treatment of AML with modest effect (Engen *et al.*, 2016), cytogenetic aberrations are markers predicting best response in azacytidine (Fenaux *et al.*, 2009) and also lenalidomide (Ades *et al.*, 2009). In practice, however, targeted therapies have been slow to translate into the clinical arena and the identification of more therapeutic targets combined with a pipeline of early phase trials will be needed to accelerate the use of personalisation of treatment in AML (Lawler & Sullivan, 2015; Hollingsworth & Biankin, 2015; Bertier *et al.*, 2016).

1.3.9 Genetics to monitor Minimal Residual Disease (MRD)

In addition to baseline prognostic factors (see sections 1.3.4 & 1.3.5), an important area in AML management is testing for response-related risk factors for monitoring disease course, to predict relapse and offer the option of early pre-emptive intervention. Monitoring minimal residual disease (MRD) in AML patients has been studied for many years and a number of reliable methods are possible, depending on the availability a suitable marker for disease evaluation. These include quantitative PCR (qRT-PCR) and flow cytometry but other methods are used, usually determined by the most sensitive method available for the patient (Hourigan & Karp, 2013; Roug *et al.*, 2014; Ommen, 2016). Early complete remission indicates the best post treatment response whilst persistent disease may indicate treatment failure and short survival. The depth of response following induction therapy provides independent prognostic information and prediction of the risk of relapse to guide consolidation therapy (Jourdan *et al.*, 2013). The genetic hallmarks of AML provide convincing targets for the detection of MRD in post-treatment samples, such as the leukaemia-specific transcripts and common gene fusions from chromosomal translocations, e.g. *PML-RARA*, *RUNX1-RUNX1T1* and *CBFB-MYH11* are typically monitored by qRT-PCR. Unfortunately, only about a third of adults carry informative gene fusions. Inclusion of other common leukaemia-specific disease markers, such as *NPM1* mutations, increases coverage to ~60% in younger adults but still, elderly patients only have an informative marker one third of the time (see Figure 1.6) (Grimwade & Freeman, 2014).

The sensitivity for MRD detection is measured relative to the level of expression of an endogenous control gene (e.g. *ABL1*); e.g. RT-qPCR assays for detection of *PML-RARA* transcripts, for example, is at least 1 in 10^4 cells and 1 in 10^5 cells is possible when the MRD target is more highly expressed than the control, providing a higher baseline level. Therefore, molecular genetics is among the most sensitive methods available making these techniques ideal for MRD assessment (Grimwade & Freeman, 2014). The desirable objective to control the leukaemic clone is a 3 log reduction (e.g. 40% to <0.4% disease cells).

Measuring disease persistence is complicated by different variables such as (Grimwade & Freeman, 2014; Hokland *et al.*, 2015; Ommen, 2016);

- Choice of optimal time points for testing
- Selection of the most informative stemline disease marker
- The presence of premalignant clones
- Different genetic drivers have different remission and relapse kinetics
- Varying persistence of abnormal cells into remission
- Varying rates of relapse
- Genotype instability and evolution of the genetic marker over time leading to change or loss of the measurable target

Nevertheless MRD monitoring has the potential to be one of the most significant mechanisms for prognostication and clinical intervention in AML. The AML19 trial incorporates a ‘monitor’ versus ‘no monitor’ randomisation, to test if the process will translate into improved outcomes (Cardiff University, 2015). NGS strategies are under evaluation to identify genetic variants (and combinations thereof) that can be used for DNA-based post-treatment MRD measurement and to use the technology as a means to detect MRD with the potential that multiple mutations can be used as markers for disease cells, obviating the reliance on specific mutational hotspots (Hokland *et al.*, 2015). This would have broader applicability than the expression fusion transcripts assays that are currently provided (Ommen, 2016).

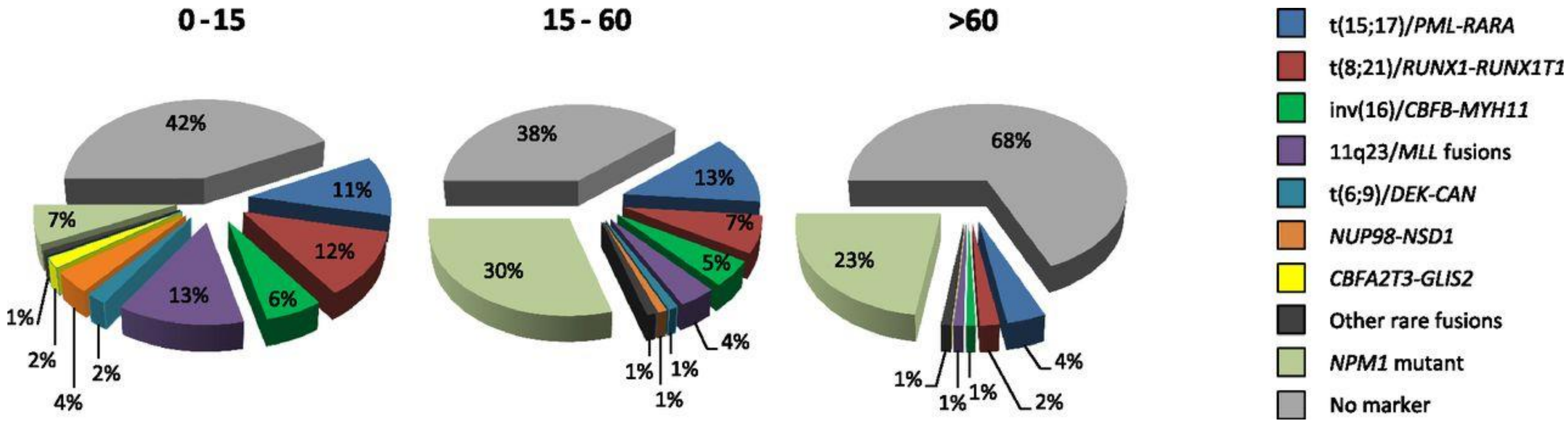


Figure: 1.6. Proportion of AML patients informative for MRD detection by RT-qPCR for leukaemia-specific MRD targets according to age (from Grimwade & Freeman, 2014).

1.4 The development of next generation sequencing for AML and other cancers

The rapid advancement in genomic testing has resulted in a large amount of literature in cancer genomics and several excellent general reviews of the subject, reflecting different authors' research interests and perspectives. Due to the pace of development and breadth of research, a thoroughly comprehensive review is difficult and is certainly beyond the scope of this thesis. The following brief review presents a summary of existing genomic techniques and their limitations, to outline a case for change. The evolution to NGS and its development in AML for diagnostic applications is described, referencing the use of NGS in other cancer types where relevant. There is a particular focus on the development of technology for the detection of different classes of genomic variation, including structural variation.

1.4.1 Traditional genomic approaches to cancer investigation; cytogenetics

Cytogenetics is the study of genetic material at the cellular level, traditionally visualised microscopically as changes in the number and structure of chromosomes, representing alterations in the DNA content of cells. Whereas next generation sequencing has become synonymous with genomics, the earliest genomic technique could be considered to be conventional cytogenetic analysis to provide a karyotype. Cytogenetics provides a global screen for gross structural abnormalities and copy number change by examination of all chromosome pairs. Conventional techniques have obvious restrictions; a fresh sample is required and successful cell culture needs to be initiated to obtain a dividing population of malignant cells so that they can be visualised in metaphase of the cell cycle. The natural presence of a varying proportion of normal (non-cancerous) cells and limitation in the number of cells available or conveniently analysable, restricts the technique's sensitivity to detect clones and significant subclones (Hook, 1977). Overgrowth by normal cells in culture can be a problem. The resolution of conventional cytogenetic analysis is limited to 5Mb DNA at best (Heim & Mitelman, 2009) and malignant cells with poor chromosome morphology are common. Chromosome structure is interpreted perceptually and requires a trained scientist to provide a subjective analysis to identify abnormalities. Unfortunately, many structural genomic rearrangements will be submicroscopic or cryptic to microscopic analysis. In other words, the driving genetic changes are often undetected and some 45% of AML show a normal karyotype (CN-AML). M-FISH or spectral karyotyping is a variant of conventional chromosome analysis using multicolour FISH with differentially labelled chromosome 'paints' for each chromosome with colours artificially assigned *in silico*. This technique provides a global screen with the same limitations as standard karyotyping, albeit with less subjectivity due to the colour differential, but with some loss of resolution at chromosome band definition and is unable to detect intra-chromosomal abnormalities. M-FISH has not been adopted for routine investigation in the UK.

Despite these shortcomings, when supplemented by targeted techniques to detect specific genetic abnormalities; such as fluorescence in situ hybridisation (FISH) with locus specific probes and polymerase chain reaction (PCR), karyotyping remains a powerful diagnostic test, which has not been supplanted by new technology for leukaemia diagnosis. Cytogenetic analysis was pioneered for the study of AML (Grimwade *et al.*, 1998) by examination of microscopically visible chromosome abnormalities and the underlying genetic abnormalities have been directing clinical decisions for two decades. A higher resolution technology which is not reliant on successful cell culture and microscopic detection would lead to a dramatic improvement in abnormality detection.

1.4.2 Microarray technology

Visualisation of chromosome architecture is not always important *per se*; it is the molecular consequences of structural and numerical rearrangements that drive disease for which the visible chromosomal changes are a surrogate marker. Alternative techniques could be used once all the relevant genetic consequences are understood. Genomic microarray technology, such as array Comparative Genomic Hybridisation (array-CGH) and single-nucleotide polymorphism (SNP) oligonucleotide microarray, use known DNA sequences deposited on to a hard surface for hybridisation with test DNA to detect the presence and concentration of sequences of interest. Array-CGH employs co-hybridisation of test (tumour) DNA and normal control DNA (Bullinger & Fröhling, 2012), whereas SNP array hybridises a single test sample to SNP DNA sequences throughout the genome (Sato-Otsubo *et al.*, 2012). Chromosomal Microarray Analysis (CMA) can detect unbalanced chromosomal abnormalities at high-resolution, including sub-microscopic abnormalities too small to be detected by conventional karyotyping. This technology has expedited technological reform in constitutional (non-cancer) cytogenetic services where, extensively, oligonucleotide microarrays have been adopted as a first-line test to detect DNA copy number defects, at high resolution, in infants and children with unexplained developmental delay and intellectual disability (Shaffer *et al.*, 2007; Miller *et al.*, 2010). This is appropriate for the testing of probands of families where genomic imbalances are likely causes of disease and the new technology reveals unprecedented fine detail in analysis, with improved detection of abnormalities. The same technology does not detect all the abnormalities required of a test for cancer, such as balanced chromosomal rearrangements, although modifications such as PCR amplification of genomic DNA translocation breakpoints regions prior to array-CGH have been piloted (Greisman *et al.*, 2011). Nevertheless, microarray technology has significant potential for applications in leukaemia diagnosis (Bullinger & Fröhling, 2012). SNP array is not only currently the most cost effective technology for SNP genotyping, e.g. by allele-specific discrimination by hybridisation, but can also detect copy number variation and recurrent copy-neutral loss of heterozygosity (cnLOH) (Sato-Otsubo *et al.*, 2012). It has gained attention for those disorders largely defined by genomic imbalance such as MDS (Tiu *et al.*, 2011; Arenillas *et al.*, 2013; Mohamedali *et al.*, 2013) and may provide utility for improved prognostication in CN-AML (Bullinger *et al.*, 2010; Parkin *et al.*, 2010; Gronseth *et al.*, 2015). SNP array is an immensely powerful technology which over a number of years has been extensively refined for diagnostic use (Vermeesch *et al.*, 2012; Cooley *et al.*, 2013). SNP array for detection of cnLOH can be used alongside other mutational profiling techniques to provide an integrated profile (Mohamedali *et al.*, 2015; Parkin *et al.*, 2015), which also demonstrates a drawback that multiple platforms are still required to provide a fully comprehensive assemblage of diagnostic information.

1.4.3 Gene Expression Profiling

Isolated and enriched test RNA can be converted to cDNA, amplified, labelled and used to hybridise to microarray for Gene Expression Profiling (GEP). Intuitively, the detection of alterations in the transcriptome, which more closely reflects proteomic changes that would affect phenotypic change, should provide signatures which characterise patients more effectively than genomic studies. In AML, the expression of individual genes such as *EVI1, BAALC, MN1* and microRNAs and recognition of alternative mRNA splicing, may offer distinct clinical utility to gene expression technology (Baldus & Bullinger, 2008; Shivarov & Bullinger, 2014). Yet, after much hype and expectation, the only multi-gene GEP classifiers in regular diagnostic use in cancer inform the use of adjuvant chemotherapy in breast cancer (Kuchel *et al.*, 2016), which are now progressing to second generation assays (Harris *et al.*, 2016). Significant challenges have dogged the transition of GEP from interesting and informative research technology to the expected transformation of clinical practice, including: Variation in tissue handling and preservation of RNA integrity; poor standardisation of assay design; selectivity in classification models for subgroup selection and analysis; variability in the analysis of complex data; effects of sample size on robustness and prediction accuracy of a prognostic gene signature; and the introduction of bias, leading to over-interpretation of clinical applicability (Ioannidis, 2005; Ioannidis, 2007a; Ioannidis, 2007b; Ioannidis *et al.*, 2009).

An early study (Valk *et al.*, 2004) led to the development of AMLProfiler™ (Skyline Diagnostics, Rotterdam, Netherlands) but there is a relative paucity of prospective studies supporting the use of GEP in AML (Verhaak *et al.*, 2009; Kohlmann *et al.*, 2010a). GEP has been applied successfully to the classification of leukaemia in the landmark Microarray Innovations in Leukaemia (MILE) study (Haferlach *et al.*, 2010). Leukaemia gene expression profiling allows accurate prediction of certain subtypes and in AML has been shown to identify major diagnostic groups defined by the expression of chimeric transcription factors. However, detection of mutations affecting signalling molecules and numerical abnormalities still requires alternative molecular methods. In short, genomic aberrations can be detected by GEP class prediction but only comparable to a range of existing genomic tests and so for regular, routine diagnostic purposes, GEP has limited advantage over conventional testing. Better standardisation of assays and an international effort to generate the large datasets required for validation of such complex combinations of expressed genes for class prediction are necessary to make further progress. Microarray-based GEP for clinical investigations have generated both unrealistic hype and excessive scepticism and it is debatable whether GEP alone will have a role in diagnosis (Theilgaard-Monch *et al.*, 2011). The technology is shifting to RNA sequencing (RNA-seq) which has encouraged a significant upsurge in transcriptome analysis, which will revitalise studies in the clinical applications for GEP and in time will probably benefit from integration of genomic data and the stratification of different *omic* analyses in clinical studies (Shivarov & Bullinger, 2014; Gerstung *et al.*, 2015).

1.4.4 Next Generation Sequencing

Next generation sequencing is at the cutting edge of genomic investigations and provided the technology for many genomic studies and an exponential increase in genomic knowledge of cancer. NGS has become almost synonymous with genomics and is undoubtedly the future of routine genetic diagnosis. NGS will be refined into compact analytical systems and provide diagnostic assays, which are being validated for clinical purposes. A diagnostic framework is being established to provide a quality system for diagnostic laboratories.

Extensive work has been performed to characterise the genes from NGS studies and to distinguish the "driver" mutations that confer a growth advantage from the background of "passenger" mutations, acquired during the lifetime of the cell but having no role in leukaemogenesis. This is not necessarily straightforward and is primarily performed by frequency of mutation in cancer genomes. Other bioinformatics approaches utilise information of structural properties of mutations and their position in mutated sub-networks of driver pathways, to infer patterns of co-occurrence or mutual exclusivity. A meta-analysis of four frequentist studies (reported in Mazzarella *et al.*, 2014) identified 21 drivers common to all four studies and a further 32 identified by a single method. Many of these have already been implicated in disease and includes tissue-specific drivers, not implicated in other neoplasms. Bioinformatics methods are becoming sophisticated and less frequent driver genes will still emerge which might require thousands of samples to catalogue genes of low frequency and confirm clinical significance. Several landmark studies but particularly the 200 cases sequenced by The Cancer Genome Atlas (The Cancer Genome Atlas Research Network, 2013), have resulted in the identification of most of the recurrent mutations in genes involved in AML development (figure 1.7).

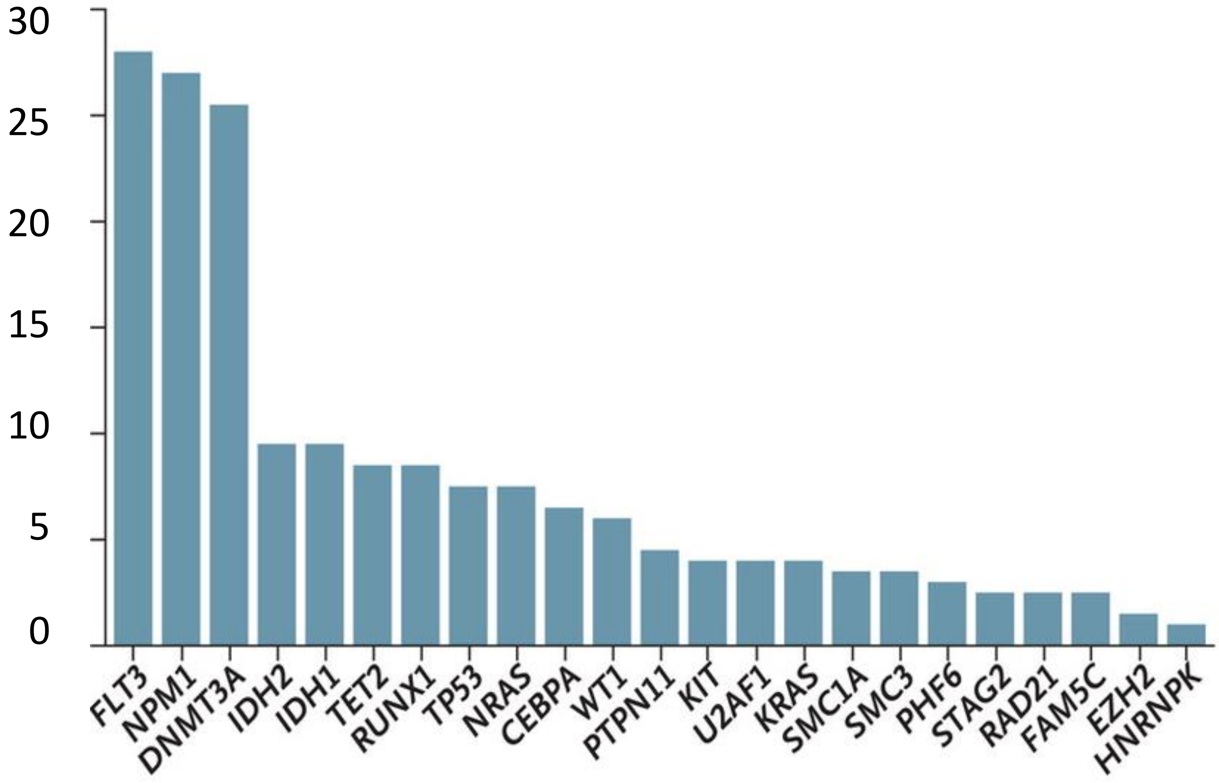


Figure 1.7. Significantly mutated genes in the TCGA dataset, showing the proportion of samples with each mutation

The development in high throughput sequencing technology enhanced researchers' ability to interrogate cancer genomes and there is a concerted effort to bring this new capability for the study of neoplastic disease into diagnostic practice, to bring universal precision medicine closer to reality. Next generation sequencing encompasses a range of different methodologies for the investigation of the genome, transcriptome or epigenome (Bentley *et al.*, 2008; Wang *et al.*, 2009; Krueger *et al.*, 2012). Over the last decade, the rapid and incremental refinement by different commercial companies has brought large scale sequencing of diagnostic standard into routine use (reviewed in Metzker, 2010; Goodwin *et al.*, 2016). The major developments have been in short read DNA sequencers, which have a restricted facility in the length of a DNA molecule that can be sequenced; read lengths of 50~150 nucleotides are typical (Goodwin *et al.*, 2016). However, the millions of simultaneous reads increase the throughput and significantly reduce the cost per nucleotide, which is several orders of magnitude less than Sanger sequencing.

De novo genome assembly uses sequencing with no prior information from a reference genome for alignment; the genome sequence is reconstructed from overlapping contigs and the resulting quality of coverage depends on the size and continuity of the data. Accurate *de novo* assembly is very effective for the characterisation of novel genomes and large scale Structural Variation but is very challenging with the typical read lengths from short-read instruments. Alternatively, NGS 'resequencing' is used extensively for human diagnostic applications, which aligns the short sequence reads to the Human Reference Genome to detect differences and predict the mutational profile of variants in the test genome (Myllykangas & Ji, 2010; Goodwin *et al.*, 2016). The detection of genomic variation is demanding due to variability of biological features and complexity of genomic sequence, and from technical limitations such as sequencing errors, limited read lengths and insert sizes, and sampling biases (e.g. in GC-rich regions). Sophisticated bioinformatic algorithms are required to make highly sensitive and specific predictions of genomic variation. This has resulted in the development of computational approaches, tailored to specific tasks, for the alignment of millions of individual short reads to detect different types of genomic rearrangements. (Ding *et al.*, 2014).

The new technology will eventually have the same (but highly refined) impact on cancer diagnosis that cytogenetics has provided, improving on the accuracy of current testing, providing information for molecular monitoring in remission (which is not currently possible for all tumours) and aiding the detection of biomarkers for targeted therapies (Cronin & Ross, 2011; Chin *et al.*, 2011; Majewski & Bernards, 2011). Importantly, using an appropriate strategy, NGS can be applied to detect single gene mutations and structural abnormalities in a single assay (Grossmann *et al.*, 2011; Graubert & Mardis, 2011) and the current range of genetic technologies will eventually converge into different applications of clonal sequencing testing, perhaps WGS. However, a significant challenge is how to develop the new genomic technology for multiple genetic abnormalities into a single practical and affordable assay, for use in the diagnostic environment to aid clinical management.

1.4.4.1 Depth of Coverage

Depth of coverage or read depth is a measure of the number of times each a specific nucleotide or genomic feature is covered by the multiple sequence reads in an experiment. Analysis detects differences in the number of reads that align to intervals in the reference genome. Assuming that reads are sampled uniformly from the genome sequence, the Lander-Waterman model (Lander & Waterman, 1988) specifies the number of reads that contain a given nucleotide is on average as;

$$\text{Coverage} = \frac{\text{Number of reads x Length of each read}}{\text{Length of the Genome}}$$

Repetitive sequences in the reference genome and biases in sequencing (e.g. different coverage of GC-rich regions) affect depth of coverage calculations. Nevertheless, depth of coverage analysis is a key consideration in calculating the efficiency of NGS and computational methods for its analysis are a key component of a NGS pipeline.

1.4.4.2 NGS Stages

NGS can be considered as a technology providing a single assay to generate a defined mutational profile. However, it comprises a number of distinct modules and the variable performance of each process can influence the outcome of the results of the sequencing assay (Daber *et al.*, 2013; Lee *et al.*, 2015). Figure 1.8 demonstrates a typical NGS workflow.

1. Library Construction – preparation of the nucleic acid target into suitable lengths with adapter ligation compatible with the sequencing equipment
2. Sequencing
3. Bioinformatics analysis
    a) Base Calling
    b) Alignment
    c) Variant Calling
    d) Variant Annotation

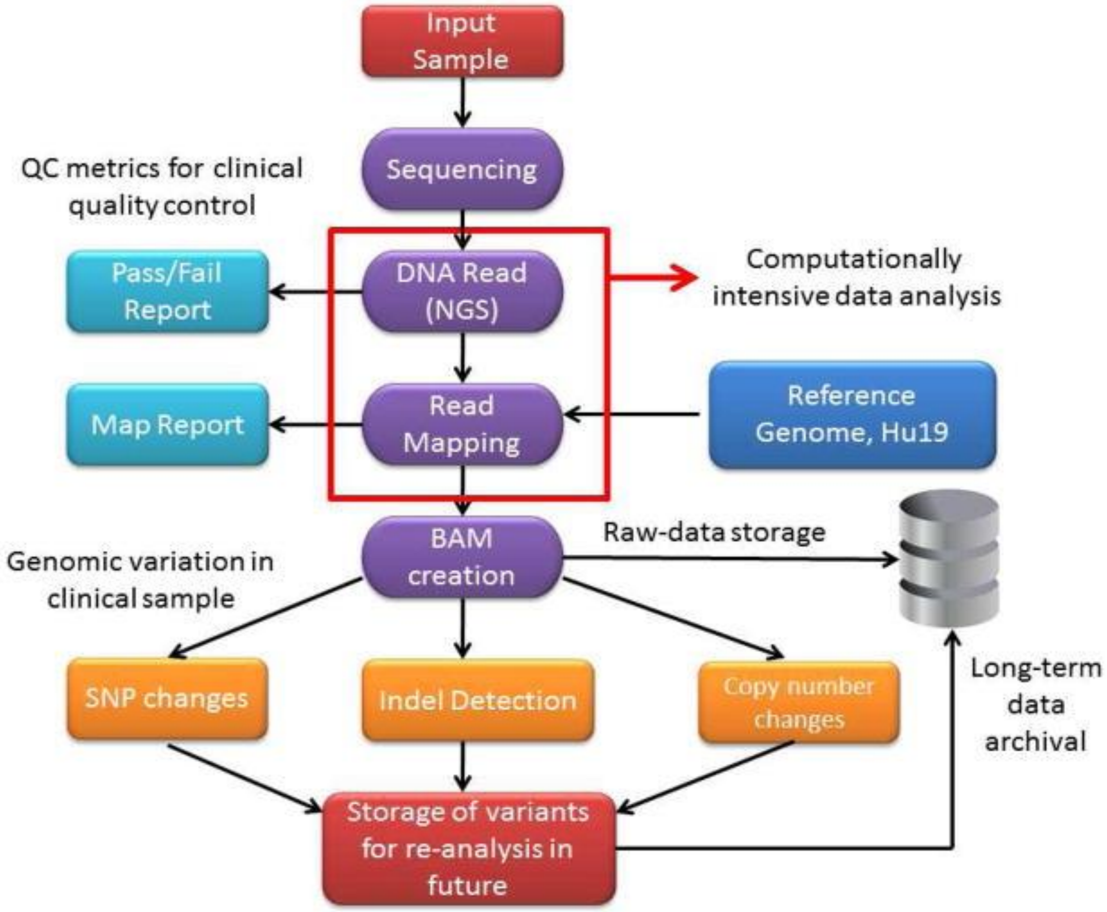


Figure 1.8. Typical workflow in a clinical next generation sequencing laboratory (Gullapalli *et al.*, 2012)

1.4.4.3 Whole Genome Sequencing (WGS)

Whole-genome sequencing (WGS) provided genome-wide, unbiased screen of all genes including the coding and non-coding regions of the genome (Ross & Cronin, 2011). WGS facilitates a comprehensive analysis of all types of genomic feature; single nucleotide changes and structural variants, including deletions, amplifications, gene fusions and loss of heterozygosity, are all readily identified. Studying the entire cancer genome permitted unprecedented global mutation detection and an understanding of cooperation between genes. Research studies of cancer, typically sequenced both the tumour and paired normal tissue from the same individual, to help distinguish acquired (somatic) sequence variants from the background of inherited polymorphisms, which was important for early cancer gene discovery studies. It is clear that this is a powerful yet straightforward technology that will have great impact on cancer diagnosis but until recently, the cost has been prohibitive for large scale use and difficult to match the requirements for diagnostic turnaround (Welch & Link, 2011).

1.4.4.4 Targeted Sequencing

Whilst WGS is undoubtedly powerful for novel gene discovery and for the investigation of the non-coding genome, the additional breadth of coverage yields limited returns relative to expenditure for diagnostic purposes. WGS presents problems for data handling and storage which can be reduced if only the genes of interest are examined, then the bioinformatics burden is reduced and the excess data is available for possible future diagnostic and research use. Alternatively, using target enrichment techniques, regions of the genome of interest can be selected prior to sequencing to streamline sequencing and simplify data processing and analysis (Mamanova *et al.*, 2010; Mertes *et al.*, 2011), making this more relevant for diagnostic use. The enrichment of specific targets is currently mandatory for clinical cancer genome sequencing. The trade-off from the reduced breadth of coverage by selective sequencing is the cost effective use of instrument capacity to accomplish better depth of sequencing. This is necessary for the detection of cancer clones and to optimise the specificity and sensitivity of assays, to derive information at sufficient depth for effective clinical utility..

The regions of interest could be a large-scale, standard target, such as the coding regions of the genome (exome sequencing) (Ng *et al.*, 2009), cDNA from the transcriptome (Fullwood *et al.*, 2009; Ruan *et al.*, 2007; Maher *et al.*, 2009) or a specific subset of genes for their cancer diagnostic relevance (e.g. cancer genes or tyrosine kinase genes, the ‘kinome’) (Loriaux *et al.*, 2008). A standardised target of relevance to diagnostics may eventually be the recognised cancer genome (Fox *et al.*, 2009; Stratton *et al.*, 2009). As well as focussing experiments to the most relevant genomic regions, target selection can overcome coverage limitations of certain whole exome approaches (Rehm, 2013).

Targeted enrichment methods fall broadly into two categories; amplicon capture, using PCR or Molecular Inversion Probes (MIP), and hybrid capture, which uses a solid surface array or in-solution capture (Mamanova *et al.*, 2010; Hagemann *et al.*, 2013). Both techniques however, may struggle to capture regions of low complexity, leading to difficulties mapping sequences back to the reference genome. Amplicon-based selection has proved successful for selection of mutational hotspots in AML (Kohlmann *et al.*, 2010b; Patel *et al.*, 2012) whilst hybrid capture has been shown to be feasible for the detection of base pair mutations and structural abnormalities for leukaemia diagnosis (Grossmann *et al.*, 2011; Duncavage *et al.*, 2012).

Target enrichment strategies for NGS have been shown to be reproducible and reliable and are becoming widely used in clinical diagnostic laboratories (Mamanova *et al.*, 2010; Meyerson *et al.*, 2010; Hagemann *et al.*, 2013). An assortment of methods and technologies have been described, most of which can now be purchased as commercial products (e.g. Fluidigm, Raindance, Haloplex amplicon capture and SureSelect, NimbleGen for hybrid capture). The fidelity of capture can vary with different methods and it is important to select an appropriate method to minimise off-target enrichment and low uniformity of capture and therefore maximise the efficiency of sequencing required to attain adequate sequence depth for all targeted regions. Different capture methods can be affected by sample quality and the presence of variants within the capture region. Scalability, throughput and ease of use are important for diagnostic practicality, whilst the size of the region for enrichment will influence the most appropriate method. Finally, the need for specialised equipment and the reagent price are also key considerations.

1.4.4.5 Next Generation Sequencing for the identification of Structural Variation

Structural variation (SV) of the genome is defined as differences in location, orientation or copy number of relatively large genomic segments, typically represented by translocations, inversions, tandem duplications, insertions, deletions, and segmental loss of heterozygosity. SV has been described as genomic variation of at least 1kb of DNA in size, however, submicroscopic rearrangements revealed by modern sequencing technology has redefined SV to >50bp, to distinguish this from the smaller indels and single nucleotide variation (Feuk *et al.*, 2006; Alkan *et al.*, 2011; Quinlan & Hall, 2012). SV contributes to the phenotypic differences among healthy individuals and is implicated in the causation of diseases, including cancer, by disrupting gene function, creating gene fusions or placing genes alongside different controlling elements. Therefore, it is crucial to systematically profile SV in the genome. The reliable detection of gene fusions resulting from chromosomal translocations, and other forms of structural variation, are critical for the accurate classification of AML and other cancer subtypes (Arber *et al.*, 2016; Döhner *et al.*, 2016).

An early experimental approach used 'shotgun' sequencing of flow-sorted derivative chromosomes followed by NGS (Illumina/Solexa) and with sufficient depth of coverage defined three different disease-associated breakpoint cluster regions, in three constitutional translocations, in patients with developmental delay. The coverage was attained by bridging the breakpoints by PCR amplification, and this procedure allowed for the determination of their exact nucleotide positions (Chen *et al.*, 2008).

1.4.4.6 Paired end technology for structural variation

A number of approaches are possible for the detection of structural abnormalities as well as single nucleotide changes. Paired end and mate pair resequencing protocols were developed to increase the effective read length of sequencing with the main incentive being the identification of SV by NGS (Raphael, 2003; Ng *et al.*, 2006; Korbel *et al.*, 2007; Fullwood *et al.*, 2009). The basic principle of paired-end mapping is to examine a short sequence read from each end of larger, linear genomic nucleic acid fragment or 'insert'. Most reads result in concordant pairs when aligned to the reference genome, where the distance between them and their orientation is as expected for the insert fragment. In contrast, mismatched paired reads indicate disruption of the spatial relationship between alignments, such as abnormal distance, different orientation or location on different chromosomes. This indicates that read pairs flank a structural breakpoint within the DNA fragment. Different classes of structural variation, such as insertions, deletions, inversions and translocations, produce a distinct mapping signature (Raphael, 2012).

A variety of methods have been used to generate paired reads using various next-generation sequencing technologies, although paired-end mapping using modern short read NGS platforms is most widely adopted and offers the most efficient technique for diagnostic applications. Sequencing libraries are constructed to a uniform length (generally 200~500bp) for paired-end libraries (compared to 1~10kb for mate-pair libraries) (Raphael, 2012). Routine NGS, therefore, has both limitations in read length and insert size. Paired-end analysis provides quantitative, digital information and the ability to sample rare events by sequencing DNA to an appropriate depth. In contrast to a complete sequencing strategy, it is possible to detect SNV, small indels and structural variation with far fewer sequence reads compared to continuous sequencing of linear DNA. However, without continuous sequencing of DNA fragments, the discordant mapping patterns from complex SV can be difficult to interpret and the chosen fragment sizes can affect the sensitivity (Quinlan & Hall, 2012). Some complex SV may remain refractory to detection or accurate characterisation. There is a significant benefit of defining specific genomic breakpoints of genomic rearrangements such as gene fusions, for the identification of disease-specific junctional fragments for the diagnosis and to provide useful molecular markers for monitoring disease course. A paired end sequencing strategy does not necessarily define the breakpoints to bp resolution.

Advanced bioinformatics algorithms are necessary for SV detection and have been developed to predict SV by finding clusters of discordant pairs, which are supported by multiple discordant paired reads, split reads, or both (Raphael, 2012; Abel & Duncavage, 2013; Liu *et al.*, 2015; Guan & Sung, 2016). Discordant reads may indicate the presence of a sequencing error, however, the recognition of clusters of discordant pairs that indicate the same true variant can discount this (Raphael, 2012). Different algorithms consider different features of structural variation and specificity is improved by computational methods which incorporate multiple signals of structural variation, such as read depth, split reads, or paired reads, into a single prediction algorithm (Sindi *et al.*, 2012). That multiple classifiers have been produced suggests that the ideal software has not been produced although there is increasing sophistication with new generation of algorithms, which upon extensive evaluation may be shown to be more accurate (Abel & Duncavage, 2013; Liu *et al.*, 2015; Guan & Sung, 2016).

1.4.4.7 NGS studies in cancer with a focus on detection of gene fusions

A number of research groups pioneered paired-end NGS for gene expression (transcriptome) sequencing (Maher *et al.*, 2009), genomic sequencing (Campbell *et al.*, 2008) or both (Fullwood *et al.*, 2009; Ruan *et al.*, 2007). Early studies using paired-end mapping studies revealed extensive high resolution maps of structural variation in the human genome (Korbel *et al.*, 2007; Kidd *et al.*, 2008) and refined the technology for the identification of somatically acquired rearrangements in cancer genomes (Raphael, 2003; Volik *et al.*, 2003; Ng *et al.*, 2006; Bignell *et al.*, 2007; Bashir *et al.*, 2008).

The first WGS of a cancer was an AML of M1 morphology, which used deep, single-end NGS (Ley *et al.*, 2008). This patient was later sequenced more comprehensively, by paired-end WGS (Mardis *et al.*, 2009). This was rapidly followed by sequencing other AML genomes (Ley *et al.*, 2010; Welch *et al.*, 2011b) and of other tumour types (reviewed by Meyerson *et al.*, 2010)**.** Of particular note were landmark studies of lung tumours (Campbell *et al.*, 2008), breast tumours (Stephens *et al.*, 2009), melanoma (Pleasance *et al.*, 2010b), and small-cell lung tumours (Pleasance *et al.*, 2010a). In particular, the study by Campbell *et al.* (2008) presented an early, technically sophisticated study in which they used genome-wide, paired-end NGS to detect somatically acquired rearrangements in two individuals with lung cancer, by comparing tumour cell line and germline DNA to the reference human genome. They characterised multiple somatic rearrangements to a base-pair resolution, including internal tandem duplications and gene fusions from interchromosomal rearrangements, some of which were also demonstrated as expressed abnormal transcripts (Heim & Mitelman, 2008). Paired-end NGS was used in an early study and demonstrated the feasibility of gene fusion detection by WGS of an acute promyelocytic leukaemia sample to show a cryptic, insertional *PML-RARA* fusion, within a diagnostically relevant timescale (seven weeks including PCR confirmation) (Welch *et al.*, 2011a). The study used paired end NGS analysed with Breakdancer software (Chen *et al.*, 2009).

Cost, analysis of complex data and storage of data are issues for routine application of WGS. If only genes of interest are examined following WGS, this reduces the bioinformatics burden and the excess data is available for possible future diagnostic and research use. However, for diagnostic purposes, clinicians require information on a defined range of genetic abnormalities with confirmed clinical significance and therefore only require testing of a specific panel of genes, making this more affordable and comprehensible (Godley, 2012). More recent studies have been able to capitalise on the streamlining of NGS methodology and the capacity of NGS technology to sequence multiple AML genomes, to provide detailed analysis of the mutational landscape of the diseases with the result that they have been extensively characterised and the common genetic defects leading to the development of AML are well understood (Patel *et al.*, 2012; Grossmann *et al.*, 2012; The Cancer Genome Atlas Research Network, 2013; Papaemmanuil *et al.*, 2016; Metzeler *et al.*, 2016) (see section 1.3, below).

1.4.4.8 The combination of paired end NGS and target enrichment

Hybrid capture has been shown to be feasible for the detection of base pair mutations and structural abnormalities for leukaemia diagnosis (Grossmann *et al.*, 2011; Duncavage *et al.*, 2012). An experimental microarray hybrid capture approach identified gene fusions involving *RUNX1,* including a novel fusion partner (Grossmann *et al.*, 2011). An early proof of principal study used hybrid capture (Agilent SureSelect target enrichment) of 20 genes and identified gene fusions in three leukaemia cell lines, as well as a *KMT2A-AF9* gene fusion and *FLT3*-ITD in a diagnostic AML specimen (Duncavage *et al.*, 2012). The research group then studied sensitivity and specificity of NGS for the detection of gene fusions in 7 *ALK* and 6 *KMT2A* rearranged tumours and validated using multiple negative cases (Abel *et al.*, 2014). The same laboratories performed similar studies on FFPE lung adenocarcinoma specimens (Spencer *et al.*, 2014) and FFPE from seven *ALK* rearranged tumours (lung adenocarcinoma and *ALK* lymphoma) and six bone marrow aspirates with known *KMT2A* gene fusions (Abel *et al.*, 2014). A recent international, multicentre collaboration developed Karyogene, a targeted resequencing and analytical platform that used gDNA in a single assay, similar to the current study, to detect nucleotide substitutions, insertions/deletions, chromosomal translocations (McKerrell *et al.*, 2016). The key to successful identification is the adoption of appropriate bioinformatics approaches for the detection of the variety of mutational types (Abel & Duncavage, 2013).

1.4.4.9 Transcriptome analysis for gene fusions

Analysis of the transcriptome by sequencing of cDNA from expressed transcripts from disease cells by RNA sequencing (RNA-seq) is gradually superseding microarray technology for gene expression studies (Wang *et al.*, 2009) and has been applied to mutation detection (Shah *et al.*, 2009). Transcriptome analysis by RNA-seq provides superior digital information to microarray GEP whilst also enabling discovery of novel splice forms, transcripts and RNA-editing. The better dynamic range of RNA-seq improves detection of low expressed transcripts (Meldrum *et al.*, 2011; Kukurba & Montgomery, 2015; Ozsolak & Milos, 2011).

RNA-seq has proven especially useful for the detection of gene fusions resulting from genomic translocations. The conventional method for identifying fusion junctions in single gene targets is reverse transcriptase PCR from expressed mRNA from disease cells. By sequencing mRNA rather than gDNA, it is possible to exploit the mRNA splicing mechanisms of the cell that join exons as single transcripts. Therefore, the intronic, non-coding sequences have been excised and sequencing can target the expressed genes only and avoid excessive sequencing of repetitive DNA from introns. An early study of RNA-seq using a novel paired-end sequencing method identified cancer-associated fusion transcripts and a novel fusion transcript, *BCAS4-BCAS3* from an unbalanced t(17;20)(q23;q13), in a breast tumour cell line (Ruan *et al.*, 2007). Whole transcriptome paired-end sequencing has revealed novel fusion transcripts of clinical significance in AML (Wen *et al.*, 2012; Masetti *et al.*, 2013) Special consideration much be given to the use of transcriptome analysis to detect structural rearrangements as a global analysis of all genes.

An advantage of RNA-seq for identifying variants compared to gDNA sequencing is that the data simultaneously provides functional information about the expressivity of genes of clinical significance, and is the most relevant application if gene expression profiling is also of interest. However, variable gene expression results in a large dynamic range of expression throughout the genome. This is not uniform across all genes and affects the genomic coverage of sequence information and the ability to accurately call variants. The dynamic range of gene expression extends over many orders of magnitude and identifying coding mutations in genes with moderate to high expression is not a problem (Mortazavi *et al.*, 2008; Conesa *et al.*, 2016). The large number of transcript copies that can be derived from a single cell, may suggest RNA-based methods could have greater sensitivity over those interrogating DNA. However, the accurate identification of somatic mutations in genes with low levels of expression or that harbour truncating somatic mutations and cause nonsense-mediated decay presents a challenge to detection. Detection of low level sequences is again dependent on read depth, and the sensitivity can be enhanced by targeted hybridisation capture of cDNA libraries, similar to that used for DNA sequencing (Levin *et al.*, 2009).

RNA-seq has proven capability to detect sequence variants and fusion products. RNA-seq should therefore reflect the sequence data garnered from WES. However, the transcriptome is an imperfect surrogate for the exome and any advantages of RNA-seq in preference to DNA for mutation detection have not been demonstrated (Meldrum *et al.*, 2011). Furthermore, mRNA of adequate quality to prepare sequencing libraries may not always be available from fresh samples or from paraffin-embedded solid tissue specimens, which is much less likely to be a problem for gDNA (Shendure & Stewart, 2009).

A significant advantage of RNA-seq is the ability to identify fusion transcripts from chimaeric fusion genes, including novel fusions (Levin *et al.*, 2009; Shah *et al.*, 2009) although not the sequence disruption from fusion genes causing upregulation without expressed chimaeric mRNA. An emerging trend for diagnostic applications, is the combination of exome and transcriptome sequencing (Hansen *et al.*, 2016), to derive the benefits of WES from gDNA and gene fusion detection from RNA-seq, thereby restricting their deficiencies but by increasing the sequencing demand and complexity of the assay.

1.5 Genomics strategy in the UK

The intense activity in the field of genomics in the last decade due to the rapid advances in genome analysis technologies has greatly improved the understanding of the molecular pathogenesis of cancer. This knowledge, in turn, will provide information for the more accurate characterisation, diagnosis, classification and monitoring of neoplastic disease of individual patients (PHG Foundation, 2011). Cancer diagnosis will benefit from the mainstream adoption of genomic testing but a significant challenge exists of how pressurised healthcare systems can establish an infrastructure to constructively integrate molecular diagnostics into the routine clinical practice, to achieve the anticipated improvement of patients' clinical outcomes.

1.5.1 The need to modernise cancer molecular pathology

Molecular pathology for cancer diagnosis has been recognised in a number of reports which represented a groundswell of opinion of the benefits for cancer diagnosis and shaped national strategies. The Royal College of Pathologists (2010) recommended a national body to approve new molecular diagnostic tests and called for coordination of testing in regional centres. The financial issues were highlighted and the report expressed the necessity to make funding for molecular diagnostics explicit in local budget planning, in Trusts' Strategies and Local Development Plans. The Inter-Specialty Committee for Molecular Pathology (ISCMP) was established to drive this development, with special consideration given to training the scientific and medical workforce in the new discipline. It was acknowledged that the utility of molecular diagnostics in haemato-oncology is better established than for solid tumours and genetic testing by karyotyping and FISH, supplemented by real-time reverse transcriptase (RT-PCR) for monitoring of minimal residual disease, has been incorporated into NICE guidance since 2003 and service organisation assessed as a cancer standard (National Institute for Health and Care Excellence, 2016). This is not to say that equitable access of testing across the country has been achieved or that significant challenges do not await the transition to the testing of panels of multiple gene targets in haematological malignancy.

The provision of companion diagnostics for medicines targeted to genetic mutations requires significant resources to make this a routine practice. It was suggested in the national Cancer Strategy (pg 38-41 Independent Cancer Taskforce, 2015) that centralised commissioning would facilitate the coordination of cancer gene testing, as a means to avoid the piecemeal adoption of testing and access to corresponding therapies. The Strategy also recognises that, to realise the full potential of novel targeted treatments, tumour diagnosis would not just require single molecular tests but multiplex panels of genetic markers. Importantly, it is also recognised that policy definitions should be sufficiently flexible and funding restrictions should not be an impediment to the transition to multi-gene panel-based tests at the appropriate time and when cost effective (Cancer Research UK, 2015). The slow progress to establish centralised commissioning for single gene companion diagnostics demonstrates the challenge for this next phase. The incorporation of routine molecular testing into the patient pathway also requires modernisation of the diagnostic infrastructure. Adoption of new workflows is necessary to provide tumour specimens of desired quality and logistical solutions to create accessibility to testing centres with the modern technology. Following from the roadmap developed by the Stratified Medicine Innovation Platform consortium of government bodies and leading charities (Technology Strategy Board, 2011), Cancer Research UK set up the Stratified Medicine Programme. This was a prototype of a coordinated, nationwide system for large scale molecular testing in cancer, providing patient sample access to testing centres, whilst promoting research into targeted therapies and creating a centralised data repository of paired genetic and clinical data (ECMC website accessed 17/11/2016). Programme 1 ran until 2013 and laid a foundation and established pathways for routine molecular genetic testing of specific mutations, in a defined set of cancer types. Programme 2 followed and is pioneering the use of NGS for pre-screening stage 3 and 4 lung cancer patients for the National Lung Matrix Trial. This multiplex assay is being developed in conjunction with the commercial sequencing company Illumina, which, with other sequencing companies, is developing products for use in routine diagnostics for the NHS as part of a new market strategy. In 2013, the Medical Research Council (MRC) also announced their ambitions for diagnostic molecular pathology in the UK and how this might be achieved (Medical Research Council, 2013). This report also highlighted the need for investment in molecular diagnostic services to increase the capacity, in order to realise the potential clinical and economic benefits of stratified medicine. The review identified the lack of a recognised developmental pathway for diagnostic testing, including shortcomings in processes for regulation, evaluation and commissioning. The divide between different stakeholding groups and sectors is a recurring theme in reports; in particular, Pathology, Research, Industry and the clinical environment are disconnected. The capability for the successful translation of innovative genomic laboratory research to clinic would be enhanced from collaboration (Human Genomics Strategy Group, 2012). In support of this, the MRC and *Engineering and Physical Sciences Research Council* (*EPSRC*) have funded six university-led Molecular Pathology 'nodes', covering various diseases including cancer, for research to collaborate with clinicians to introduce new molecular diagnostics, to determine clinical validity and utility, and embed this into quality diagnostic provision, to enable disease stratification. The Open Targets initiative was developed as a partnership between academia (Wellcome Trust Sanger Institute and the European Bioinformatics Institute) and pharma (GSK and Biogen) to use NGS to identify candidate mutations and determine their biological validity as therapeutic targets (Open Targets, 2017).

1.5.2 The 100,000 Genomes Project

The vehicle to deliver the UK governmental strategy has emerged as the 100,000 Genomes Project (100KGP), which was announced in December 2012. Genomics England, a wholly owned company of the Department of Health, was set up in partnership with the sequencing industry to deliver the Project. The 100,000 genomes to be sequenced is planned to include 50,000 genomes from around 25,000 cancer patients and combining this data with their medical records. An important resource for clinical research will be established, which, it is hoped, will help inform the precise diagnosis of cancer and personalisation of medicine. It is debatable, however, whether so many genome sequences are within budget and when normal comparative genomes, multiple tissue samples from the same patient, different family members and a diverse range of diseases are tested (including rare diseases, multiple subtypes of cancer and infectious diseases) the repository of information on each disease group will be limited. The Project has made an encouraging progress with 31,730 whole genomes sequenced to date (as of 1st July 2017; source Genomics England, 2017 website) with the full target to be achieved by the end of 2017. It is evident, however, that scientific data generation is not the key aim of the project. The main objective is for genomic medicine to transform the way all healthcare is delivered and particularly, to use this new paradigm as justification to modernise outmoded pathology systems.

1.5.3 Genomic Services Reconfiguration in England

The network of GMCs were thought to provide the template for a new network of UK regional genetic centres, although this was not explicit in the early communications regarding Genetics Services Reconfiguration (NHS England, 2015). The project will re-procure NHS genetic testing in November 2017, to restructure laboratories into a national network of NHS regional genomics centres in seven geographical pods. The new services will be aligned with the Academic Health Science Networks, established genomics research infrastructure, including medical schools, universities and research institutions such as the Wellcome Trust Sanger Institute and the MRC Institute of Genetics and Molecular Medicine, whilst also linking genomic services to local pathology provision at the clinical interface. A hierarchy will be defined, therefore, from the GESC, a maximum of seven Genomics Central Laboratory Hubs (GCLH) and subcontracted designated local hubs, from which local molecular pathology services will be provided. The proposed infrastructure will bring together clinicians, diagnostic laboratories and research units into ‘clinical interpretation partnerships,’ to review new information and “ensure the legacy of the 100KGP,” uniting different disciplines and accelerating translational research into clinical practice. Reconfiguration has been preoccupying genetic laboratories for more than two years whilst in development and the details of the Invitation to Tender are awaited with interest in 2017.

1.5.4 Challenges for the new testing paradigm

Whilst the use of NGS is accelerating and is undoubtedly the direction for genetic testing, the technology is currently used in only a small proportion of patients, mainly for the diagnosis of Mendelian conditions and as companion diagnostics for few cancer treatments.

A potential problem with strategies for accelerating genomics into healthcare is the assumption that WGS is ready to transfer for all applications. Whilst comprehensive sequencing is beneficial for genetic discovery studies and may prove to be cost effective for medical genetics, where multiple gene profiling will have long-term future application for the patient and their families, this is not the case for the molecular pathology of cancer, where rapid diagnosis and prompt assignment of therapy is a key requirement. For the near future, cancer genetics services must accommodate conventional testing by simpler, cost effective methods, such as cytogenetics, FISH and single gene PCR and small gene panels. Sequencing technology is progressing in several directions, and varying applications of NGS will be required by different pathology specialties (Cree, 2016). Novel applications of NGS specific to cancer, such as 'liquid biopsy' (using circulating cells, nucleic acids and proteins for diagnosis), and NGS to monitor the efficacy of treatment is gaining evidence to support wider adoption and requires special consideration.

A potential bottleneck is that many diagnostic tests are ready for validation as pharmacogenomic applications but that entry into practice is prolonged due to difficulties identifying robust drug targets in complex cancer genomes and the length of time required for drug development. A changing focus will be to provide screening of the cancer genome, rather than sequential single tests, to provide downstream access to the small number of approved therapies and also early phase trails of experimental medicines.

The main obstacle to mainstream adoption of sequencing is no longer technical; it is in the interpretation of large amounts of complex data. There will be a growing requirement to assess complex signatures, possibly from multiple molecular modalities provided by genomic, transcriptomic and proteomic platforms, which collectively will characterise the individual make-up of a cancer. This is probably most advanced in AML (The Cancer Genome Atlas Research Network, 2013; Papaemmanuil *et al.*, 2016). The evidence base is accumulating to give scientists the resources to assess the clinical utility of sequence variants, such as the large, internationally coordinated research programmes, The Cancer Genome Atlas (Tomczak *et al.*, 2015) and the International Cancer Genome Consortium (ICGC) (Hudson *et al.*, 2010), which led to the identification of recurrent somatic mutations that drive oncogenesis in selected cancer types, through the comparative sequencing of tumour tissue and the constitutional genome. The Catalogue Of Somatic Mutations In Cancer (COSMIC) is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer (Forbes *et al.*, 2015) and also provides the Cancer Gene Census, a continuously updated comprehensive catalogue of curated genes that are causally implicated in cancer.

The evolution of the genomics sector brings with it the requirement to have scientific and technical staff at the frontline, with the necessary skills and expertise to deliver the transformation. The necessity to support a modern workforce is recognised in the 5 Year Forward View (NHS England, 2014). The importance of genomics is recognised and the need to invest in relevant training of the current and future workforce is supported by the Health Education England (HEE) long term strategy (Health Education England, 2015) to support the successful integration of genomics into mainstream care pathway.

The shortage of genomic scientific and bioinformatic skills is recognised (Monitor Deloitte, 2015). Bioinformatics expertise is a key constraint for development of the discipline and is desperately needed to underpin genomic and genetic testing. To meet this challenge for data analysis and interpretation, new education programmes have been commissioned and are under development including: new scientist training programmes (STP) in molecular pathology, genomics and bioinformatics; higher specialist scientist training (HSST) posts in genetics and molecular pathology with an approved curriculum in molecular pathology leading to Fellowship of the RCPath, in order to develop the future consultant clinical scientist workforce, and the introduction of a modular MSc in genomic medicine (Delon & Scott, 2016). There will be additional opportunities for translational research stemming from the 100,000 Genomes Project research fellowships for PhD students and post-doctoral scientists, thus further enhancing the skills base and developing future research leaders in pathology and increasing capacity in data analysis (Medical Research Council, 2013). The UK Government has agreed to support genomics education and the workforce transformation project with an allocation of £25 million (Delon & Scott, 2016).

Whilst some traditional assays will be transferrable immediately to new platforms, the utility and cost efficiency of some new assays may take years to emerge with the persistence of traditional testing methods. One technology will not dominate for many years, and, most likely, a hybrid approach to address practical, clinical, and financial pressures will prevail (Schnepp *et al.*, 2015). The integration of genomics in cancer diagnosis practice will take time and will require a phased adaptation of the technology, gradually transferring single gene tests to panels, validating multi-gene profiles by NGS and introducing WGS when interpretation is deemed feasible and cost efficient.

Changes in practice need to be closely aligned with a health economic assessment of their impact on patient pathways. Cancer care is costly: genomic testing can be used in decision making to rule out chemotherapy or other treatment options that would not be effective for the care of an individual patient. The benefits of increasing personalisation of medicine will be to use genomic testing to help eliminate ineffective or possible harmful treatment options and determine appropriate care will benefit the patient while reducing healthcare utilisation and costs. The effect of personalised medicine on NHS costs is still a matter of debate, but there is no question that for advanced cancer, most individualised treatment will add to costs and the prolongation of life will add to the overall cost of healthcare. Such developments therefore need to be introduced in concert with strategies to improve early diagnosis of cancer and other disorders, when the chance of cure is high and less expensive (Cree, 2016). Deep uncertainty in a time of unpredictability and uncertainty at a time when the NHS is suffering its "worst financial crisis in history" and according to the Forward View the NHS will need to make efficiency savings of £22bn by 2020/21 to live within the planned budget (Royal College of Physicians, 2016; Murray *et al.*, 2014; Lafond *et al.*, 2016).

Genome sequencing also requires a 'big data' approach to analysis and interpretation which is already apparent in many research facilities; presenting a challenge to storage capacity and transfer of this data across networks. Computer facilities will require to be upgraded, coordinated and delivered effectively, which the NHS is not renowned for. That this will be patient identifiable data through linked demographics and personal DNA sequence data, including whole genome sequences for use with outputs from central facilities, local pathology labs or even point of care diagnostic devices, data security will be a significant issue. The project will test the serious practical challenges of integrating and safeguarding genomic data within an expansive but under-resourced health service and this requires the further development of the medico-legal and ethical framework for the use of genomic data.

Finally, it may be necessary to exercise caution and reframe our expectations and acknowledge that translational rather than transformational change is required; this is not so much a “revolution as an evolution”.

1.6 Project aims and what it adds to scientific field

The project was a Proof of Principle (POP) study to demonstrate the feasibility of performing targeted NGS to detect a range of types of genetic abnormalities in AML, for the detection of all known, clinically relevant mutations in a single experiment. This examined the practicality of merging multiple existing diagnostic tests and newly discovered mutation profiles into a single assay, thereby being able to replace existing tests and unifying diverse workflows. In particular, a number of studies have demonstrated the utility of exome sequencing for the detection of single nucleotide variants (SNVs) and small indels, which have been transferred to the diagnostic laboratory environment. WGS can reveal large indels and gene fusions but the combination of target selection and the analytical pipelines required to detect large scale structural variation is underdeveloped. The project entailed the design of a novel NGS assay and used a custom-designed selection of targets for sequencing genomic DNA, including for the detection of gene fusions and other structural variants, which represented a unique aspect of this project in AML. The project compared the performance of the new assay for detection of genomic abnormalities with standard methods of cytogenetics, FISH and molecular testing and with an orthogonal NGS assay using a commercial kit for SNV detection, based on different target capture technology and sequencing chemistry.

The main aims of the project were;

1. **To demonstrate the principle that a novel genome sequencing assay for the diagnosis of AML was feasible combining key features;**
    - Use **target enrichment** for genes of clinical relevance, to optimise use of capacity of currently available sequencers and facilitate the management of costs.
    - Use **genomic DNA** as the sole test material, to streamline the assay and provide a stable nucleic acid as the analyte, facilitating universal accessibility and reproducible performance. Investigate the potential for using gDNA for the detection of different

variant types, including large duplications and gene fusions, which was considered technically demanding.

- Perform NGS **without use of control,** to reflect the typical diagnostic workflow. A sample of normal cells would not be readily available from an AML patient at diagnosis and therefore, there would be no option to use constitutional germline DNA as normal control, to distinguish somatic mutations from germline variants.
- **Target genomic variation of different types** to detect recurrent genetic changes in AML that are clinically relevant and abnormalities that are likely to be actionable in the near future. In particular, **detect structural variation,** including gene fusions from translocations and inversions, to reproduce the output of different conventional technologies, including cytogenetics.

2. **Evaluate the feasibility of translation of the NGS assay to a routine diagnostic genetics laboratory.**
   - Examine the principle that this could be transferrable to the diagnostic laboratory.
   - Reproduce the diagnostic and prognostic findings from existing methods, to accurately_detect the types of genetic abnormalities encountered in conventional AML diagnosis.
   - Consider how the stringent principles required in a diagnostic laboratory and rigorous accreditation standards can be met.
   - Demonstrate how the conventional workflow can be improved and, by extending the range of conventional testing to the highest, single nucleotide resolution, it should be possible to increase the sensitivity of the tests by sequencing to an appropriate depth.

A key objective for the project was to not only take into account the technical aspects of the test but to understand the challenges of the technique and the practicality for mainstream adoption of the technology in a specialised pathology laboratory, such as the one at The Christie. This will provide evidence of incorporating NGS in routine practice and the problems of adopting complex technology into routine diagnostic environments and developing the infrastructure for NGS in mainstream healthcare systems, with particular emphasis on challenges presented by bioinformatics, data generation and storage. Commentary will be provided on the prospects of transformational change to the diagnostic laboratory infrastructure and how this can be managed.

3. Enhancement of the Patient pathway

- To evaluate the potential of the new targeted genome sequencing assay in a clinical context and how this may improve AML management.
- In particular, how the successful genomic screen of a wider range of genetic targets, not part of the standard workflow, could facilitate reclassification of patients by their mutational and cytogenetic profile. Demonstrate how the incorporation of new markers and combination thereof may lead to better disease classification; improve prognostication to emerging standards; provide targets for personalisation of treatments.
- Evaluate the potential clinical utility by comparing the new NGS assay with standard testing regime, from samples from the test cohort of clinical cases.

The hypothesis is that within 35 representative cases, it should be possible to demonstrate that there is a change in diagnostic classification or prognostic stratification as a result of the additional findings. Therefore, from experimentation, it will be possible to understand the performance of the new assay in terms of technical validity, clinical validity and its potential clinical utility, prior to refinement and formal validation for use in the clinical setting, consistent with protocols for validation and professional guidance. Whilst focussing on AML, the adoption of a similar strategy applicable to all cancer types will be considered.

2.0 Methods

2.1 Sample selection and processing

Bone marrow aspirate specimens are routinely referred to the Oncology Cytogenetics department of The Christie NHS Foundation Trust for cytogenetic analysis to aid the diagnosis of AML. Cytogenetic samples are sent in a Universal container in transport medium (RPMI-1640 Culture Medium, 10% Foetal Bovine Serum (FBS), sodium heparin solution 40IU/ml, penicillin-streptomycin 50U/ml; (Gibco™, ThermoFisher Scientific, Waltham, Massachusetts, USA). Separate specimens in potassium EDTA were sent to different departments for molecular genetics studies. Following a manual cell count using a haemocytometer, a volume of sample containing 10^7 mononuclear cells (MNCs) were removed for each of two cytogenetic cultures (see below). For next generation sequencing, samples strictly surplus to diagnostic requirements were collected over a period of eight months (March 2013 to October 2013). 35 bone marrow samples and one peripheral blood sample were selected for the project. MNCs were separated by density gradient centrifugation within 24 hours, using Ficoll-Paque Plus™ (GE Healthcare UK Limited, Little Chalfont, Buckinghamshire) and cryopreserved at -80^{o}C. Once a diagnosis of AML was confirmed by morphology, immunophenotyping and/or by identification of a clonal, diagnostic cytogenetic marker, and ratified at Multi-Disciplinary Team meeting, DNA was extracted with QIAamp DNA Blood Mini Kit™ (Qiagen, Hilden, Germany) using the standard protocol. Extracted DNA was stored at -20^{o}C until sequencing. DNA quality was initially estimated using a NanoDrop Lite spectrophotometer (Thermo Scientific™, ThermoFisher Scientific, Waltham, MA, USA) and re-quantified using a Qubit® 3.0 fluorometer with Qubit® dsDNA BR Assay Kit (Invitrogen™, ThermoFisher Scientific), immediately prior to next generation sequencing.

2.2 Ethical Approval

Samples were received in the Oncology Cytogenetics department following informed consent for genetic testing. Advice from National Research Ethics Service (Ref 04/26/57) (NRES; now part of the Health Research Authority) suggested that the project was a genetic service evaluation and development, not considered to be research requiring review by an

NHS Research Ethics Committee. Management permission was obtained from Research & Development Division of the Christie NHS Foundation Trust (“R&D approval”) and the project registered with the Trust as Method Development under protocol (HT014). The Academic Ethics Committee of the Faculty of Health, Psychology and Social Care of Manchester Metropolitan University approved formal ethics application for study (Ethics Application 1186). Despite not requiring formal review by an NHS Research Ethics Committee, experiments were conducted in accordance with core ethical principles and strictly conforming to Information Governance policies of the Pathology department and the Christie NHS Foundation Trust. Informed consent was obtained for genetic testing for diagnostic purposes. The confidentiality of participants was respected and samples were anonymised, meeting requirements in processing identifiable data under the Data Protection Act 2000.

2.3 Routine genetic testing

2.3.1 Cytogenetics

Samples for conventional cytogenetic analysis were processed and analysed according to standard protocols of the Oncology Cytogenetics laboratory, which are based on universal methods (Czepulkowski, 2001) with the following amendments. Two overnight 10 ml suspension cell cultures using unselected MNCs (10^6 cells/ml) were set up in complete culture medium containing RPMI 1640 supplemented with 20% fetal bovine serum and reagents to final concentrations; L-Glutamine (2mM) and Penicillin-Streptomycin 50 U/ml (Gibco™, ThermoFisher Scientific). One of the cultures has Colcemid® (KaryoMAX™, ThermoFisher Scientific) added for the final 45 minutes of culture (0.1 µg/ml) and the other has a cocktail of 5-bromo-2-deoxyuridine (BrDU) (Sigma-Aldrich, Dorset, UK) (16mg/ml) and 1/10th concentration Colcemid® added overnight. Both cultures are treated with 75 mM hypotonic solution of potassium chloride and are fixed in 3:1 methanol/glacial acetic acid, as per protocol. Chromosome preparations are made using GTL banding method (pg 14 *ISCN 2016; An International System for Human Cytogenomic Nomenclature* 2016). A minimum of 20 cells were analysed and reported using standard nomenclature (*ISCN 2016; An International System for Human Cytogenomic Nomenclature* 2016).

2.3.2 FISH

Fluorescence *in situ* Hybridisation (FISH) was used to supplement conventional cytogenetic testing to investigate the molecular involvement of genes indicated by disease subtype (cell morphology and immunophenotype) or to confirm presence of a specific chromosomal abnormality detected by conventional cytogenetic analysis. FISH was performed on fixed cells from cytogenetic cultures following manufacturers' protocols (Cytocell, Oxford Gene Technology, Begbroke, Oxfordshire), with the exception that the pre-treatment stages were eliminated.

2.3.3 PCR for NPM1 and FLT3-ITD mutations

Routine diagnostic molecular genetic testing for *NPM1* and *FLT3* was performed at the Molecular Diagnostics Centre, Manchester Royal Infirmary using standard protocols, as follows. A number of cases were not tested at diagnosis and were tested retrospectively at the time of the project. Screening for *NPM1* gene mutations was performed using a melting curve PCR assay with primers and fluorescent probes hybridising to the mutated region, as previously reported (Schnittger *et al.*, 2005). All cases that showed an abnormal melting curve were analysed by Sanger sequencing, as described. Detection of *FLT3*-ITD mutation was performed with fragment analysis by PCR amplification of DNA with primers flanking the juxtamembrane coding sequence at exons 11 and 12 and the intervening intron. A 328bp fragment is produced from wild-type (WT) alleles and an additional band indicated *FLT3*-ITD. These cases were retested using semi-quantitative PCR and the level of mutant FLT3 expressed as a percentage of the total signal (Kottaridis *et al.*, 2001).

2.4 Next Generation Sequencing

2.4.1 Target Selection

42 target genes were selected for the project (see Section 1.3 ‘Genetic, cytogenetic and genomic influences in AML’), including all exons from the set of 30 genes known to be frequently recurrently mutated in AML for detection of SNV only (Figure 1.7), the positions of typical junctions of *FLT3* internal tandem duplications and *KMT2A* partial tandem duplications, and exons and introns covering known breakpoints in genes of recurrent gene fusions in AML, including the breakpoint cluster regions in both partner genes of *CBFB*/*MYH11*, *PML*/*RARA*, *DEK*/*NUP214*, *RUNX1*/*RUNX1T1*, two genes with multiple fusion partners, *KMT2A* and *NUP98*, as well as two common partners of *KMT2A* (*MLLT1* and *MLLT3*). All exons of RUNX1, as well as breakpoint cluster regions of the RUNX1-RUNX1T1 gene fusion were included (see Table 2.2 below). Genomic coordinates were extracted for HGP19 locations from the Genome Browser tool suite **(http://genome.ucsc.edu/cgi-bin/hgTables)** (Kent *et al.*, 2002) using Table Browser (Karolchik *et al.*, 2004), with the following criteria; Clade = Mammal, Genome = Human, Assembly = **GRCh37/hg19 (Feb. 2009), Group = Genes and Gene Prediction Tracks, track = RefSeq Genes (**NCBI RefSeq release 45) and Table = refGene (see Figure 2.1).

Under the “Region” header, using the “Define Regions” button, the list of target genes was uploaded and submitted. Filters were not applied. Output was selected in browser extensible data (BED) format. On the following screen ‘Output refGene as BED’ page, the desired features were selected as either ‘coding exons’ or ‘whole gene’, depending on whether exons only or exons and introns were required for specific genes (Figure 2.2). The exon coordinates for all possible splice variants were selected, to ensure capture of all coding regions. To limit the number of required baits and eliminate unnecessary sequencing, introns outside of the expected breakpoint regions were manually deleted from the resulting BED files. The ‘Merge BED files (mergeBed)’ function in Galaxy (Galaxy Version 2.22.1) was used to merge overlapping regions. The data was loaded as a track back into the UCSC genome browser as a visual check that the correct regions were selected.

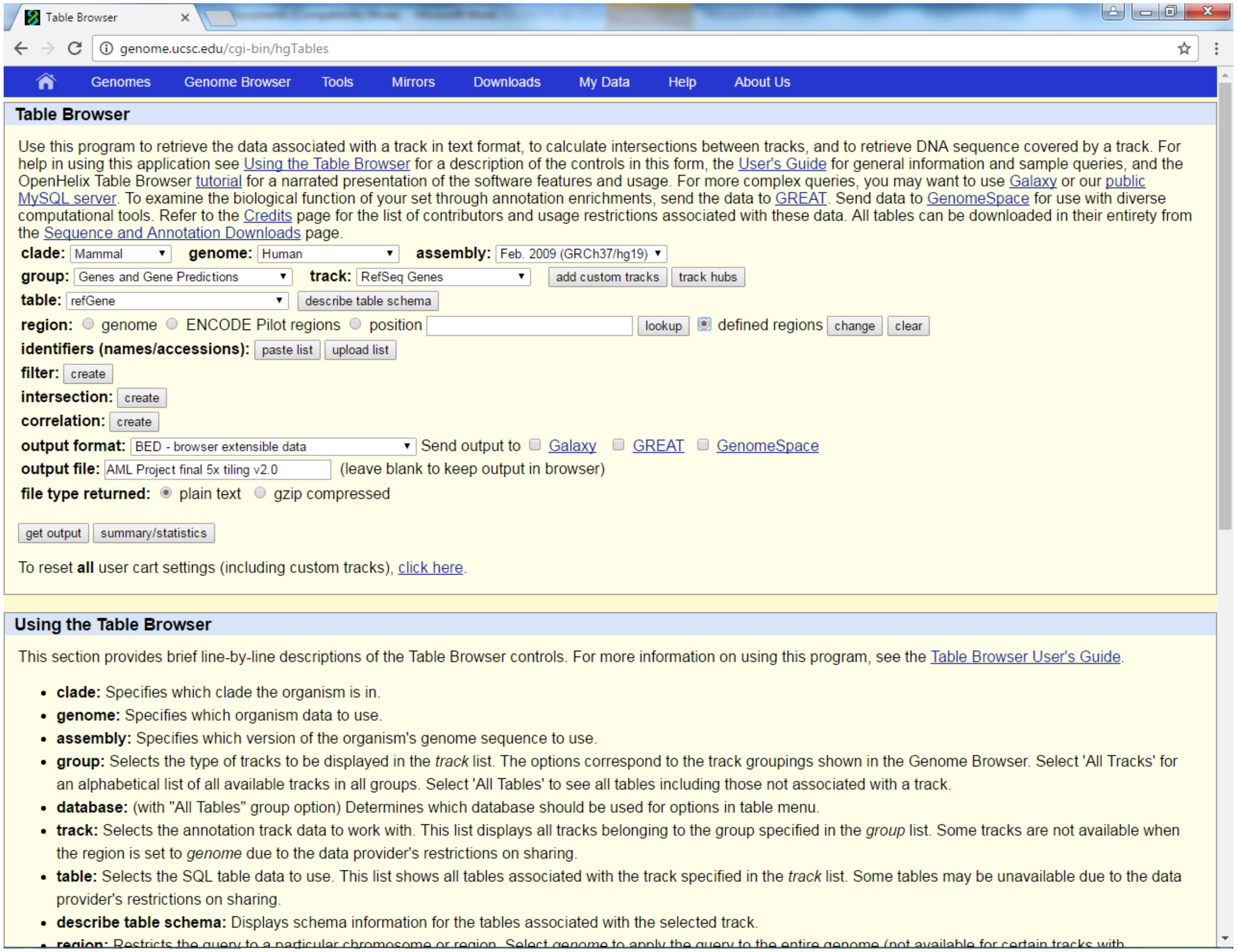


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Predictions **track:** RefSeq Genes
table: refGene
region: ○ genome ○ ENCODE Pilot regions ○ position ● defined regions
identifiers (names/accessions):
filter:
intersection:
correlation:
output format: BED - browser extensible data Send output to □ Galaxy □ GREAT □ GenomeSpace
output file: AML Project final 5x tiling v2.0 (leave blank to keep output in browser)
file type returned: ● plain text ○ gzip compressed

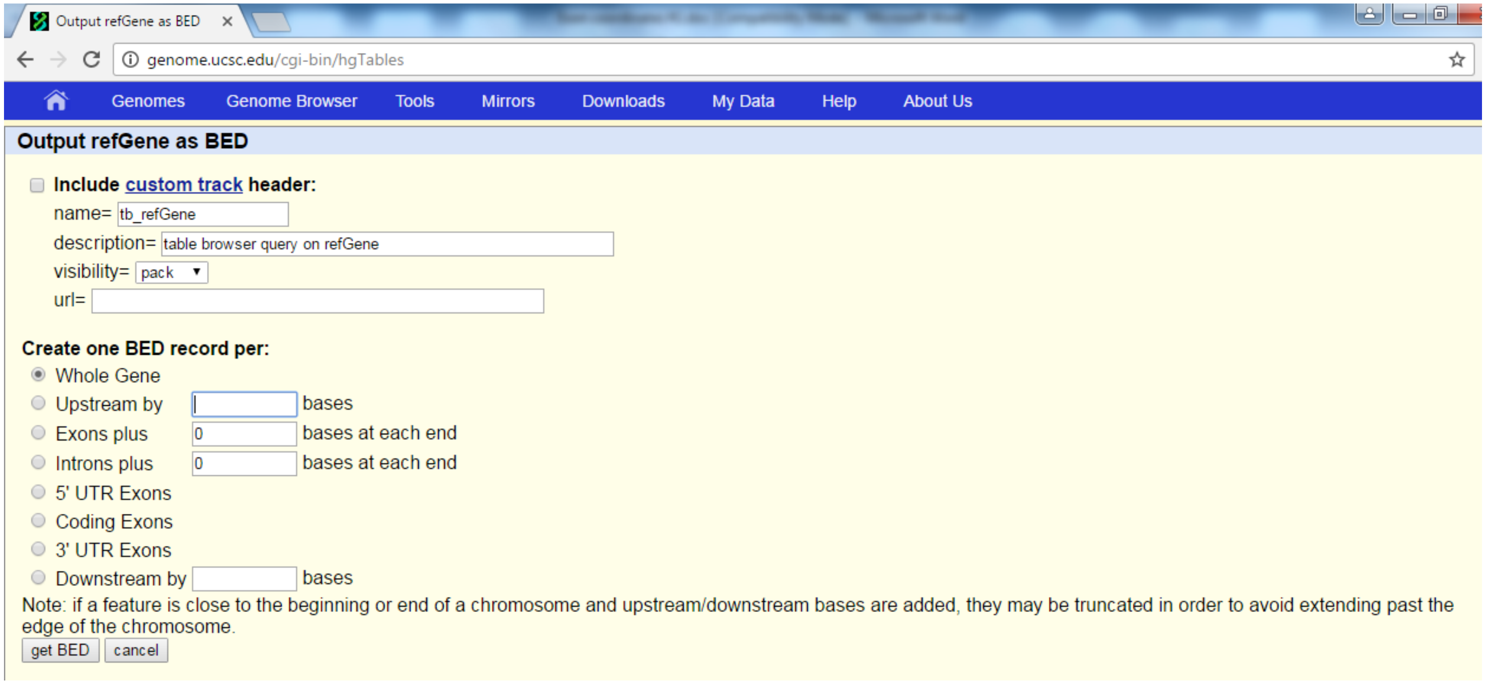
To reset **all** user cart settings (including custom tracks), click here.

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

- **clade:** Specifies which clade the organism is in.
- **genome:** Specifies which organism data to use.
- **assembly:** Specifies which version of the organism's genome sequence to use.
- **group:** Selects the type of tracks to be displayed in the *track* list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all available tracks in all groups. Select 'All Tables' to see all tables including those not associated with a track.
- **database:** (with "All Tables" group option) Determines which database should be used for options in table menu.
- **track:** Selects the annotation track data to work with. This list displays all tracks belonging to the group specified in the *group* list. Some tracks are not available when the region is set to *genome* due to the data provider's restrictions on sharing.
- **table:** Selects the SQL table data to use. This list shows all tables associated with the track specified in the *track* list. Some tables may be unavailable due to the data provider's restrictions on sharing.
- **describe table schema:** Displays schema information for the tables associated with the selected track.
- **region:** Restricts the query to a particular chromosome or region. Select *genome* to apply the query to the entire genome (not available for certain tracks with

Figure 2.1. UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables)



Output refGene as BED

□ **Include custom track header:**

name= tb_refGene

description= table browser query on refGene

visibility= pack

url=

Create one BED record per:

● Whole Gene

○ Upstream by [ ] bases

○ Exons plus 0 bases at each end

○ Introns plus 0 bases at each end

○ 5' UTR Exons

○ Coding Exons

○ 3' UTR Exons

○ Downstream by [ ] bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Figure 2.2; UCSC Output refGene as BED (http://genome.ucsc.edu/cgi-bin/hgTables)

The BED file of the custom gene coordinates for the target regions was used to create the Agilent SureSelect custom, cRNA biotinylated oligonucleotide bait library (Agilent Technologies, Santa Clara, CA, USA) to capture relevant sequences. This was designed using Agilent SureDesign software (release version 3.5.2), Agilent's proprietary algorithm to design optimal 120bp bait coverage and synthesise the custom capture reagents for the gene panel. The masking option used was "least stringent", to avoid removal of repetitive regions in introns. To compensate, "Boosting" was set to "Maximise Performance" (see Table 2.1). To utilise the capacity of the library kit, the tiling density was set to 5x so that five different baits covered each base and allowed baits to overlap by 100bp to optimise DNA capture of repetitive, GC-rich regions. In the design, target regions were extended by 10 bases at 3' end and 10 bases at 5' end, to ensure adequate coverage at exon and target boundaries, as the majority of intronic single nucleotide variants listed as disease-causing mutations are within this distance and is the recommended minimal region of interest for the design of new targeted assays (Association for Clinical Genetic Science, 2015).

Table 2.1. Summary of SureSelect DNA design for custom AML sequencing panel

Species	**H. sapiens (H. sapiens, hg19, GRCh37, February 2009)**
Target Summary	490 Target IDs resolved to 490 targets comprising 490 regions. 0 Target IDs were not found. Region Size: 591.402 kbp
Probe Summary	Total Probes: 36572 Total Probes Size: 591.579 kbp Coverage: 98.84292%
Target Parameters	Databases: RefSeq, Ensembl, CCDS, Gencode, VEGA, SNP, CytoBand Region: Entire Transcribed Region Region Extension: 10 bases from 3' end and 10 bases from 5' end. Allow Synonyms: No
Probe Tiling Parameters	Tiling density: 5x Masking: Least Stringent Boosting: MaximizePerformance

Table 2.2. Genes selected for AML NGS sequencing project

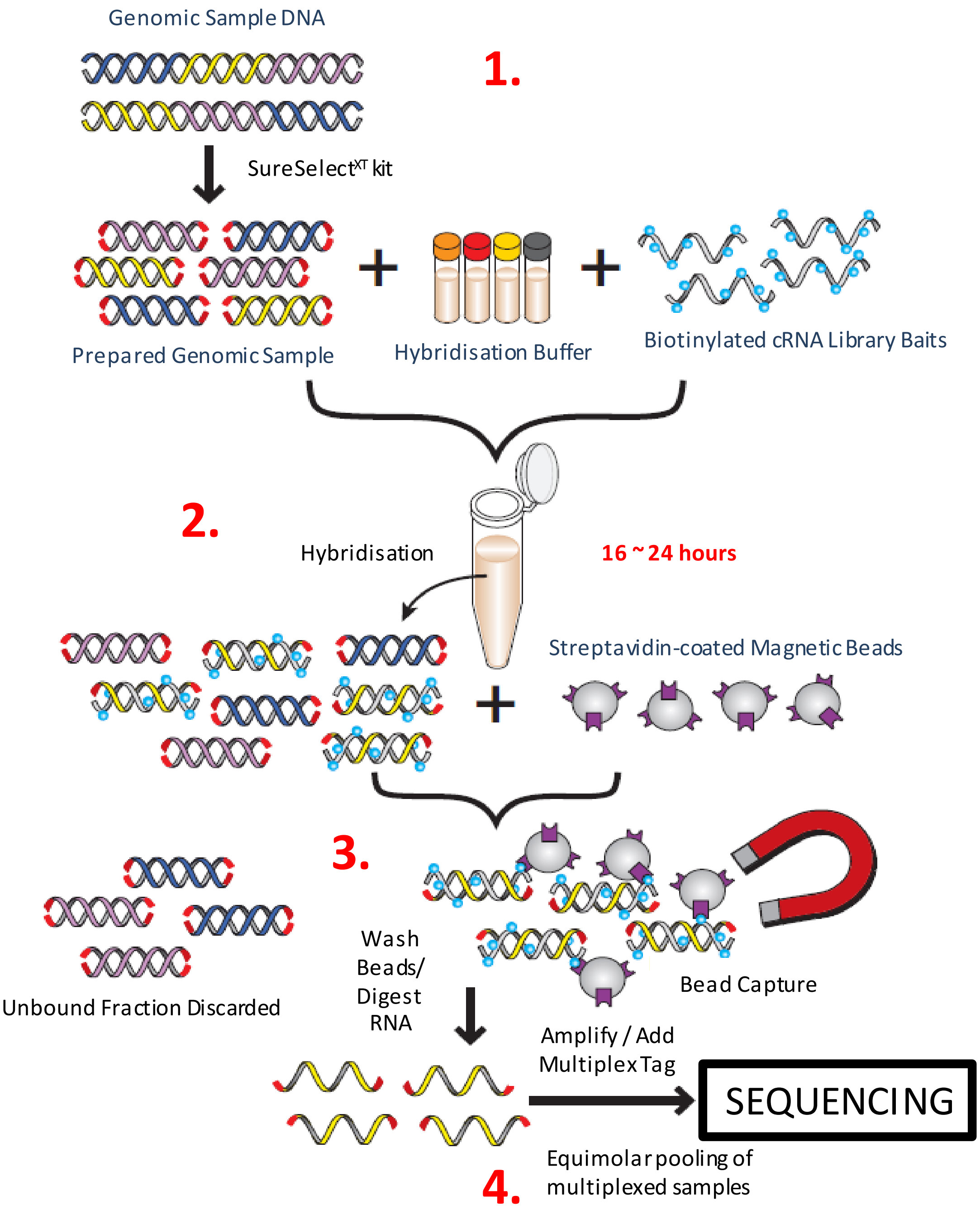
Gene	Chromosome	Target	Human Feb.2009 (GRCh37/hg19) Assembly		Regions	Size (bp)	Coverage (%)
			Start	End			
ASXL1	chr20	all exons	30946137	31027132	13	7685	100.00
BCOR	chrX	all exons	39910489	40036592	17	7329	100.00
CBFB	chr16	intron 5	67116243	67132612	1	16370	97.61
CBL	chr11	exons 1 - 15	119076976	119169260	15	2876	100.00
CBL	chr11	exon 16	119170195	119178869	1	8675	99.53
CEBPA	chr19	exon 1	33790830	33793480	1	2651	100.00
CUX1	chr7	all exons	101459174	101927260	36	16576	100.00
DEK	chr6	intron 2 - intron 3 inc	18258294	18264073	1	5780	97.68
DNMT3A	chr2	all exons	25455820	25565469	27	6475	100.00
EGR1	chr5	all exons	137801171	137805014	2	3176	100.00
EZH2	chr7	all exons	148504454	148581451	23	3762	100.00
FAM5C	chr1	all exons	190066787	190446769	9	3097	100.00
FLT3	chr13	exon 20	28592604	28592726	1	123	100.00
FLT3	chr13	exon 14 - exon 15 inc	28608024	28608351	1	328	100.00
GATA2	chr3	all exons	128198255	128369729	19	6569	100.00
IDH1	chr2	all exons	209100941	209119877	12	3021	100.00
IDH2	chr15	all exons	90627201	90645796	12	2096	100.00
KIT	chr4	all exons	55524085	55606891	22	5798	100.00
KMT2A	chr11	intron 2 - exon 13 inc	118339490	118360844	1	21355	99.11
KRAS	chr12	all exons	25357713	25403875	6	6009	100.00
MLLT1	chr19	exon 1 - intron 6 inc	6218053	6304958	1	86906	97.22
MLLT3	chr9	intron 4 - intron 8 inc	20354879	20448120	1	93242	99.45
MYH11	chr16	intron 28 - intron 34 inc	15814170	15826565	1	12396	96.10
NPM1	chr5	all exons	170814698	170837898	12	2020	100.00
NRAS	chr1	all exons	115247075	115259525	7	4594	100.00
NUP214	chr9	exon 4	134006224	134007982	1	1759	100.00
NUP98	chr11	intron 8 - intron 14 inc	3746450	3789810	1	43361	96.97
PHF6	chrX	all exons	133507332	133562832	14	8750	100.00
PML	chr15	intron 3	74315751	74317197	1	1447	100.00
PML	chr15	exon 6 - intron 6 inc	74325497	74326818	1	1322	100.00
PTPN11	chr12	all exons	112856526	112947727	17	7072	100.00
RAD21	chr8	all exons	117858163	117887115	14	4030	100.00
RARA	chr17	intron 2	38487649	38504567	1	16919	97.09
RPS14	chr5	all exons	149823782	149829329	7	1245	100.00
RUNX1	chr21	exons 6 - exon 9 inc	36160088	36206908	4	5664	100.00

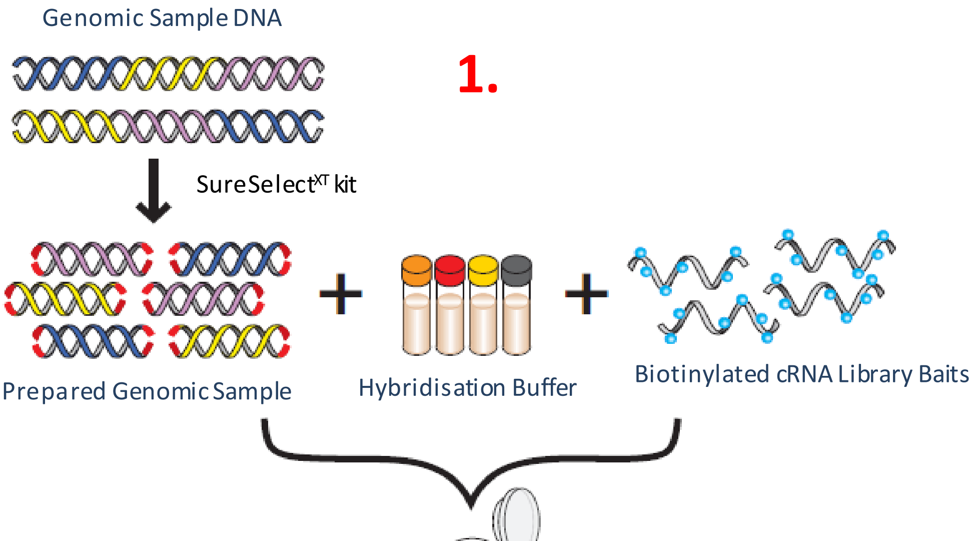
RUNX1	chr21	intron 6	36206899	36231770	1	24872	99.42
RUNX1	chr21	exons 1 - 5	36421455	36421605	7	2701	100.00
RUNX1T1	chr8	intron 1	93029592	93115084	1	85493	99.16
SETBP1	chr18	all exons	42260128	42648485	8	11132	100.00
SF3B1	chr2	all exons	198256688	198299827	27	6779	100.00
SMC1A	chrX	all exons	53401060	53449687	26	10540	100.00
SMC3	chr10	all exons	112327439	112364402	29	4692	100.00
STAG2	chrX	all exons	123094400	123236515	38	7440	100.00
TET2	chr4	all exons	106067022	106200970	13	19127	100.00
TP53	chr17	exons 5 - 11	7571710	7578564	9	2387	100.00
TP53	chr17	intron 4 - exon 5	7578361	7578821	1	461	98.26
TP53	chr17	exons 1 -4	7579302	7590878	6	1037	100.00
U2AF1	chr21	all exons	44513056	44527698	9	1200	100.00
WT1	chr11	all exons	32409312	32457091	12	3638	100.00
		* inc denotes inclusive. All interval is targeted					

Table 2.2 shows the gene targets selected for the project, including gene name, chromosome location and the start and end positions of the DNA sequence (according to Human Genome Assembly (GRCh37/hg19). 'Regions' denotes the number of separate regions that were included in the design and the size of the genomic target in base pairs. Coverage denotes the proportion of the region covered by target baits. 'inc' in the Target column indicates 'inclusive,' that all intervening sequences are targeted not isolated exons.

2.4.2 DNA Library Construction and Hybrid Capture

1 µg of genomic DNA was sheared in a Covaris S2 ultrasonicator (Covaris, Woburn, MA, USA) to an average size of 150-200 bp. The resulting DNA was then end repaired and ligated to Illumina adapters (Illumina, San Diego, CA, USA) using the manufacturer's protocol. Multiplexed libraries were prepared using the SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library kit and a custom 42 gene panel SureSelect Target enrichment Library kit (Agilent). Molecular barcode index sequences were ligated to sample's DNA to permit multiplexing of all thirty six specimens and sequencing in the same flow cells. Libraries were purified to remove small fragments of DNA and unligated adapters from the mix using AMPure magnetic beads (Agencourt, Brea, CA, USA). DNA targets were pulled down by solution phase hybridisation capture using the custom biotinylated RNA oligo pools in the custom kit designed in SureDesign, which was performed according to manufacturer's instructions for the kit *SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library* (Agilent Technologies). Streptavidin-coated paramagnetic beads were added and allowed to bind the biotinylated capture probes (see Figure 2.3). An external magnetic field was then applied and unbound DNA removed. The bound, captured DNA was eluted from the magnetic beads by digestion of the cRNA capture probes and purified. Successful capture was confirmed and library quality was checked on a high sensitivity DNA chip using the Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). Libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina (Kapa Biosystems, Wilmington, MA, USA).


Genomic Sample DNA
1.
SureSelectXT kit
Prepared Genomic Sample
Hybridisation Buffer
Biotinylated cRNA Library Baits
2.
Hybridisation
16 ~ 24 hours
Streptavidin-coated Magnetic Beads
3.
Unbound Fraction Discarded
Wash Beads/ Digest RNA
Bead Capture
Amplify / Add Multiplex Tag
SEQUENCING
Equimolar pooling of multiplexed samples
4.



. 1. Library

sh, elution

sion).

2.4.3 DNA Sequencing

Libraries were pooled (5 µL per library) and quantified by PCR to determine molarity for loading onto the NextSeq flow cell to achieve optimal cluster density (170~220 k/m^2). 1.8 pM pooled libraries were denatured, diluted and loaded onto a reagent cartridge according to standard Illumina protocol. The enriched target DNA was then amplified using universal primers targeting the paired-end adapters and clusters generated. 2x 150bp paired end sequencing was performed on the NextSeq 500 desktop sequencer (Illumina, San Diego, CA, USA) with the 36 samples multiplexed on a single NextSeq 500 High Output setting (300 cycles), using a NextSeq 500/550 High Output v2 kit (Illumina). The sample sequencing was run across multiple lanes to reduce the possibility of lane and position bias. Eight FASTQ files (Cock *et al.*, 2010) were generated for each specimen; two for each paired reads across four lanes of the flow cell. The FASTQ files were automatically uploaded to BaseSpace® (Illumina) cloud for storage of genomics data.

2.5 Data Analysis

2.5.1 Run Parameters and Quality Control

The sequencing metrics variables were recorded and used as evidence of run quality to inform confidence in data interpretation. This included;

- The number of reads before and after trimming, including the numbers of trimmed and untrimmed reads and the resulting mean read length.
- The number of reads discarded as being below the Quality threshold and number of duplicates identified
- Error rate
- The number of remaining unique reads used to calculate the proportion of reads aligned to target regions and reads mapping to multiple regions
- Proportion of target regions captured and covered in the sequencing reads
- Number of reads mapped and the percentage of target covered at the minimum coverage required

- Average and median read depth was calculated for target regions. The number of bases and regions covered at various read depths was recorded, as an indication of the average depth and uniformity of coverage. This is an estimate of the quality of sequencing to sample the complexity of data across the target regions.
- Range of insert size
- Average base call Quality scores for each position as a Phred-based value on a log scale related to the base calling error probabilities $(P)^2$

The alignment algorithm and alignment settings (seed length, mismatch tolerance, mismatch penalties, gap penalties and gap extension penalties) were recorded

Initial data analysis was performed within the Illumina online BaseSpace genomics analysis platform, BaseSpace®, which provides a computing environment with common analytical tools within its suite of software to integrate sequencing, data storage and analytical workflow. Initial quality metrics were assessed by running FASTQC analysis (v1.0.0) for each sample in the sequencing run using BaseSpace® Apps.

2.5.2 Analysis of data for SNVs and short indels

2.5.2.1 SureCall Pre-alignment Settings

Pre-alignment filters were adjusted in configurable Settings in SureCall and saved as a specific workflow for the analyses, for the exclusion of reads or bases from downstream analysis. The threshold for end trimming was set to 5 bp and reads between 30%~100% of original read length were retained (default setting). FASTQ files were enabled with a minimum read depth of 20. The variant score threshold was 0.3 and minimum Q Score (Quality Score) was set to 30, which is considered the standard in NGS, equivalent to the probability of an incorrect base call 1 in 1000 times or that the probability of a correct base call is 99.9%. The minimum variant call quality score used was 100. The SNPPET parameters Minimum Mutant allele frequency adjusted to 0.01 (from 0.10) for inclusion of greater number of high quality reads at low frequency, to be inclusive and subject low frequency reads to downstream analysis to assess the performance of the assay to detect subclonal variants. A minimum of 10 reads were required to support the variant call.

2.5.2.2 SureCall Method

Data was analysed using Agilent SureCall software (v3.5.1.46) (Agilent Technologies), which includes widely accepted, open source algorithms, augmented with tools specific to Agilent assays for optimisation for use with Agilent target selection products. SureCall was used for quality control (QC) metric calculation, assembly, visualisation and contextualisation of sequencing results. The multiple raw FASTQ files from the Illumina Next Seq for each sample were downloaded from Basespace and each set of 8 FASTQ files uploaded to Surecall as a Single Sample analysis, to detect insertions and deletions in the individual samples. Residual adaptor sequences were removed from each read using cutadapt (Martin, 2011). Sequence reads were aligned to (GRCh37/hg19 Feb. 2009 assembly) using the Burrows-Wheeler Aligner (BWA-MEM) (Li, 2013) and a BAM file was produced for each sample.(Li, 2013) and a BAM file was produced for each sample. SNPPET, an algorithm designed by Agilent to enhance detection of low allele frequency variants, was preferred as an alternative to SAMtools for variant calling and alignment viewing as well as sorting, indexing, data extraction and format conversion within SureCall for the mosaic leukaemia cancer samples, with a range of VAF. Incorporating regions from the BED file of the SureSelect analysis design, SNPPET used information from the coordinate sorted BAM file with additional Region Padding to extend the ends of the analysable covered region (the sequencable region with first bases removed) in the sequencing data by 100bp. Duplicate reads were removed and the quality of reads was recalibrated.

2.5.2.3 Annotation and Filtering of Variants to diagnostic reporting standards

The practice of variant calling is developing in the UK and guidelines are available for general practice (Association for Clinical Genetic Science, 2013) and general guidelines for NGS (Association for Clinical Genetic Science, 2015; Rehm *et al.*, 2013). Oncology samples present additional challenges and this area of practice is emerging (Dienstmann *et al.*, 2014; Lee *et al.*, 2015; Sukhai *et al.*, 2015; Strom, 2016). These sources were used to devise the following strategy for variant calling for the AML project. It was necessary to use various variant calling algorithms, with different functionality, to display and annotate sample mutations (see Table 2.3). Text files of all variant calls for SNV and short indel analysis were downloaded from SureCall and were imported into Microsoft Excel™ for manual examination and interpretation. The analysis of variant call files from other software is described below. Attempt was made to eliminate the false positive calls inherent in most NGS data (Clark *et al.*, 2011; Wall *et al.*, 2014) (see Technical Validation below). Having established that the data calls are likely to represent true variants in a patient's DNA sample, it was then necessary to detect clinically relevant changes from background passenger changes and germline polymorphisms (see Clinical Validation below). The variants that neither cannot be eliminated as insignificant nor confirmed as clinically significant remain as Variants of Uncertain Significance (VUS). The following strategy for Variant Selection was devised *a priori,* as a standardised approach to interrogate exons for nucleotide substitutions and short insertion-deletion variants using SureCall. Examination of regions for structural variation, including introns for gene fusions and complex rearrangements, by specific software is described in other sections below.

2.6 Rationale for Variant Selection

A number of features of data quality were reported in SureCall and recorded as part of the sequencing audit trail, which were taken into account in technical validation of variant calls. This included the following (Association for Clinical Genetic Science, 2015);

2.6.1 Base call quality score (Q Score)

A base Quality score (Q Score) is a measure of the reliability of each base call and it is generated during sequencing and is assigned to each base call for every sequencing cycle during a sequencing run. Illumina Q Scores are computed measures of quality predictor values, such as intensity profiles and signal-to-noise ratios, compared to an empirically pre-calibrated quality model (Q table) (Illumina, 2011; Illumina, 2014b).(Illumina, 2011; Illumina, 2014b). A Q Score of less than 20 (equivalent to a P-value of 0.01) is considered low quality and will be removed from consideration (Strom, 2016) and generally scores less than 30 were filtered from downstream analysis.

2.6.2 P-value

Variant calls with a *P*-value of >0.01, as calculated within SureCall, were eliminated from further consideration.

2.6.3 Read Depth

The minimum depth of coverage was established for 'true' read depth as a measure of the number of independent, overlapping sequence alignments at a locus of interest. Coverage was determined for unique reads only, following removal of any duplicate reads which probably represented copies generated by amplification during the library preparation. A minimum read depth of 100x was considered desirable for single nucleotide detection, to avoid missing low level variants and to filter sequencing artefacts. However, this was reduced for multiple reads confirming the same complex variant and the target was adjusted. Where a definite pathogenic mutation was identified and confirmed, regions of sequence not meeting the required read depth were accepted and described as low coverage (Association for Clinical Genetic Science, 2015; Strom, 2016; Rehm *et al.*, 2013).

2.6.4 Variant read number and Variant Allele Frequency (VAF)

10 independent reads supporting the presence of a single base variant was considered the number below which would indicate a false positive signal, to accommodate a higher number of false reads expected with increasing read depth (Strom, 2016; Wall *et al.*, 2014). Variant allele frequency was calculated as the proportion of mutated reads versus total number of reads covering that base. It was decided to be permissive with variant read proportion and to allow the a VAF threshold of 1%, providing higher specificity (and reduced sensitivity) and submit these variants for scrutiny of their clinical significance, as a measure of test sensitivity (see below), to test the assay in its intended use as routine testing without confirmatory control sample or parallel testing.

2.6.5 Variant Quality Scores (QUAL)

Variant quality scores (QUAL) are similar to base quality score and are transformed log scale Phred-like scores generated during the variant calling step and are included in the Variant Call File (VCF) for an analysis (Strom, 2016). QUAL is an estimate of the confidence that the variant caller correctly identified that a given genome position displays variation in at least one sample; a score of 90 is equivalent to P value of $1x10^{-9}$. The threshold for the project was set at 100.

2.6.6 Strand Bias

Strand bias is a measure of likelihood that variant reads are detected equally on + and – DNA strands and deviation from equivalence indicates that that the variant is an alignment artefact. Forward and reverse read data was manually examined and considered in variant calling.

2.7 Clinical Validation of Variants

Having excluded technical artefacts which would be likely false positive variants with insubstantial credentials to be considered as genuine sequence changes, it was then necessary to attempt to filter non-pathogenic variation non-pathogenic in AML, including germline single nucleotide polymorphisms (SNPs) and more challengingly, somatic passenger mutations of no clinical effect. Some variants will be common and immediately recognisable as recurrent driver mutations and clearly pathogenic. It was necessary to apply certain criteria to reduce subjectivity for downstream filtering and provide high confidence variants in clinically relevant information leading to annotation (from RefSeq database), leading to simulation of diagnostic standard reports. For those variants that are not clearly pathogenic, several means were used of interpreting clinical significance, to provide high confidence.

2.7.1 Filtering benign germline polymorphisms (SNPs)

The presence of a variant in an unaffected individual at population risk was used as evidence of non-pathogenicity. Benign germline polymorphisms were defined as variants with a population frequency of ≥1% (in dbSNP, a large normal population screening databases) were excluded as probable germline SNPs. All variants passing initial filtering were manually interrogated and those suspicious for germline polymorphisms were excluded (with VAF of 50% or 100%, reflecting germline heterozygous or homozygous state respectively). To screen for lower frequency germline variants, all those detected at <1% frequency were manually interrogated and retained for further analysis, if they had a corresponding COSMIC database entry with multiple reports in malignancy, particularly in haematopoietic tissue of a confirmed somatic variant.

2.7.2 Filtering non-pathogenic variation

Variant calls were further assessed to eliminate variants of no significance in AML. The association of the variant and its frequency in AML (and possibly or other tumours) was assessed by examination of the COSMIC database and multiple entries in haematopoietic and lymphoid tissue showed co-segregation with AML and were considered evidence as evidence of pathogenicity. Apparent somatic variants not found in haematological neoplasia but with apparent functionality in other cancer subtypes were scrutinised further. Other cancer databases and online tools for these variant annotation tasks were consulted as necessary, such as:

- **ClinVar - https://www.ncbi.nlm.nih.gov/clinvar/**
- **My Cancer Genome (https://www.mycancergenome.org/**
- **cBioPortal - http://www.cbioportal.org/**
- **HGMD - http://www.hgmd.cf.ac.uk/ac/index.php**
- **OMIM -https://www.ncbi.nlm.nih.gov/omim**

The use of locus specific databases (LSDB) was limited for this study but TP53 database would be consulted if necessary. Manual curation of primary PubMed citations, professional guidelines; and comprehensive clinical, genetic, biochemical, and functional database searches, and clinical trials of possible targeted therapies were checked as appropriate. Although the main aim of the project is not to mine for novel variants of ambiguous clinical significance, *in silico* prediction of oncogenicity using tools for predicting the biochemical dysfunction of the variant were considered such as SIFT (and as necessary others such as Provean, PolyPhen-2, MutationTaster, CADD) to substantiate evidence for damaging/deleterious or potentially driver mutations. Variants were described in accordance with the Human Genome Variation Society (HGVS) recommendations (Human Genome

Variation Society (HGVS), 2016) or the coding sequence derived from Reference Sequence (RefSeq) (National Center for Biotechnology Information (NCBI), 2016).

2.7.3 Identification of actionable cancer-associated somatic mutations

Once benign or likely benign variants were excluded, three categories of actionability were used although for clinical purposes this could be refined (as defined by a new classification system – see Table 3.5) (Lin *et al.*, 2017):

1. Known actionable (direct impact on patient care)
2. Potentially actionable (somatic mutation with biological relevance not previously reported)
3. Variant of unknown significance (VUS) (likely somatic mutation with uncertain effect).

2.8 Variant Annotation using SureCall

SureCall used several tools to provide input for the mutation classification. SureCall also identified additional information, including the chromosomal location of the mutation with Human Genome Variation Society (HGVS) nomenclature. Each mutation was evaluated by the software based on its location, amino acid change, and effect on protein function (SIFT) (Hu & Ng, 2013) and impact on structure and function of the protein using the Polymorphism Phenotyping v2 (PolyPhen-2) tool. Further information regarding the mutation was then aggregated from various public sources, including National Center for Biotechnology Information (NCBI) Database of Genomics Structure Variation (dbVar), COSMIC (Catalogue of Somatic Mutations in Cancer), PubMed, and Locus-Specific Databases. After collecting the various inputs for classification, the proprietary mutation classifier evaluated the significance of the mutation following default guidelines. Each variant was then independently examined by triaging each mutation and reviewing supporting evidence. Mutation calls were examined in the built-in Integrative Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2013) to confirm regional coverage, visualize read alignments and confirm variant calls.

2.8.1 Analysis of data for SNVs and short indels using GATK HaplotypeCaller

FASTQ reads from each sample were aligned to the GRCh37 human genome using Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009). SAM alignment files were then converted to sorted, indexed BAM format using Picard tools from the Broad Institute (Broad Institute, 2017). BAM files for each sample from four lanes were merged using Picard into a single unified BAM file, so that full depth of coverage could be estimated and that duplicates for the whole sample could be removed. Indel-realignment and Base Quality Score Recalibration (BQSR) were performed using GATK 3.6 (McKenna *et al.*, 2010), as recommended in guidance (DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). Extra effort realignment was performed in order to take advantage of high depth. Variants were called using the joint genotyping procedure which allows low frequency variation in one sample to be rescued statically by evidence from another. Following alignment and variant calling, hard filtering was performed, which used a set of pre-defined hard cut-offs for variation to be included, producing filtered VCF files for all samples containing the variants passing the filters. Following filtration procedures, filtered VCF files were used as the input to Ensembl Variant Effect Predictor (VEP) (McLaren *et al.*, 2016).

2.8.2 Indel and Structural Variations Calling using Pindel

Pindel (version 0.2.5b9) (Ye *et al.*, 2009) was used to detect breakpoints of large deletions, medium sized insertions, inversions, tandem duplications and other structural variants at single-base resolution, within the targeted regions from the merged BAM files from. The event finding algorithm was set to maximum size, to detect the largest possible events. Of the Parameters which affect which structural variants are reported, the '--report_interchromosomal_events' option was activated in attempt to detect gene fusions. The output from Pindel was in the form of a text file for each type of structural event, in a specific format, including in many cases a text-drawn diagram of how the reads are stacked up as evidence for particular events (User Manual from The Genome Institute at Washington University School of Medicine, 2014).

2.8.3 Gene fusions and Structural Variation Calling using LUMPY

LUMPY (version 0.2.13) (Layer *et al.*, 2014) was used to detect gene fusions and other structural variation (SV). LUMPY provides a mechanism based upon probability distributions of an SV breakpoint that allows evidence from multiple alignment signals to be simultaneously integrated into a single detection algorithm, which includes discordant read-pairs and split reads, as well as other signals such as read depth and prior knowledge of the SV breakpoint. LUMPY aligns the discordant read-pairs and determines a pair of intervals upstream or downstream of the mapped reads as possible breakpoint positions; the evidence from the alignment signals is mapped to the breakpoint intervals and the probabilities of overlapping intervals and clustered breakpoints are integrated. Any clustered breakpoint region that contains sufficient evidence is returned as a predicted SV. LUMPY was downloaded from (https://github.com/arq5x/lumpy-sv) and lumpyexpress was run on the BWA / Picard produced BAM files. The individual BAM files from each of four sequencing where ran separately in order for lumpyexpress to accurately detect split and discordant reads. The output from LUMPY is in the form of a VCF file v.4.2 (Samtools, 2015) which were converted to text files and imported into Microsoft Excel. The number of reads was encoded into the genotype field as 'SU,' 'PE.' and 'SR' in the final column of the VCF output. The minimum number of supporting reads for an event to be reported in LUMPY was 4. Different types of events with multiple supporting reads were examined and particularly, the presence of gene fusions were examined as represented in the INFO field as SVTYPE=BND (a break end), which were also present as paired events with pairs of IDs on adjacent lines.

Table 2.3. List of Analysis Software accessed for project

Pipeline step	Software	Software Suite	Function
Adapter trimming	FASTQ Toolkit 2.0.0	Basespace	Removal of residual adapter sequences from reads
	Cutadapt	Surecall	Removal of residual adapter sequences from reads
Quality control	FastQC	Basespace	Quality metrics for sequencing Highlight sequencing errors
Mapping/Read Alignment	BWA-MEM	SureCall	Aligns FASTQ files from samples to the reference genome. Create index. Read alignment generate alignment metrics. Duplicate removal
	Haplotype Caller	GATK	Read alignment generate alignment metrics. Duplicate removal.
Pre-processing	Picard	In BWA, Basespace	Mark duplicates
	Samtools	In BWA, Basespace	Convert SAM to BAM, fix read pairing information and flags, sort BAM by coordinates. Merge and index BAM files, Calculate alignment statistics.
Post-alignment QC	FastQC	Basespace	QC info on alignments on merged BAM files
	BamQC		QC info on alignments on merged BAM files
Variant calling	BWA-MEM	SureCall	Detection of SNV and small indels
	Variant Quality Score Recalibration (VQSR)	GATK	Variant filter SNV calling
	Pindel	-	Detection of large scale indels
	Lumpy	-	Identification of discordant paired end reads or split end reads to detect gene fusions

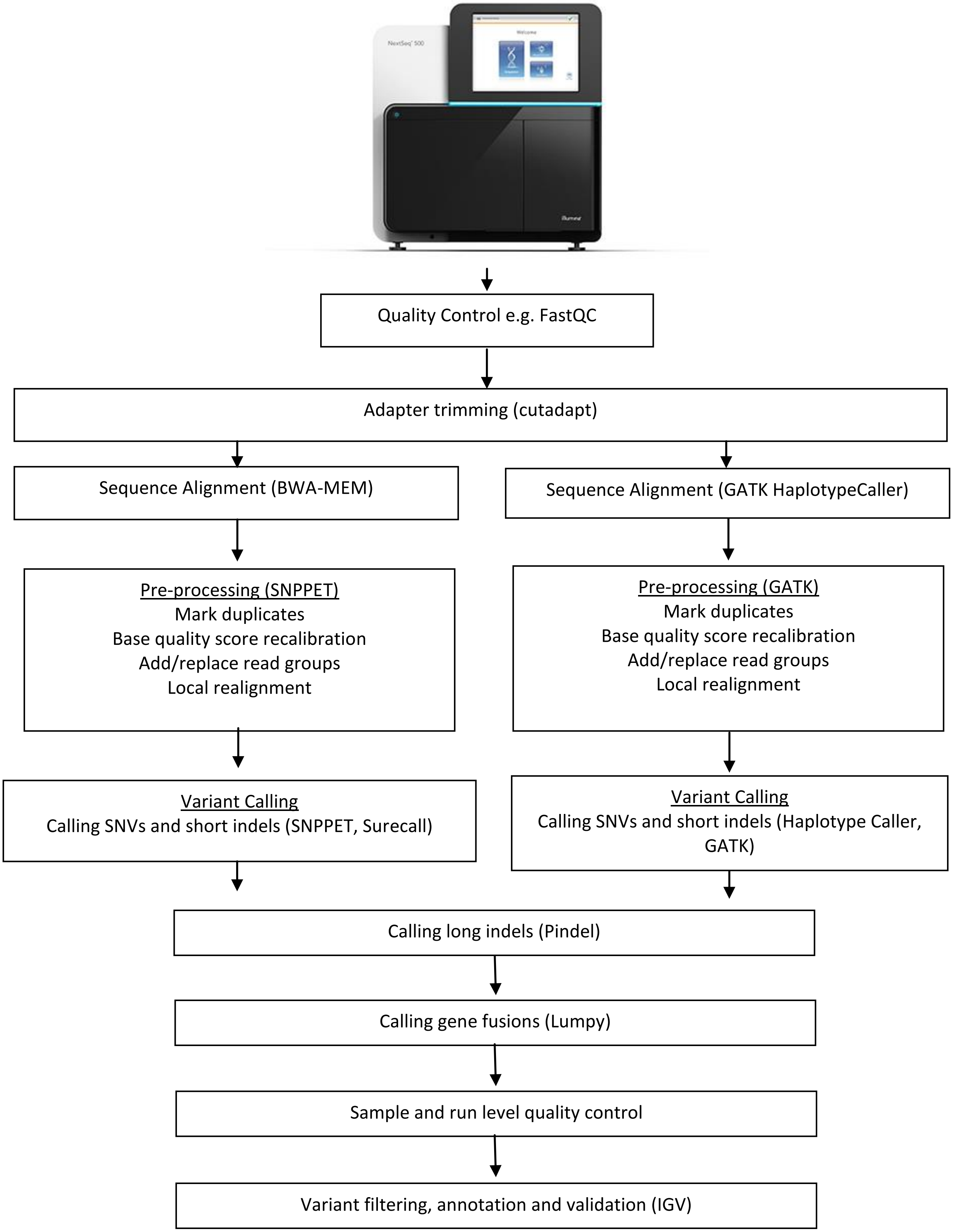

NextSeq 500
Quality Control e.g. FastQC
Adapter trimming (cutadapt)
Sequence Alignment (BWA-MEM)
Sequence Alignment (GATK HaplotypeCaller)
Pre-processing (SNPPET)
Mark duplicates
Base quality score recalibration
Add/replace read groups
Local realignment
Pre-processing (GATK)
Mark duplicates
Base quality score recalibration
Add/replace read groups
Local realignment
Variant Calling
Calling SNVs and short indels (SNPPET, Surecall)
Variant Calling
Calling SNVs and short indels (Haplotype Caller, GATK)
Calling long indels (Pindel)
Calling gene fusions (Lumpy)
Sample and run level quality control
Variant filtering, annotation and validation (IGV)

Figure 2.4. Analysis Pipeline of Software used for the project and data flow

2.9 Orthogonal sequencing using Ion Torrent

AML project samples with sufficient DNA and expected to be most informative were selected for sequencing using Ion Personal Genome Machine™ (PGM™) System (Thermo Fisher Scientific Inc.), for comparison of the somatic mutations detected by the main project. 10ng of DNA (see Table 2.4 and Table 3.7 in Results) was used to create a targeted sequencing library using Ion AmpliSeq™primers for the AML Community Panel (Ion AmpliSeq™ AML Research Panel), targeting the coding regions of known mutations in 21 commonly mutated genes involved in AML. A modified protocol, AmpliSeq™ AML Protocol Modifications for 4-Pool Panel (https://ioncommunity.thermofisher.com/docs/DOC-8939 accessed 30/03/2016), was used to enhance GC rich regions that are difficult to amplify consistently. The library was prepared from resulting amplicons using Ion AmpliSeq™ Library Kit 2.0. Automated template preparation for 200 base-read libraries was performed using an automated Ion OneTouch™ 2 System with Ion PGM™ Template OT2 200 reagent kit, yielding barcoded libraries using Ion Xpress™Barcode Adapters 1-96 Kits. Six barcoded samples were loaded on to each of five Ion 318™ Chips (kit v2.0), the recommended, optimal capacity to detect enabling 4~5.5 million reads per chip at target >95% 500x base coverage. Sequencing was performed in April and May 2015 on an Ion Personal Genome Machine™ (PGM™) System semiconductor-based sequencer, which uses digital reads of polymerase-driven base incorporation. Standard Ion PGM™ Hi-Q™ Sequencing Kit reagents were used.

Sequencing data was uploaded to Ion Reporter™ Software v5.0 (https://ionreporter.thermofisher.com/ir/ last accessed 29/09/2015), which comprises a suite of bioinformatics tools to provide a workflow from data import, detection of SNPs, indels, and annotation of variants. BAM files were generated by use of the standard pipeline. Detects and annotates low frequency variants (SNPs, Indels) from targeted DNA libraries from the Ion AmpliSeq™AML Cancer Research Panel. Target regions AML hg19, Custom filters COSMIC (v67), 0.0 <= PValue <= 0.01, Filtered Coverage >= 250, 100 <= Allele Read-Count <= 100000, dbSNP.

Table 2.4. Genes covered in the Ion AmpliSeq™ AML Cancer Research Panel.

This utilises 237 amplicons to analyse 19 genes implicated in AML

ASXL1*	FLT3*	KIT*	RUNX1*
BRAF*	GATA2	KRAS*	TET2
CBL*	IDH1*	NPM1*	TP53
CEBPA	IDH2*	NRAS*	WT1*
DNMT3A	JAK2*	PTPN11*	
			*Hotspot regions

3.0 Results

3.1 Patient population

Samples were collected from patients referred with a possible diagnosis of AML, and that also had sufficient surplus cells after diagnostic testing. After the diagnosis of AML had been confirmed, DNA was extracted and thirty-six samples with DNA of adequate quality and quantity for the range of sequencing studies were used. Five samples were rejected for having a poor yield of DNA. Two samples were later confirmed to be pre-treatment, diagnostic bone marrow aspirate samples from the same patient (13.1141 and 13.1485). There were 20 male and 15 female patients. The cases were representative of the typical age range in AML; the median age of the patient cohort was 64 (range 14-84). There were two adolescent patients (ages 14 and 16) and 20 patients were aged 60 or over (see Figure 3.1).

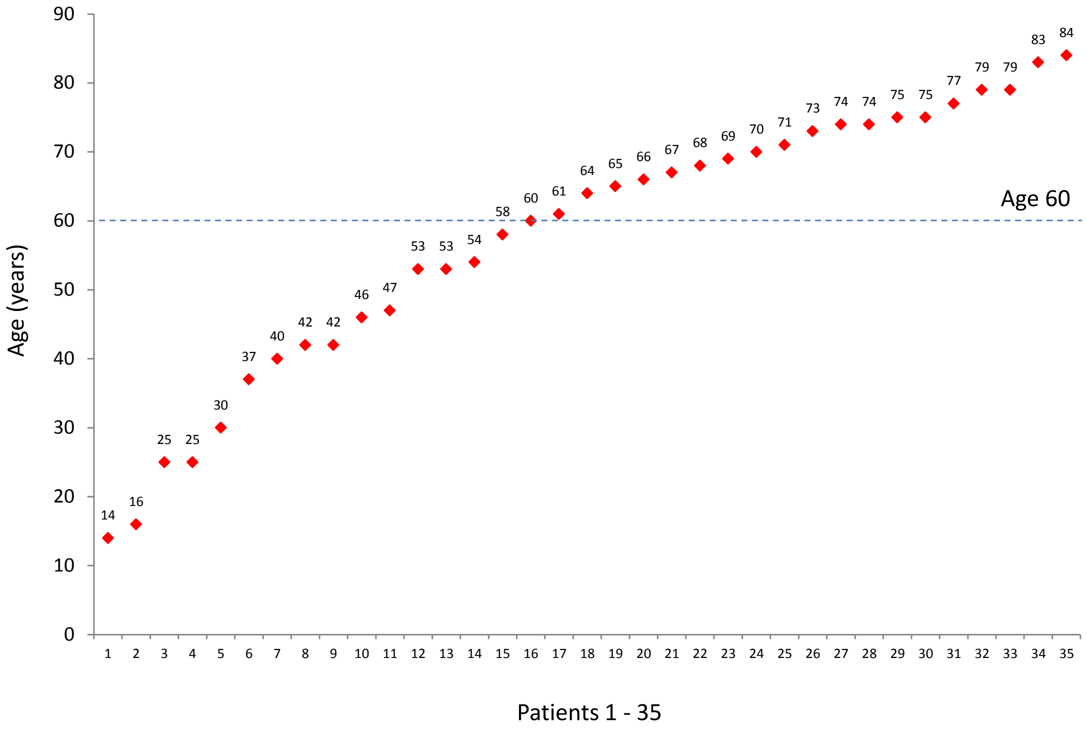


Figure 3.1. Age distribution of participants in the AML sequencing project

3.2 Cytogenetic and standard molecular testing

Thirty-five out of thirty-six samples were successfully cytogenetically analysed; one sample failed to yield sufficient analysable metaphases. The majority of patients (68%) showed intermediate risk cytogenetics; 18 showed a normal karyotype (CN-AML) and 5 patients with intermediate risk abnormalities, including various non-specific abnormalities common to myeloid neoplasia. Sample 23 was found to have t(11;19)(q23;p13.3) with *KMT2A-KMT2AT1* by FISH. Sample 13 showed a *NUP98* gene rearrangement by FISH, apparently resulting from a large pericentric inversion of chromosome 11, inv(11)(p15q22).

Six patients had favourable cytogenetics, two had acute promyelocytic leukaemia with t(15;17) and *PML-RARA* gene fusion. CBF-AML was diagnosed in four patients, three with inv(16)/t(16;16) and *CBFB* gene rearrangement and one with t(8;21) with *RUNX1-RUNX1T1*. Specific gene fusions were confirmed by FISH. The remaining 6 samples showed karyotypes with adverse prognostic features, two of which were from the same patient and showed karyotypes which were both complex (≥4 abnormalities) and monosomal. As would be expected, the APL and CBF-AML patients were of lower median age (27.5 years) compared to remaining patients (NK-AML and other abnormal) (67 years). The remaining 6 samples showed karyotypes with adverse prognosis, two of which were from the same patient and showed karyotypes which were both complex (≥4 abnormalities) and monosomal karyotypes.

Thirteen *NPM1* exon 12 mutations (36%) and eleven *FLT3-ITD* (31%) were detected by conventional molecular genetic studies. Six patients were *NPM1* mutated alone (*NPM1*+/*FLT3*wt) (17%), four patients *FLT3*-ITD alone (*NPM1*wt/*FLT3*-ITD+) (11%), and seven were double mutated (*NPM1*+/*FLT3*-ITD+) (19%). All showed normal karyotypes with the exception of Sample 18 who was *NPM1*+/*FLT3*wt with a non-specific abnormal karyotype, apparently involving rearrangement of chromosome 5. In Patient 5, who showed double mutation, *NPM1*+/*FLT3*-ITD+, the conventional cytogenetics failed and negative results were obtained with the panel of FISH tests. Of the four cases with *NPM1*wt/*FLT3*-ITD genotype, two also showed translocations by cytogenetics; t(8;21) and t(15;17) (summarised in Table 3.1 and details in Table 3.2).

Table 3.1. Summary of clinicopathologic characteristics of 35 patients with AML

Features	Number (%)
Gender	
Male	20 (57%)
Female	15 (43%)
Age years (median/range)	62.5 (14-84)
Cytogenetics	
Favourable	
A*PML*/t(15;17)	2 (5.5%)
CBF-AML / inv(16)/t(16;16)	3 (8.3%)
CBF-AML / t(8;21)	1 (2.7%)
Intermediate	23 (63.9%)
Normal karyotype	18 (50%)
Other abnormalities	5 (13.9%)
Adverse	5 (13.9%)
Complex	1 (2.7%)
Molecular Genetics	
***NPM1+/FLT3*wt**	6 (16.7%)
***NPM1*wt/*FLT3*-ITD**	4 (11.1%)
***NPM1+/FLT3*-ITD**	7 (19.4%)

Table 3.2 - Patient samples and routine cytogenetic and molecular genetic profiles

Samples 10 and 14 are pre-treatment diagnostic samples from the same patient

Sample No.	**Sex**	**Age**	**Karyotype**	**FISH positive results (%)**	**Other FISH tested (negative results)**	***NPM1***	***FLT3***
1	F	75	46,XX,del(5)(q15q33)[4]/46,XX[6]	Deletion of EGR1 at 5q31 (56%)		neg	neg
2	F	67	46,XX,idic(7)(q11.2)[4]/46,XY[6]	Deletion 7q detected (55%)		neg	neg
3	M	83	46,XY[20]		*KMT2A*	neg	neg
4	F	70	46,XX[20]		BCR-ABL1	neg	pos
5	F	74	Failed		5q31, 7q31, *TP53*, *KMT2A*	pos	pos
6	F	84	46,XX[20]			neg	neg
7	F	53	46,XX[20]		*PML/RARA*, *KMT2A*	pos	neg
8	M	79	46,XY[20]			pos	neg
9	M	42	46,XY,t(15;17)(q24.1;q21.1)[9]/46,XY[1]	*PML-RARA* gene rearrangement (94%).		neg	neg
10	M	60	41~47,XY,del(1)(q41q42),add(2)(p13),-5,-7,+8,+11,add(13)(q14), der(13;21)(q11;q11),add(17)(p11.2),-18,-20,-21,+22,+1~5mar[cp9]/46,XY[1]		*CBFB*, *RUNX1/RUNX1T1*	neg	neg
11	M	65	46,XY[20]			neg	neg
12	M	58	46,XY[20]		*PML-RARA*, *RARA*	pos	neg
13	M	64	47,XX,+inv(11)(p15q22)[10]	*NUP98* gene rearrangement (77%)		neg	neg
14	M	60	41~46,XY,del(1)(q41q42),-5,-7,+8,add(13)(q14),add(17)(p11.2),-18,-20,-21,+1~2mar[cp5]/46,XY[5]			neg	neg
15	M	61	46,XY[20]			pos	neg
16	M	25	46,XY,inv(16)(p13q22)[10]	*CBFB* gene rearrangement (80%)	*PML/RARA*	neg	neg
17	F	47	46,XX[20]		*PML/RARA*, *RARA*, BCR/ABL1	pos	pos
18	M	37	48,XY,+del(5)(q?12q?33),+8,der(20)t(1;20)(q12;q13.3)[8]/46,XY[2]			pos	neg

Table 3.2 (continued)

Sample No.	Sex	Age	Karyotype	FISH positive results (%)	Other FISH tested (negative results)	*NPM1*	*FLT3*
19	F	16	46,XX,t(15;17)(q24.1;q21.1)[10]	*PML-RARA* gene rearrangement (91%)		neg	pos
20	M	73	46,XY[20]			pos	pos
21	F	66	46,XX[20]		*KMT2A*	pos	pos
22	M	77	46,XX[20]		*CBFB*	pos	neg
23	F	69	46,XX,t(11;19)(q23;p13.3),del(12)(p11p13)[10]	*KMT2A-KMT2A*T1 gene rearrangement (82%), deletion of ETV6 (94%).		neg	neg
24	M	25	45,X,-Y,t(8;21)(q22;q22.3)[10]	*RUNX1-RUNX1T1* gene rearrangement (95%)		neg	pos 35%
25	M	74	46,XY,der(9)t(9;11)(q34;q13)[4]/46,XY,der(9)del(9)(q13q34)t(9;11)(q34;q13)[6]	Gain of *KMT2A* (70%)	*KMT2A*, BCR-ABL1	neg	neg
26	F	54	49,XX,+6,+7,+8[9]/46,XX[1]		*PML/RARA*, *RARA*, BCR-ABL1, *KMT2A*	neg	neg
27	M	42	46,XY,del(9)(q?13q?22)[3]/46,XY[28]		MECOM, *KMT2A*, *RUNX1/RUNX1T1*	neg	neg
28	M	75	46,XY[20]			neg	pos
29	M	53	46,XY[20]		*KMT2A*	pos	pos
30	F	30	46,XX,t(16;16)(p13;q22)[17]/46,XX[3]	*CBFB-MYH11* gene rearrangement (48%)		neg	neg
31	M	68	46,XY[20]			pos	pos
32	M	46	46,XY[20]		*KMT2A*	pos	pos
33	M	40	46,XY,inv(16)(p13q22),i(22)(q10)[5]/47,XY,inv(16)(p13q22),+22[5]	*CBFB* gene rearrangement (66%)		neg	neg
34	F	71	46,XX[20]			neg	neg
35	F	79	45,XX,-7[6]/47,XX,+8[4]			neg	neg
36	F	14	46,XX[20]		*KMT2A*, *NUP98*	neg	neg

3.3 Illumina sequencing of the custom NGS panel

3.3.1 Run level performance of custom panel sequencing

The libraries from 36 DNA samples were sequenced as described in the Methods (section 2.3 and 2.4) to produce paired end reads with a single index tag (8bp). 150 informative cycles for each read cluster were obtained and the yield (Gbp) is shown in Table 3.3. The percentage of the sample that aligned to the control PhiX genome was used to calculate the error rate, including at specific cycle points during the sequencing run (Illumina, 2014a). The signal intensity at cycle 20 (as a proportion of intensity at cycle 1) and the percentage of reads with Q Score ≥ 30) is shown in Table 3.3. Overall, 83.7% of base calls achieved a Q Score of ≥ 30 (87.1% of Read 1 and 79.7% of Read 2) (see Figure 3.2), with an estimated error rate of 1.39% (1.08% of Read 1 and 1.69% of Read 2).

Table 3.3. Run Metrics Summary

	Cycles	Yield (Gbp)	Aligned to PhiX genome control (%)	Error Rate (%)	Intensity Cycle 20/1*100	≥Q30 (%)
Read 1	151	69.49	0.89	1.08	4,292	87.42
Read 2	151	69.45	0.85	1.69	4,847	80.00
Non-Index Reads Total	302	138.93	0.87	1.39	4,569	83.71
Totals	310	142.18	0.87	1.39	4.055	83.95

The total number of reads for the 36 sample runs was 512,880,334, with 463,258,793 of the indexed read clusters passing quality filtering. The indexing QC results for the run showed that there were no unexpected results for a sample and that all indexed samples were properly represented. The mean number of reads per sample was 12.39 million (standard deviation 1.04 million, range 10.46~14.87m). Indicating that there was even coverage across different lanes of the flow cell and good representation of reads from each sample (maximum and minimum % reads per sample per lane = 3.2241 and 2.2495 respectively) (see Table 3.4). The run metrics from Ion Torrent sequencing are presented in Appendix 3 (see section 7.3).

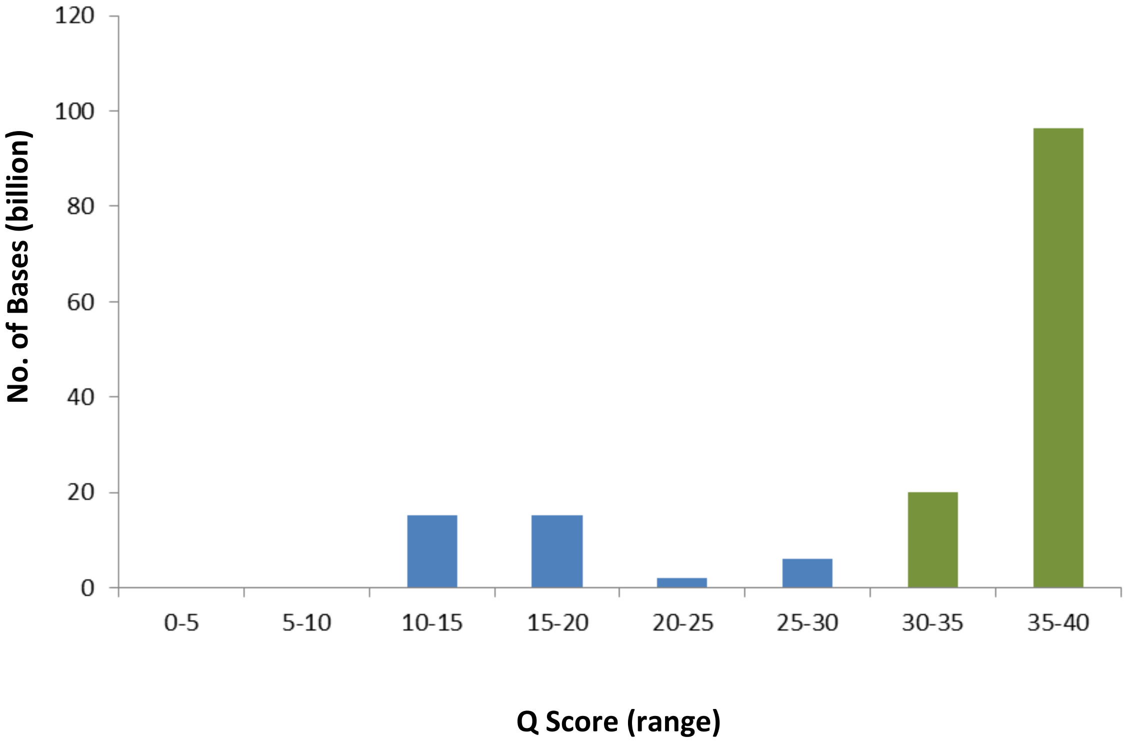


Figure 3.2. Histogram showing distribution of Quality Scores (total read 1 and 2) showing proportion of reads ≥ 30 (green)

Table 3.4. Index Performance Summary

Flow Cell Lane	Total Reads	Reads Passing Filter	Identified Reads Passing Filter (%)	Coefficient of Variation	Reads per sample Identified Minimum%	Reads per sample Identified Maximum%
1	131999850	120908797	96.4733	0.0847	2.2513	3.2241
2	130027750	119339052	96.3578	0.0842	2.2495	3.2057
3	126349863	112689372	96.2354	0.0836	2.2676	3.2099
4	124502871	110321572	95.9765	0.0836	2.2688	3.2011

3.3.2 Sample level post-alignment quality statistics from SureCall BWA-MEM

Quality Control (QC) reports for each sequence alignment were generated from SureCall, downloaded and assimilated into an Excel spreadsheet. A mean of 22,909,567 reads (SD 1,169,853) were present in untrimmed FASTQ files and 22,900,409 (SD 1,169,321) after trimming. The resulting average read length was 143 (SD 0.41). A mean of 10,667,341 reads were trimmed and 12,233,068 remained untrimmed. 9,158 reads were discarded. 3,443,288 (SD 275,847) were duplicate reads, ranging from 13.81% and 16.27% of the total trimmed reads. This is expected and acceptable for this sequencing run.

36.97~38.79% mapped to target regions, which was expected due to capturing multiple regions with repetitive elements. 4.95-5.37% of reads mapped to multiple locations and 0.10% of reads of covered regions had no coverage. The mean number of analysable reads in covered regions was 6,816,275 (SD 295,846) (36.29~38.08%). The overall average of and median read depths in analysable target regions was 1,650 (SD 70.64) and 1,805 (SD 74.03) respectively. 99.9% of analysable target bases had at least 10 reads, 98.96% at least 100 reads, 93.39% at least 500 reads and 82.22% at least 1,000 reads. All analysable target regions had at least 10 reads, 98.85% at least 100 reads, 94.66% at least 500 reads and 88.84% at least 1,000 reads. These figures suggested a run of good quality, as demonstrated by Average Target-Coverage plot and Covered Region-Coverage plot from SureCall QC reports.

3.3.3 Sample level quality statistics from FASTQC and BAM QC reports

FASTQ files were analysed using FastQC and BAMQC using analysis modules in Illumina Basespace, to provide further quality control checks on raw sequence data from the NGS pipeline. Sample level, post-alignment quality control reports were generated automatically by SureCall. These contain a modular set of analyses, based on different parameters, to highlight issues with the data prior to further analysis. Formal reports are presented with summary graphs and tables. There were no problems or biases detected in the sequencer or in the starting library material.

3.3.4 SureCall suite of software for sequence alignment and variant annotation

A mean of 855 variants per sample (range 725~989) were identified using the SureCall settings described in Methods (see Sections 2.7 and 2.8). Using the filtration criteria described in section 2.6, all variant calls showed extremely low P-values, indicating high confidence in the validity of the reads. All reads met the criteria for Variant Quality Score (mean 229.5, SD 51.4, range 103~255) outlined in section 2.5. A number of SNVs with small variant read number and low VAF showed reads of dubious significance. In particular, a series of 12 *U2AF1* homozygous SNV calls (at 21:44513243 G>A) (VAF 0.0205-0.0532) were removed despite having a COSMIC reference. Also, nine homozygous *SETBP1* exon 4 tetranucleotide insertions (all at 18:42456664) and three *SMC1A* exon 2 substitutions with significant strand bias, particularly at low read depth or VAF, were eliminated from further consideration. The strategy detected a number of reads of VAF of less than 5%. These were not detected by Ion Torrent PGM sequencing, and are therefore not validated and remain classified as Variants of Uncertain Significance. It is possible that they are subclonal variants that may emerge as significant disease clones later in disease course, but it is likely they are artefacts of sequencing and will not be considered as clinically significant at this time. *NPM1* and *CEPBA* indels showed multiple low VAF variants at similar loci and were assumed to be misreads of the same source sequence and the apparently aberrant reads were removed (see Indel section below). Low level variants were retained if they had a COSMIC annotation suggesting recurrence as a potentially pathogenic variant in a haematological tumour.

3.3.5 Variant Annotation from SureCall, Pindel and Lumpy alignments

Table 3.6 shows all variants that are classified as pathogenic or likely to be of clinical significance. These variants had not been detected previously from conventional testing at diagnosis. The indels and gene fusions are also included in Table 3.6, some of which were detected previously, if specifically tested or identified in the karyotype.

A total of 196 variants were retained after the initial filtering process. Each variant was then reviewed in detail and characterised using COSMIC, published literature and other on line resources as necessary. Variants were annotated according to recently published guidelines (Li *et al.*, 2017) and were curated as Tier I or Tier II variants based on Levels of evidence A – D (see Table 3.5; Categories of Evidence for Somatic Variant Interpretation). Variants of Uncertain Significance (VUS) and Benign variants were excluded from further consideration and not included in the variant Table (see Table 3.6). Variant passing assessment filters and annotated as strong or potential clinical significance (Tier I or Tier II) from the three sets of data from the analytical pipeline were assimilated (Table 3.6). 143 genetic abnormalities were retained after variant annotation described in section 2.5.2.3. This included 46 Level A variants (Tier 1 Variants of Strong Clinical Significance). 7 *IDH1/2* known mutations (2x *IDH1* R132, 4x *IDH2* R140Q and 1x *IDH2* R172K) were interpreted as Level C evidence for diagnostic significance and putative therapeutic targets. The remaining 90 variants were assigned as Level D markers that may assist diagnosis, alone or with other mutations (Li *et al.*, 2017). All seven gene fusions found by standard techniques were also detected by the experimental NGS. Other recurrent abnormalities included thirteen *NPM1* exon 11 tetranucleotide insertions, ten *FLT3*-ITD, four *FLT3*-TKD and fifteen *CEBPA* mutations, five were heterozygous single mutations, two were homozygous mutation and six were double mutation in three patients. The two samples from the same patient (samples 10 & 14) showed the same *TP53* exon 8 mutation.

Eleven variants were called at a VAF of less than 10%. Three variants were not covered in the Ion Torrent panel and so could not be confirmed. 4 mutations were identified by both NGS methods, with the smallest VAF of 2.4%; however, the level detected by Ion Torrent was reported as 7.5%. An *ASXL1* SNV at VAF 0.1% could not be confirmed by Ion

Torrent. Double *KIT* mutations (in patient 13.1760) at 1.8% and 1.5% were also not confirmed by Ion Torrent, but were found in a typical setting, secondary to *CBFB-MYH11* fusion.

Overall, 105 out of 112 variants (93.8%) that were sequenced by both NGS systems were in agreement. Two mutations (1x *DNMT3A* at VAF 0.18 and 1x *RUNX1* at VAF 0.15) could not be confirmed by Ion Torrent. The *DNMT3A* was an insertion of 22 nucleotides in exon 14 and RUNX1 a single base pair in exon 6. Two *CEBPA* variants were not called by Ion Torrent, in sample 27, who was also showed to have a 6bp insertion. The sequencing anomalies and *CEPBA* sequencing will be discussed. The Variant allele Frequency (VAF) of selected variants ranged from 0.0105 – 0.969, with a median of 0.444 and mode 0.45 – 49 (see Figure 3.4).

Table 3.5. Categories of clinical and/or experimental evidence for Somatic Variant Interpretation (adapted from Li *et al.*, 2017)

Tier	Category	Therapeutic	Diagnosis	Prognosis
I	Level A	1. Biomarkers that predict response or resistance to FDA-approved therapies for a specific type of tumour 2. Biomarkers included in professional guidelines that predict response or resistance to therapies for a specific type of tumour	Biomarkers included in professional guidelines as diagnostic for a specific type of tumour	Biomarkers included in professional guidelines as prognostic for a specific type of tumour
	Level B	Biomarkers that predict response or resistance to therapies for a specific type of tumour based on well-powered studies with consensus from experts in the field	Biomarkers of diagnostic significance for a specific type of tumour based on well-powered studies with consensus from experts in the field	Biomarkers of prognostic significance for a specific type of tumour based on well-powered studies with consensus from experts in the field
II	Level C	1. Biomarkers that predict response or resistance to therapies approved by the FDA or professional societies for a different type of tumour 2. Biomarkers that serve as inclusion criteria for clinical trials	Biomarkers of diagnostic significance based on the results of multiple small studies	Biomarkers of prognostic significance based on the results of multiple small studies
	Level D	Biomarkers that show plausible therapeutic significance based on preclinical studies	Biomarkers that may assist disease diagnosis themselves or along with other biomarkers based on small studies or a few case reports	Biomarkers that may assist disease prognosis themselves or along with other biomarkers based on small studies or a few case reports
III	Variants of Unknown Significance	Not observed at a significant allele frequency in the general or specific subpopulation databases, or pan-cancer or tumour specific variant databases No convincing published evidence of cancer association		
IV	Benign or Likely Benign Variants	Observed at significant allele frequency in the general or specific subpopulation databases. No existing published evidence of cancer association		

Table 3.6. Mutations, indels and gene fusions called by the analysis pipeline

Sample No.	Impacted Gene	HGVS(Genomic)	HOM/HET	Position	Allele Frequency	No. of Variant Alleles	Filtered Read Depth (per sample)	AA	Chromosome	Exon ID	AMP Category	AMP Tier	Ion Torrent
1	NRAS	NC_000001.10:g.115256529T>G	HOM	115256529	0.152	197	1299	Q61P	1	3	D	II	✓
2	DNMT3A	NC_000002.11:g.25466797C>T	HET	25466797	0.603	620	1028	V636M	2	16	D	II	✓
	IDH2	NC_000015.9:g.90631838C>T	HET	90631838	0.461	428	929	R172K	15	4	C	II	✓
3	CEBPA (smx2)		HOM	33793252	0.799	295	369		19	1	A	I	✓
	TET2	NC_000004.11:g.106193748C>T	HET	106193748	0.443	430	970	R1404*	4	10	D	II	✓
	TET2		HET	106156042	0.465	862	1853	S315	4	3	D	II	✓
4	FLT3-ITD	ins 57	HET	28608248		108			13	14	A	I	n/a
	DNMT3A	NC_000002.11:g.25457242C>T	HET	25457242	0.429	471	1097	R882H	2	23	D	II	✓
	TET2	NC_000004.11:g.106180777A>T	HET	106180777	0.444	575	1295	R1269*	4	7	D	II	✓
	BCOR		HET	39932538	0.443	537	1212		X	4	D	II	n/a
5	NPM1	TCTG (Type A) insertion	HET	170837543	0.51	241	473		5	11	A	I	✓
	FLT3-ITD	ins 54	HET	28608220		895			13	14	A	I	n/a
	TET2		HET	106155748	0.486	858	1766	S217	4	3	D	II	✓
	TET2	NC_000004.11:g.106180795G>T	HET	106180795	0.469	549	1171	G1275W	4	7	D	II	✓
6	CEBPA (dm)		HET	33792981	0.514	89	173		19	1	A	I	✓
	CEBPA		HET	33792731	0.405	30	74	-183TR	19	1	A	I	✓
	IDH2	NC_000015.9:g.90631934C>T	HET	90631934	0.418	410	981	R140Q	15	4	C	II	✓
	ASXL1		HET	31022441	0.272	237	872	G641	20	12	D	II	✓
	STAG2		HET	123210249	0.376	600	1595	A867	X	26	D	II	✓
7	NPM1	TGTG insertion	HET	170837543	0.504	288	571		5	11	A	I	✓

	NRAS	NC_000001.10:g.115258744C>T	HET	115258744	0.261	568	2180	G13D	1	2	D	II	✓
	SMC3	NC_000010.10:g.112356176G>T	HOM	112356176	0.0656	131	1998	G662C	10	19	D	II	n/a
	CEBPA sm	NC_000019.9:g.33793394C>T	HET	33792731	0.294	75	255	G11D	19	1	D	II	✓
	DNMT3A		HOM	25467472	0.166	268	1614	-346TTTTATC	2	14	D	II	X
	FLT3	NC_000013.10:g.28592612T>C	HOM	28592612	0.0959	175	1824	R845G	13	20	D	II	✓
8	NPM1	TGCA insertion	HET	170837545	0.522	275	527		5	11	A	I	✓
	IDH2	NC_000015.9:g.90631934C>T	HET	90631934	0.47	471	1003	R140Q	15	4	C	II	✓
	CEBPA sm		HET	33792731	0.296	45	152	-183TR	19	1	D	II	✓
9	PML-RARA	t(15;17)(q24.1;q21.1)		74326159::38493875		759					A	I	n/a
	KRAS	NC_000012.11:g.25380276T>C	HET	25380276	0.439	849	1936	Q61R	12	3	D	II	✓
10	TP53	NC_000017.10:g.7577108C>A	HET	7577108	0.614	583	949	C238F	17	8	A	I	✓
11	DNMT3A	NC_000002.11:g.25469528A>C	HET	25469528	0.533	602	1129	F414V	2	10	D	II	✓
	DNMT3A	NC_000002.11:g.25457249T>C	HET	25457249	0.47	555	1181	M880V	2	23	D	II	✓
	IDH2	NC_000015.9:g.90631934C>T	HET	90631934	0.444	477	1075	R140Q	15	4	C	III	✓
	PTPN11	NC_000012.11:g.112888165G>C	HET	112888165	0.426	831	1952	D61H	12	3	D	II	✓
	SMC1A		HOM	53432200	0.855	796	931	E679-	X	12	D	II	n/a
	CEBPA sm		HET	33792555	0.422	588	1394		19	1	D	II	✓
12	NPM1	TATT insertion	HET	170837543	0.502	298	594		5	11	A	II	✓
	IDH1	NC_000002.11:g.209113113G>A	HET	209113113	0.476	503	1056	R132C	2	4	C	II	✓
	NRAS	NC_000001.10:g.115258748C>T	HET	115258748	0.444	1014	2286	G12S	1	2	D	II	✓
	CEBPA sm		HET	33792357	0.444	1096	2466		19	1	D	II	✓
	ASXL1		HOM	31022441	0.0105	11	1051	G641	20	12	D	II	X
13	NUP98-DDX10	inv(11)(p15q22)		3758130::108549638		230					D	II	n/a
	NRAS	NC_000001.10:g.115258744C>T	HET	115258744	0.292	626	2147	G13D	1	2	D	II	✓
	TET2		HET	106197269	0.435	705	1622	H1868	4	11	D	II	✓
14	TP53	NC_000017.10:g.7577108C>A	HET	7577108	0.375	637	1700	C238F	17	8	A	I	✓
15	NPM1	TCTG (Type A) insertion	HET	170837543	0.52	263	506		5	11	A	I	✓

	TET2		HET	106196829	0.464	896	1932	L1721W	4	11	D	II	✓
	IDH1		HET	209108317	0.46	858	1864	V178I	2	6	Benign	IV	✓
	DNMT3A		HET	25457242	0.414	439	1060	R882H	2	23	D	II	✓
	PTPN11		HET	112888189	0.343	651	1896	E69K	12	3	D	II	✓
	TET2		HET	106156187	0.487	931	1911	P363L	4	3	D	II	✓
	TET2		HET	106180852	0.473	696	1473		4	7	D	II	✓
16	CBFB-MYH11	inv(16)(p13q22)		15814603::67116255		541					A	I	n/a
	NRAS	NC_000001.10:g.115256529T>C	HOM	115256529	0.108	134	1243	Q61R	1	3	D	II	✓
	KRAS		HOM	25398287	0.034	59	1735	-11E	12	2	D	II	✓
	KIT		HOM	55589778	0.0158	27	1714	R420	4	8	VUS	III	X
	KIT		HOM	55589765	0.0181	27	1494	R420	4	8	VUS	III	X
17	NPM1	TCTG (Type A) insertion	HET	170837543	0.574	303	528		5	11	A	I	✓
	ASXL1	NC_000020.10:g.31023821G>T	HET	31023821	0.485	847	1746	E1101D	20	12	D	II	✓
	IDH2	NC_000015.9:g.90631934C>T	HET	90631934	0.459	428	933	R140Q	15	4	C	II	✓
	FLT3-ITD	ins 30	HOM	28608235	0.0464	73	1573	K634NEYDLKWEVPR	13	14	A	II	n/a
	TET2	NC_000004.11:g.106155199C>T	HET	106155199	0.496	877	1768	L34F	4	3	VUS	III	✓
18	NPM1	GTAG insertion	HOM	170837546	0.232	139	599		5	11	A	I	✓
19	PML-RARA	t(15;17)(q24.1;q21.1)		74315996::38488278		1097					A	I	n/a
	FLT3-ITD	ins 24 – 72		28608237-80		133			13	14	A	I	n/a
20	NPM1	TCTG (Type A) insertion	HET	170837543	0.471	252	535		5	11	A	I	✓
	DNMT3A	NC_000002.11:g.25457242C>T	HET	25457242	0.389	461	1186	R882H	2	23	D	II	✓
	FLT3-ITD	ins 36 – 87		28608231/52					13	14	A	I	✓
21	NPM1	ACTG insertion	HET	170837543	0.473	288	609		5	11	A	I	✓
	FLT3-ITD	ins 51~57		28608261					13	14	A	I	✓
	TET2	NC_000004.11:g.106196819G>T	HET	106196819	0.457	1085	2376	V1718L	4	11	D	II	✓
	DNMT3A	NC_000002.11:g.25467449C>A	HET	25467449	0.438	698	1592	G543C	2	14	D	II	✓
22	NPM1	TCTG (Type A) insertion	HOM	170837543	0.128	72	563		5	11	A	I	✓

	DNMT3A	NC_000002.11:g.25457242C>T	HET	25457242	0.378	453	1197	R882H	2	23	D	II	✓
	TET2	NC_000004.11:g.106155620C>A	HET	106155620	0.505	1074	2125	P174H	4	3	D	II	✓
	TET2		HET	106162500	0.368	778	2116	Q1138	4	4	D	II	✓
	TET2	NC_000004.11:g.106156180C>T	HOM	106156180	0.213	489	2294	Q361*	4	3	D	II	✓
	TET2		HOM	106157162	0.195	424	2174	D688	4	3	D	II	✓
23	KMT2A-MLLT1	t(11;19)(q23;p13.3)		118359297::6277438		642					A	II	n/a
	PTPN11	NC_000012.11:g.112888168T>G	HET	112888168	0.352	782	2219	Y62D	12	3	D	II	✓
	SMC1A	NC_000023.10:g.53432008C>T	HOM	53432008	0.107	98	917	R711Q	X	13	D	II	✓
24	RUNX1-RUNX1T1	t(8;21)(q22;q22.3)		93076976::36213332		897					A	I	n/a
	FLT3-ITD	ins 12	HET	28608260	0.296	461	1555		13	14	A	I	n/a
25	JAK2	g.5073770V>F	HET	5073770	0.111	114	1290		9	14	A	I	✓
	TET2		HET	106164895	0.517	391	756		4	6	D	II	✓
	TET2		HET	106158438	0.53	1011	1907		4	3	D	II	✓
	KRAS	NC_000012.11:g.25378562C>G	HET	25378562	0.289	594	2052	A146P	12	4	D	II	✓
26	DNMT3A	NC_000002.11:g.25505372G>A	HET	25505372	0.41	335	818	S129L	2	4	D	II	✓
	NRAS	NC_000001.10:g.115258748C>A	HOM	115258748	0.161	306	1900	G12C	1	2	D	II	✓
27	CEBPA dm		HET	33793002	0.528	149	282		19	1	A	I	X
	CEBPA	NC_000019.9:g.33793004A>C	HET	33793004	0.438	135	308	F106C	19	1	A	I	X
	CEBPA		HET	33792731	0.261	29	111	-183TR	19	1	A	I	✓
	WT1	NC_000011.9:g.32417947G>A	HOM	32417947	0.968	1124	1161	R639*	11	7	D	II	✓
	ASXL1		HOM	31022908	0.461	613	1329		20	12	D	II	✓
28	TET2	NC_000004.11:g.106158509G>C	HET	106158509	0.488	959	1966	G1137A	4	3	D	II	✓
	TET2	NC_000004.11:g.106156776T>A	HET	106156776	0.466	996	2138	Y559*	4	3	D	II	✓
	RUNX1	NC_000021.8:g.36259161C>A	HET	36259161	0.453	266	587	K110N	21	4	A	I	✓
	RUNX1		HOM	36252939	0.246	489	1990	S114LIGVA	21	2	A	I	✓
	RUNX1	NC_000021.8:g.36164601G>A	HOM	36164601	0.175	88	503	P398L	21	6	A	I	X
	DNMT3A	NC_000002.11:g.25457242C>A	HET	25457242	0.443	461	1040	R882L	2	23	D	II	✓

	ASXL1		HET	31022441	0.378	345	913		20	12	D	II	✓
	FLT3	NC_000013.10:g.28592629T>C	HOM	28592629	0.201	385	1911	D839W	13	14	D	II	✓
	PHF6	NC_000023.10:g.133549137G>A	HET	133549137	0.636	589	926	R274Q	X	8	D	II	n/a
	SETBP1	NC_000018.9:g.42643337C>T	HET	42643337	0.428	360	842	P1489S	18	6	D	II	n/a
	KMT2A-PTD	29.619 kbp		118326943::118356561					11		D	II	n/a
	FLT3-ITD	ins 69		28608227					13	14	A	I	n/a
29	NPM1	TCTG (Type A) insertion	HET	170837543	0.603	334	554		5	11	A	I	✓
	FLT3-ITD	ins 36		28608266					13	14	A	I	n/a
	DNMT3A	NC_000002.11:g.25457243G>A	HET	25457243	0.463	537	1160	R693C	2	23	D	II	✓
	FLT3	NC_000013.10:g.28592642C>A	HOM	28592642	0.151	317	2097	D835Y	13	20	D	II	✓
	PTPN11	NC_000012.11:g.112926887G>A	HOM	112926887	0.0369	76	2059	G503R	12	13	D	II	✓
	SMC1A	NC_000023.10:g.53436051C>T	HOM	53436051	0.949	391	412	R496H	X	9	D	II	n/a
	FAM5C	NC_000001.10:g.190068080G>T	HET	190068080	0.455	679	1492	R457S	1	8	D	II	n/a
	KRAS	NC_000012.11:g.25398284C>A	HOM	25398284	0.0242	43	1780	G12	12	2	D	II	✓
30	CBFB-MYH11	t(16;16)(p13;q22)		15815105::67127201			1588				A	I	✓
31	NPM1	TGCA insertion	HET	170837545	0.494	344	697		5	11	A	I	✓
	FLT3-ITD	Ins 24	HOM	28608259	0.107	180	1683	I626IVDFREYEE	13	14	A	I	n/a
	IDH1	NC_000002.11:g.209113112C>T	HOM	209113112	0.0554	71	1281	R132H	2	4	C	II	✓
32	NPM1	GCCA insertion	HET	170837546	0.505	380	753		5	11	A	II	✓
	DNMT3A	NC_000002.11:g.25457242C>T	HET	25457242	0.487	538	1105	R882H	2	23	D	II	✓
	CEBPA sm		HET	33792731	0.304	38	125	-183TR	19	1	D	II	✓
	FLT3-ITD	ins 42		28608263		152			13	14	A	I	n/a
33	CBFB-MYH11	inv(16)(p13q22)		15814942::67118135		1114					A	I	n/a
	RUNX1	NC_000021.8:g.36164646G>A	HET	36164646	0.436	246	564	S410L	21	9	D	II	✓
34	CEBPA smx2		HOM	33793252	0.778	309	397		1	1	A	I	✓
	RUNX1		HET	36252855	0.512	662	1294		21	9	D	II	✓
	TET2		HET	106197114	0.512	701	1369		4	11	D	II	✓

	ASXL1		HET	31022419	0.51	469	920		20	12	D	II	✓
	SETBP1	NC_000018.9:g.42643142A>G	HET	42643142	0.476	728	1531	K1424E	18	6	D	II	n/a
	ASXL1	NC_000020.10:g.31022422C>A	HET	31022422	0.435	327	752	A635E	20	12	D	II	✓
35	TET2		HET	106190836	0.49	689	1407		4	9	D	II	✓
	TET2	NC_000004.11:g.106164913C>T	HET	106164913	0.447	257	575	R1261C	4	6	D	II	✓
	DNMT3A	NC_000002.11:g.25463511T>C	HET	25463511	0.488	440	901	Y724C	2	18	D	II	✓
	DNMT3A	NC_000002.11:g.25463287G>A	HET	25463287	0.466	428	918	R736C	2	19	D	II	✓
	U2AF1	NC_000021.8:g.44524456G>A	HET	44524456	0.456	718	1576	S34F	21	2	D	II	n/a
	KMT2A-PTD	9.827 kbp		118341773::118351603							D	II	n/a
36	CEBPA dm		HET	33792381	0.472	1135	2403	K194-	19	1	A	I	✓
	CEBPA		HET	33793227	0.458	330	720		19	1	A	I	✓
	EZH2	NC_000007.13:g.148526829C>T	HOM	148526829	0.0551	90	1633	G159R	7	5	D	II	n/a
	EZH2	NC_000007.13:g.148512610G>A	HET	148512610	0.55	560	1018	Q512*	7	13	D	II	n/a
	GATA2	NC_000003.11:g.128202732G>C	HET	128202732	0.462	425	919	R330G	3	4	D	II	n/a
	SMC3	NC_000010.10:g.112333474A>G	HET	112333474	0.44	606	1377	N34S	10	3	D	II	n/a

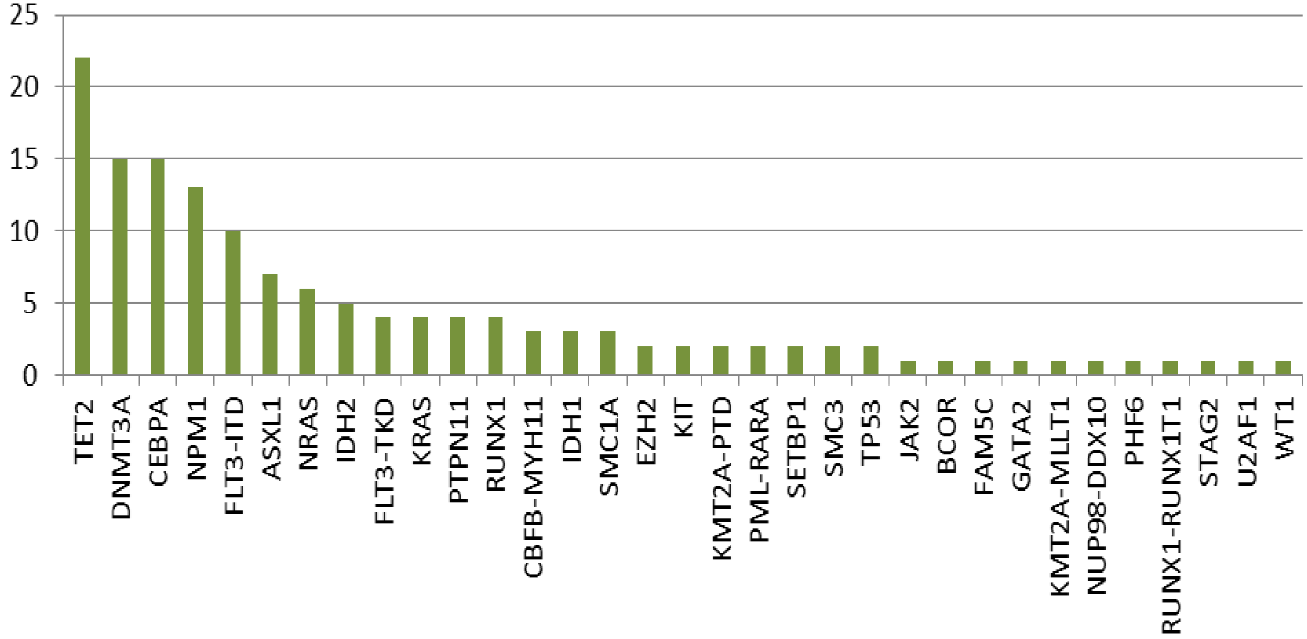


Figure 3.3. The frequency of mutated genes and gene fusions detected by the variant calling pipeline and the number of samples with each mutation.

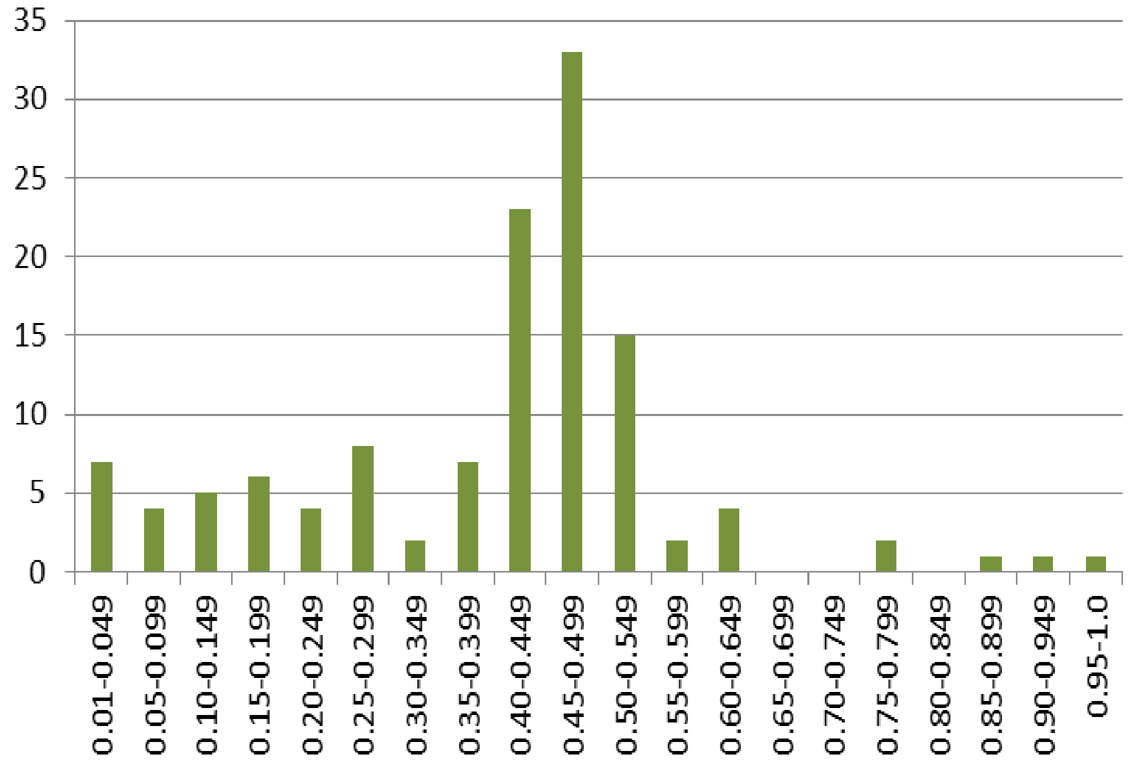


Figure 3.4. Distribution of Variant Allele Fraction (VAF) of Single Nucleotide Variants called by SureCall

3.3.6 Comparison of SureCall to GATK Haplotyper and Ion Torrent AML Ampliseq

There was broad but not complete agreement between the alternative bioinformatic pipelines used for SNV and short indel detection, BWA-MEM (used in SureCall for the experiment) and GATK Haplotyper hard-filtered data. The most significant anomaly was an *NRAS* variant in sample 13.0816 (NC_000001.10:g.115256529T>G @VAF 0.152) identified by SureCall but not GATK. This was confirmed to be present by Ion Torrent and is therefore a result of misalignment or inappropriate filtering by GATK. Generally, the low level variants at <3%, allowed by SureCall, were not detected by GATK, in its default setting to avoid sequencing errors being called as variants.

Sequencing of the samples was also performed on an Ion Torrent PGM (PGM), using the AML Ampliseq panel, as a comparison of the experimental method to alternative (amplicon) DNA library preparation and sequencing (semi-conductor) technology, to aid validation of the results. The AML Ampliseq panel covers mutation hotspots or entire coding regions in 19 disease-associated genes, limiting the comparison to the most clinically relevant variants and excludes *FLT3*-ITD (see Table 3.7). The PGM detected a *JAK2* V617F mutation (chr9:5073770) in sample 25, which was not part of our panel but would be included in a more comprehensive diagnostic panel. This mutation was confirmed by PCR testing in a separate diagnostic work-up and will be included in further analysis. A median of 26 variants per sample were detected by the PGM pipeline pre-filtering (range 21-27).

Table 3.7 Gene locations covered by the Ion Torrent AML Ampliseq panel

Gene	Chromosome	Hotspot	Human Feb. 2009 (GRCh37/hg19) Assembly		Target (bp)	Missed	Coverage %	Amplicons
			Start	**End**				
ASXL1	chr20	exon 12	31022215	31025161	2947	0	100	28
BRAF	chr7	exon 15	140453131	140453141	11	0	100	1
CBL	chr11	exon 8	119148856	119149027	172	0	100	2
		exon 9	119149200	119149443	244	0	100	3
CEBPA	chr19	all coding	33792224	33793340	1117	0	100	9
DNMT3A	chr2	all coding	25457128	25536873	3,619	0	100	42
FLT3	chr13	exon 21	28589830	28589838	9	0	100	1
		exon 20	28592604	28592657	54	0	100	1
		exon 16	28602340	28602342	3	0	100	1
GATA2	chr3	all coding	128199842	128205894	1,643	0	100	20
IDH1	chr2	exon 4	209113073	209113404	332	0	100	3
IDH2	chr15	exon 4	90631799	90631999	201	0	100	2
JAK2	chr9	exon 14	5073678	5073805	128	0	100	1
KIT	chr4	exon 8	55589730	55589884	155	0	100	1
		exon 10	55593364	55593510	147	0	100	3
		exon 11	55593562	55593728	167	0	100	3
		exon 17	55599216	55599378	163	0	100	2
KRAS	chr12	exon 3	25380148	25380366	219	0	100	2
		exon 2	25398188	25398338	151	0	100	2
NPM1	chr5	exon 11	170837511	170837589	79	0	100	1
NRAS	chr1	exon 3	115256401	115256619	219	0	100	2
		exon 2	115258651	115258801	151	0	100	1
PTPN11	chr12	exon 3	112888102	112888336	235	0	100	3
		exon 13	112926808	112926999	192	0	100	2
RUNX1	chr21	exon 8	36171578	36171779	202	0	100	3
		exon 7	36206687	36206918	232	0	100	3
		exon 6	36231751	36231895	145	0	100	3
		exon 5	36252834	36253030	197	0	100	2
		exon 4	36259120	36259413	294	0	100	3
		exon 3	36265202	36265280	79	0	100	1
TET2	chr4	all coding	106155080	106197696	6,369	0	100	59
TP53	chr17	exon 12	7572907	7573028	122	7	94.26	2

		all coding	7573907	7578574	1146	0	100	19
		exon 4	7579282	7579610	329	15	95.44	4
		exon 3	7579680	7579741	62	18	70.97	1
		exon 2	7579819	7579932	114	18	84.21	1
WT1	chr11	exon 9	32413498	32413630	133	0	100	2
		exon 7	32417783	32417973	191	0	100	2

The different bioinformatics pipelines yielded different variant calls and a number from Ion Torrent sequencing, selected by Ion Reporter software, were filtered by the custom protocol as normal variation. Of the most significant genes that passed filtering by Sure Call, *CEBPA,* and *FLT3*-ITD are discussed below.

3.3.7 Pindel for ITD classification

Pindel (Ye *et al.*, 2009) was used for the detection of large indels and showed no false positive calls, compared to standard testing. 13 *NPM1* exon 11 mutants were included in the dataset and all were identified by next generation sequencing by both SureCall and Pindel programmes and Ion Torrent sequencing (see Table 3.8). The different alignment algorithms and platforms identified mutations at the same breakpoints but there was a difference in the tetranucleotide sequences; the Pindel alignment and Ion Torrent sequencer agreeing on all sequences which were discordant with SureCall in 3 samples. 9 out of 13 *NPM1*+ cases were of the common Type A TCTG insertion and therefore the subtype was misclassified in three cases by SureCall.

CEBPA indels were detected in 9 patients There was general agreement between the methodologies for detection of *CEBPA* indels, with 6 producing identical mutations and two samples with heterozygous single mutations (3 and 34) yielding a 2 base pair difference in insertion position. Sample 37 showed different mutations with different methods, a GCGGGT insertion detected by SureCall and Ion Torrent (but not Pindel) and a 39bp insertion which was detected by Pindel and no others. A single bp insertion and base substitution was detected only by SureCall. These appear to be a double mutation which shows the limitation of different methodologies to detect variants of different sizes (see also *FLT3*-ITD below).

All eleven *FLT3*-ITD mutations were correctly identified by Pindel but only three of the smaller ones (between 12-30bp) were detected by SureCall and the larger ones (30-87bp) were not. Multiple reads were generated for three samples (19, 20 and 21) which overlapped and probably represent sequencing misrepresentation of the same large insertions or actual subclonal variation. This cannot be resolved by this study and an alternative technique such as Sanger sequencing would be required. Pindel detected two *KMT2A*-PTD in sample 28 between introns 1 and 10, resulting in a 29.619 kbp duplication of exons 2-10 and in sample 35 between introns 2 and 6, resulting in a 9.827 kbp duplication of exons 3-6 (see Table 3.11). This was also detected by *Lumpy* (with slightly different intron 6 breakpoint (11:118351603). The fusion breakpoint could be retrospectively detected in the SureCall data and visualisable in IGV (see Figures 3.3a and 3.3b).

Table 3.8. *NPM1* exon 11 tetranucleotide insertions (NM_002520.ex.11)

							No. Supporting Reads		
Sample no.	**Pindel sequence**	**SureCall sequence**	**Ion Torrent**	**Insertion type**	**Zygosity**	**Breakpoint**	**Pindel**	**SureCall**	**Ion Torrent**
5	TCTG	TCTG	TCTG	Type A	HET	170837543	107	241	866
7	TCTG	TGTG	TCTG	Type A	HET	170837543	137	288	957
8	TGCA	TGCA	TGCA		HET	170837545	122	275	907
12	TCTG	TATT	TCTG	Type A	HET	170837543	151	298	838
15	TCTG	TCTG	TCTG	Type A	HET	170837543	113	263	
17	TCTG	TCTG	TCTG	Type A	HET	170837543	147	303	
18	GTAG	GTAG	GTAG		HOM	170837546	44	139	406
20	TCTG	TCTG	TCTG	Type A	HET	170837543	105	252	
21	TCTG	ACTG	TCTG	Type A	HET	170837543	150	288	
22	TCTG	TCTG	TCTG	Type A	HOM	170837543	37	72	
29	TCTG	TCTG	TCTG	Type A	HET	170837543	160	334	
31	TGCA	TGCA	TGCA		HET	170837545	140	344	848
32	GCCA	GCCA	GCCA		HET	170837546	179	380	968

Table 3.9. *CEBPA* variants

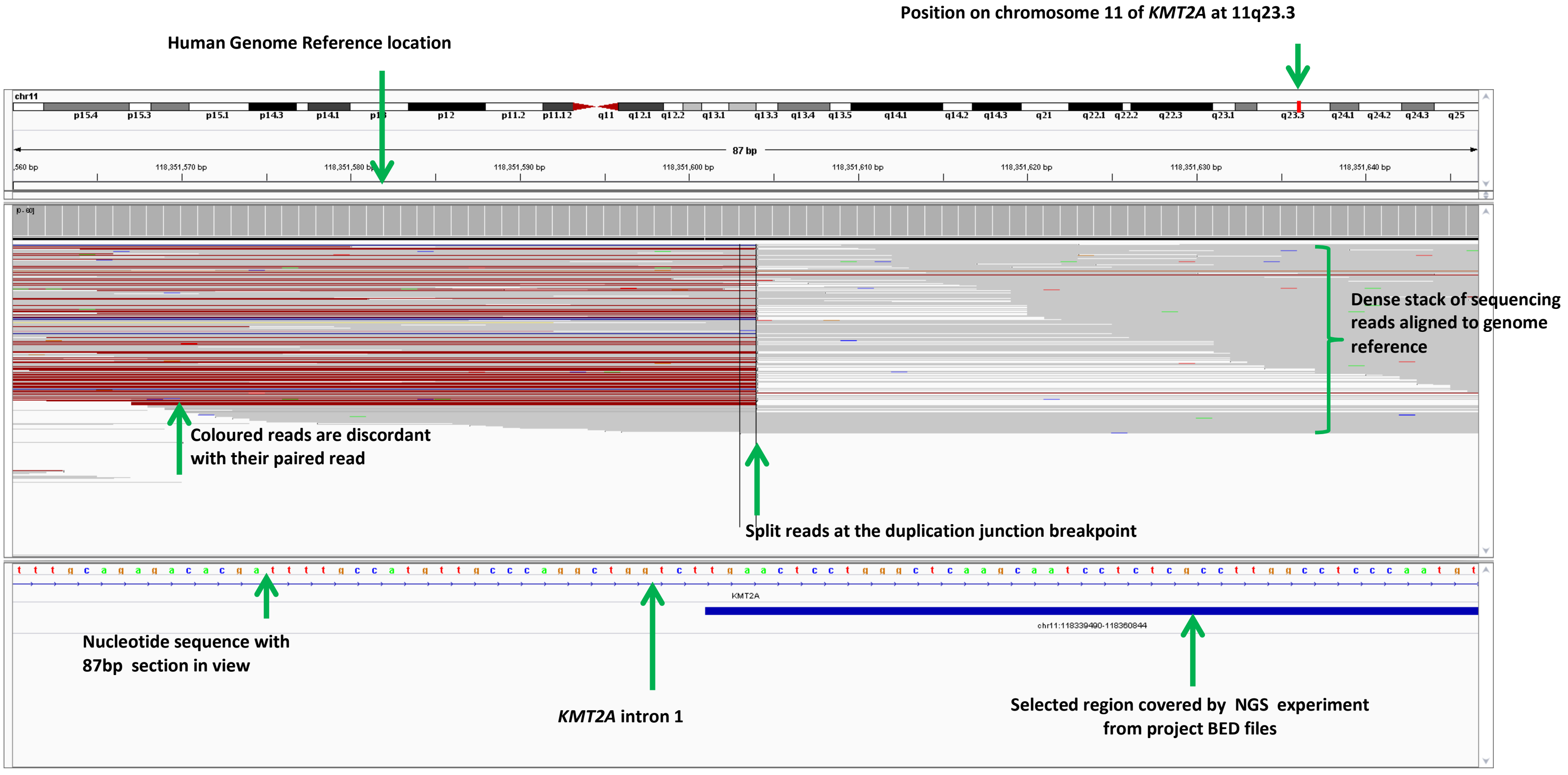
Sample No.	No. Nucleotides	Insertion	Zygosity	Pindel Breakpoint	SureCall Breakpoint	Ion Torrent	SureCall No. Alleles	Pindel No. Alleles
3	ins 1	G	HOM	33793252	33793252	33793252	295	145
6	ins 6	GCGGGT	HET	33792731	33792731	n/a	30	28
6	del 8	CGCGGGCG	HET	33792981	33792981	n/a	89	
7	ins 6	GCGGGT	HET	33792731	33792731	33792729	45	36
11	ins 1	C	HET	33792555	33792555	33792555	588	400
12	ins 1	A	HET	33792357	33792357	33792355	1096	650
27	Ins 39	GTCACTGGTCAGC TCCAGCACCTTCT GCTGCGTCTCCAC	**n/a**	33792360	-	-	-	519
27	ins 6	GCGGGT	HET	-	33792731	33792729	29	-
27	ins 2	AG	HET	-	33793002	-	149	-
27	SNP	A>C	HET	-	33793004	-	135	-
32	ins 6	GCGGGT	HET	33792731	33792731	33792731	38	24
34	ins 1	G	HOM	33793252	33793252	33793252	309	162
36	del 3	CTT	HET	-	33792381	33792381	1135	-
36	del 1	G	HET	-	33793227	33793227	330	-

Table 3.10. *FLT3*-ITD variants

Sample No.	Nucleotides inserted	Sequence inserted	Breakpoint	Detected SureCall	SureCall bp	Pindel Supporting Reads
4	ins 57	ATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGG	28608248	X		108
5	ins 54	AACTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTG	28608220	X		183
17	ins 30	CTTGGAAACTCCCATTTGAGATCATATTCA	28608235	✓	28608235	226
19	ins 66	CTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCC	28608243	X		57
	ins 72	TGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCC	28608237	X		40
	ins 30	ACGTAGAAGTACTCATTATCTGAGGAGCCG	28608280	X		24
	ins 24	TCTCTGAAATCAACGTAGAAGGGG	28608268	X		12
20	ins 36	TTCTCTTGGAAACTCCCATTTGAGATCATATTCATA	28608231	X		161
	ins 87	GAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCCGGTCACCTGTACCATCTGTAGCTGGCTTTC	28608252	X		14

Table 3.10. *FLT3*-ITD variants (cont.)

Sample No.	Nucleotides inserted	Sequence inserted	Breakpoint	Detected SureCall	SureCall bp	Pindel Supporting Reads
21	ins 57	TTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCCGGTCACCTG	28608261	X		185
	ins 51	GAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGA	28608252	X		33
24	ins 12	ATTCATATTCTA	28608260	✓	28608260	385
28	ins 69	AATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCGG	28608227	X		12
29	ins 36	ATTCTCTGAAATCAACGTAGAAGTACTCATTATCTG	28608266	X		6
31	ins 24	TATTCATATTCTCTGAAATCAACG	28608259	✓	28608259	389
32	ins 42	CATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGG	28608263	X		152


chr11
p15.4 p15.3 p15.1 p14.3 p14.1 p13 p12 p11.2 p11.12 q11 q12.1 q12.2 q13.1 q13.3 q13.4 q13.5 q14.1 q14.2 q14.3 q21 q22.1 q22.2 q22.3 q23.1 q23.3 q24.1 q24.2 q24.3 q25
87 bp
,560 bp
118,351,570 bp
118,351,580 bp
118,351,590 bp
118,351,600 bp
118,351,610 bp
118,351,620 bp
118,351,630 bp
118,351,640 bp
[0 - 60]
t t t g c a g a g a c a c g a t t t t g c c a t g t t g c c c a g g c t g g t c t t g a a c t c c t g g g c t c a a g c a a t c c t c t c g c c t t g g c c t c c c a a t g t
KMT2A
chr11:118339490-118360844

Position on chromosome 11 of *KMT2A* at 11q23.3

Human Genome Reference location

Dense stack of sequencing reads aligned to genome reference

Coloured reads are discordant with their paired read

Split reads at the duplication junction breakpoint

Nucleotide sequence with 87bp section in view

KMT2A intron 1

Selected region covered by NGS experiment from project BED files

Figure 3.3a Integrated Genome Viewer screenshot of region of chromosome 11 (chr11:118,351,560-118,351,645) showing track of read alignments in "squished" view, centred at 11:118,351,603. The read alignments are coloured by insert size and show the reads with discordant pairs due to *KMT2A*-PTD coloured brown, with duplication junction at 11: 11:118,351,603.

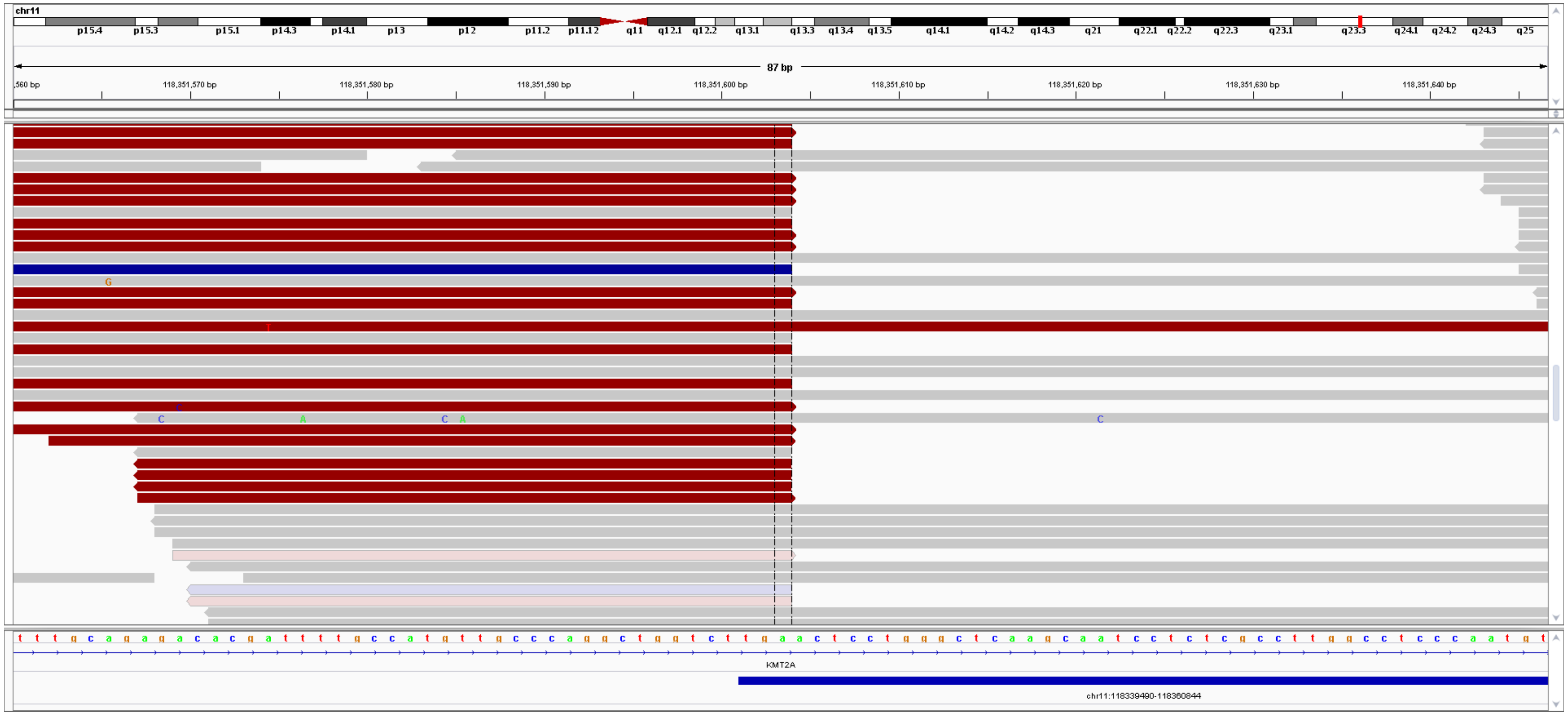


Figure 3.3b Integrated Genome Viewer screenshot of region of chromosome 11 (chr11:118,351,560-118,351,645) showing partial track of read alignments in "expanded" view, centred at 11:118,351,603. The read alignments are coloured by insert size and show the reads with discordant pairs due to *KMT2A*-PTD coloured brown, with duplication junction at 11: 11:118,351,603.

Table 3.11. ***KMT2A*** **partial tandem duplications detected by Lumpy alignment of targeted Next Generation Sequencing**

Sample No.	Conventional Result	Gene	Exons duplicated	Position	Position	Supporting Reads
28	Normal	*KMT2A*	2~10	118326943 (/49)*	118356561	111
35	-7,+8	*KMT2A*	3~6	118341773 (/72)*	118351603	175

*Numbers in brackets in breakpoint Position column represent last two digits difference in breakpoint designation in a smaller proportion of reads.

3.3.8 LUMPY for detection of gene fusions

The bioinformatics algorithm *Lumpy* was used to detect structural variation, using points of evidence from DNA sequence, including read pair and split reads (Layer *et al.*, 2014). Seven gene fusions were identified by conventional testing (see section 3.2 and Table 3.2 above) and all were detected by this next generation sequencing approach of targeting each partner gene in the translocations. This included two *PML-RARA* fusions, three *CBFB-MYH11* fusions, one *RUNX1-RUNX1T1*, and a *KMT2A-KMT2A*T1 (see Table 3.12). The fusion breakpoint could be retrospectively detected in both chromosomal breakpoints in the SureCall data and visualisable in IGV (see Figures 3.4 a and b). Slight variation in breakpoint designation was apparent between *Lumpy* alignments from different lanes of the sequencer that were not merged before alignment. Two alternative sets of molecular breakpoints are offered for inv(16) in Sample 16 and t(8;21) in Sample 24. Nevertheless, it was possible to define chromosomal rearrangements, with close to base pair accuracy, in the majority of cases or to identify the breakpoints to a narrow range. A deletion of chromosome 11 is also apparent in 13.3005, in addition to inv(16). No gene fusions were detected in any other genes tested for gene fusion in the panel.

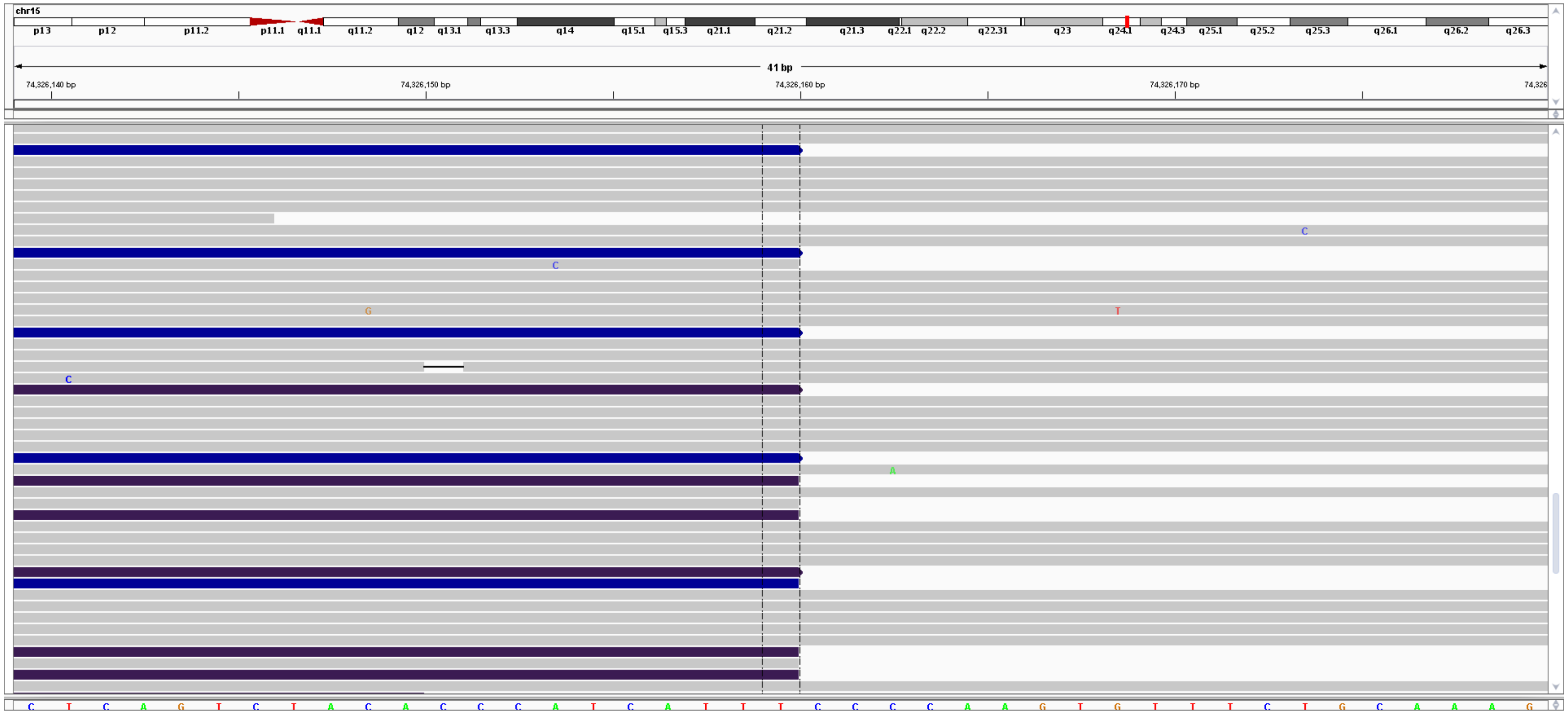
Case 13.1420 was known to have a *NUP98* gene rearrangement by FISH, apparently resulting from a large pericentric inversion of chromosome 11, a rare but recognised abnormality in AML. The partner gene of *NUP98* could be inferred from this rearrangement but had not been identified. A strategy to target all possible partner genes is costly and impractical, particularly for promiscuous genes like *NUP98*. *NUP98* was targeted as part of the panel and chimaeric reads resulting from gene fusion were captured and sequencing extended into the unidentified sequences. *DDX10* at 11q22.3 was identified as the fusion partner and the stack of reads can be shown in IGV (see Figures 3.5a and b and 3.6 a and b).

Table 3.12. Gene fusions detected by Lumpy alignment of targeted Next Generation Sequencing

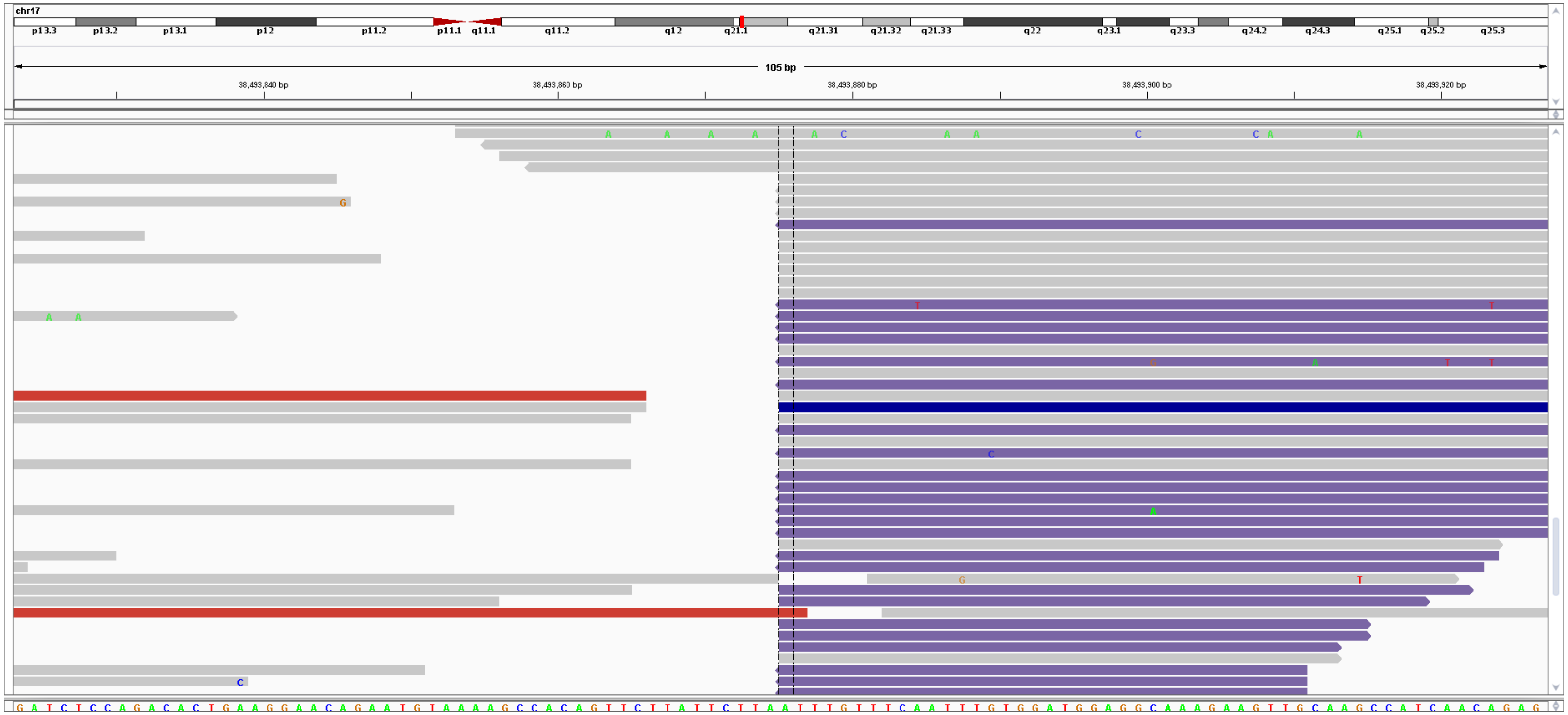
Sample No.	Conventional Result	Genes	Chr #1	Position	Chr #2	Position	Supporting Reads
9	t(15;17)(q24.1;q21.1)	*PML-RARA*	15	74326159	17	38493875	759
16	inv(16)(p13q22)	*CBFB-MYH11*	16	15814603	16	67116255	541
			16	15814963	16	67116382 (/79)*	374
19	t(15;17)(q24.1;q21.1)	*PML-RARA*	15	74315996 (/95)*	17	38488278 (/73)*	1097
23	t(11;19)(q23;p13.3)	*KMT2A-MLLT1*	11	118359297	19	6277438	640
24	t(8;21)(q22;q22.3)	*RUNX1-RUNX1T1*	8	93076976	21	36213332	897
			8	93077238	21	36183258	385
30	t(16;16)(p13;q22)	*CBFB-MYH11*	16	15815105 (/03)*	16	67127201	1588
33	inv(16)(p13q22)	*CBFB-MYH11*	16	15814942	16	67118135	1114
			11	119148585	11	119149432	565
13	inv(11)(p15q22)	*NUP98*-DDX10	11	3758130 (/29)*	11	108549638	230

Two alternative set of molecular breakpoints are offered for inv(16) in 13.1760 and t(8;21) in 13.2412. A deletion of chromosome 11 is also apparent in 13.3005, in addition to inv(16).

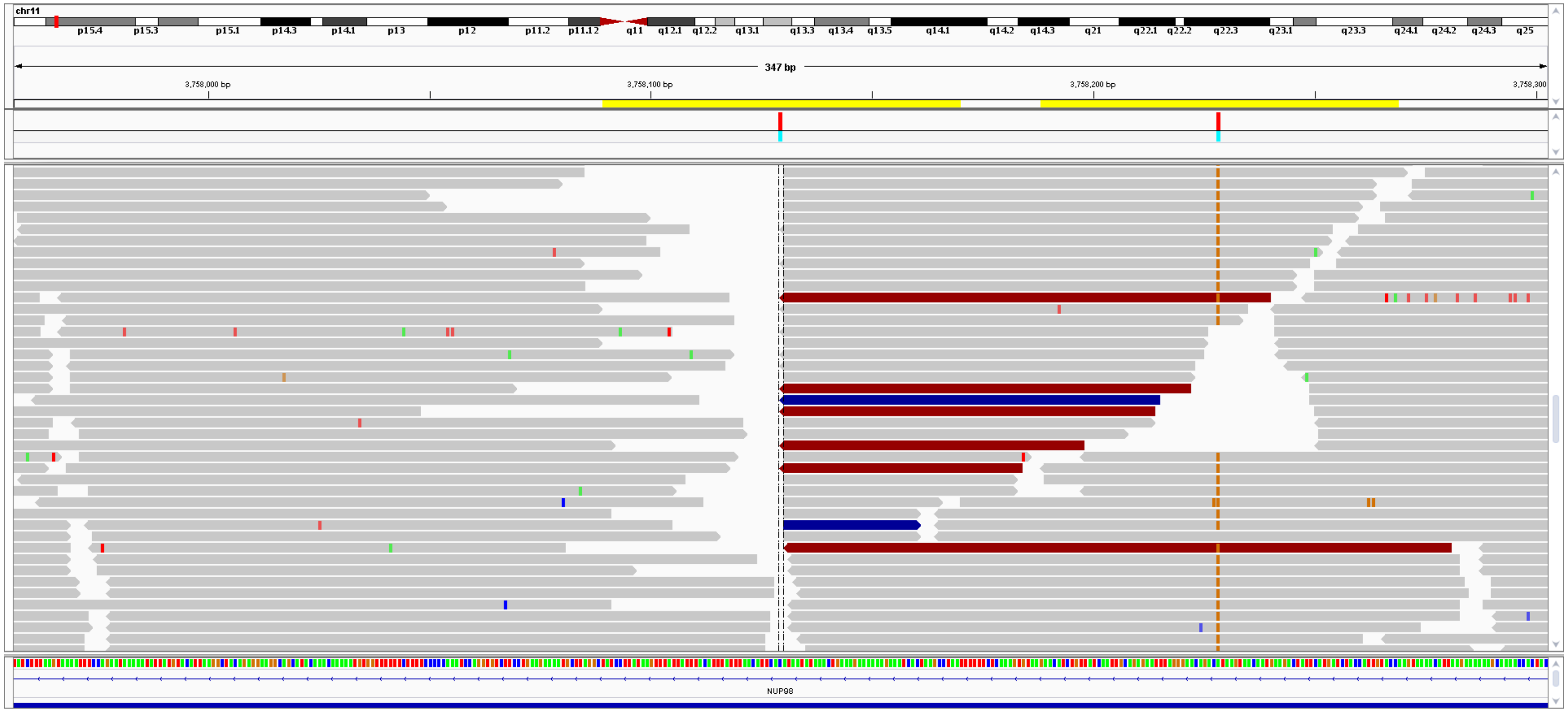
*Numbers in brackets in breakpoint Position column represent last two digits difference in breakpoint designation in a smaller proportion of reads.


chr15
p13 p12 p11.2 p11.1 q11.1 q11.2 q12 q13.1 q13.3 q14 q15.1 q15.3 q21.1 q21.2 q21.3 q22.1 q22.2 q22.31 q23 q24.1 q24.3 q25.1 q25.2 q25.3 q26.1 q26.2 q26.3
41 bp
74,326,140 bp
74,326,150 bp
74,326,160 bp
74,326,170 bp
74,326
C
G
T
C
A
C T C A G T C T A C A C C C A T C A T T T C C C C A A G T G T T T C T G C A A A G

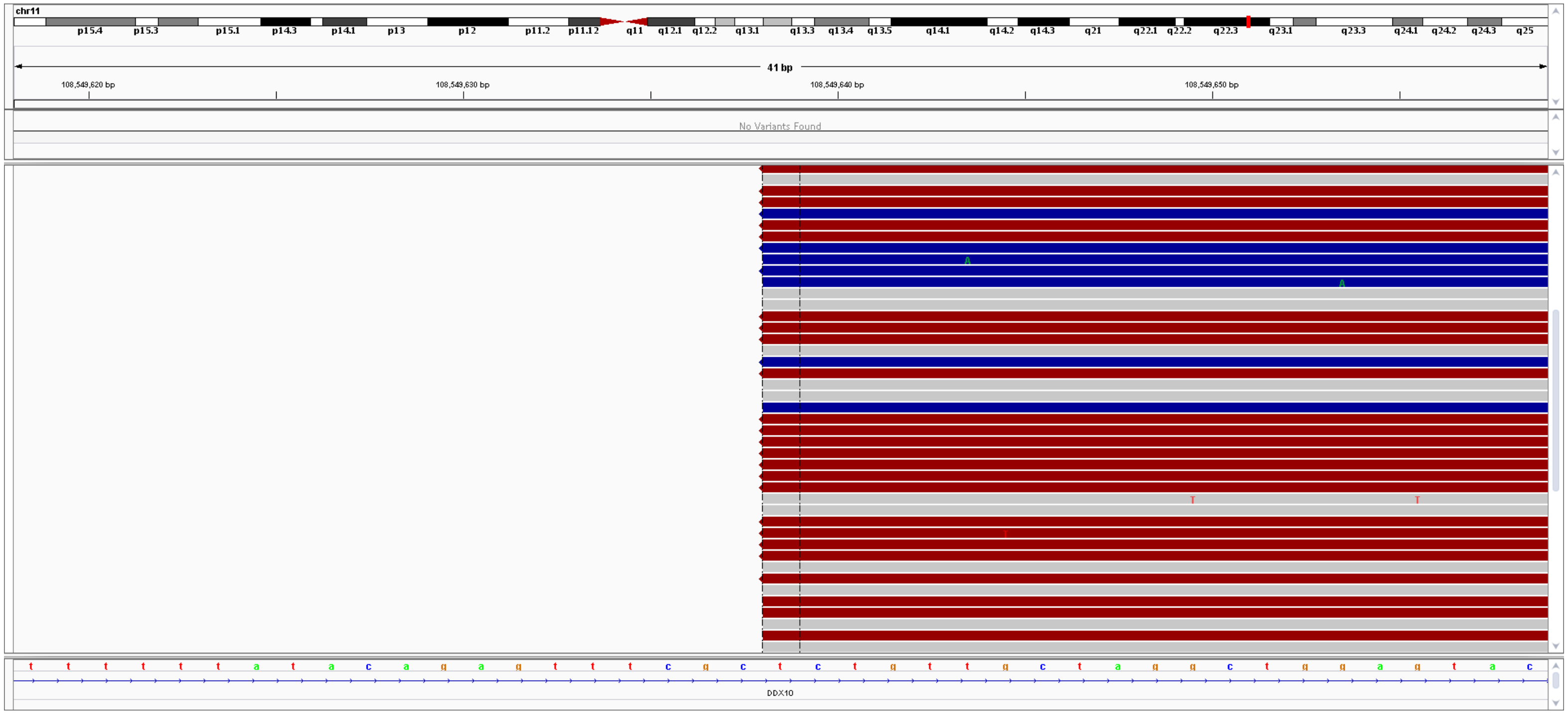
Figures 3.4a. Integrated Genome Viewer screenshots of regions of (a) chromosome 15 (chr15:74,326,139-74,326,179) and (b) **chromosome 17** (chr17:38,493,855-38,493,895) showing partial track of read alignments in "expanded" view, centred at breakpoints in intron 6 in *PML* (15: 74326159) and intron 2 in *RARA* (17:38493875) respectively. The read alignments are coloured by insert size and show the reads with discordant pairs due to *PML*-*RARA* gene fusion in case no. 13.1050.


chr17
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.1
q11.2
q12
q21.1
q21.31
q21.32
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.2
q25.3
105 bp
38,493,840 bp
38,493,860 bp
38,493,880 bp
38,493,900 bp
38,493,920 bp
G A T C T C C A G A C A C T G A A G G A A C A G A A T G T A A A A G C C A C A G T T C T T A T T C T T A A T T T G T T T C A A T T T G T G G A T G G A G G C A A A G A A G T T G C A A G C C A T C A A C A G A G

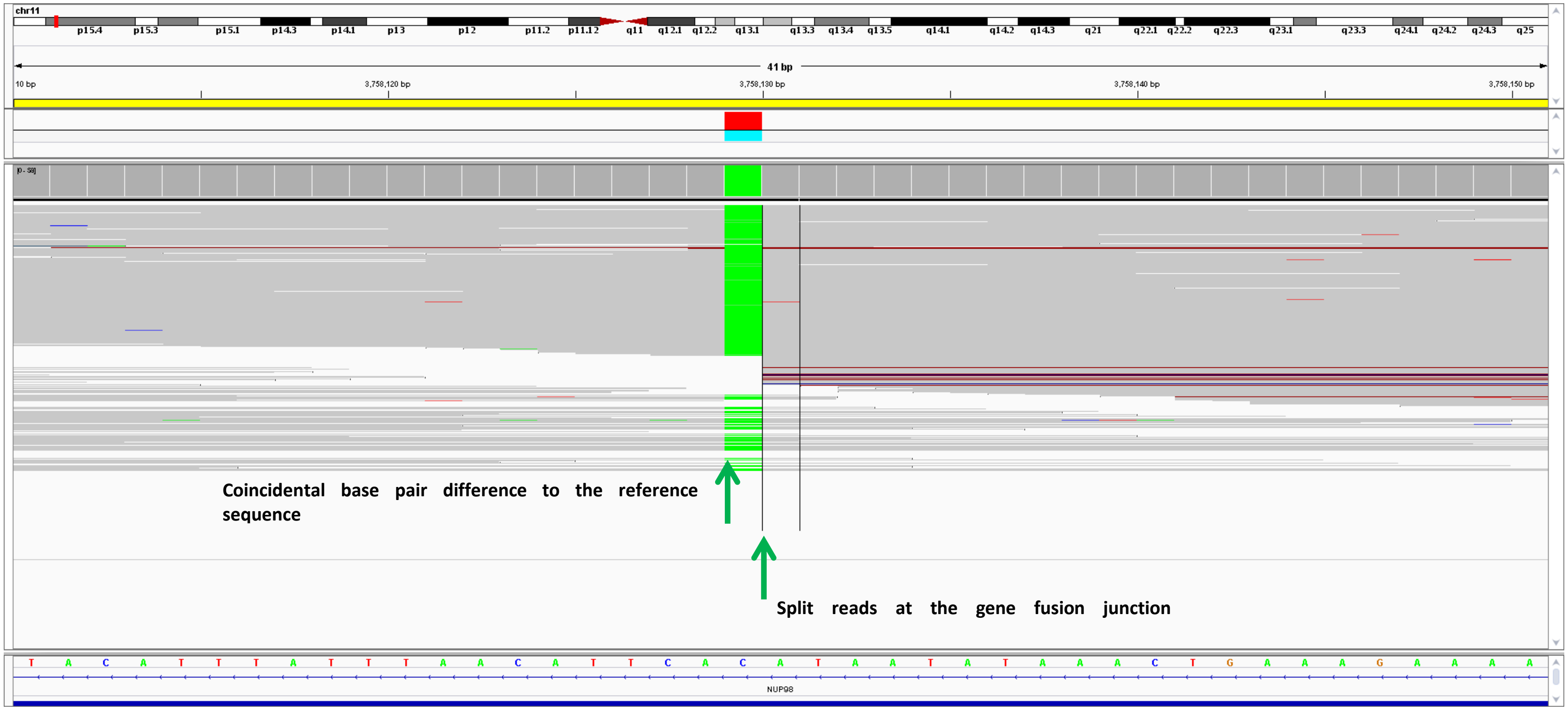
Figures 3.4b. Integrated Genome Viewer screenshots of regions of (a) chromosome 15 (chr15:74,326,139-74,326,179) and (b) chromosome 17 (chr17:38,493,855-38,493,895) showing partial track of read alignments in "expanded" view, centred at breakpoints in intron 6 in *PML* (15: 74326159) and intron 2 in *RARA* (17:38493875) respectively. The read alignments are coloured by insert size and show the reads with discordant pairs due to *PML-RARA* gene fusion in case no. 13.1050.



Figures 3.5a. Integrated Genome Viewer screenshots from case no. 13.1420 of regions at (a) chromosome 11 (from chr11:3,757,956-3,758,301) centred at breakpoint in *NUP98* in intron 12 at 11:3758130 and (b) view of chromosome 11 (chr11:108,549,618-108,549,658) centred at the second breakpoint in chromosome 11 in intron 5 in DDX10 (at 11:108549638). Diagram shows partial stack of reads in "expanded" view, with reads enriched in DDX10 due to preferential selection. The read alignments are coloured by insert size and show the reads with discordant pairs due to *NUP98*-DDX10 gene fusion.


chr11
p15.4 p15.3 p15.1 p14.3 p14.1 p13 p12 p11.2 p11.12 q11 q12.1 q12.2 q13.1 q13.3 q13.4 q13.5 q14.1 q14.2 q14.3 q21 q22.1 q22.2 q22.3 q23.1 q23.3 q24.1 q24.2 q24.3 q25
41 bp
108,549,620 bp
108,549,630 bp
108,549,640 bp
108,549,650 bp
No Variants Found
A
A
T
T
t t t t t t a t a c a g a g t t t c g c t c t g t t g c t a g g c t g g a g t a c
DDX10

Figures 3.5b. Integrated Genome Viewer screenshots from case no. 13.1420 of regions at (a) chromosome 11 (from chr11:3,757,956-3,758,301) centred at breakpoint in *NUP98* in intron 12 at 11:3758130 and (b) view of chromosome 11 (chr11:108,549,618-108,549,658) centred at the second breakpoint in chromosome 11 in intron 5 in DDX10 (at 11:108549638). Diagram shows partial stack of reads in "expanded" view, with reads enriched in DDX10 due to preferential selection. The read alignments are coloured by insert size and show the reads with discordant pairs due to *NUP98*-DDX10 gene fusion.


chr11
p15.4 p15.3 p15.1 p14.3 p14.1 p13 p12 p11.2 p11.12 q11 q12.1 q12.2 q13.1 q13.3 q13.4 q13.5 q14.1 q14.2 q14.3 q21 q22.1 q22.2 q22.3 q23.1 q23.3 q24.1 q24.2 q24.3 q25
41 bp
10 bp
3,758,120 bp
3,758,130 bp
3,758,140 bp
3,758,150 bp
[0 - 58]
T A C A T T T A T T T A A C A T T C A C A T A A T A T A A A C T G A A A G A A A A
NUP98

Coincidental base pair difference to the reference sequence

Split reads at the gene fusion junction

Figure 3.6a. chr11:3,758,110-3,758,150) centred at breakpoint of *NUP98* in intron 12 at 11:3758130 and (b) view of chromosome 11 (chr11:108,549,465-108,549,810) centred at the second breakpoint of chromosome 11 in intron 5 in DDX10 (at 11:108549638, also showing C>A base substitution at 11:3758129 in green). Diagram shows partial stack of reads in “squished” view, with reads enriched in DDX10 due to preferential selection. The read alignments are coloured by insert size and show the reads with discordant pairs due to *NUP98*-DDX10 gene fusion.

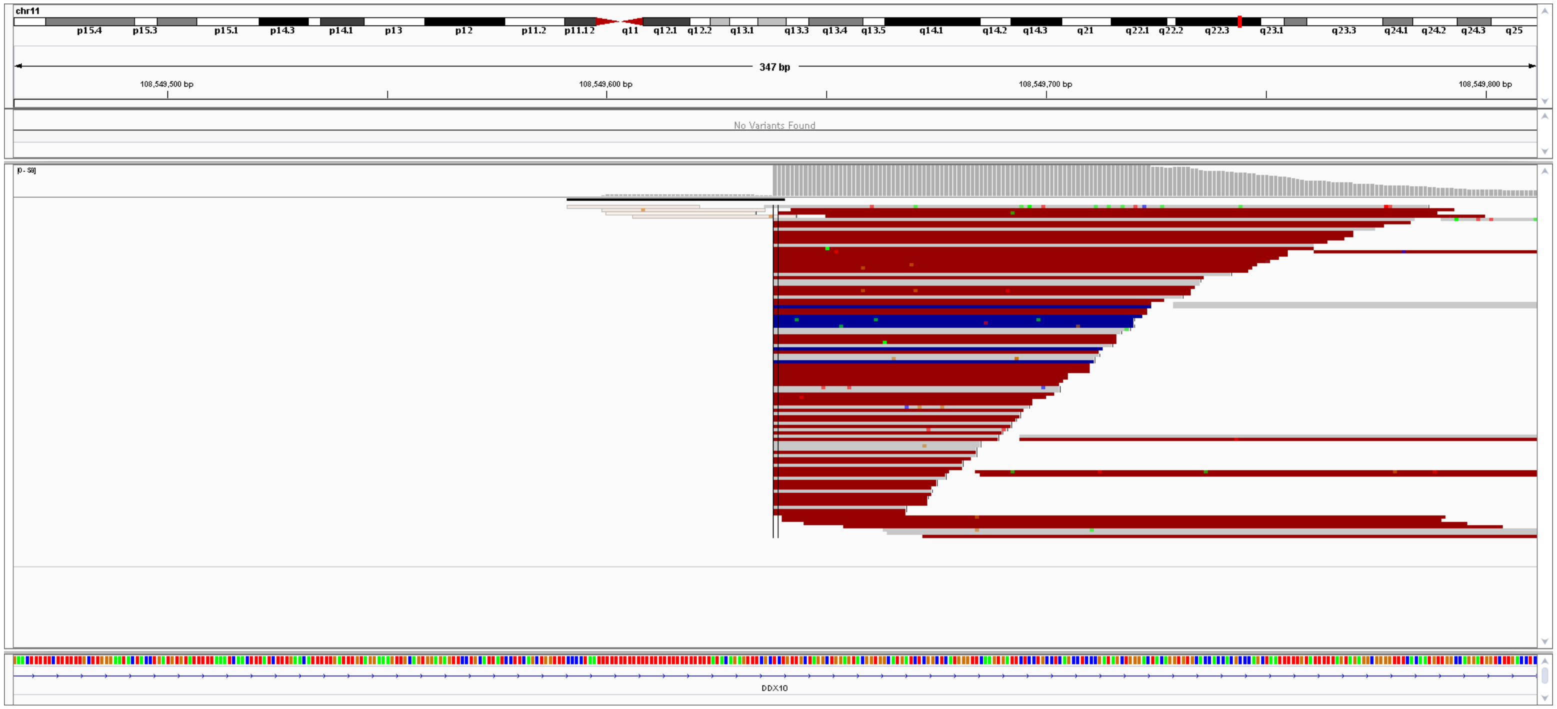


Figure 3.6b. Integrated Genome Viewer screenshots from case no. 13.1420 of regions in (a) chromosome 11 (from chr11:3,758,110-3,758,150) centred at breakpoint of *NUP98* in intron 12 at 11:3758130 and (b) view of chromosome 11 (chr11:108,549,465-108,549,810) centred at the second breakpoint of chromosome 11 in intron 5 in DDX10 (at 11:108549638, also showing C>A base substitution at 11:3758129 in green). Diagram shows partial stack of reads in “squished” view, with reads enriched in DDX10 due to preferential selection. The read alignments are coloured by insert size and show the reads with discordant pairs due to *NUP98*-DDX10 gene fusion.

3.4 Functional classification of diseases, complementation groups and mutual exclusivity

The mutations, indels and gene fusions that were newly found from the variant annotation protocol that were classified as pathogenic or likely pathogenic were examined to assess their value for more accurate prognostication, and potential for better treatment stratification. The prominent genetic class or cytogenetic profile were collated in Table 3.13 and grouped according to conventional ELN prognostic criteria, colour-coded according to disease risk (Döhner *et al.*, 2016). The additional genomic information was added and used to calculate the new prognostic group from a recent genomic-based classification and prognostic stratification (Papaemmanuil *et al.*, 2016). The *a priori* ELN grouping was compared to the new grouping and the change recorded (see Table 3.13).

The cohort of 35 AML cases is representative of the range of disease subtypes (see Tables 3.1 and 3.2) and incidence of mutations is as expected (see Figure 3.3). All patients presented with detectable recognisable driver gene fusions or mutations. Four patients had no class-defining mutations and two patients qualified in more than one group (Table 3.13). The largest distinct diagnostic subgroup diagnostic were *NPM1* mutated group (13 patients; 37%), one of whom also qualified as a second diagnostic category by having del(5q) by cytogenetics. Seven of these patients also harboured *FLT3*-ITD, 7 with *DNMT3A*, 3 *IDH1* and 2 *IDH2*. The other main mutational profiles were represented in approximately equal numbers; *CEBPA* double mutation (3), *TP53*/aneuploidy (4), and the Chromatin modifier/Spliceosome group (3). Of the patients defined by chromosomal rearrangements leading to gene fusions, there were two *PML-RARA*, four CBF-AML (3 with *MYH11-CBFB* and 1 *RUNX1-RUNX1T1*), and one each of *KMT2A-MLLT1* and *NUP98-DDX10* (see Table 3.12).

It is predicted from the additional information with emerging clinical significance of mutational profiling that 8 out of 35 patients' prognostic would be modified. Three patients with *CEBPA* double mutation were reclassified as favourable as these mutations were not part of routine testing prior to this study. Three patients with chromatin modifier/spliceosome mutations were reclassified as unfavourable (Papaemmanuil *et al.*, 2016). The *KMT2A-MLLT1* fusion was reclassified as Intermediate. Therefore, 7 patients moved from Intermediate to Favourable (3) or Adverse (4). Two *NPM1* patients could be reclassified to Intermediate as in one study, the favourable prognosis of *NPM1* is not retained in elderly patients (> 60 years) (Metzeler *et al.*, 2016). Patient 7 is the single case where mutation profiling has the most powerful impact, changing the complexion of the disease from a favourable outlook (due to *NPM1* prognosis) to a high risk disease associated with *NPM1/NRAS/DNMT3A* mutation combination (Metzeler *et al.*, 2016).

Multi-gene mutational profiling is starting to create a picture of cooperating mutations in CBF-AML and explain the outcome heterogeneity in this group (Duployez *et al.*, 2016). The adverse effect of *KIT* mutation on *CBFB-MYH11* rearranged leukaemia has been previously reported. It is suggested that the *NRAS* and *KRAS* mutations detected in case Sample 16, in the absence of *KIT* or *FLT3* mutations are associated with a favourable prognosis.

3.5 Influence of genomic profiling on the outcomes of the AML patients

Crude outcome data were collected from death statistics only and therefore Overall Survival (OS) could be calculated (see Table 3.14 below). This was compared to genomic profile alone, without access to additional clinical information and therefore with no knowledge of confounding factors, such as comorbidities, treatment related mortality, or the decisions to treat and treatment modalities. There are insufficient data for rigorous statistical analysis or to draw any definite conclusions; they are included for indicative purposes only.

However, the data is interesting and shows distinct trends and improved prediction from genomic profile data and independently from other clinical and laboratory data, demonstrated the potential of genomic profiling. 24 out of 35 outcomes were as expected based on genomic profile. 10 out of 13 patients with favourable genomics were alive at the time of analysis. 3 patients with favourable genomics died very early, possibly for reasons unrelated to genomic risk factors. 7 out of 8 patients with adverse genomics died within 12 months, however, probably also including uncorrelated to prognostic influence. 1 patient was an unexpected long term survivor. Of 8 patients who changed prognosis due to additional genomic information, 6 outcomes appeared to reflect the expected outcome more accurately. Two patients with favourable outcome were alive at the time of data collection.

Three patients with an adverse genomics profile (change from Intermediate) died rapidly (20, 84 and 258).

A one-way Analysis of Variance (ANOVA) with Tukey's multiple comparisons test was performed to test the performance of ELN and genomic prognostic systems to stratify patient risk (see Figure 3.7). The genomic profiling showed a significant difference between (1) favourable and intermediate and (2) favourable and adverse. However, by ELN criteria only the favourable and intermediate groups were statistically different. This appears mainly attributable to the reclassification of ELN Intermediate risk patients to favourable and adverse in the genomic data.

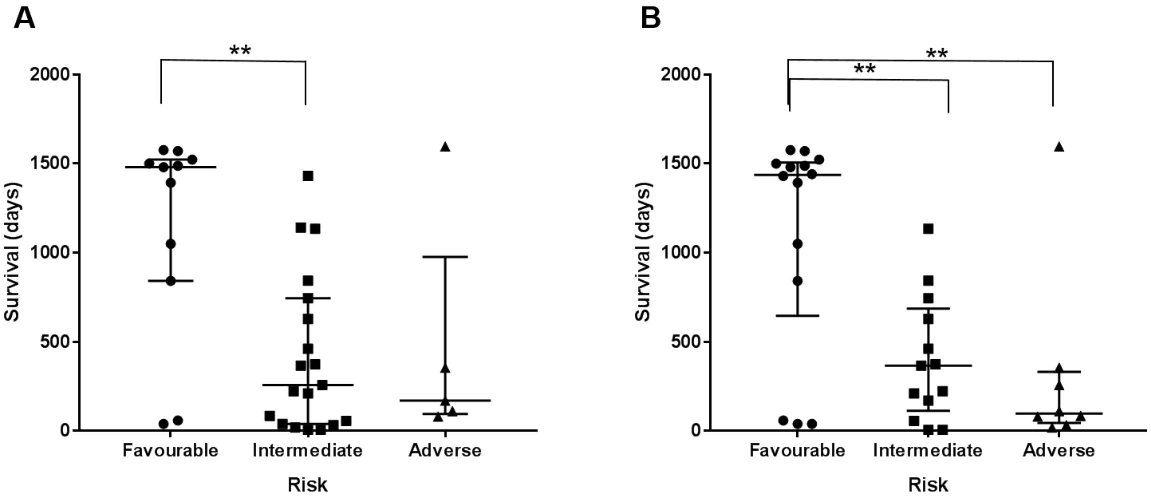


Figure 3.7. Scatter plot showing the median overall survival of patients in the study with interquartile range. (A) Overall survival of patients stratified using ELN criteria. (B) Overall survival of patients stratified using sequencing data from current study. Groups were compared using one way ANOVA with Tukey's multiple comparisons test **p<0.01.

Table 3.13. Main genomic features in 35 patients in the project, prior ELN risk status and new genomic classification

Sample	Prior Main Genetic Features	Prior ELN Risk Group*	Genes with Variants of Clinical Significance	Papaemmanuil (2016) group**	Risk Change
1	del(5q)	Adv	NRAS	*TP53*/aneuploidy	→
2	7q abn, *NPM1-/FLT3-*	Int	*DNMT3A*, *IDH2*^{R172K}	*TP53*/aneuploidy	→
3	CN-AML *NPM1-/FLT3-*	Int	CEBPA(smx2), *TET2*	biallelic *CEBPA*	↑
4	CN-AML *NPM1-/FLT3*-ITD	Int	*DNMT3A*, *TET2*, *BCOR*, *FLT3*-ITD	Chromatin/Spliceosome	↓
5	Failed cytogenetics *NPM1+/FLT3*-ITD	Int	*NPM1*, *FLT3*-ITD, *TET2*	*NPM1+* (*FLT3+*)	→
6	CN-AML *NPM1-/FLT3-*	Int	*CEBPA*(dm), *IDH2*^{R140Q}, *ASXL1*, *STAG2*	Chromatin/Spliceosome	↓
7	CN-AML *NPM1+/FLT3-*	Fav	*NPM1*, *DNMT3A*, NRAS, *SMC3*, *CEBPA*(sm), *FLT3*-TKD	*NPM1* (NRAS/*DNMT3A*)	→
8	CN-AML *NPM1+/FLT3-*	Fav	*NPM1*, *IDH2*^{R140Q}, *CEBPA*(sm)	*NPM1* (*IDH*)	→
9	t(15;17)/*PML-RARA*	Fav	*KRAS*	t(15;17) *PML-RARA*	→
10 & 14	Complex karyotype	Adv	*TP53*	*TP53*/aneuploidy	→
11	CN-AML *NPM1-/FLT3-*	Int	*DNMT3A*(x2), *IDH2*^{R140Q}, *PTPN11*, *SMC1A*, *CEBPA*(sm)	No class-defining lesions	→
12	CN-AML *NPM1+/FLT3-*	Fav	*NPM1*, *IDH1*^{R132C}, *NRAS*, *CEBPA* (sm), *ASXL1*	*NPM1* (IDH)	→
13	inv(11) with *NUP98* rearranged	Int	*NRAS*, *TET2*	No class-defining lesions	→
15	CN-AML *NPM1+/FLT3-*	Fav	*NPM1*, *DNMT3A*, *TET2*, *PTPN11*	*NPM1*	→
16	CBF-AML with inv(16)/*CBFB-MYH11*	Fav	*NRAS*, *KRAS*, *(KIT)*	CBF-AML with inv(16)/*CBFB-MYH11*	→
17	CN-AML *NPM1+/FLT3*-ITD	Int	*NPM1*, *FLT3*-ITD, *TET2*, *IDH2*^{R140L}, *ASXL1*	*NPM1* (*FLT3*-ITD)	→
18	del(5q) *NPM1+/FLT3-*	Adv	*NPM1*	*NPM1*/aneuploidy	→
19	*PML-RARA*, *FLT3*-ITD	Fav	*FLT3*-ITD	*PML-RARA* (*FLT3*-ITD)	→
20	CN-AML *NPM1+/FLT3*-ITD	Int	*NPM1*, *DNMT3A*, *FLT3*-ITD	*NPM1* (*FLT3*-ITD)	→

21	CN-AML *NPM1*+/*FLT3*-ITD	Int-I	*NPM1*, *TET2*, *DNMT3A*, *FLT3*-ITD	*NPM1* (*FLT3*-ITD)	→
22	CN-AML *NPM1*+/*FLT3*-	Fav	*NPM1*, *DNMT3A*, *TET2* (x2)	*NPM1*	→
23	t(11;19) *KMT2A*-MLLT1	Adv	*PTPN11, SMC1A*	*KMT2A-KMT2A*T1	↑
24	CBF-AML with t(8;21)/*RUNX1-RUNX1T1* *FLT3*-ITD	Fav	*FLT3*-ITD	CBF-AML with *RUNX1-RUNX1T1*	→
25	Other cyto	Int	$JAK2^{V617F}$, *TET2*x2, *KRAS*	No class-defining lesions	→
26	Other cyto	Int	*DNMT3A, NRAS*	No class-defining lesions	→
27	Other cyto	Int	*CEBPA*(dm), *WT1*, *ASXL1*	biallelic *CEBPA*	↑
28	CN-AML *NPM1*-/*FLT3*-ITD	Int	*TET2*x2, *RUNX1*x2, *DNMT3A*, *ASXL1*x2, *FLT3*-TKD, *FLT3*-ITD, *KMT2A*-PTD, *PHF6*, *SETBP1*	Chromatin/Spliceosome	↓
29	CN-AML *NPM1*+/*FLT3*-ITD	Int	*NPM1*, *FLT3*-ITD, *DNMT3A*, *FLT3*-TKD, *PTPN11, SMC1A, FAM5C, KRAS*	*NPM1* (*FLT3*-ITD)	→
30	CBF-AML with inv(16)/*CBFB-MYH11*	Fav	No variants detected	CBF-AML with inv(16)/*CBFB-MYH11*	→
31	CN-AML *NPM1*+/*FLT3*+	Int	*NPM1*, $IDH1^{R132H}$, *FLT3*-ITD	*NPM1* (*FLT3*-ITD)	→
32	CN-AML *NPM1*+/*FLT3*-ITD	Int	*NPM1*, *DNMT3A*, *FLT3*-ITD, *CEBPA*(sm)	*NPM1* (*DNMT3A* & *FLT3*-ITD)	↓
33	CBF-AML with inv(16)/*CBFB-MYH11*	Fav	*RUNX1*	CBF-AML with inv(16)/*CBFB-MYH11*	→
34	CN-AML *NPM1*-/*FLT3*-	Int	*CEBPA*(smx2), *RUNX1*, *TET2*, *ASXL1*, *SETBP1*	Meeting criteria for ≥2 genomic subgroups	→
35	Monosomy 7	Adv	*TET2*, *DNMT3A*, *U2AF1*, *KMT2A*-PTD	*TP53*/aneuploidy	→
36	CN-AML *NPM1*-/*FLT3*-	Int	*CEBPA*(dm), *EZH2*x2, *GATA2*, *SMC3*	biallelic *CEBPA*	↑

Key to abbreviations

NPM1- (NPM1 wild type), *NPM1*+ (NPM1 tetranucleotide insertion), *FLT3*- (*FLT3* wild type), *FLT3*-ITD (*FLT3* internal tandem duplication), *FLT3*-TKD (*FLT3* Tyrosine kinase domain mutation), *KMT2A*-PTD (*KMT2A* partial tandem duplication), *CEBPA*(dm) (CEBPA double heterozygote mutation), CEBPA(smx2) (CEBPA homozygous mutation), *CEBPA*(sm) (CEPPA single mutation), CBF-AML (Core Binding Factor AML), CN-AML (Cytogenetically Normal AML)

*ELN Risk Group derived from (Döhner *et al.*, 2016)

**Papaemmanuil (2016) group from (Papaemmanuil *et al.*, 2016)

Table 3.14 Survival and change of risk status in 35 patients used from the project.

Sample	Prior ELN Risk Group	Risk Change	Survival (OS days) (days)	Actual	O vs E
1	Adv	→	1596	F	X
2	Int-II	→	1135	I	✓
3	Int-I	↑	40	A	X
4	Int-I	↓	33	A	✓
5	Int-I	→	6	A	X
6	Int-I	↓	258	A	✓
7	Fav	→	1576	F	✓
8	Fav	→	59	A	X
9	Fav	→	1570	F	✓
10 & 14	Adv	→	111	A	✓
11	Int-I	→	374	I	✓
12	Fav	→	40	A	X
13	Int-II	→	629	I	✓
15	Fav	→	1522	F	✓
16	Fav	→	1500	F	✓
17	Int-I	→	745	I	✓
18	Adv	→	81	A	✓
19	Fav	→	1480	F	✓
20	Int-I	→	211	A	X
21	Int-I	→	461	i	✓
22	Fav	→	1050	I	X
23	Adv	↑	170	A	X
24	Fav	→	842	F	✓
25	Int-II	→	56	A	X
26	Int-II	→	6	A	X
27	Int-II	↑	1431	F	✓
28	Int-I	↓	20	A	✓
29	Int-I	→	843	F	✓
30	Fav	→	1488	F	✓
31	Int-I	→	366	I	✓
32	Int-I	↓	84	A	✓
33	Fav	→	1393	F	✓
34	Int-I	→	223	A	X
35	Adv	→	355	A	✓
36	Int-I	↑	1441	F	✓

4.0 Discussion

4.1 Introduction; the potential for routine use of NGS in AML diagnosis

Next-generation sequencing facilitated an era of intense study in cancer genomics by providing the technology for genome-wide examination of the acquired and germline abnormalities in cancer. This provided new insight into how these changes initiate the disease, contribute to phenotypes and influence clinical behaviour. Rapid evolution of the technology in the last decade has resulted in improved speed and better resolution of DNA sequencing, at dramatically lower costs. The technology is now being used in a routine diagnostic setting for many applications. AML is the most intensively studied type of cancer and its somatic genome has been defined in detail (The Cancer Genome Atlas Research Network, 2013; Mazzarella *et al.*, 2014; Metzeler *et al.*, 2016; Papaemmanuil *et al.*, 2016). The genomic heterogeneity of AML has been shown to reflect the diversity in haematological and clinical features. Genetic studies are the basis for conventional risk-adapted therapy (Döhner *et al.*, 2016) but the newfound variety of underlying genetic defects adds to the determination of clinical outcome; recently, AML has been categorised into 11 molecularly-defined subgroups that correlate with pathogenesis and disease prognosis (Figure 1.4) (Papaemmanuil *et al.*, 2016). AML is an aggressive cancer and is particularly difficult to treat. It typically occurs in the elderly and most patients relapse after initial remission. It is widely predicted that the detailed understanding of the mutational landscape will be used to modernise treatment, to more accurately stratify the disease risk, direct conventional treatment, and offer the prospect to personalise therapy (Ofran & Rowe, 2013; Meyer & Levine, 2014; Roug *et al.*, 2014).

The aim of the project was to develop a streamlined NGS assay for the genomic diagnosis of AML, to demonstrate the principle that this could be transferrable to the diagnostic laboratory and to evaluate its potential to improve AML management. The project used a novel design to detect all clinically relevant genetic changes in AML and recurrent abnormalities that are likely to be actionable in the near future. It investigated the potential for using genomic DNA for the detection of variants, including duplications and gene fusions, which was previously considered technically demanding. AML is currently sub-classified using several disparate laboratory methods with limited resolution. Key objectives were to reproduce the diagnostic and prognostic findings from existing methods, to accurately detect the types of genetic abnormalities encountered in conventional AML diagnosis. Additional molecular abnormalities of growing importance to AML diagnosis were identified and the sensitivity of testing was increased. The project demonstrated how a single, specific NGS assay could achieve these objectives and replace multiple techniques, with the potential to rationalise laboratory workflows. The genome-wide screen of 42 genes, including genetic targets not part of the standard workflow, permitted reclassification of patients to the genomic classification of Papaemmanuil et al (2016) and it was possible to demonstrate how a benefit to patient management can be derived from its implementation.

The new assay would need to achieve a speed of turnaround appropriate to clinical demands and provide reliability and reproducibility for the upmost confidence in diagnostic results. In practice, it would need to meet the stringent principles required in a diagnostic laboratory and rigorous accreditation standards (Jennings *et al.*, 2009; Mattocks *et al.*, 2010). To reflect the typical diagnostic workflow, a sample of normal cells would not be readily available and there would be no option to use constitutional DNA as normal control from the same individual, to distinguish somatic mutations from germline variants. By detecting all types of genomic rearrangements to the highest, single nucleotide resolution, it should be possible to increase the sensitivity of the tests by sequencing to an appropriate depth. To provide this capacity on currently available sequencers, it is necessary to be selective about the sequencing of genomic targets, which also facilitates management of costs. The feasibility of the implementation of the technique will be discussed.

4.2 Improved detection of clinically relevant genomic profiles

The project shows that this targeted sequencing strategy and analytical pipeline can efficiently identify all major categories of somatic mutations found in AML. Recognisable driver mutations, indels and/or gene fusions were detected in all patients by the new genomic profile. The sequencing panel detected disease-defining lesions, conventionally detected by cytogenetics and FISH; two *PML-RARA,* three *CBFB-MYH11,* one *RUNX1-RUNX1T1,* and one *KMT2A-MLLT1* gene fusions resulting from chromosomal translocations or inversions. A *NUP98-DDX10* fusion, resulting from inv(11)(p15q22), was correctly identified by the targeting of *NUP98* only; its partner gene, *DDX10*, was not selected directly, but was identified and characterised by virtue of the sequencing of chimaeric DNA fragments containing *NUP98*. The analytical pipeline correctly detected thirteen *NPM1* exon 11 tetranucleotide insertions and ten *FLT3*-ITDs which were part of normal molecular genetics work-up for the cases. *JAK2* was not included in the panel design at this time and so one patient with *JAK2* V617F was not detected but was identified using the parallel sequencing method and confirmed during routine molecular genetic testing. A total of 112 other variants from 36 samples were annotated and passed filtering as known or likely pathogenicity. Many of these were specific with known hotspots, including *CEBPA, IDH1/2* mutations and *FLT3*-TKD. The remaining genes were recurrent in myeloid disease, the majority of which showed relevant annotations in the COSMIC database (Wellcome Trust Sanger Institute, 2016). Some variants were included without a specific COSMIC reference but passed filtering as having a suggestive functionality and likely pathogenicity. Often these were supported by studies where any coding location was sequenced and therefore no sublocation of the SNV defined, but were considered to contribute to AML development. Functionality studies are required to delineate their clinical significance further.

The sole case of complex karyotype showed *TP53* exon 8 mutation, consistent with the genomic instability conferred by lack of functional p53. *TP53* mutation has been used as a surrogate marker for these high risk patients previously, in a molecular genetic classification of AML (Grossmann *et al.*, 2012). However, 30% of AML patients with complex karyotype do not have TP53 mutations and it needs to be confirmed that other genes may account for this observation. Whilst *TP53* is an essential molecular marker for disease stratification, the significance of other mutations in this important prognostic group needs to be identified for a relevant molecular classification to be devised. Of the nucleotide substitutions and indels in 19 genes that were also sequenced by the Ion Torrent semi-conductor NGS method, there was broad agreement; the majority of point mutations were detected by both NGS protocols. Discrepancies were detected between systems and were thought to be products of sequencing and alignment errors, which are still common and expected with any NGS platform. The significance of these is discussed below.

4.3 Clinical significance of genomic evidence on AML management

The new genomic profile not only identified significant lesions previously detected by cytogenetics and molecular genetics, at an increased resolution and sensitivity, but also was able to redefine prognostic categories in a significant proportion of patients, according to a proposed genomic prognostic system (Papaemmanuil *et al.*, 2016). This offers the prospect of improved disease stratification and better options of standard treatment. Previously, eighteen out of thirty-five patients were CN-AML, five of which showed no evidence of *NPM1* or *FLT3*-ITD mutations and so could not be resolved further by standard molecular testing. Nineteen patients were of Intermediate ELN risk. Under the new system, four patients had no class-defining mutations and two patients qualified in more than one group (Table 3.13). In total, 8 patients were reclassified. Four patients changed from Intermediate to Adverse, three of which entered the new 'Chromatin/Spliceosome' category. If eligible for intensive treatment, this high risk group would indicate allogeneic SCT as post remission therapy, and affect the decision-to-treat in the elderly, although age alone is not the sole predictor of treatment-related mortality, due to better health status and improved supportive care (Döhner *et al.*, 2016). Three patients changed from Intermediate to Favourable, all three by detection of biallelic *CEBPA* mutations (two compound heterozygotes and 1 homozygote). Two patients were long term survivors (the other died very rapidly, presumably unrelated to genetic prognostic factors). This emphasises the need for routine *CEBPA* mutation screening, which was not part of routine diagnostic testing. Favourable patients typically receive single cytotoxic agent as consolidation (e.g. intermediate dose Ara-C, IDAC) and would avoid allogeneic SCT and high dose cytarabine, under typical circumstances. One patient changed from High Risk to Intermediate, due to redesignation of their *KMT2A* translocation in the new scheme.

The crude survival data from the project, based on 35 patients only without access to any other modifying factors, can only be considered for indicative purposes but it was interesting by showing better discrimination of patient outcome by genomic profiling, mainly by redesignation of seven patients from the Intermediate category. It should be noted that the genomic classification used an extended panel of 111 genes for screening, of which only the common 30 genes were tested in this study. By inclusion of more infrequently mutated genes, only rarely will positive results add to the diagnostic profile but occasionally better distinction of prognostic groups would be possible. The predictive power of the new genomic classification requires validation in independent clinical trials. Other well-powered studies take a different perspective and could add to this evidence; e.g. patients with *DNMT3A* mutations, well-represented in this study, have been reported to have a poor prognosis in other studies, particularly in younger patients (Metzeler *et al.*, 2016).

AML is an active field of new drug investigation in early-phase, experimental trials, usually in refractory patients, and in clinical studies (Döhner *et al.*, 2016). Genomic targets provide an important indication of possible efficacy of novel therapies to oncogenic proteins or broader spectrum agents, such as protein kinase inhibitors and epigenetic modulators. Patients already screened as part of routine genetic testing could access *FLT3* inhibitors which have been under trial and may not only impede the *FLT3*-ITD but also *FLT3*-TKD, of which there were 4 in this patient group, two concurrently with *FLT3*-ITD (Grunwald & Levis, 2015; Stein & Tallman, 2016). There is early understanding of the *KMT2A* gene fusions as a possible target for epigenetic therapies (Placke *et al.*, 2014; Chen & Armstrong, 2015). Whilst effective inhibitors of the common *NPM1* mutations are elusive, alternative therapies may emerge that use the disease hallmark as a predictive marker (El Hajj *et al.*, 2015; Martelli *et al.*, 2015). The *IDH1/2* mutations, of which there were eight in this cohort that would not have been detected previously, are also under investigation in early phase trials (Wang *et al.*, 2013; Stein & Tallman, 2016; Dombret & Gardin, 2016). The subclonal (but unconfirmed) *KIT* mutation, which was found secondary to *CBFB-MYH11* in case 16, could be a target for dasatinib (or other TKI) if this clone progressed (Paschka *et al.*, 2013). The knowledge that AML is comprised of subclones that harbour mutations which may be differently susceptible to existing and novel therapies suggests that their accurate identification at first presentation is increasingly important.

4.4 Clonal heterogeneity and cooperating mutations

This NGS experiment demonstrates the remarkable genetic heterogeneity in AML even with a limited panel of genes and in only thirty-six samples. Such multi-gene mutational profiling is able to depict patterns of cooperating mutations and mutual exclusivity (Figure 1.5) (The Cancer Genome Atlas Research Network, 2013; Papaemmanuil *et al.*, 2016; Metzeler *et al.*, 2016). AML is clonal and has been shown to evolve in both linear and branching patterns, with mutations occurring in non-random temporal order, often leading to a complex network of subclones and forming an intricate subclonal architecture (Anderson *et al.*, 2011; Welch, 2014).

Genetic defects can be considered to be 'initiation' or 'progression' events; initiators occur at an early stage and will therefore be present in the founding clone and in all descendants of the leukaemic cells. Balanced translocations involving transcription factor gene fusions, e.g. t(15;17), t(8;21), and inv(16), such as cases 9, 16, 19, 24 & 33 in this study, are primary, initiating events found in the founding clone (Grimwade *et al.*, 2016). Mutations in genes affecting the epigenome, such as *DNMT3A* and *TET2* are also more likely to occur earlier in leukaemogenesis (Corces-Zimmerman *et al.*, 2014; Shlush *et al.*, 2014) and have also been found in pre-leukaemic clones of haematopoietic stem cells that provide a selective growth advantage but are insufficient alone to initiate leukaemic transformation (Corces-Zimmerman & Majeti, 2014; Grove & Vassiliou, 2014). These mutations in the study can all be shown to have high VAF, consistent with being initiating mutations (Table 3.6 and Figure 3.4).

Progression mutations are secondary events that enhance the proliferative capacity of the cell and may occur in distinct subclones and only be present in a proportion of the leukaemic cells. *NPM1* mutations were considered early events (Jan *et al.*, 2012; The Cancer Genome Atlas Research Network, 2013), however, they have been shown to often occur with *DNMT3A* mutations, are not always present in all cells, and can be lost in clonal development in some studies (Kronke *et al.*, 2013; Corces-Zimmerman & Majeti, 2014). Chromosome abnormalities such as trisomy 8 and nucleotide variants in genes involved in activated signalling, such as *FLT3, RAS, KIT and WT1,* are also found exclusively as late events and frequently occur in subclones. They are observed either to emerge or to be lost at relapse (Corces-Zimmerman *et al.*, 2014; Welch, 2014). These are often found with lower VAF consistent with emerging subclones in many cases in these samples. Primary abnormalities account for mutations with VAF around 0.5, consistent with heterozygous mutations present in the majority of cells in the sample (Fig. 3.4). In contrast, mutations in genes with lower VAF allele frequencies, indicate they are present only in a subpopulation of the cells. Therefore multiple genetic events are required to promote development of AML and they act in cooperation in leukaemogenesis and have complementary modes of action (The Cancer Genome Atlas Research Network, 2013;

Papaemmanuil *et al.*, 2016; Metzeler *et al.*, 2016). The recognition of clonal heterogeneity within AML will become important in our understanding of the efficacy of therapeutic interventions. Apparent clonal diversity and subclonal resistance to chemotherapy and the choice of appropriate single biomarkers as therapeutic targets and MRD monitoring, will all be significant considerations. We are moving into an age whereby further insight into AML will be obtained by an understanding of functional synergism and complex interrelations between gene mutations, with an increasing importance of genome-wide cancer studies rather than testing for single genes in isolation.

4.5 Performance of the NGS assay for AML diagnosis

4.5.1 Detection of Somatic Mutations

NGS technology has now progressed sufficiently that multiple studies have now been performed and many diagnostic laboratories are experimenting with allele-specific assays to identify hotspot mutations in genes. For validation, it is impractical to have positive controls for mutations in each region or to confirm all alterations detected independently. Diagnostically, this would not be required. However, to demonstrate confidence in a new assay, the conventional approach is to assess the overall performance of the sequencing platform using different samples with a range of mutation types (Singh *et al.*, 2013). In this study, the experimental custom panel using Illumina NextSeq and SureSelect analysis was compared to a different sequencing technology and mutation hotspot panel; Ion Torrent PGM/Ion AML Ampliseq. A proportion of genes in the experimental panel were not covered by the standard AML Ampliseq but evidence from the covered genes offered a direct comparison and provided an understanding of the performance of the approach and would inform future work.

Somatic mutations in the dominant leukaemic clone were identified in all cases studied using sequence alignment/configuration with SureCall corroborated by other methods. The discrepancies were two lower VAF calls (of 0.18 and 0.15) that were not detected by Ion Torrent, and discordance between sequencing the *CEBPA* mutations in sample 27. There were no abnormalities found by Ion Torrent that were not detected by the study, suggesting 100% sensitivity and false positive calls appear unlikely. Despite the rapid advances in NGS development to clinical standards, anomalies in sequencing are still apparent in all studies. Standard NGS analysis software typically makes insertion calls based on data obtained during the initial read mapping and alignment process, which is thought to be the source of the errors observed in these data, because of the difficulty associated with always aligning short reads consistently.

In line with the developing nature of this technology, improvements are still to be gained in bioinformatics and there is need for standardisation of practice. In a recent Pre-Pilot external quality assurance round for AML mutation detection, variability in all aspects of the process was observed including gene panel composition, variants detected, variant nomenclature and pathogenicity classification (UKNEQAS, 2017). In particular, nineteen out of twenty-three participants detected an *IDH2*^{R172K} variant, five out of twenty participants detected a *KRAS*^{G12R} variant of low VAF and only nine out of twenty three participants detected an *FLT3*-ITD, due to attempting to use inappropriate variant calling software. This was a trial assessment but shows the current stage of development of NGS for mutational screening. By applying relevant bioinformatic software for variant detection, this assay performed well and appears would perform favourably compared to other systems. It is understood that NGS is still at an early phase of its development for AML diagnosis but this will improve with further use and refinement of the sequencing technology, improvement in bioinformatics approaches and with more personal experience in variant calling, to permit expert review and to recognise nonsensical calls.

4.5.2 Detection of *FLT3* and *CEBPA* insertions

The inclusion of regions of genomic insertions is essential for a comprehensive mutation profile in AML, due to the prognostic significance of *NPM1*, *FLT3*-ITD and *KMT2A*-PTD. This analytical pipeline was not used to detect CNV; detection of ITD/PTD was achieved in this experiment by targeted sequencing of the entire span of the known regions to look for signature junction fragments. This avoided reliance on copy number estimates of duplicated regions and extending sequencing into normal regions as normal controls to validate duplications, which is complex and bioinformatically demanding (Conte *et al.*, 2013).

The favourable *NPM1* exon 11 insertions are typically of uniform size and, generally, do not present a challenge for sequencing; most sequence alignment software is designed to detect small indels as well as SNV and short reads containing small insertions retain sufficient homology in the regions flanking the insertion to permit alignment whereas large insertions do not. *NPM1* is detected reliably in most AML NGS studies, without using special procedures (e.g. Luthra *et al.*, 2014) and is sufficiently sensitive to be used to detect MRD (Salipante *et al.*, 2014). All thirteen *NPM1* mutations in our samples, that were detected by standard PCR, were also identified by the two NGS platforms and multiple sequence mapping programmes. The software used in SureCall for sequence alignment miscalled the tetranucleotide sequence in three out of eleven cases; the reason for accurate detection of the duplicate but wrong assignment of the sequence is unclear but is probably an issue of reporting by the software and will be investigated further. Whilst not currently of significance, allelic heterogeneity is a known feature in NPM1 mutation which may prove to be important and needs to be accurately represented (Falini *et al.*, 2007; Alpermann *et al.*, 2016).

FLT3-ITD and *CEBPA* present different challenges and are notorious for being difficult to detect by NGS. *FLT3*-ITD, in particular, has highly variable length (up to 300bp) and allelic burden. SureCall performed poorly for detection of *FLT3*-ITD but this was expected as it is designed for detecting SNVs and short insertions only. Pindel was chosen for *FLT3*-ITD detection and eleven mutations, ranging from 12 – 87bp, were detected by this NGS approach, which was in complete concordance with the results from standard molecular tests (Spencer *et al.*, 2013). Other reports suggest that Pindel is limited for determining large *FLT3*-ITD, which were not detected in our patients (Bolli *et al.*, 2015; Au *et al.*, 2016). Alternative detection methods (Schnittger *et al.*, 2014; Kim *et al.*, 2016) and special bioinformatics approaches for ITDs (Kadri *et al.*, 2015; Rustagi *et al.*, 2016) are being tested. Allelic fraction was not calculated by our conventional testing and so could not be compared. The nine samples with *CEBPA* appeared to be reliably called, with the exception of case 27 with the largest insertion (39 bp by this method) which showed multiple variant alleles but only one (heterozygous) call by Ion Torrent, which is assumed to have been a false by this method. Platforms such as Ion Torrent relying on Single Nucleotide Addition (SNA) strategies suffer predominantly indel errors (Dohm *et al.*, 2008; Schirmer *et al.*, 2015) and homopolymer read errors occur with homopolymers longer than 6~8bp (Goodwin *et al.*, 2016).

4.5.3 Detection of gene fusions and structural variation

A novel feature of this project was to evaluate the use of a highly targeted assay to detect known, clinically relevant, structural alterations by combining testing genomic DNA with hybrid capture of genomic features, thereby detecting multiple types of genomic abnormalities in AML, in a single workflow. NGS of structural variation breakpoints is feasible and has shown to work with WGS, however, this was only recently demonstrated in combination with target enrichment (Abel *et al.*, 2014). The strategy required coverage of known genomic breakpoints within exons and introns in genes of recurrent gene fusions and tandem duplications. By definition this necessitated selection and sequencing of repetitive intronic regions where breakpoints typically reside, which favoured a hybrid capture approach, as opposed to target enrichment by amplification. Solution-phase hybrid capture can capture larger regions, has fewer PCR-introduced artefacts and has better within run and inter-run reproducibility.

Chromosomal breakpoints involved in structural variation arise from DNA double-strand breaks (DSBs) and ineffective DNA repair. The genomic distribution is non-random and typically occurs in repetitive regions and may be facilitated by repeat elements (Bacolla *et al.*, 2016). However, the relative contribution of different mechanisms to the generation of chromosomal translocations and other types of structural variants is unclear (Yang *et al.*, 2013; Ghezraoui *et al.*, 2014; Ottaviani *et al.*, 2014). The low complexity of regions involved in chromosomal rearrangements presents an issue for target enrichment and sequencing of these regions. The breakpoint cluster regions in both partner genes of *CBFB*/*MYH11*, *PML*/*RARA*, *DEK*/*NUP214*, *RUNX1*/*RUNX1T1*, two genes with multiple fusion partners, *KMT2A* and *NUP98*, as well as two common partners of *KMT2A* (*MLLT1* and *MLLT3*) were included. All exons of *RUNX1*, as well as breakpoint cluster regions of the *RUNX1-RUNX1T1* gene fusion were included. This was successful and sufficed to demonstrate the feasibility of sequencing translocation breakpoints for this Proof-of-Principle experiment.

The method of target enrichment for DNA selection used a custom designed hybrid capture Agilent SureSelect panel (Gnirke *et al.*, 2009) prior to deep sequencing. This enabled the simultaneous identification of specific mutations and structural variants in genes described above (section 4.2 and 4.3). Target coordinates needed to be thoroughly checked to ensure all regions are adequately covered, before uploading the

BED file into the Agilent Sure Design project design tool (Bolli *et al.*, 2015). A reagent kit is then provided which contains the custom set of biotinylated RNA 120mer oligonucleotides which is used to isolate the specific DNA fragments of interest to form the library of genomic DNA for sequencing.

To optimise target enrichment for the type of library required, different adjustments were applied to the design in SureDesign. Repeat masking was set to the lowest stringency (at most 20 bp overlap) to include a standard set of repetitive genomic regions in the Bait Tiling process. Masking of repeat sequences was not applied. Different iterations of the design were performed, reducing the stringency (to moderate then least stringent) in an attempt to optimise coverage. By including regions of low complexity within non-coding regions, there is a risk that a larger number of poor quality, non-specific baits were generated and so ‘Boosting’ was set to ‘maximum performance’ to make genomic fragment pull-down consistent and balanced; to increase the number of replicate copies of both orphan baits and GC-rich baits.

5x Bait Tiling was chosen to obtain better coverage and offer the best opportunity of capturing the breakpoints of rearrangements when tiling across intronic regions where the exact breakpoints are not known. Increased tiling of baits increases the cost of kit but in this experiment, the size capacity of the library was not reached in the design tool and increasing the frequency was not cost prohibitive. Theoretically, baits are staggered and overlap starting every 24bp so that five baits cover each base in each interval, however, the exact density of tiling may be less particularly at the extreme 5' and 3' ends of each interval. Generally, the higher level of tiling means that a target sequence that is represented at the end of one bait is represented in different positions of alternate baits that cover the same region.

Overall coverage of the design by baits was 98.84% with all exonic regions showing 100% coverage (see Table 2.2). However, the coverage of targets which included introns varied from 96.10% for *MYH11* and 99.53% for *KMT2A* although the design report classed all regions with >90% as high coverage. The ‘Regions not covered’ were reported by the design software and were found in the alltracks.bed file under missed regions. These regions were viewed in USCS Genome Browser and it was noticeable that coverage was not evenly distributed and that gaps were present in bait coverage, for example *CBFB* intron 5 was sequenced for inv(16); 16 regions missed with a size of up to 304 base pairs (chr16:67129543 – 67129846). A region of 490 bp within *RARA* intron 2 had no probe coverage (see Figure 4.1). This creates the potential for some genomic sequences to miss selection by bait pull-down and therefore miss representation in the sequencing. To mitigate this, where possible, the breakpoint cluster regions of both partner genes in a known gene fusion were targeted. Also, as DNA sequencing potential is driven by DNA fragment shear size and not by bait size or proximity, sequences can be obtained from regions without baits by sequencing into these areas with overhanging DNA fragments, e.g. a 500bp DNA fragment will not all be covered by the 120mer bait that pulls it down, but the entire fragment will be available for sequencing. A core principle of the project was that coverage would extend into adjacent regions. This principle also supports the sequencing of split reads from chimaeric gene fusions with unknown partners (see below). Fragment length was part of a standard workflow and not quantified for this experiment prior to sequencing but 500bp fragments are typical. It is possible to enhance coverage beyond the bait site by increasing fragment size. Anecdotally, it might be possible to pull down fragments of 800 ~ 1,200bp to enhance the sensitivity of the assay, which would be a modification for future tests. Theoretically, a randomly sheared 500 bp fragment, anchored by a 120mer bait could extend into 380 bp of region that is not covered by baits, as could a fragment pulled down by an adjacent bait on the other side of the coverage gap. Therefore, regions of several hundred bases could be salvaged, with an issue being their alignment and recognition. This might rely on sufficient complexity of sequence being present in the paired read from the DNA fragment extending into the uncharted regions. The sensitivity of this strategy would need to be tested on more samples.

By capturing non-coding sequences, repetitive areas become the majority of sequencing, which could reduce representation of other targeted regions and limit the depth of sequencing. The project achieved good read depth achieved for translocation detection. Another consequence was that reads with repeat sequences were aligned to multiple areas of the genome; 61-63% of off-target alignment resulted. This did not appear to affect the variant prediction of gene fusions for the project. It is possible, however, that breakpoints may not be recognised amidst regions of low DNA specificity and a high degree of redundant sequencing was observed and which reduced efficiency.

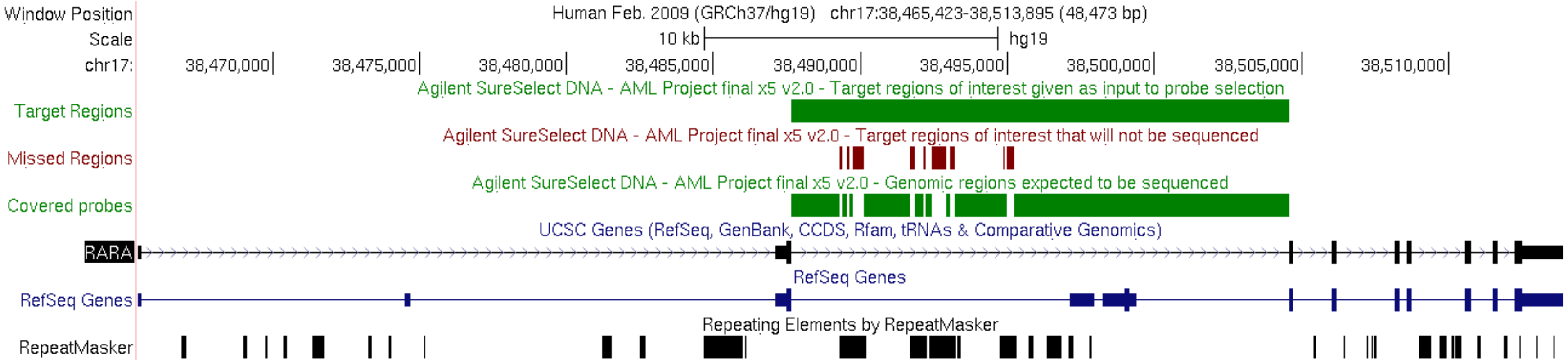

Window Position
Human Feb. 2009 (GRCh37/hg19) chr17:38,465,423-38,513,895 (48,473 bp)
Scale
10 kb
hg19
chr17:
38,470,000
38,475,000
38,480,000
38,485,000
38,490,000
38,495,000
38,500,000
38,505,000
38,510,000
Agilent SureSelect DNA - AML Project final x5 v2.0 - Target regions of interest given as input to probe selection
Target Regions
Agilent SureSelect DNA - AML Project final x5 v2.0 - Target regions of interest that will not be sequenced
Missed Regions
Agilent SureSelect DNA - AML Project final x5 v2.0 - Genomic regions expected to be sequenced
Covered probes
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RARA
RefSeq Genes
RefSeq Genes
Repeating Elements by RepeatMasker
RepeatMasker

Figure 4.1. USCC Genome Browser view tracks of RARA intron 2. The diagram shows genome position with coordinates (Window Position in black, top track). The 'Target Regions' track (green) is the selected region from the design BED file to cover the known breakpoints in *RARA* gene fusions. The lower green track, 'Covered Probes' are the probes available from Agilent bait library to cover the target region. The red 'Missed Regions' track are the gaps in probe coverage. The *RARA* and RefSeq Genes tracks show the *RARA* gene structure. RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences; this track shows the repeat elements that would be screen by the programme. Bait design was optimised to allow repeat sequences by having no masking (see text).

4.6 Bioinformatics

The project was performed in a diagnostic environment with no prior experience of use of NGS and therefore relied entirely on access to external software for bioinformatic analysis. Diagnostic bioinformatic pipelines usually contain a mix of software that has been developed in-house and externally (Association for Clinical Genetic Science, 2016). A bioinformatic pipeline was planned for the project and multiple options were examined before arriving at the version that was accessible and worked, as described in Table 2.3 and Figure 2.4. Open source, external, non-commercial software that has been published following peer review (with detailed documentation, support network and version control) is readily available on line. However, the download, installation and implementation of the software required specialist expertise and knowledge of computer language coding to run the programmes. Without the skill set of a trained bioinformatician, setting up a diagnostic pipeline was particularly challenging, at this stage of the technology's development. Highly developed and tested software with a graphical user interface is highly desirable in this context. Furthermore, a software package, with settings different to those recommended or the default, and the bioinformatics pipeline would need to be validated before introduction into clinical service, by comparison with the validation dataset (Association for Clinical Genetic Science, 2016). Validation of a bioinformatics pipeline was not possible for the project and is the main restriction from widespread dissemination of routine sequencing. Sequencing technology providers have belatedly understood this critical need for diagnostic grade workflow and commercial software is being developed.

Different aspects of the analysis required different analytical software and no single analytical framework can be used in all cases. Agilent SureSelect, a leading hybrid capture system (Hedges *et al.*, 2011; Chilamakuri *et al.*, 2014; Rykalina *et al.*, 2014), was chosen for selection of genomic targets. Agilent SureCall suite of software was therefore accessible for data analysis and was optimised for use with Agilent reagents. SureCall provided a user-friendly analysis tool that incorporated the widely accepted, open source libraries and algorithms, augmented with tools specific to Agilent assays. Sequence read pre-processing, quality control, and read alignment to the reference the genome (BWA-MEM, BWA) were performed using the SureCall package. Single Sample analysis was used to detect mutations and insertions or deletions (indels) in individual samples, using SNPPET, an in-house Agilent algorithm developed specifically for the detection of low allele frequency variants. The software suite provided accurate detection of variants with supporting features to aid the annotation. This software was perfectly adequate to access data processing software for preparation sequence data and was useful for use by a novice. The graphical user interface and incorporation of IGV was particularly helpful. Other software was tried, such as the suite of tools provided in Illumina Basespace but was less intuitive for the novice user and because Agilent reagents were used, access was not available to Illumina's BaseSpace Variant Studio™. SureCall is available for research use only and would need to be extensively, independently validated for diagnostic purposes. SureCall adequately provided initial data alignment, using publicly accessible BWA-MEM (Li, 2013).

4.6.1 Single versus Paired analysis

The project attempted to analyse the samples without the requirement for a matched normal sample, to match the typical diagnostic work-up, whereby normal tissue is not readily available at diagnosis from leukaemia patients. The differentiation of germline and somatic variants is potentially a significant challenge and the approach employed for the project in the filtration of SNVs and indels was designed to minimise the likelihood of misreporting inherited variants. This was successful although it introduces more subjectivity into variant determination than if germline DNA was co-analysed, with a risk that more variants are discarded with uncertain significance and rare (uncurated) pathogenic mutations are missed. This process is helped by a structured protocol for variant annotation and the recent publication of guidelines for the Assignment of variant significance in cancer (Li *et al.*, 2017). The value of paired normal germline DNA as comparator (pre-treatment or remission samples) was demonstrated by the *GATA2* germline data, where paired analysis was used and the controls simplified the distinction between germline and somatic variants. However, erroneous calls are still possible with normal control samples and all variants were scrutinised manually for significance, to reduce the possibility that clinically important mutations could be missed due to mutations persisting in the remission sample, or pre-leukaemic sample in the case of our study (McKerrell *et al.*, 2016).

4.7 Variant annotation and the most useful determinants of pathogenicity

SureCall provide several tools for mutation classification and interpreted the chromosomal location of the mutation with Human Genome Variation Society (HGVS) nomenclature. Each variant was examined for quality and confidence measures, as described in Methods (section 2.6 and Table 3.5), before using links to external databases for further *in silico* analysis. Each variant was evaluated by the software based on its location, amino acid change, and effect on protein function; Sorting Intolerant From Tolerant (SIFT) (Hu & Ng, 2013), was particularly informative, and impact on structure and function of the protein using the Polymorphism Phenotyping v2 (PolyPhen-2) tool. Further information regarding the mutation was then aggregated from various public sources, including Catalogue of Somatic Mutations in Cancer (COSMIC), PubMed, and Locus-Specific Databases. After collecting the various inputs for classification, the proprietary mutation classifier evaluated the significance of the mutation following default guidelines. Each mutation was triaged to categorise the predicted mutations by reviewing supporting evidence. Variant calls passing filtering were examined in the built-in Integrative Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2013) to confirm regional coverage, visualize read alignments and confirm variant calls.

Typically, no single criterion was used to exclude pathogenicity but some were more useful determinants of pathogenicity than others. Generally, the sequencer generated performance scores were adequate and quality scores generated upon alignment were satisfactory. Filtering did not eliminate variants on these criteria alone. A small number of reads with low VAF and/or low coverage were excluded by strand bias; the surviving variants with low VAF are described above. It was very difficult to filter variants by VAF alone as VAF nearing 1.0 or 0.5 resembled germline homozygous or heterozygous fraction allele fraction and caution was exerted in the use of VAF to define a somatic variant due to the mononuclear cells enriched prior to DNA extraction.

Single Nucleotide Polymorphism Database (dbSNP) was consulted for all variants, to exclude previously reported variants that occur naturally in the human population. It is known that such repositories include somatic mutations, and indiscriminate filtering can often remove variants with known pathogenicity (e.g. *IDH2* ^{R140Q}) (Jung *et al.*, 2013; Kenna *et al.*, 2013). Benign germline polymorphisms were defined as variants which are common in the human population, with a population frequency of ≥1% (in dbSNP, a large normal population screening databases) were excluded as probable inherited polymorphisms germline SNPs; this approach eliminated variants which otherwise looked to be convincing cancer genes. All related information was scrutinised in ambiguous annotations and were examined in multiple sources.

An entry in the COSMIC database, with reports in haematopoietic tissue of a confirmed somatic variant but caution was exerted with more ambiguous records. All COSMIC annotations were examined in detail for their relevance. An entry as a somatic variant in an unrelated tumour was considered poor quality evidence for pathogenicity and cross-reference with other information was necessary. Caution that germline SNPs are also present in COSMIC, derived from unpaired data analysis and it is possible that acquired mutants may be present in dbSNP (Kenna *et al.*, 2013; Jung *et al.*, 2013). All related information was scrutinised in ambiguous annotations and multiple points of evidence considered.

4.8 Experimental design and future considerations

4.8.1 Patient cohort

The Oncology Cytogenetics service offers a regional service to all DGHs, specialised tertiary hospitals and paediatric services for leukaemia genetic diagnosis. The case mix is considered representative of the range of patient demographic and AML subtypes. Cells were collected from samples with cells that were surplus to diagnostic requirements. This may introduce limited bias from preferential selection of patients with cellular bone marrows. However, hypercellularity of the involved bone marrow sample is a typical feature of the disease when testing the involved tissue and this is not thought to be an issue. Five samples were rejected for having poor yield of DNA but this was partly due to training issues in DNA extraction. The cases were representative of the typical age range in AML; the median age of the patient cohort was 64 (range 14~84). AML with recurrent genetic abnormalities were included in the group, including 4 CBF-AML (11%), 2 APML (5.5%) and 1 11q23/*KMT2A* (3%). Approximately, the expected numbers of *NPM1* mutated (%) and *FLT3*-ITD (%) were included for this small sample and so random occurrence will explain deviation from expected. 50% of cases were NK-AML and 64% of cases overall were of intermediate prognosis. Morphological subtypes were inconsistently reported from multiple diagnostic centres; in the absence of recurrent genetic abnormalities, most reports made the diagnosis of 'AML' only or 'AML, not otherwise specified' without specifying the morphological subtype. This probably reflects the questionable significance of morphological subtype alone (Swerdlow *et al.*, 2008). Multiple different morphological types were present but without re-interpretation of flow cytometry immunophenotyping, disease subtype could not be analysed further. Similarly, disease history was not commented on in morphology reports and so background of myelodysplasia to identify 'AML with myelodysplasia-related changes' and prior disease history to diagnose 'therapy-related myeloid neoplasm' was unreliable although the cohort contained each of these categories; this final diagnosis would be made at MDT, with full facts from the case presentation. Crude survival data (date of diagnostic marrow - date deceased) was obtained from the *NHS Summary Care Record* (where not obtainable from local database) to inform the influence of genetic testing within the remit of the ethical approval. These records were not scrutinised further and therefore detailed cause of death is not known and will confound some data.

4.8.2 Choice of target genes

It was necessary to define a narrow range of diagnostically relevant and actionable mutations which a comprehensive genomic assay would need to identify. Genetic targets were selected for the assay based on known current clinical utility and likelihood that genes would have future applications, based on current impact on disease classification, efficacy for risk-stratification and potential as biomarkers for emerging clinical applications in the individualisation of treatment (Grimwade *et al.*, 2010; Döhner *et al.*, 2010). To capture sufficient genes with recurrent single nucleotide variation, a list was generated of known genes and revised, taking into account emerging evidence from a range of multiple sources, particularly the 23 ‘significantly mutated’ genes from The Cancer Genome Atlas (The Cancer Genome Atlas Research Network, 2013) and a larger cohort of AML cases with genes listed by frequency and mutation type with an applied false discovery rate calculations to rank them by significance (Broad Institute, 2011). Other sources where mutation profiling of AML has inferred clinical significance in specific genes were also considered (Grossmann *et al.*, 2012; Patel *et al.*, 2012; Conte *et al.*, 2013; Rinke *et al.*, 2013; Hou *et al.*, 2014; Kihara *et al.*, 2014). A panel of 30 genes, generally of the highest frequency resulted, including the 10 most frequently mutated in CN-AML. Many lacked obvious clinical validity at the time of design but were recurrent and sufficiently common in multiple datasets that they may achieve significance by their frequency in future, independently or as part of an informative mutational profile. The least common abnormalities were present in the ‘long tail’ of genes of low frequency and an arbitrary cut-off was finally made for practicality and involved subjective assessment but was satisfactory for this POP study. All exons from these 30 genes, for detection of SNV only, were sequenced. The selection of specific exons of regions to capture was a particular challenge due to varying nomenclature of genetic sub-regions, from different sources in the literature, changes over time, with different version of the Human Genome reference and with exons based on nomenclature of different expressed transcripts. All exons were sequenced from these genes as a result, to ensure that all relevant exons were targeted and for simplicity of defining gene coordinates. A small amount of redundant sequencing resulted, without significantly increasing the sequencing burden (compared to sequencing intervening introns), e.g. the exonic sequencing of 30 genes for SNV only represented 405kbp whereas, the 15 genes with introns was 439kbp. The sequencing showed excellent detection and sequencing depth of all regions selected. The main implication of capturing repetitive sequences was that there was additional sequencing necessary to cover the large intronic regions of structural variant breakpoints. A disproportionate amount of sequencing was necessary to detect translocation breakpoints, particular as both partner genes are also sequenced. With the addition of more genes to the panel for structural variant detection, these then become the bulk of the sequencing, including sequencing a lot of repeats and proportionately little of other targeted regions, or needing to sequence more to get the depth of coverage on the targets.

The benefits of sequencing both genes in a translocation, including the non-oncogenic partner (e.g. *MLLT3, MLLT1* and *RUNX1T1*), was thought to increase the likelihood of capturing the relevant breakpoint. Introns and exons were sequenced encompassing the positions of typical junctions of *FLT3* internal tandem duplications and *KMT2A* partial tandem duplications were sequenced. This worked successfully. Targeting *NUP98* only in the panel also detected *DDX10* as its partner gene. However, by sequencing the oncogenic partner within a pair of fused genes, there is an option to reduce costs with no loss of sensitivity. Sensitivity performance would need to be assessed with more cases.

Several genes were not included in the original panel to restrict costs and because they would be unlikely to be detected for the study but they would need to be included for a comprehensive gene set for diagnosis. This includes *BCR-ABL1* now that this is a provisional diagnostic entity, and is a prognostic marker that also indicates alternative therapy (Arber *et al.*, 2016; Döhner *et al.*, 2016). *RBM15-MKL1* is a rare gene fusion resulting from t(1;22) but might be considered necessary if paediatric disease is to be included. *RPN1-EVI1/MECOM* presents a particularly difficult challenge in this regard; *EVI1/MECOM* translocations are important diagnostic and prognostic subgroups. However, *EVI1/MECOM* is activated from distance with breakpoints spanning a large region upstream and downstream of the *EVI1* gene (Chapter 5, AML pg 56 from Heim & Mitelman, 2009). The cost of sequencing extensive genomic regions was prohibitive for this research, particularly by the price structure of the bait design kit. This will remain lest the production of the targeted panel is reduced or parts of the assay are replaced by standard technique such as FISH.

A number of genes that are uncommonly mutated but achieve significance for other reasons would probably now need be included in a revised set for contemporary diagnosis of AML. Due to our understanding advancing as the result of some landmark reports into the significance of AML genomics, to organised patients into functionally relevant categories, more class-defining genetic abnormalities would need to be included (Metzeler *et al.*, 2016; Papaemmanuil *et al.*, 2016). (Metzeler tested 68 genes and Papaemmanuil 111 genes plus cytogenetic subgroups). The sequencing methodology is modular and target regions can be redefined to include any newly discovered gene mutations without significant changes to laboratory protocols and with only marginal increases in costs, with the main cost in this experiment being redesign of custom target baits.

There is an increasing demand for testing all genes from recognised functional groups, such as cohesion complex, spliceosome, signalling proteins and histone-modifying proteins. The most common genes from these categories are represented in the test gene panel. However, other genes within these groups, which are individually rare, were researched in comprehensive studies. As a result, they are now considered mutually exclusive with other members of the group and typically one representative mutation is adequate for AML pathogenesis. Although rare, genes within these groups are now considered interchangeable and cumulatively comprise a group of clinically and relevant genomic alterations and interrogation of the larger set of cancer-related genes is required for most accurate classification. For example, untested cohesion complex genes *STAG1* and *SMC5* and spliceosome genes *SF1, SF3A1, SF3B1, SRSF2, U2AF2, U2AF35, U2AF65, ZRSR2* and *PRPF40B* may need to be included for a comprehensive profile for AML diagnosis.

4.8.3 CNV

Copy number abnormalities and zygosity changes are key determinants of prognosis in many cancers including AML and MDS. For a comprehensive test for all types of genomic abnormalities, CNV and CN-LOH would be necessary to obtain a complete picture of genomic changes and replace the battery of existing conventional tests. Copy number analysis of a large number of polymorphic SNPs evenly spaced and consistent coverage would allow identification of genome-wide CNV and CN-LOH which has been shown to be informative – cost relative to information gained (Gronseth *et al.*, 2015). In current diagnostic practice, large-scale genomic gains and losses are detected using karyotyping or FISH, but more subtle changes go undetected, as does CN-LOH. CNV analysis was not performed for these samples due to added complexity of experimental design and that markers would have to be designed to comprehensively cover the genome at high density, thereby increasing complexity of design and amount of sequencing required. Coverage was very consistent across different samples and so by normalisation of gene coverage relative to each sample coverage, copy number status would have been possible without matched normal DNA (Bolli *et al.*, 2015; McKerrell *et al.*, 2016). Only few genomic targets were present on each chromosome and it is not anticipated that this would be informative.

4.8.4 Minimal residual disease

All AML patients will harbour acquired genetic abnormalities in their leukaemia cells, and NGS assays, such as the experimental technique, provide testing an extended set of genetic markers for the disease and offer the opportunity for the systematic assessment of minimal residual disease (MRD), applicable to all patients. The identification of clonotypic mutations at the molecular level breakpoints offers the potential for detection of MRD by gDNA at the cellular correlated to absolute cell numbers with advantages over gene expression. The recent development of microfluidics-based systems, (e.g. digital PCR) provides the potential to develop assays with higher sensitivity (Hindson *et al.*, 2013) and personalisation of marker detection for a greater number of patients. However, this also presents challenges for bespoke design of individual assays, standardisation, validation of assays to accreditation standards, and interpretation of the clonal dynamics of a wide range of genetic lesions, which are known to vary in their relapse kinetics. The timescale of identification of a signature profiles and subsequent design of a testing regime for monitoring translation for its universal translation to clinical setting would be demanding. Consensus PCR primer sets would not be possible for any but the common recurrent mutations and gene fusions. However, there is large potential for the design of primers for specific qPCR reactions which can be constructed for patient-specific assays for DNA-based post-treatment MRD measurement, with broader applicability rather than the expression fusion transcripts assays using RNA that are currently necessary (Ommen, 2016).

4.9 Feasibility of introducing NGS protocol into routine practice

The successful trial of this novel NGS shows the principle that the assay is technically feasible and has the potential to offer benefit to patients for the diagnosis of AML. This is encouraging that upon further trial, it could proceed to show analytical validity and clinical utility, to merit its validation for use in the medical laboratory (Mattocks *et al.*, 2010; Association for Clinical Genetic Science, 2015). It is therefore worthy to consider the feasibility of its introduction into routine practice. The technical challenges are described above and would be overcome by gaining more experience in the use of the NGS platform and variant annotation. The performance of the sequencers themselves is no longer an issue for deep sequencing and variants with low allele fraction can be confidently called.

The cost of generating sequencing data is also no longer a restriction; the costs of individual, targeted assays are reasonably priced and are in the same order of magnitude of other genetic and cytogenetic tests. The capital cost is not inconsiderable; taking into account automation, IT infrastructure and data storage capacity, as well as the cost of sequencers. However, this initial outlay is not insurmountable; the cost is transparent and therefore can be planned for. Many reliable analysers are available, from different companies, to meet the needs of different applications and are scalable to meet the needs of the laboratory and their price is no longer prohibitive (Goodwin *et al.*, 2016).

The need to achieve diagnostic standards for routine laboratory practice presents obstacles not typically encountered in research. For the molecular diagnosis of AML, a rapid diagnosis is necessary for the prompt assignment of therapy. The small amount of DNA used in this project can be transferred easily and does not rely on labile RNA preservation and transportation. The analysis of genomes in the clinical context might take several days of sample preparation and sequencing and the complex data, several more days of expert analysis. This brings with it the requirement to have scientific and technical staff with the necessary skills and expertise to deliver the new services. However, from the experience of this project, the timescale for data analysis from individual samples and the clinical interpretation of genome sequencing data with emerging frameworks to guide analysis is not a problem. Once the technique is established, it would be possible to run such assays on high or lower-throughput sequencers, to analyse up to 20 samples once or twice weekly and achieve clinically relevant turnaround times of less than 14 days, thus integrating comfortably into a diagnostic service (McKerrell *et al.*, 2016). Batching tests to make best use of sequencer capacity may be a problem at first. It is likely, however, for the near future, cancer genetics services must accommodate testing by simpler, cost effective methods, such as cytogenetics, FISH and PCR for rapid testing where necessary. Extensive testing would be necessary, for the validation of the new assay. An enormous amount of work is necessary to develop and validate the assay as diagnostic grade tests for clinical use, to standardise the methodology and demonstrate accurate, consistent and reliable performance for UKAS accreditation and for EQA. A significant change management programme would be necessary to deliver the full transformation of genomic services.

Genome sequencing requires a ‘big data’ approach to analysis and interpretation, storage capacity and transfer of this data across networks. This is available in many research facilities but not in many NHS establishments. Data security will be a significant issue, to protect patient identifiable data with linked demographics and personal DNA sequence data. There is a serious practical challenge of the extensive use of genomic data in an expansive but under-resourced health service (Delon & Scott, 2016).

The bioinformatics challenge is significant and presented the largest problems for this project. A multitude of freely available open source and commercial software is readily available for use but requires powerful computer hardware for its use, significant data storage capacity on a protected server, and expertise to choose the most appropriate software to set up and run. Complex data still requires the development and optimisation of mathematical algorithms to identify signatures characteristic of different AML subtypes (Medical Research Council, 2013). Complex bioinformatics is beyond the capability of small laboratories. For the time being, access to trained personnel is absolutely necessary and bioinformatics expertise is a key constraint. The balance between the capacity requirement of bioinformatics and interpretive genetic scientists is not clear; as algorithms are developed and the sector matures the necessary skill mix is likely to change (Delon & Scott, 2016).

Genomic testing by NGS can realistically aid therapeutic decision making, to determine best standard or novel treatment options and to eliminate ineffective or possible harmful therapy. A number of challenges are still apparent for this to reach its full potential. Genomic profiling must be built into new care models and diagnostic pathways. Scientific evidence of clinical utility is necessary to optimise treatment, which ultimately requires the genomic data to be incorporated into randomised controlled trials. The evidence base is gradually accumulating to give scientists the resources to assess the clinical utility of sequence variants and the technology is maturing to be able to deliver these goals. Changes in practice need to be closely aligned with a health economic assessment of their impact on laboratory infrastructure and on patient pathways; the cost of personalised medicine is significant. However, genomics is primed for progression to routine practice and the potential benefits of increasing personalisation of medicine will be significant.

5.0 Conclusion

Cytogenetic and molecular genetic investigations are routinely carried out in AML. They aid diagnosis and are critical for prognostic stratification, to optimise therapy and enhance survival of patients. NGS technology can detect a large number of genetic variants, defining numerous abnormalities at high resolution, with patient-specific profiles. Advances in the understanding of the genomics of AML have defined a landscape of recurrent gene mutations. NGS-based assays will be employed to optimise the genetic profiling of AML, to enhance the efficacy of conventional therapies and maximise the appropriate use of available targeted therapies. This will allow patients with activating mutations in specific genes, which are not currently part of routine testing, to be identified for clinical trials and expand therapeutic possibilities for AML patients. It is anticipated that NGS will become a standard of care in AML within the next few years but optimised assays need to be devised and extensive testing is required for validation of these new platforms.

This project describes a new molecular diagnostic strategy that enables extensive characterisation of the AML genome at diagnosis, which improves partitioning of patients into diagnostic and prognostic groups. The assay shows good performance for highly targeted NGS using genomic DNA, to detect all clinically relevant types of genomic abnormalities, including a broader range of abnormalities than multiple current clinical assays. The capability to identify gene fusions, to breakpoint resolution, adds to the understanding of the use of targeted sequencing for chromosomal translocation detection. This method appears to have the potential to be a powerful and reliable assay and that it could be refined for use in the diagnostic laboratory, with the potential to rationalise multidisciplinary workflows. The cost is comparable to standard testing by using targeted sequencing, facilitating optimal use of analyser capacity. The method justifies further investigation and consideration to develop as a validated assay, for clinical service.

In conclusion, these results suggest that this targeted NGS method, for AML genomic profiling, offers promise to reliably detect actionable genetic alterations across a panel of diverse types of cancer genes. This provides an alternative strategy to other panels for AML, currently in development. If deployed in the clinical diagnostic arena, its implementation has the potential to enhance the ability to identify AML patients at high risk of relapse as well as those that would benefit from molecularly directed therapies, to improve clinical outcomes. This may ultimately impact clinical practice by offering a categorical means to identify alterations of genes and pathways, targeted by existing and emerging drugs, thereby accelerating the practice of personalised cancer medicine.

6.0 References

6. References

Abbas, S., Lugthart, S., Kavelaars, F. G., Schelen, A., Koenders, J. E., Zeilemaker, A., van Putten, W. J. L., Rijneveld, A. W., Löwenberg, B. and Valk, P. J. M. (2010) 'Acquired mutations in the genes encoding IDH1 and IDH2 both are recurrent aberrations in acute myeloid leukemia: prevalence and prognostic value.' *Blood*, 116(12), Sep, pp. 2122-2126.

Abdel-Wahab, O. (2012) 'Molecular genetics of acute myeloid leukemia: clinical implications and opportunities for integrating genomics into clinical practice.' *Hematology*, 17 Suppl 1, Apr, 2012/04/25, pp. S39-42.

Abdel-Wahab, O., Patel, J. and Levine, R. L. (2011) 'Clinical Implications of Novel Mutations in Epigenetic Modifiers in AML.' *Hematology/Oncology Clinics of North America*, 25(6) pp. 1119-1133.

Abel, H. J. and Duncavage, E. J. (2013) 'Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches.' *Cancer Genetics*, 206(12) pp. 432-440.

Abel, H. J., Al-Kateb, H., Cottrell, C. E., Bredemeyer, A. J., Pritchard, C. C., Grossmann, A. H., Wallander, M. L., Pfeifer, J. D., Lockwood, C. M. and Duncavage, E. J. (2014) 'Detection of Gene Rearrangements in Targeted Clinical Next-Generation Sequencing.' *The Journal of Molecular Diagnostics*, 16(4) pp. 405-417.

Abramson, R. (2016) *Overview of Targeted Therapies for Cancer.*: [Online] [Accessed https://www.mycancergenome.org/content/molecular-medicine/overview-of-targeted-therapies-for-cancer/

Adams, J. and Nassiri, M. (2015) 'Acute Promyelocytic Leukemia: A Review and Discussion of Variant Translocations.' *Arch Pathol Lab Med*, 139(10), Oct, 2015/09/29, pp. 1308-1313.

Ades, L., Boehrer, S., Prebet, T., Beyne-Rauzy, O., Legros, L., Ravoet, C., Dreyfus, F., Stamatoullas, A., Pierre Chaury, M., Delaunay, J., Laurent, G., Vey, N., Burcheri, S., Mbida, R.-M., Hoarau, N., Gardin, C. and Fenaux, P. (2009) 'Efficacy and safety of lenalidomide in intermediate-2 or high-risk myelodysplastic syndromes with 5q deletion: results of a phase 2 study.' *Blood*, 113(17), April 23, 2009, pp. 3947-3952.

Akiki, S., Dyer, S. A., Grimwade, D., Ivey, A., Abou-Zeid, N., Borrow, J., Jeffries, S., Caddick, J., Newell, H., Begum, S., Tawana, K., Mason, J., Velangi, M. and Griffiths, M. (2013) 'NUP98-NSD1 fusion in association with FLT3-ITD mutation identifies a prognostically relevant subgroup of pediatric acute myeloid leukemia patients suitable for monitoring by real time quantitative PCR.' *Genes, Chromosomes and Cancer*, 52(11) pp. 1053-1064.

Albiero, E., Madeo, D., Bolli, N., Giaretta, I., Di Bona, E., Martelli, M. F., Nicoletti, I., Rodeghiero, F. and Falini, B. (2007) 'Identification and functional characterization of a cytoplasmic nucleophosmin leukaemic mutant generated by a novel exon-11 NPM1 mutation.' *Leukemia*, 21(5), May, pp. 1099-1103.

Alkan, C., Coe, B. P. and Eichler, E. E. (2011) 'Applications of Next-Generation Sequencing; Genome structural variation discovery and genotyping.' *Nature Reviews Genetics*, 12(5), May, pp. 363-375.

Alpermann, T., Schnittger, S., Eder, C., Dicker, F., Meggendorfer, M., Kern, W., Schmid, C., Aul, C., Staib, P., Wendtner, C.-M., Schmitz, N., Haferlach, C. and Haferlach, T. (2016) 'Molecular subtypes of NPM1 mutations have different clinical profiles, specific patterns of accompanying molecular mutations and varying outcomes in intermediate risk acute myeloid leukemia.' *Haematologica*, 101(2) pp. e55-e58.

Anderson, K., Lutz, C., van Delft, F. W., Bateman, C. M., Guo, Y., Colman, S. M., Kempski, H., Moorman, A. V., Titley, I., Swansbury, J., Kearney, L., Enver, T. and Greaves, M. (2011) 'Genetic variegation of clonal architecture and propagating cells in leukaemia.' *Nature*, 469(7330), Jan 20, 2010/12/17, pp. 356-361.

Anelli, L., Pasciolla, C., Zagaria, A., Specchia, G. and Albano, F. (2017) 'Monosomal karyotype in myeloid neoplasias: a literature review.' *Onco Targets Ther*, 10 2017/05/04, pp. 2163-2171.

Arber, D. A., Brunning, R. D., Orazi, A., Bain, B. J., Porwit, A., Vardiman, J. W., Le Beau, M. M. and Greenberg, P. L. (2008) 'Acute myeloid leukaemia with myelodysplasia-related changes.' *In* Swerdlow, S. H., Campo, E., Harris, N. L., Jaffe, E. S., Pilieri, S. A., Stein, H., Thiele, J. and Vardiman, J. W. (eds.) *WHO Classification of tumours of haematopoietic and lymphoid tissues*
4th ed., Lyon: IARC, pp. pg 124-126.

Arber, D. A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., Bloomfield, C. D., Cazzola, M. and Vardiman, J. W. (2016) 'The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia.' *Blood*, 127(20), May 19, 2016/04/14, pp. 2391-2405.

Arenillas, L., Mallo, M., Ramos, F., Guinta, K., Barragan, E., Lumbreras, E., Larrayoz, M. J., De Paz, R., Tormo, M., Abaigar, M., Pedro, C., Cervera, J., Such, E., Jose Calasanz, M., Diez-Campelo, M., Sanz, G. F., Hernandez, J. M., Luno, E., Saumell, S., Maciejewski, J., Florensa, L. and Sole, F. (2013) 'Single nucleotide polymorphism array karyotyping: a diagnostic and prognostic tool in myelodysplastic syndromes with unsuccessful conventional cytogenetic testing.' *"Genes, Chromosomes and Cancer"*, 52(12), Dec, 2013/10/15, pp. 1167-1177.

Arney, K. (2016) *Herding Hemingway's Cats.* 1st ed., Bloomsbury Sigma. London: Bloomsbury Publishing Plc.

Association for Clinical Genetic Science. (2013) Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics In: Association for Clinical Genetic Science.

Association for Clinical Genetic Science. (2015) Practice guidelines for Targeted Next Generation Sequencing Analysis and Interpretation. Best practice guidelines. (22/02/2016).

Association for Clinical Genetic Science. (2016) Guidelines for development and validation of software, with particular focus on bioinformatics pipelines for processing NGS data. Best practice guidelines.

Au, C. H., Wa, A., Ho, D. N., Chan, T. L. and Ma, E. S. K. (2016) 'Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms.' *Diagnostic Pathology*, 11, Jan,

Bacher, U., Haferlach, C., Schnittger, S., Kohlmann, A., Kern, W. and Haferlach, T. (2010) 'Mutations of the TET2 and CBL genes: novel molecular markers in myeloid malignancies.' *Annals of Hematology*,

Bacher, U., Weissmann, S., Kohlmann, A., Schindela, S., Alpermann, T., Schnittger, S., Kern, W., Haferlach, T. and Haferlach, C. (2012) 'TET2 deletions are a recurrent but rare phenomenon in myeloid malignancies and are frequently accompanied by TET2 mutations on the remaining allele.' *British Journal of Haematology*, 156(1), Jan, 2011/10/25, pp. 67-75.

Bacolla, A., Tainer, J. A., Vasquez, K. M. and Cooper, D. N. (2016) 'Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences.' *Nucleic Acids Res*, 44(12), Jul 08, 2016/04/17, pp. 5673-5688.

Baldus, C. D. and Bullinger, L. (2008) 'Gene Expression With Prognostic Implications in Cytogenetically Normal Acute Myeloid Leukemia.' *Seminars in Oncology*, 35(4) pp. 356-364.

Balgobind, B. V., Raimondi, S. C., Harbott, J., Zimmermann, M., Alonzo, T. A., Auvrignon, A., Beverloo, H. B., Chang, M., Creutzig, U., Dworzak, M. N., Forestier, E., Gibson, B., Hasle, H., Harrison, C. J., Heerema, N. A., Kaspers, G. J. L., Leszl, A., Litvinko, N., Nigro, L. L., Morimoto, A., Perot, C., Pieters, R., Reinhardt, D., Rubnitz, J. E., Smith, F. O., Stary, J., Stasevich, I., Strehl, S., Taga, T., Tomizawa, D., Webb, D., Zemanova, Z., Zwaan, C. M. and van den Heuvel-Eibrink, M. M. (2009) 'Novel prognostic subgroups in childhood 11q23/MLL-rearranged acute myeloid leukemia: results of an international retrospective study.' *Blood*, 114(12), September 17, 2009, pp. 2489-2496.

Barnes, B. and Dupre, J. (2008) *Genomes and what to make of them.* Chicago, Ill.: University of Chicago Press ; Bristol : University Presses Marketing [distributor].

Basecke, J., Whelan, J. T., Griesinger, F. and Bertrand, F. E. (2006) 'The MLL partial tandem duplication in acute myeloid leukaemia.' *British Journal of Haematology*, 135(4), Nov, pp. 438-449.

Bashir, A., Volik, S., Collins, C., Bafna, V. and Raphael, B. J. (2008) 'Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer.' *PLoS Comput Biol*, 4(4), Apr, 2008/04/12, p. e1000051.

Bennett, J. M., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D. A., Gralnick, H. R. and Sultan, C. (1976) 'Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group.' *British Journal of Haematology*, 33(4), Aug, 1976/08/01, pp. 451-458.

Bennett, J. M., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D. A. G., Gralnick, H. R. and Sultan, C. (1985) 'Proposed Revised Criteria for the Classification of Acute Myeloid Leukemia A Report of the French-American-British Cooperative Group.' *Annals of Internal Medicine*, 103(4) pp. 620-625.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry.' *Nature*, 456(7218), Nov 06, 2008/11/07, pp. 53-59.

Bertier, G., Carrot-Zhang, J., Ragoussis, V. and Joly, Y. (2016) 'Integrating precision cancer medicine into healthcare-policy, practice, and research challenges.' *Genome Medicine*, 8, Oct, Bienz, M., Ludwig, M., Leibundgut, E. O., Mueller, B. U., Ratschiller, D., Solenthaler, M., Fey, M. F. and Pabst, T. (2005) 'Risk assessment in patients with acute myeloid leukemia and a normal karyotype.' *Clinical Cancer Research*, 11(4), Feb 15, 2005/03/05, pp. 1416-1424.

Bignell, G. R., Santarius, T., Pole, J. C., Butler, A. P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., Campbell, P., Quail, M., Plumb, B., Matthews, L., McLay, K., Edwards, P. A., Rogers, J., Wooster, R., Futreal, P. A. and Stratton, M. R. (2007) 'Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution.' *Genome Res*, 17(9), Sep, 2007/08/07, pp. 1296-1303.

Bodor, C., Renneville, A., Smith, M., Charazac, A., Iqbal, S., Etancelin, P., Cavenagh, J., Barnett, M. J., Kramarzova, K., Krishnan, B., Matolcsy, A., Preudhomme, C., Fitzgibbon, J. and Owen, C. (2014) 'Germ-line GATA2 p.THR354MET mutation in familial myelodysplastic syndrome with acquired monosomy 7 and ASXL1 mutation demonstrating rapid onset and poor survival.' *Haematologica-the Hematology Journal*, 97(6), Jun, pp. 890-894.

Bolli, N., Manes, N., McKerrell, T., Chi, J., Park, N., Gundem, G., Quail, M. A., Sathiaseelan, V., Herman, B., Crawley, C., Craig, J. I., Conte, N., Grove, C., Papaemmanuil, E., Campbell, P. J., Varela, I., Costeas, P. and Vassiliou, G. S. (2015) 'Characterization of gene mutations and copy number changes in acute myeloid leukemia using a rapid target enrichment protocol.' *Haematologica*, 100(2), Feb, 2014/11/09, pp. 214-222.

Breems, D. A. and Löwenberg, B. (2011) 'Acute myeloid leukemia with monosomal karyotype at the far end of the unfavorable prognostic spectrum.' *Haematologica-the Hematology Journal*, 96(4), Apr, pp. 491-493.

Broad Institute. (2011) *Acute Myeloid Leukemia: Mutation Analysis (MutSig).* [Online] [Accessed on 26/06/2014] http://gdac.broadinstitute.org/runs/analyses__2012_02_17/reports/cancer/LAML/mutation/significance/nozzle.html

Broad Institute. (2017) *Picard*. [Online] [Accessed on 25/04/2017] http://broadinstitute.github.io/picard/

Bullinger, L. and Fröhling, S. (2012) 'Array-Based Cytogenetic Approaches in Acute Myeloid Leukemia: Clinical Impact and Biological Insights.' *Seminars in Oncology*, 39(1) pp. 37-46.

Bullinger, L., Döhner, K. and Döhner, H. (2017) 'Genomics of Acute Myeloid Leukemia Diagnosis and Pathways.' *J Clin Oncol*, 35(9), Mar 20, 2017/03/16, pp. 934-946.

Bullinger, L., Kronke, J., Schon, C., Radtke, I., Urlbauer, K., Botzenhardt, U., Gaidzik, V., Cario, A., Senger, C., Schlenk, R. F., Downing, J. R., Holzmann, K., Döhner, K. and Döhner, H. (2010) 'Identification of acquired copy number alterations and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis.' *Leukemia*, 24(2), Feb, 2009/12/18, pp. 438-449.

Burnett, A. K. (2012) 'Treatment of acute myeloid leukemia: are we making progress?' *Hematology Am Soc Hematol Educ Program*, 2012 2012/12/13, pp. 1-6.

Butler, D. (2010) 'Human genome at ten: Science after the sequence.' *Nature*, 465(7301), Jun 24, 2010/06/26, pp. 1000-1001.

Campbell, P., Stephens, P., Pleasance, E., O'meara, S., Li, H., Santarius, T., Stebbings, L., Leroy, C., Edkins, S., Hardy, C., Teague, J., Menzies, A., Goodhead, I., Turner, D., Clee, C., Quail, M., Cox, A., Brown, C., Durbin, R., Hurles, M., Edwards, P., Bignell, G., Stratton, M. and Futreal, P. (2008) 'Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.' *Nat Genet*, 40 pp. 722 - 729.

Cancer Research UK. (2015) *Molecular diagnostic provision in the NHS in England* August 2015. Cancer Research UK.

Carbuccia, N., Trouplin, V., Gelsi-Boyer, V., Murati, A., Rocquain, J., Adelaide, J., Olschwang, S., Xerri, L., Vey, N., Chaffanet, M., Birnbaum, D. and Mozziconacci, M. J. (2010) 'Mutual exclusion of ASXL1 and NPM1 mutations in a series of acute myeloid leukemias.' *Leukemia*, 24(2), Feb, 2009/10/30, pp. 469-473.

Cardiff University. (2015) *AML 19*. [Online] [Accessed on 17/04/2016] http://medicine.cf.ac.uk/HCTU/our-trials/aml-19/

Chen, C. W. and Armstrong, S. A. (2015) 'Targeting DOT1L and HOX gene expression in MLL-rearranged leukemia and beyond.' *Exp Hematol*, 43(8), Aug, 2015/06/30, pp. 673-684.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L. and Mardis, E. R. (2009) 'BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.' *Nat Meth*, 6(9) pp. 677-681.

Chen, W., Kalscheuer, V., Tzschach, A., Menzel, C., Ullmann, R., Schulz, M., Erdogan, F., Li, N., Kijas, Z., Arkesteijn, G., Pajares, I., Goetz-Sothmann, M., Heinrich, U., Rost, I., Dufke, A., Grasshoff, U., Glaeser, B., Vingron, M. and Ropers, H. (2008) 'Mapping translocation breakpoints by next-generation sequencing.' *Genome Res*, 18 pp. 1143 - 1149.

Chilamakuri, C. S. R., Lorenz, S., Madoui, M. A., Vodak, D., Sun, J. C., Hovig, E., Myklebost, O. and Meza-Zepeda, L. A. (2014) 'Performance comparison of four exome capture systems for deep sequencing.' *BMC Genomics*, 15, Jun, p. 13.

Chin, L., Andersen, J. N. and Futreal, P. A. (2011) 'Cancer genomics: from discovery science to personalized medicine.' *Nature Medicine*, 17(3) pp. 297-303.

Cho, Y. U., Jang, S., Seo, E. J., Park, C. J., Chi, H. S., Kim, D. Y., Lee, J. H., Lee, J. H., Lee, K. H., Koh, K. N., Im, H. J., Seo, J. J., Park, S. H., Park, Y. M. and Lee, J. K. (2015) 'Preferential occurrence of spliceosome mutations in acute myeloid leukemia with preceding myelodysplastic syndrome and/or myelodysplasia morphology.' *Leukemia & Lymphoma*, 56(8), Aug, pp. 2301-2308.

Chong, C. E., Venugopal, P., Stokes, P. H., Lee, Y. K., Brautigan, P. J., Yeung, D. T. O., Babic, M., Engler, G. A., Lane, S. W., Klingler-Hoffmann, M., Matthews, J. M., D'Andrea, R. J., Brown, A. L., Hahn, C. N. and Scott, H. S. (2017) 'Differential effects on gene transcription and hematopoietic differentiation correlate with GATA2 mutant disease phenotypes.' *Leukemia*, Jun 23, 2017/06/24,

Chou, W. C., Chou, S. C., Liu, C. Y., Chen, C. Y., Hou, H. A., Kuo, Y. Y., Lee, M. C., Ko, B. S., Tang, J. L., Yao, M., Tsay, W., Wu, S. J., Huang, S. Y., Hsu, S. C., Chen, Y. C., Chang, Y. C., Kuo, K. T., Lee, F. Y., Liu, M. C., Liu, C. W., Tseng, M. H., Huang, C. F. and Tien, H. F. (2011) 'TET2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics.' *Blood*, 118(14), Oct, pp. 3803-3810.

Churpek, J. E., Pyrtel, K., Kanchi, K. L., Shao, J., Koboldt, D., Miller, C. A., Shen, D., Fulton, R., O'Laughlin, M., Fronick, C., Pusic, I., Uy, G. L., Braunstein, E. M., Levis, M., Ross, J., Elliott, K., Heath, S., Jiang, A., Westervelt, P., DiPersio, J. F., Link, D. C., Walter, M. J., Welch, J., Wilson, R., Ley, T. J., Godley, L. A. and Graubert, T. A. (2015) 'Genomic analysis of germ line and somatic variants in familial myelodysplasia/acute myeloid leukemia.' *Blood*, 126(22), Nov 26, 2015/10/24, pp. 2484-2490.

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Euskirchen, G. and Butte, A. J. (2011) 'Performance comparison of exome DNA sequencing technologies.' *Nat Biotechnol.*, 29

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010) 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.' *Nucleic Acids Research*, 38(6), April 1, 2010, pp. 1767-1771.

Collin, M., Dickinson, R. and Bigley, V. (2015) 'Haematopoietic and immune defects associated with GATA2 mutation.' *British Journal of Haematology*, 169(2), Apr, pp. 173-187.

Collins, F. S. (2010a) 'Has the revolution arrived?' *Nature*, 464(7289), Apr 01, 2010/04/03, pp. 674-675.

Collins, F. S. (2010b) *The Language of Life; DNA and the Revolution in Personalized Medicine.* London: Profile Books Ltd.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. and Mortazavi, A. (2016) 'A survey of best practices for RNA-seq data analysis.' *Genome Biology*, 17, 01/26, p. 13.

Conte, N., Varela, I., Grove, C., Manes, N., Yusa, K., Moreno, T., Segonds-Pichon, A., Bench, A., Gudgin, E., Herman, B., Bolli, N., Ellis, P., Haddad, D., Costeas, P., Rad, R., Scott, M., Huntly, B., Bradley, A. and Vassiliou, G. S. (2013) 'Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture.' *Leukemia*,

Cooley, L. D., Lebo, M., Li, M. M., Slovak, M. L., Wolff, D. J. and Working Grp Amer Coll Med, G. (2013) 'American College of Medical Genetics and Genomics technical standards and guidelines: microarray analysis for chromosome abnormalities in neoplastic disorders.' *Genetics in Medicine*, 15(6), Jun, pp. 484-494.

Coombs, C. C., Tavakkoli, M. and Tallman, M. S. (2015) 'Acute promyelocytic leukemia: where did we start, where are we now, and the future.' *Blood Cancer J*, 5, Apr 17, 2015/04/18, p. e304.

Corces-Zimmerman, M. R. and Majeti, R. (2014) 'Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis.' *Leukemia*, 28(12), Dec, 2014/07/10, pp. 2276-2282.

Corces-Zimmerman, M. R., Hong, W.-J., Weissman, I. L., Medeiros, B. C. and Majeti, R. (2014) 'Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission.' *Proceedings of the National Academy of Sciences*, 111(7), February 18, 2014, pp. 2548-2553.

Cornelissen, J. J. and Blaise, D. (2016) 'Hematopoietic stem cell transplantation for patients with AML in first complete remission.' *Blood*, 127(1), 2016-01-07 00:00:00, pp. 62-70.

Cortés-Lavaud, X., Landecho, M. F., Maicas, M., Urquiza, L., Merino, J., Moreno-Miralles, I. and Odero, M. D. (2015) 'GATA2 Germline Mutations Impair GATA2 Transcription, Causing Haploinsufficiency: Functional Analysis of the p.Arg396Gln Mutation.' *The Journal of Immunology*, 194(5), March 1, 2015, pp. 2190-2198.

Cree, I. (2016) 'Overview of Molecular Pathology.' *The Bulletin of the Royal College of Pathologists* (173), January 2016, pp. 6-9.

Creutzig, U., Zimmermann, M., Reinhardt, D., Rasche, M., von Neuhoff, C., Alpermann, T., Dworzak, M., Perglerova, K., Zemanova, Z., Tchinda, J., Bradtke, J., Thiede, C. and Haferlach, C. (2016) 'Changes in cytogenetics and molecular genetics in acute myeloid leukemia from childhood to adult age groups.' *Cancer*, 122(24), Dec 15, 2016/08/17, pp. 3821-3830.

Crispino, J. D. and Horwitz, M. S. (2017) 'GATA factor mutations in hematologic disease.' *Blood*, 129(15), Apr 13, 2017/02/10, pp. 2103-2110.

Cronin, M. and Ross, J. S. (2011) 'Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology.' *Biomarkers in Medicine*, 5(3), Jun, pp. 293-305.

Czepulkowski, B. (2001) 'Basic techniques for the preparation and analysis of chromosomes from bone marrow and leukaemic blood.' *In* Rooney, D. E. (ed.) *Human Cytogenetics: malignancy and acquired abnormalities*. 3rd ed., Oxford: Oxford University Press, p. 2~26.

Daber, R., Sukhadia, S. and Morrissette, J. J. (2013) 'Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets.' *Cancer Genet*, 206(12), Dec, 2014/02/18, pp. 441-448.

Damm, F., Markus, B., Thol, F., Morgan, M., Gohring, G., Schlegelberger, B., Krauter, J., Heuser, M., Bernard, O. A. and Ganser, A. (2014) 'TET2 mutations in cytogenetically normal acute myeloid leukemia: clinical implications and evolutionary patterns.' *"Genes, Chromosomes and Cancer"*, 53(10), Oct, 2014/06/06, pp. 824-832.

de Pater, E., Kaimakis, P., Vink, C. S., Yokomizo, T., Yamada-Inagawa, T., van der Linden, R., Kartalaei, P. S., Camper, S. A., Speck, N. and Dzierzak, E. (2013) 'Gata2 is required for HSC generation and survival.' *J Exp Med*, 210(13), Dec 16, 2013/12/04, pp. 2843-2850.

Delon, I. and Scott, M. (2016) 'Genomic transformation in UK pathology: clinical applications, challenges and workforce adaption.' *The Bulletin of the Royal College of Pathologists* (174), April 2016, pp. 94-98.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R. and Hartl, C. (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data.' *Nature Genetics.*, 43

Devillier, R., Gelsi-Boyer, V., Brecqueville, M., Carbuccia, N., Murati, A., Vey, N., Birnbaum, D. and Mozziconacci, M.-J. (2012) 'Acute myeloid leukemia with myelodysplasia-related changes are characterized by a specific molecular pattern with high frequency of ASXL1 mutations.' *American Journal of Hematology*, 87(7) pp. 659-662.

Dickinson, R. E., Griffin, H., Bigley, V., Reynard, L. N., Hussain, R., Haniffa, M., Lakey, J. H., Rahman, T., Wang, X. N., McGovern, N., Pagan, S., Cookson, S., McDonald, D., Chua, I., Wallis, J., Cant, A., Wright, M., Keavney, B., Chinnery, P. F., Loughlin, J., Hambleton, S., Santibanez-Koref, M. and Collin, M. (2011) 'Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency.' *Blood*, 118(10), Sep 08, 2011/07/19, pp. 2656-2658.

Dickinson, R. E., Milne, P., Jardine, L., Zandi, S., Swierczek, S. I., McGovern, N., Cookson, S., Ferozepurwalla, Z., Langridge, A., Pagan, S., Gennery, A., Heiskanen-Kosma, T., Hämäläinen, S., Seppänen, M., Helbert, M., Tholouli, E., Gambineri, E., Reykdal, S., Gottfredsson, M., Thaventhiran, J. E., Morris, E., Hirschfield, G., Richter, A. G., Jolles, S., Bacon, C. M., Hambleton, S., Haniffa, M., Bryceson, Y., Allen, C., Prchal, J. T., Dick, J. E., Bigley, V. and Collin, M. (2014) 'The evolution of cellular deficiency in GATA2 mutation.' *Blood*, 123(6), 2014-02-06 00:00:00, pp. 863-874.

Dienstmann, R., Dong, F., Borger, D., Dias-Santagata, D., Ellisen, L. W., Le, L. P. and Iafrate, A. J. (2014) 'Standardized decision support in next generation sequencing reports of somatic cancer variants.' *Molecular Oncology*, 8(5) pp. 859-873.

Dietel, M., Johrens, K., Laffert, M. V., Hummel, M., Blaker, H., Pfitzner, B. M., Lehmann, A., Denkert, C., Darb-Esfahani, S., Lenze, D., Heppner, F. L., Koch, A., Sers, C., Klauschen, F. and Anagnostopoulos, I. (2015) 'A 2015 update on predictive molecular pathology and its role in targeted cancer therapy: a review focussing on clinical relevance.' *Cancer Gene Ther*, 22(9), Sep, 2015/09/12, pp. 417-430.

Ding, L., Wendl, M. C., McMichael, J. F. and Raphael, B. J. (2014) 'Expanding the computational toolbox for mining cancer genomes.' *Nat Rev Genet*, 15(8), Aug, 2014/07/09, pp. 556-570.

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) 'Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.' *Nucleic Acids Res*, 36(16), Sep, 2008/07/29, p. e105.

Döhner, H. and Gaidzik, V. I. (2011) 'Impact of genetic features on treatment decisions in AML.' *Hematology Am Soc Hematol Educ Program*, 2011 2011/12/14, pp. 36-42.

Döhner, H., Estey, E. H., Amadori, S., Appelbaum, F. R., Buchner, T., Burnett, A. K., Dombret, H., Fenaux, P., Grimwade, D., Larson, R. A., Lo-Coco, F., Naoe, T., Niederwieser, D., Ossenkoppele, G. J., Sanz, M. A., Sierra, J., Tallman, M. S., Löwenberg, B. and Bloomfield, C. D. (2010) 'Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet.' *Blood*, 115(3), January 21, 2010, pp. 453-474.

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., Dombret, H., Ebert, B. L., Fenaux, P., Larson, R. A., Levine, R. L., Lo-Coco, F., Naoe, T., Niederwieser, D., Ossenkoppele, G. J., Sanz, M., Sierra, J., Tallman, M. S., Tien, H.-F., Wei, A. H., Löwenberg, B. and Bloomfield, C. D. (2016) 'Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel.' *Blood*,

Döhner, K., Tobis, K., Ulrich, R., Fröhling, S., Benner, A., Schlenk, R. F. and Döhner, H. (2002) 'Prognostic significance of partial tandem duplications of the MLL gene in adult patients 16 to 60 years old with acute myeloid leukemia and normal cytogenetics: a study of the Acute Myeloid Leukemia Study Group Ulm.' *Journal of Clinical Oncology*, 20(15), Aug 1, 2002/08/01, pp. 3254-3261.

Döhner, K., Schlenk, R. F., Habdank, M., Scholl, C., Rücker, F. G., Corbacioglu, A., Bullinger, L., Fröhling, S. and Döhner, H. (2005) 'Mutant nucleophosmin (NPM1) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: interaction with other gene mutations.' *Blood*, 106(12), Dec 1, 2005/07/30, pp. 3740-3746.

Dombret, H. and Gardin, C. (2016) 'An update of current treatments for adult acute myeloid leukemia.' *Blood*, 127(1), 2016-01-07 00:00:00, pp. 53-61.

Duncavage, E. J., Abel, H. J., Szankasi, P., Kelley, T. W. and Pfeifer, J. D. (2012) 'Targeted next generation sequencing of clinically significant gene mutations and translocations in leukemia.' *Modern Pathology*,

Duployez, N., Marceau-Renaut, A., Boissel, N., Petit, A., Bucci, M., Geffroy, S., Lapillonne, H., Renneville, A., Ragu, C., Figeac, M., Celli-Lebras, K., Lacombe, C., Micol, J.-B., Abdel-Wahab, O., Cornillet, P., Ifrah, N., Dombret, H., Leverger, G., Jourdan, E. and Preudhomme, C. (2016) 'Comprehensive mutational profiling of core binding factor acute myeloid leukemia.' *Blood*, 127(20) pp. 2451-2459.

ECMC. *Precision Medicine at CRUK: Stratified Medicine Programme*. Experimental Cancer Medicine Centre Network,. [Online] [Accessed on 17/11/2016] http://www.ecmcnetwork.org.uk/stratified-medicine-programme

El Hajj, H., Dassouki, Z., Berthier, C., Raffoux, E., Ades, L., Legrand, O., Hleihel, R., Sahin, U., Tawil, N., Salameh, A., Zibara, K., Darwiche, N., Mohty, M., Dombret, H., Fenaux, P., de The, H. and Bazarbachi, A. (2015) 'Retinoic acid and arsenic trioxide trigger degradation of mutated NPM1, resulting in apoptosis of AML cells.' *Blood*, 125(22), May 28, 2015/03/25, pp. 3447-3454.

Engen, C. B., Hajjar, E. and Gjertsen, B. T. (2016) 'Development of Personalized Molecular Therapy for Acute Myeloid Leukemia.' *Curr Pharm Biotechnol*, 17(1) 2015/10/01, pp. 20-29.

Erba, H. P. *Finding the optimal combination therapy for the treatment of newly diagnosed AML in older patients unfit for intensive therapy*. (2015) 183-191.http://www.sciencedirect.com/science/article/pii/S0145212614003798

Falini, B., Nicoletti, I., Martelli, M. F. and Mecucci, C. (2007) 'Acute myeloid leukemia carrying cytoplasmic/mutated nucleophosmin (NPMc(+) AML): biologic and clinical features.' *Blood*, 109(3), Feb, pp. 874-885.

Falini, B., Mecucci, C., Tiacci, E., Alcalay, M., Rosati, R., Pasqualucci, L., La Starza, R., Diverio, D., Colombo, E., Santucci, A., Bigerna, B., Pacini, R., Pucciarini, A., Liso, A., Vignetti, M., Fazi, P., Meani, N., Pettirossi, V., Saglio, G., Mandelli, F., Lo-Coco, F., Pelicci, P.-G. and Martelli, M. F. (2005) 'Cytoplasmic Nucleophosmin in Acute Myelogenous Leukemia with a Normal Karyotype.' *New England Journal of Medicine*, 352(3) pp. 254-266.

Fasan, A., Haferlach, C., Alpermann, T., Jeromin, S., Grossmann, V., Eder, C., Weissmann, S., Dicker, F., Kohlmann, A., Schindela, S., Kern, W., Haferlach, T. and Schnittger, S. (2014) 'The role of different genetic subtypes of CEBPA mutated AML.' *Leukemia*, 28(4) pp. 794-803.

Fenaux, P., Mufti, G. J., Hellstrom-Lindberg, E., Santini, V., Finelli, C., Giagounidis, A., Schoch, R., Gattermann, N., Sanz, G., List, A., Gore, S. D., Seymour, J. F., Bennett, J. M., Byrd, J., Backstrom, J., Zimmerman, L., McKenzie, D., Beach, C. and Silverman, L. R. (2009) 'Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study.' *Lancet Oncol*, 10(3), Mar, 2009/02/24, pp. 223-232.

Feuk, L., Carson, A. R. and Scherer, S. W. (2006) 'Structural variation in the human genome.' *Nat Rev Genet*, 7(2) pp. 85-97.

Fisher, J. B., McNulty, M., Burke, M. J., Crispino, J. D. and Rao, S. (2017) 'Cohesin Mutations in Myeloid Malignancies.' *Trends Cancer*, 3(4), Apr, 2017/06/20, pp. 282-293.

Fisher, R., Pusztai, L. and Swanton, C. (2013) 'Cancer heterogeneity: implications for targeted therapeutics.' *British Journal of Cancer*, 108(3), Feb 19, 2013/01/10, pp. 479-485.

Fitzgibbon, J., Smith, L. L., Raghavan, M., Smith, M. L., Debernardi, S., Skoulakis, S., Lillington, D., Lister, T. A. and Young, B. D. (2005) 'Association between acquired uniparental disomy and homozygous gene mutation in acute myeloid leukemias.' *Cancer Research*, 65(20), Oct 15, 2005/10/19, pp. 9152-9154.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U. and Campbell, P. J. (2015) 'COSMIC: exploring the world's knowledge of somatic mutations in human cancer.' *Nucleic Acids Research*, 43(D1), January 28, 2015, pp. D805-D811.

Fox, E. J., Salk, J. J. and Loeb, L. A. (2009) 'Cancer Genome Sequencing-An Interim Analysis.' *Cancer Research*, 69(12), Jun, pp. 4948-4950.

Fröhling, S., Schlenk, R. F., Breitruck, J., Benner, A., Kreitmeier, S., Tobis, K., Döhner, H. and Döhner, K. (2002) 'Prognostic significance of activating FLT3 mutations in younger adults (16 to 60 years) with acute myeloid leukemia and normal cytogenetics: a study of the AML Study Group Ulm.' *Blood*, 100(13), Dec 15, 2002/10/24, pp. 4372-4380.

Fröhling, S., Schlenk, R. F., Stolze, I., Bihlmayr, J., Benner, A., Kreitmeier, S., Tobis, K., Döhner, H. and Döhner, K. (2004) 'CEBPA Mutations in Younger Adults With Acute Myeloid Leukemia and Normal Cytogenetics: Prognostic Relevance and Analysis of Cooperating Mutations.' *Journal of Clinical Oncology*, 22(4), February 15, 2004, pp. 624-633.

Fujiwara, T., Fukuhara, N., Funayama, R., Nariai, N., Kamata, M., Nagashima, T., Kojima, K., Onishi, Y., Sasahara, Y., Ishizawa, K., Nagasaki, M., Nakayama, K. and Harigae, H. (2014) 'Identification of acquired mutations by whole-genome sequencing in GATA-2 deficiency evolving into myelodysplasia and acute leukemia.' *Annals of Hematology*, 93(9), Sep, 2014/05/02, pp. 1515-1522.

Fullwood, M. J., Wei, C. L., Liu, E. T. and Ruan, Y. J. (2009) 'Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses.' *Genome Research*, 19(4) pp. 521-532.

Gaidzik, V. I., Bullinger, L., Schlenk, R. F., Zimmermann, A. S., Rock, J., Paschka, P., Corbacioglu, A., Krauter, J., Schlegelberger, B., Ganser, A., Spath, D., Kundgen, A., Schmidt-Wolf, I. G., Gotze, K., Nachbaur, D., Pfreundschuh, M., Horst, H. A., Döhner, H. and Döhner, K. (2011) 'RUNX1 mutations in acute myeloid leukemia: results from a comprehensive genetic and clinical analysis from the AML study group.' *J Clin Oncol*, 29(10), Apr 01, 2011/02/24, pp. 1364-1372.

Gaidzik, V. I., Teleanu, V., Papaemmanuil, E., Weber, D., Paschka, P., Hahn, J., Wallrabenstein, T., Kolbinger, B., Kohne, C. H., Horst, H. A., Brossart, P., Held, G., Kundgen, A., Ringhoffer, M., Gotze, K., Rummel, M., Gerstung, M., Campbell, P., Kraus, J. M., Kestler, H. A., Thol, F., Heuser, M., Schlegelberger, B., Ganser, A., Bullinger, L., Schlenk, R. F., Döhner, K. and Döhner, H. (2016) 'RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features.' *Leukemia*, 30(11), Nov, 2016/11/03, pp. 2160-2168.

Gale, R. E., Green, C., Allen, C., Mead, A. J., Burnett, A. K., Hills, R. K. and Linch, D. C. (2008) 'The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia.' *Blood*, 111(5), Mar 1, 2007/10/25, pp. 2776-2784.

Ganapathi, K. A., Townsley, D. M., Hsu, A. P., Arthur, D. C., Zerbe, C. S., Cuellar-Rodriguez, J., Hickstein, D. D., Rosenzweig, S. D., Braylan, R. C., Young, N. S., Holland, S. M. and Calvo, K. R. (2015) 'GATA2 deficiency-associated bone marrow disorder differs from idiopathic aplastic anemia.' *Blood*, 125(1), Jan 01, 2014/11/02, pp. 56-70.

Genomics England. (2017) *The 100,000 Genomes Project by numbers*. [Online] [Accessed on 07/07/2017] https://www.genomicsengland.co.uk/the-100000-genomes-project-by-numbers/

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Della Porta, M. G., Jadersten, M., Dolatshad, H., Verma, A., Cross, N. C. P., Vyas, P., Killick, S., Hellstrom-Lindberg, E., Cazzola, M., Papaemmanuil, E., Campbell, P. J. and Boultwood, J. (2015) 'Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes.' *Nature Communications*, 6, Jan,

Ghezraoui, H., Piganeau, M., Renouf, B., Renaud, J. B., Sallmyr, A., Ruis, B., Oh, S., Tomkinson, A. E., Hendrickson, E. A., Giovannangeli, C., Jasin, M. and Brunet, E. (2014) 'Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining.' *Mol Cell*, 55(6), Sep 18, 2014/09/10, pp. 829-842.

Gilliland, D. G. and Griffin, J. D. (2002) 'The roles of FLT3 in hematopoiesis and leukemia.' *Blood*, 100(5), Sep 1, 2002/08/15, pp. 1532-1542.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S. and Nusbaum, C. (2009) 'Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.' *Nature Biotechnology*, 27(2), Feb, 2009/02/03, pp. 182-189.

Godley, L. A. (2012) 'Profiles in Leukemia.' *New England Journal of Medicine*, 366(12) pp. 1152-1153.

Goemans, B. F., Zwaan, C. M., Miller, M., Zimmermann, M., Harlow, A., Meshinchi, S., Loonen, A. H., Hahlen, K., Reinhardt, D., Creutzig, U., Kaspers, G. J. and Heinrich, M. C. (2005) 'Mutations in KIT and RAS are frequent events in pediatric core-binding factor acute myeloid leukemia.' *Leukemia*, 19(9), Sep, 2005/07/15, pp. 1536-1542.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) 'Coming of age: ten years of next-generation sequencing technologies.' *Nat Rev Genet*, 17(6), 06//print, pp. 333-351.

Graubert, T. A. and Mardis, E. R. (2011) 'Genomics of Acute Myeloid Leukemia.' *Cancer Journal*, 17(6), Nov-Dec, pp. 487-491.

Green, C. L., Evans, C. M., Hills, R. K., Burnett, A. K., Linch, D. C. and Gale, R. E. (2010) 'The prognostic significance of IDH1 mutations in younger adult patients with acute myeloid leukemia is dependent on FLT3/ITD status.' *Blood*, 116(15), Oct, pp. 2779-2782.

Green, C. L., Evans, C. M., Zhao, L., Hills, R. K., Burnett, A. K., Linch, D. C. and Gale, R. E. (2011) 'The prognostic significance of IDH2 mutations in AML depends on the location of the mutation.' *Blood*, 118(2), Jul, pp. 409-412.

Green, E. D., Watson, J. D. and Collins, F. S. (2015) 'Human Genome Project: Twenty-five years of big biology.' *Nature*, 526(7571), Oct 1, 2015/10/04, pp. 29-31.

Greisman, H. A., Hoffman, N. G. and Yi, H. S. (2011) 'Rapid High-Resolution Mapping of Balanced Chromosomal Rearrangements on Tiling CGH Arrays.' *The Journal of Molecular Diagnostics*, 13(6) pp. 621-633.

Griffiths, P. and Stotz, K. (2013) *Genetics and Philosophy; an Introduction.* Cambridge Introductions to Philosophy and Biology. Cambridge: Cambridge University Press.

Grimwade, D. and Freeman, S. D. (2014) 'Defining minimal residual disease in acute myeloid leukemia: which platforms are ready for "prime time"?' *Blood*, 124(23), Nov 27, 2014/07/23, pp. 3345-3355.

Grimwade, D., Ivey, A. and Huntly, B. J. (2016) 'Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance.' *Blood*, 127(1), Jan 7, 2015/12/15, pp. 29-41.

Grimwade, D., Hills, R. K., Moorman, A. V., Walker, H., Chatters, S., Goldstone, A. H., Wheatley, K., Harrison, C. J., Burnett, A. K. and National Cancer Research Institute Adult Leukaemia Working Group. (2010) 'Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials.' *Blood*, 116(3), July 22, 2010, pp. 354-365.

Grimwade, D., Walker, H., Oliver, F., Wheatley, K., Harrison, C. J., Harrison, G., Rees, J., Hann, I., Stevens, R., Burnett, A. K. and Goldstone, A. (1998) 'The Importance of Diagnostic Cytogenetics on Outcome in AML: Analysis of 1,612 Patients Entered Into the MRC AML 10 Trial.' *Blood*, 92(7), October 1, 1998, pp. 2322-2333.

Gronseth, C. M., McElhone, S. E., Storer, B. E., Kroeger, K. A., Sandhu, V., Fero, M. L., Appelbaum, F. R., Estey, E. H. and Fang, M. (2015) 'Prognostic significance of acquired copy-neutral loss of heterozygosity in acute myeloid leukemia.' *Cancer*, On line version of Article published online: 29 MAY 2015 29/05/2015,

Grossmann, V., Kohlmann, A., Klein, H. U., Schindela, S., Schnittger, S., Dicker, F., Dugas, M., Kern, W., Haferlach, T. and Haferlach, C. (2011) 'Targeted next-generation sequencing detects point mutations, insertions, deletions and balanced chromosomal rearrangements as well as identifies novel leukemia-specific fusion genes in a single procedure.' *Leukemia*, Grossmann, V., Schnittger, S., Kohlmann, A., Eder, C., Roller, A., Dicker, F., Schmid, C., Wendtner, C. M., Staib, P., Serve, H., Kreuzer, K. A., Kern, W., Haferlach, T. and Haferlach, C. (2012) 'A novel hierarchical prognostic model of AML solely based on molecular mutations.' *Blood*, Aug 20, 2012/08/24,

Grove, C. S. and Vassiliou, G. S. (2014) 'Acute myeloid leukaemia: a paradigm for the clonal evolution of cancer?' *Disease Models & Mechanisms*, 7(8), Aug, pp. 941-951.

Gruber, T. A., Larson Gedman, A., Zhang, J., Koss, C. S., Marada, S., Ta, H. Q., Chen, S. C., Su, X., Ogden, S. K., Dang, J., Wu, G., Gupta, V., Andersson, A. K., Pounds, S., Shi, L., Easton, J., Barbato, M. I., Mulder, H. L., Manne, J., Wang, J., Rusch, M., Ranade, S., Ganti, R., Parker, M., Ma, J., Radtke, I., Ding, L., Cazzaniga, G., Biondi, A., Kornblau, S. M., Ravandi, F., Kantarjian, H., Nimer, S. D., Döhner, K., Döhner, H., Ley, T. J., Ballerini, P., Shurtleff, S., Tomizawa, D., Adachi, S., Hayashi, Y., Tawa, A., Shih, L. Y., Liang, D. C., Rubnitz, J. E., Pui, C. H., Mardis, E. R., Wilson, R. K. and Downing, J. R. (2012) 'An Inv(16)(p13.3q24.3)-encoded CBFA2T3-GLIS2 fusion protein defines an aggressive subtype of pediatric acute megakaryoblastic leukemia.' *Cancer Cell*, 22(5), Nov 13, 2012/11/17, pp. 683-697.

Grunwald, M. R. and Levis, M. J. (2015) 'FLT3 Tyrosine Kinase Inhibition as a Paradigm for Targeted Drug Development in Acute Myeloid Leukemia.' *Seminars in Hematology*, 52(3), Jul, 2015/06/27, pp. 193-199.

Guan, P. and Sung, W.-K. (2016) 'Structural variation detection using next-generation sequencing data: A comparative technical review.' *Methods*, 102, 6/1/, pp. 36-49.

Gullapalli, R. R., Lyons-Weiler, M., Petrosko, P., Dhir, R., Becich, M. J. and LaFramboise, W. A. (2012) 'Clinical integration of next-generation sequencing technology.' *Clin Lab Med*, 32(4), Dec, 2012/10/20, pp. 585-599.

Guttmacher, A. E. and Collins, F. S. (2003) 'Welcome to the genomic era.' *New England Journal of Medicine*, 349(10), Sep, pp. 996-998.

Haferlach, C., Dicker, F., Herholz, H., Schnittger, S., Kern, W. and Haferlach, T. (2008) 'Mutations of the TP53 gene in acute myeloid leukemia are strongly associated with a complex aberrant karyotype.' *Leukemia*, 22(8), Aug, 2008/06/06, pp. 1539-1541.

Haferlach, C., Alpermann, T., Schnittger, S., Kern, W., Chromik, J., Schmid, C., Pielken, H. J., Kreuzer, K.-A., Haffkes, H.-G. and Haferlach, T. (2012) 'Prognostic value of monosomal karyotype in comparison to complex aberrant karyotype in acute myeloid leukemia: a study on 824 cases with aberrant karyotype.' *Blood*, 119(9), March 1, 2012, pp. 2122-2125.

Haferlach, C., Mecucci, C., Schnittger, S., Kohlmann, A., Mancini, M., Cuneo, A., Testoni, N., Rege-Cambrin, G., Santucci, A., Vignetti, M., Fazi, P., Martelli, M. P., Haferlach, T. and Falini, B. (2009) 'AML with mutated NPM1 carrying a normal or aberrant karyotype show overlapping biologic, pathologic, immunophenotypic, and prognostic features.' *Blood*, 114(14), Oct 1, 2009/05/12, pp. 3024-3032.

Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Kronnie, G. T., Bene, M. C., De Vos, J., Hernandez, J. M., Hofmann, W. K., Mills, K. I., Gilkes, A., Chiaretti, S., Shurtleff, S. A., Kipps, T. J., Rassenti, L. Z., Yeoh, A. E., Papenhausen, P. R., Liu, W. M., Williams, P. M. and Foa, R. (2010) 'Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report From the International Microarray Innovations in Leukemia Study Group.' *Journal of Clinical Oncology*, 28(15), May, pp. 2529-2537.

Hagemann, I. S., Cottrell, C. E. and Lockwood, C. M. (2013) 'Design of targeted, capture-based, next generation sequencing tests for precision cancer therapy.' *Cancer Genetics*, 206(12) pp. 420-431.

Hahn, C. N., Brautigan, P. J., Chong, C. E., Janssan, A., Venugopal, P., Lee, Y., Tims, A. E., Horwitz, M. S., Klingler-Hoffmann, M. and Scott, H. S. (2015) 'Characterisation of a compound in-cis GATA2 germline mutation in a pedigree presenting with myelodysplastic syndrome/acute myeloid leukemia with concurrent thrombocytopenia.' *Leukemia*, 29(8), Aug, pp. 1795-1797.

Hahn, C. N., Chong, C. E., Carmichael, C. L., Wilkins, E. J., Brautigan, P. J., Li, X. C., Babic, M., Lin, M., Carmagnac, A., Lee, Y. K., Kok, C. H., Gagliardi, L., Friend, K. L., Ekert, P. G., Butcher, C. M., Brown, A. L., Lewis, I. D., To, L. B., Timms, A. E., Storek, J., Moore, S., Altree, M., Escher, R., Bardy, P. G., Suthers, G. K., D'Andrea, R. J., Horwitz, M. S. and Scott, H. S. (2011) 'Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia.' *Nat Genet*, 43(10), Sep 04, 2011/09/06, pp. 1012-1017.

Hamburg, M. A. and Collins, F. S. (2010) 'The Path to Personalized Medicine.' *New England Journal of Medicine*, 363(11), 2010, pp. 1092-1092.

Hanahan, D. and Weinberg, Robert A. (2011) 'Hallmarks of Cancer: The Next Generation.' *Cell*, 144(5) pp. 646-674.

Hansen, M. C., Herborg, L. L., Hansen, M., Roug, A. S. and Hokland, P. (2016) 'Combination of RNA- and exome sequencing: Increasing specificity for identification of somatic point mutations and indels in acute leukaemia.' *Leuk Res*, 51, Dec, 2016/11/09, pp. 27-31.

Harris, L. N., Ismaila, N., McShane, L. M., Andre, F., Collyar, D. E., Gonzalez-Angulo, A. M., Hammond, E. H., Kuderer, N. M., Liu, M. C., Mennel, R. G., Van Poznak, C., Bast, R. C. and Hayes, D. F. (2016) 'Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline.' *Journal of Clinical Oncology*, 34(10), Apr, pp. 1134-+.

Harris, T. J. R. and McCormick, F. (2010) 'The molecular pathology of cancer.' *Nat Rev Clin Oncol*, 7(5) pp. 251-265.

Harrison, C. J., Hills, R. K., Moorman, A. V., Grimwade, D. J., Hann, I., Webb, D. K., Wheatley, K., de Graaf, S. S., van den Berg, E., Burnett, A. K. and Gibson, B. E. (2010) 'Cytogenetics of childhood acute myeloid leukemia: United Kingdom Medical Research Council Treatment trials AML 10 and 12.' *J Clin Oncol*, 28(16), Jun 01, 2010/05/05, pp. 2674-2681.

Health Education England. (2015) 'Framework 15; Health Education England Strategic Framework 2014 - 2029 (Update 2015).' [Online]. [Accessed on 01/01/2017] https://www.hee.nhs.uk/sites/default/files/documents/HEE%20Strategic%20Framework%20%202015%20Refresh%20Final%20document.pdf

Hedges, D. J., Guettouche, T., Yang, S., Bademci, G., Diaz, A., Andersen, A., Hulme, W. F., Linker, S., Mehta, A., Edwards, Y. J. K., Beecham, G. W., Martin, E. R., Pericak-Vance, M. A., Zuchner, S., Vance, J. M. and Gilbert, J. R. (2011) 'Comparison of Three Targeted Enrichment Strategies on the SOLiD Sequencing Platform.' *Plos One*, 6(4), Apr, p. 8.

Heim, S. and Mitelman, F. (2008) 'Molecular screening for new fusion genes in cancer.' *Nat Genet*, 40(6), Jun, 2008/05/30, pp. 685-686.

Heim, S. and Mitelman, F. (2009) *Cancer Cytogenetics.* 3rd edition ed.: Wiley Blackwell.

Hindson, C. M., Chevillet, J. R., Briggs, H. A., Gallichotte, E. N., Ruf, I. K., Hindson, B. J., Vessella, R. L. and Tewari, M. (2013) 'Absolute quantification by droplet digital PCR versus analog real-time PCR.' *Nat Methods*, 10(10), Oct, 2013/09/03, pp. 1003-1005.

Hirst, M. (2013) 'Epigenomics: sequencing the methylome.' *Methods Mol Biol*, 973 2013/02/16, pp. 39-54.

Hoelder, S., Clarke, P. A. and Workman, P. (2012) 'Discovery of small molecule cancer drugs: Successes, challenges and opportunities.' *Molecular Oncology*, 6(2), 4//, pp. 155-176.

Hokland, P., Ommen, H. B., Mule, M. P. and Hourigan, C. S. (2015) 'Advancing the Minimal Residual Disease Concept in Acute Myeloid Leukemia.' *Seminars in Hematology*, 52(3), Jul, 2015/06/27, pp. 184-192.

Hollingsworth, S. J. (2015) 'Precision medicine in oncology drug development: a pharma perspective.' *Drug Discov Today*, 20(12), Dec, 2015/10/21, pp. 1455-1463.

Hollingsworth, S. J. and Biankin, A. V. (2015) 'The Challenges of Precision Oncology Drug Development and Implementation.' *Public Health Genomics*, 18(6) 2015/11/12, pp. 338-348.

Hollink, I. H., van den Heuvel-Eibrink, M. M., Arentsen-Peters, S. T., Pratcorona, M., Abbas, S., Kuipers, J. E., van Galen, J. F., Beverloo, H. B., Sonneveld, E., Kaspers, G. J., Trka, J., Baruchel, A., Zimmermann, M., Creutzig, U., Reinhardt, D., Pieters, R., Valk, P. J. and Zwaan, C. M. (2011) 'NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern.' *Blood*, 118(13), Sep 29, 2011/08/05, pp. 3645-3656.

Holme, H., Hossain, U., Kirwan, M., Walne, A., Vulliamy, T. and Dokal, I. (2012) 'Marked genetic heterogeneity in familial myelodysplasia/acute myeloid leukaemia.' *Br J Haematol*, 158(2), Jul, 2012/04/27, pp. 242-248.

Hood, L. and Rowen, L. (2013) 'The Human Genome Project: big science transforms biology and medicine.' *Genome Med*, 5(9) 2013/09/18, p. 79.

Hook, E. B. (1977) 'Exclusion of chromosomal mosaicism: tables of 90%, 95% and 99% confidence limits and comments on use.' *American Journal of Human Genetics*, 29(1) pp. 94-97.

Hou, H. A., Lin, Y. C., Kuo, Y. Y., Chou, W. C., Lin, C. C., Liu, C. Y., Chen, C. Y., Lin, L. I., Tseng, M. H., Huang, C. F., Chiang, Y. C., Liu, M. C., Liu, C. W., Tang, J. L., Yao, M., Huang, S. Y., Ko, B. S., Hsu, S. C., Wu, S. J., Tsay, W., Chen, Y. C. and Tien, H. F. (2015) 'GATA2 mutations in patients with acute myeloid leukemia-paired samples analyses show that the mutation is unstable during disease evolution.' *Annals of Hematology*, 94(2), Feb, pp. 211-221.

Hou, H. A., Lin, C. C., Chou, W. C., Liu, C. Y., Chen, C. Y., Tang, J. L., Lai, Y. J., Tseng, M. H., Huang, C. F., Chiang, Y. C., Lee, F. Y., Kuo, Y. Y., Lee, M. C., Liu, M. C., Liu, C. W., Lin, L. I., Yao, M., Huang, S. Y., Ko, B. S., Hsu, S. C., Wu, S. J., Tsay, W., Chen, Y. C. and Tien, H. F. (2014) 'Integration of cytogenetic and molecular alterations in risk stratification of 318 patients with de novo non-M3 acute myeloid leukemia.' *Leukemia*, 28(1), Jan, pp. 50-58.

Hourigan, C. S. and Karp, J. E. (2013) 'Minimal residual disease in acute myeloid leukaemia.' *Nat Rev Clin Oncol*, 10(8) pp. 460-471.

Hsu, A. P., Johnson, K. D., Falcone, E. L., Sanalkumar, R., Sanchez, L., Hickstein, D. D., Cuellar-Rodriguez, J., Lemieux, J. E., Zerbe, C. S., Bresnick, E. H. and Holland, S. M. (2013) 'GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome.' *Blood*, 121(19), May 09, 2013/03/19, pp. 3830-3837, S3831-3837.

Hsu, A. P., Sampaio, E. P., Khan, J., Calvo, K. R., Lemieux, J. E., Patel, S. Y., Frucht, D. M., Vinh, D. C., Auth, R. D., Freeman, A. F., Olivier, K. N., Uzel, G., Zerbe, C. S., Spalding, C., Pittaluga, S., Raffeld, M., Kuhns, D. B., Ding, L., Paulson, M. L., Marciano, B. E., Gea-Banacloche, J. C., Orange, J. S., Cuellar-Rodriguez, J., Hickstein, D. D. and Holland, S. M. (2011) 'Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome.' *Blood*, 118(10), Sep 08, 2011/06/15, pp. 2653-2655.

Hu, J. and Ng, P. C. (2013) 'SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins.' *PLoS One*, 8(10) 2013/11/07, p. e77940.

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T. S., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M., Knoppers, B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O., Joly, Y., Kato, K., Kennedy, K. L., Nicolas, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter, P., Biankin, A. V., Chabannon, C., Chin, L., Clement, B., de Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Hudson, T. J., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., Shibata, T., van de Vijver, M., Futreal, P. A., Aburatani, H., Bayes, M., Botwell, D. D., Campbell, P. J., Estivill, X., Gerhard, D. S., Grimmond, S. M., Gut, I., Hirst, M., Lopez-Otin, C., Majumder, P., Marra, M., McPherson, J. D., Nakagawa, H., Ning, Z., Puente, X. S., Ruan, Y., Shibata, T., Stratton, M. R., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman, P. T., Bader, G. D., Boutros, P. C., Campbell, P. J., Flicek, P., Getz, G., Guigo, R., Guo, G., Haussler, D., Heath, S., Hubbard, T. J., Jiang, T., Jones, S. M., Li, Q., Lopez-Bigas, N., Luo, R., Muthuswamy, L., Ouellette, B. F., Pearson, J. V., Puente, X. S., Quesada, V., Raphael, B. J., Sander, C., Shibata, T., Speed, T. P., Stein, L. D., Stuart, J. M., Teague, J. W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D. A., Wu, H., Zhao, S., Zhou, G., Stein, L. D., Guigo, R., Hubbard, T. J., Joly, Y., Jones, S. M., Kasprzyk, A., Lathrop, M., Lopez-Bigas, N., Ouellette, B. F., Spellman, P. T., Teague, J. W., Thomas, G., Valencia, A., Yoshida, T., Kennedy, K. L., Axton, M., Dyke, S. O., Futreal, P. A., Gerhard, D. S., Gunter, C., Guyer, M., Hudson, T. J., McPherson, J. D., Miller, L. J., Ozenberger, B., Shaw, K. M., Kasprzyk, A., Stein, L. D., Zhang, J., Haider, S. A., Wang, J., Yung, C. K., Cros, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Bobrow, M., Chalmers, D. R., Hasel, K. W., Joly, Y., Kaan, T. S., Kennedy, K. L., Knoppers, B. M., Lowrance, W. W., Masui, T., Nicolas, P., Rial-Sebbag, E., Rodriguez, L. L., Vergely, C., Yoshida, T., Grimmond, S. M., Biankin, A. V., Bowtell, D. D., Cloonan, N., deFazio, A., Eshleman, J. R., Etemadmoghadam, D., Gardiner, B. B., Kench, J. G., Scarpa, A., Sutherland, R. L., Tempero, M. A., Waddell, N. J., Wilson, P. J., McPherson, J. D., Gallinger, S., Tsao, M. S., Shaw, P. A., Petersen, G. M., Mukhopadhyay, D., Chin, L., DePinho, R. A., Thayer, S., Muthuswamy, L., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu, Y., Ji, J., Zhang, X., Chen, F., Hu, X., Zhou, G., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Yang, H., Lathrop, M., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., Egevard, L., Prokhortchouk, E., Banks, R. E., Uhlen, M., Cambon-Thomsen, A., Viksna, J., Ponten, F., Skryabin, K., Stratton, M. R., Futreal, P. A., Birney, E., Borg, A., Borresen-Dale, A. L., Caldas, C., Foekens, J. A., Martin, S., Reis-Filho, J. S., Richardson, A. L., Sotiriou, C., Stunnenberg, H. G., Thoms, G., van de Vijver, M., van't Veer, L., Calvo, F., Birnbaum, D., Blanche, H., Boucher, P., Boyault, S., Chabannon, C., Gut, I., Masson-Jacquemier, J. D., Lathrop, M., Pauporte, I., Pivot, X., Vincent-Salomon, A., Tabone, E., Theillet, C., Thomas, G., Tost, J., Treilleux, I., Calvo, F., Bioulac-Sage, P., Clement, B., Decaens, T., Degos, F., Franco, D., Gut, I., Gut, M., Heath, S., Lathrop, M., Samuel, D., Thomas, G., Zucman-Rossi, J., Lichter, P., Eils, R., Brors, B., Korbel, J. O., Korshunov, A., Landgraf, P., Lehrach, H., Pfister, S., Radlwimmer, B., Reifenberger, G., Taylor, M. D., von Kalle, C., Majumder, P. P., Sarin, R., Rao, T. S., Bhan, M. K., Scarpa, A., Pederzoli, P., Lawlor, R. A., Delledonne, M., Bardelli, A., Biankin, A. V., Grimmond, S. M., Gress, T., Klimstra, D., Zamboni, G., Shibata, T., Nakamura, Y., Nakagawa, H., Kusada, J., Tsunoda, T., Miyano, S., Aburatani, H., Kato, K., Fujimoto, A., Yoshida, T., Campo, E., Lopez-Otin, C., Estivill, X., Guigo, R., de Sanjose, S., Piris, M. A., Montserrat, E., Gonzalez-Diaz, M., Puente, X. S., Jares, P., Valencia, A., Himmelbauer, H., Quesada, V., Bea, S., Stratton, M. R., Futreal, P. A., Campbell, P. J., Vincent-Salomon, A., Richardson, A. L., Reis-Filho, J. S., van de Vijver, M., Thomas, G., Masson-Jacquemier, J. D., Aparicio, S., Borg, A., Borresen-Dale, A. L., Caldas, C., Foekens, J. A., Stunnenberg, H. G., van't Veer, L., Easton, D. F., Spellman, P. T., Martin, S., Barker, A. D., Chin, L., Collins, F. S., Compton, C. C., Ferguson, M. L., Gerhard, D. S., Getz, G., Gunter, C., Guttmacher, A., Guyer, M., Hayes, D. N., Lander, E. S., Ozenberger, B., Penny, R., Peterson, J., Sander, C., Shaw, K. M., Speed, T. P., Spellman, P. T., Vockley, J. G., Wheeler, D. A., Wilson, R. K., Hudson, T. J., Chin, L., Knoppers, B. M., Lander, E. S., Lichter, P., Stein, L. D., Stratton, M. R., Anderson, W., Barker, A. D., Bell, C., Bobrow, M., Burke, W., Collins, F. S., Compton, C. C., DePinho, R. A., Easton, D. F., Futreal, P. A., Gerhard, D. S., Green, A. R., Guyer, M., Hamilton, S. R., Hubbard, T. J., Kallioniemi, O. P., Kennedy, K. L., Ley, T. J., Liu, E. T., Lu, Y., Majumder, P., Marra, M., Ozenberger, B., Peterson, J., Schafer, A. J., Spellman, P. T., Stunnenberg, H. G., Wainwright, B. J., Wilson, R. K. and Yang, H. (2010) 'International network of cancer genome projects.' *Nature*, 464(7291), Apr 15, 2010/04/16, pp. 993-998.

HUGO Gene Nomenclature Committee (HGCN) Database. In: HUGO Gene Nomenclature Committee (HGNC), E. O.-H., European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

Human Genome Variation Society (HGVS). (2016) *Sequence Variant Nomenclature*. [Online] [Accessed on 17/09/2016] http://varnomen.hgvs.org/recommendations/DNA/

Department of Health. (2012) *Building on our inheritance; Genomic technology in healthcare. A report by the Human Genomics Strategy Group.* COI for the Department of Health. (Human Genomics Strategy Group Report)

Hyde, R. K. and Liu, P. P. (2011) 'GATA2 mutations lead to MDS and AML.' *Nat Genet*, 43(10), Sep 28, 2011/10/01, pp. 926-927.

Illumina, I. (2011) 'Quality Scores for Next-Generation Sequencing.' *Technical Note: Sequencing.* [Online]. [Accessed http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_Q-Scores.pdf

Illumina, I. (2014a) 'Using a PhiX Control for HiSeq® Sequencing Runs.' [Online]. [Accessed on 28/10/2016] http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/hiseq-phix-control-v3-technical-note.pdf

Illumina, I. (2014b) 'Understanding Illumina Quality Scores.' *Technical Note: Informatics.* [Online]. [Accessed on 06/10/2016] http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

Independent Cancer Taskforce. (2015) *Achieving world-class cancer outcomes: a strategy for England 2015-2020.* July 2015.

Inoue, D., Bradley, R. K. and Abdel-Wahab, O. (2016) 'Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis.' *Genes & Development*, 30(9), May 1, 2016, pp. 989-1001.

Ioannidis, J. P. A. (2005) 'Microarrays and molecular research: noise discovery?' *Lancet*, 365(9458), Feb, pp. 454-455.

Ioannidis, J. P. A. (2007a) 'Is molecular profiling ready for use in clinical decision making?' *Oncologist*, 12(3) pp. 301-311.

Ioannidis, J. P. A. (2007b) 'Molecular evidence-based medicine - Evolution and integration of information in the genomic era.' *European Journal of Clinical Investigation*, 37(5), May, pp. 340-349.

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X. Q., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E. and van Noort, V. (2009) 'Repeatability of published microarray gene expression analyses.' *Nature Genetics*, 41(2), Feb, pp. 149-155.

ISCN 2016; An International System for Human Cytogenomic Nomenclature (2016) McGowan-Jordan, J., Simons, A. and Schmid, M. (eds.) Basel: Karger.

Jan, M., Snyder, T. M., Corces-Zimmerman, M. R., Vyas, P., Weissman, I. L., Quake, S. R. and Majeti, R. (2012) 'Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia.' *Sci Transl Med*, 4(149), Aug 29, 2012/08/31, p. 149ra118.

Jennings, L., Van Deerlin, V. M. and Gulley, M. L. (2009) 'Recommended Principles and Practices for Validating Clinical Molecular Pathology Tests.' *Archives of Pathology and Laboratory Medicine*, 133(5), 2013/09/23, pp. 743-755.

Johnson, K. D., Hsu, A. P., Ryu, M. J., Wang, J., Gao, X., Boyer, M. E., Liu, Y., Lee, Y., Calvo, K. R., Keles, S., Zhang, J., Holland, S. M. and Bresnick, E. H. (2012) 'Cis-element mutated in GATA2-dependent immunodeficiency governs hematopoiesis and vascular integrity.' *J Clin Invest*, 122(10), Oct, 2012/09/22, pp. 3692-3704.

Joshi, P., Halene, S. and Abdel-Wahab, O. (2017) 'How do messenger RNA splicing alterations drive myelodysplasia?' *Blood*, 129(18), May, pp. 2465-2470.

Jourdan, E., Boissel, N., Chevret, S., Delabesse, E., Renneville, A., Cornillet, P., Blanchet, O., Cayuela, J. M., Recher, C., Raffoux, E., Delaunay, J., Pigneux, A., Bulabois, C. E., Berthon, C., Pautas, C., Vey, N., Lioure, B., Thomas, X., Luquet, I., Terre, C., Guardiola, P., Bene, M. C., Preudhomme, C., Ifrah, N. and Dombret, H. (2013) 'Prospective evaluation of gene mutations and minimal residual disease in patients with core binding factor acute myeloid leukemia.' *Blood*, 121(12), Mar 21, 2013/01/17, pp. 2213-2223.

Jung, H., Bleazard, T., Lee, J. and Hong, D. (2013) 'Systematic investigation of cancer-associated somatic point mutations in SNP databases.' *Nat Biotechnol*, 31(9), Sep, 2013/09/12, pp. 787-789.

Kadri, S., Zhen, C. J., Wurst, M. N., Long, B. C., Jiang, Z. F., Wang, Y. L., Furtado, L. V. and Segal, J. P. (2015) 'Amplicon Indel Hunter Is a Novel Bioinformatics Tool to Detect Large Somatic Insertion/Deletion Mutations in Amplicon-Based Next-Generation Sequencing Data.' *Journal of Molecular Diagnostics*, 17(6), Nov, pp. 635-643.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. and Kent, W. J. (2004) 'The UCSC Table Browser data retrieval tool.' *Nucleic Acids Research*, 32(suppl 1), January 1, 2004, pp. D493-D496.

Kayser, S., Zucknick, M., Doehner, K., Krauter, J., Koehne, C.-H., Horst, H. A., Held, G., von Lilienfeld-Toal, M., Wilhelm, S., Rummel, M., Germing, U., Goetze, K., Nachbaur, D., Schlegelberger, B., Goehring, G., Spaeth, D., Morlok, C., Teleanu, V., Ganser, A., Doehner, H., Schlenk, R. F. and German-Austrian, A. M. L. S. G. (2012) 'Monosomal karyotype in adult acute myeloid leukemia: prognostic impact and outcome after different treatment strategies.' *Blood*, 119(2), Jan 12, pp. 551-558.

Kazenwadel, J., Secker, G. A., Liu, Y. J., Rosenfeld, J. A., Wildin, R. S., Cuellar-Rodriguez, J., Hsu, A. P., Dyack, S., Fernandez, C. V., Chong, C. E., Babic, M., Bardy, P. G., Shimamura, A., Zhang, M. Y., Walsh, T., Holland, S. M., Hickstein, D. D., Horwitz, M. S., Hahn, C. N., Scott, H. S. and Harvey, N. L. (2012) 'Loss-of-function germline GATA2 mutations in patients with MDS/AML or MonoMAC syndrome and primary lymphedema reveal a key role for GATA2 in the lymphatic vasculature.' *Blood*, 119(5), Feb 02, 2011/12/08, pp. 1283-1291.

Kenna, K. P., McLaughlin, R. L., Hardiman, O. and Bradley, D. G. (2013) 'Using reference databases of genetic variation to evaluate the potential pathogenicity of candidate disease variants.' *Hum Mutat*, 34(6), Jun, 2013/03/01, pp. 836-841.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler and David. (2002) 'The Human Genome Browser at UCSC.' *Genome Research*, 12(6), June 1, 2002, pp. 996-1006.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R. and Eichler, E. E. (2008) 'Mapping and sequencing of structural variation from eight human genomes.' *Nature*, 453(7191), May 01, 2008/05/03, pp. 56-64.

Kihara, R., Nagata, Y., Kiyoi, H., Kato, T., Yamamoto, E., Suzuki, K., Chen, F., Asou, N., Ohtake, S., Miyawaki, S., Miyazaki, Y., Sakura, T., Ozawa, Y., Usui, N., Kanamori, H., Kiguchi, T., Imai, K., Uike, N., Kimura, F., Kitamura, K., Nakaseko, C., Onizuka, M., Takeshita, A., Ishida, F., Suzushima, H., Kato, Y., Miwa, H., Shiraishi, Y., Chiba, K., Tanaka, H., Miyano, S., Ogawa, S. and Naoe, T. (2014) 'Comprehensive analysis of genetic alterations and their prognostic impacts in adult acute myeloid leukemia patients.' *Leukemia*, Feb 3, 2014/02/04,

Kim, D., Hong, Y., Koh, Y., Yoon, S. S., Sun, C. H., Ahn, K. S., Lee, S., Yun, H. and Lee, S. (2016) 'Improved sensitive detection method of FLT3 (FMS-like tyrosine kinase) internal tandem duplication (ITD) mutation using next-generation sequencing technology and nested PCR.' *Cancer Research*, 76, Jul,

Kohlmann, A., Grossmann, V. and Haferlach, T. (2012) 'Integration of Next-Generation Sequencing Into Clinical Practice: Are We There Yet?' *Seminars in Oncology*, 39(1) pp. 26-36.

Kohlmann, A., Bullinger, L., Thiede, C., Schaich, M., Schnittger, S., Döhner, K., Dugas, M., Klein, H. U., Döhner, H., Ehninger, G. and Haferlach, T. (2010a) 'Gene expression profiling in AML with normal karyotype can predict mutations for molecular markers and allows novel insights into perturbed biological pathways.' *Leukemia*, 24(6), Jun, 2010/04/30, pp. 1216-1220.

Kohlmann, A., Grossmann, V., Klein, H.-U., Schindela, S., Weiss, T., Kazak, B., Dicker, F., Schnittger, S., Dugas, M., Kern, W., Haferlach, C. and Haferlach, T. (2010b) 'Next-Generation Sequencing Technology Reveals a Characteristic Pattern of Molecular Mutations in 72.8% of Chronic Myelomonocytic Leukemia by Detecting Frequent Alterations in TET2, CBL, RAS, and RUNX1.' *Journal of Clinical Oncology*, 28(24), August 20, 2010, pp. 3858-3865.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M. and Snyder, M. (2007) 'Paired-end mapping reveals extensive structural variation in the human genome.' *Science*, 318(5849), Oct 19, 2007/09/29, pp. 420-426.

Kottaridis, P. D., Gale, R. E., Frew, M. E., Harrison, G., Langabeer, S. E., Belton, A. A., Walker, H., Wheatley, K., Bowen, D. T., Burnett, A. K., Goldstone, A. H. and Linch, D. C. (2001) 'The presence of a FLT3 internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy: analysis of 854 patients from the United Kingdom Medical Research Council AML 10 and 12 trials.' *Blood*, 98(6), Sep 15, 2001/09/06, pp. 1752-1759.

Krauth, M. T., Alpermann, T., Bacher, U., Eder, C., Dicker, F., Ulke, M., Kuznia, S., Nadarajah, N., Kern, W., Haferlach, C., Haferlach, T. and Schnittger, S. (2015) 'WT1 mutations are secondary events in AML, show varying frequencies and impact on prognosis between genetic subgroups.' *Leukemia*, 29(3), Mar, pp. 660-667.

Kronke, J., Bullinger, L., Teleanu, V., Tschurtz, F., Gaidzik, V. I., Kuhn, M. W., Rücker, F. G., Holzmann, K., Paschka, P., Kapp-Schworer, S., Spath, D., Kindler, T., Schittenhelm, M., Krauter, J., Ganser, A., Gohring, G., Schlegelberger, B., Schlenk, R. F., Döhner, H. and Döhner, K. (2013) 'Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia.' *Blood*, 122(1), Jul 4, 2013/05/25, pp. 100-108.

Krueger, F., Kreck, B., Franke, A. and Andrews, S. R. (2012) 'DNA methylome analysis using short bisulfite sequencing data.' *Nat Methods*, 9(2), Jan 30, 2012/02/01, pp. 145-151.

Kuchel, A., Robinson, T., Comins, C., Shere, M., Varughese, M., Sparrow, G., Sahu, A., Saunders, L., Bahl, A., Cawthorn, S. J. and Braybrooke, J. P. (2016) 'The impact of the 21-gene assay on adjuvant treatment decisions in oestrogen receptor-positive early breast cancer: a prospective study.' *British Journal of Cancer*, 114(7), Mar, pp. 731-736.

Kukurba, K. R. and Montgomery, S. B. (2015) 'RNA Sequencing and Analysis.' *Cold Spring Harbor protocols*, 2015(11), 04/13, pp. 951-969.

Lafond, S., Charlesworth, A. and Roberts, A. (2016) A perfect storm: an impossible climate for NHS providers' finances? (30/12/2016): The Health Foundation.

Lagunas-Rangel, F. A. and Chavez-Valencia, V. (2017) 'FLT3-ITD and its current role in acute myeloid leukaemia.' *Med Oncol*, 34(6), Jun, 2017/05/05, p. 114.

Lam, H. Y., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., Gerstein, M. B., Butte, A. J., Ji, H. P. and Snyder, M. (2011) 'Performance comparison of whole-genome sequencing platforms.' *Nat Biotechnol*, 30(1), Dec 18, 2011/12/20, pp. 78-82.

Lander, E. S. and Waterman, M. S. (1988) 'Genomic mapping by fingerprinting random clones: a mathematical analysis.' *Genomics*, 2(3), Apr, 1988/04/01, pp. 231-239.

Lavallée, V.-P., Gendron, P., Boucher, G., Lemieux, S., Armstrong, R. N., Boivin, I., Sauvageau, G. and Hébert, J. (2015) 'Mutational and Transcriptomic Landscape of AML with Core-Binding Factor Rearrangements.' *Blood*, 126(23) pp. 802-802.

Lawler, M. and Sullivan, R. (2015) 'Personalised and Precision Medicine in Cancer Clinical Trials: Panacea for Progress or Pandora's Box?' *Public Health Genomics*, 18(6) pp. 329-337.

Layer, R. M., Chiang, C., Quinlan, A. R. and Hall, I. M. (2014) 'LUMPY: a probabilistic framework for structural variant discovery.' *Genome Biology*, 15(6) p. R84.

Lazarus, H. M. and Litzow, M. R. (2012) 'AML cytogenetics: the complex just got simpler.' *Blood*, 120(12), Sep 20, 2012/09/22, pp. 2357-2358.

Lee, L. A., Arvai, K. J. and Jones, D. (2015) 'Annotation of Sequence Variants in Cancer Samples: Processes and Pitfalls for Routine Assays in the Clinical Laboratory.' *J Mol Diagn*, 17(4), Jul, 2015/05/16, pp. 339-351.

Levin, J. Z., Berger, M. F., Adiconis, X., Rogov, P., Melnikov, A., Fennell, T., Nusbaum, C., Garraway, L. A. and Gnirke, A. (2009) 'Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts.' *Genome Biol*, 10(10) 2009/10/20, p. R115.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D. E., Koboldt, D. C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L. W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J. R., Minx, P., Gordon, D., Chinwalla, A., Zhao, Y., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W. D., Nagarajan, R., Walter, M. J., Link, D. C., Graubert, T. A., DiPersio, J. F. and Wilson, R. K. (2008) 'DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.' *Nature*, 456(7218), Nov 6, 2008/11/07, pp. 66-72.

Ley, T. J., Ding, L., Walter, M. J., McLellan, M. D., Lamprecht, T., Larson, D. E., Kandoth, C., Payton, J. E., Baty, J., Welch, J., Harris, C. C., Lichti, C. F., Townsend, R. R., Fulton, R. S., Dooling, D. J., Koboldt, D. C., Schmidt, H., Zhang, Q. Y., Osborne, J. R., Lin, L., O'Laughlin, M., McMichael, J. F., Delehaunty, K. D., McGrath, S. D., Fulton, L. A., Magrini, V. J., Vickery, T. L., Hundal, J., Cook, L. L., Conyers, J. J., Swift, G. W., Reed, J. P., Alldredge, P. A., Wylie, T., Walker, J., Kalicki, J., Watson, M. A., Heath, S., Shannon, W. D., Varghese, N., Nagarajan, R., Westervelt, P., Tomasson, M. H., Link, D. C., Graubert, T. A., DiPersio, J. F., Mardis, E. R. and Wilson, R. K. (2010) 'DNMT3A Mutations in Acute Myeloid Leukemia.' *New England Journal of Medicine*, 363(25), Dec, pp. 2424-2433.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.' [Online]. [Accessed on 31/07/2016] http://arxiv.org/pdf/1303.3997v2.pdf

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform.' *Bioinformatics*, 25(14), Jul 15, 2009/05/20, pp. 1754-1760.

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., Tsimberidou, A. M., Vnencak-Jones, C. L., Wolff, D. J., Younes, A. and Nikiforova, M. N. (2017) 'Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists.' *J Mol Diagn*, 19(1), Jan, 2016/12/21, pp. 4-23.

Lin, P. H., Li, H. Y., Fan, S. C., Yuan, T. H., Chen, M., Hsu, Y. H., Yang, Y. H., Li, L. Y., Yeh, S. P., Bai, L. Y., Liao, Y. M., Lin, C. Y., Hsieh, C. Y., Lin, C. C., Lin, C. H., Lien, M. Y., Chen, T. T., Ni, Y. H. and Chiu, C. F. (2017) 'A targeted next-generation sequencing in the molecular risk stratification of adult acute myeloid leukemia: implications for clinical practice.' *Cancer Med*, Jan 10, 2017/01/11,

Liu, B. A., Conroy, J. M., Morrison, C. D., Odunsi, A. O., Qin, M. C., Wei, L., Trump, D. L., Johnson, C. S., Liu, S. and Wang, J. M. (2015) 'Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives.' *Oncotarget*, 6(8), Mar, pp. 5477-5489.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. and Pallen, M. J. (2012) 'Performance comparison of benchtop high-throughput sequencing platforms.' *Nat Biotech*, 30(5) pp. 434-439.

Loriaux, M. M., Levine, R. L., Tyner, J. W., Fröhling, S., Scholl, C., Stoffregen, E. P., Wernig, G., Erickson, H., Eide, C. A., Berger, R., Bernard, O. A., Griffin, J. D., Stone, R. M., Lee, B., Meyerson, M., Heinrich, M. C., Deininger, M. W., Gilliland, D. G. and Druker, B. J. (2008) 'High-throughput sequence analysis of the tyrosine kinome in acute myeloid leukemia.' *Blood*, 111(9), May 1, 2008, pp. 4788-4796.

Lu, Y., Shen, Y., Warren, W. and Walter, R. (2016) *Next Generation Sequencing - Advances, Applications and Challenges.* Kulski, J. K. (ed.): InTech.

Lubking, A., Vosberg, S., Konstandin, N. P., Dufour, A., Graf, A., Krebs, S., Blum, H., Weber, A., Lenhoff, S., Ehinger, M., Spiekermann, K., Greif, P. A. and Cammenga, J. (2015) 'Young woman with mild bone marrow dysplasia, GATA2 and ASXL1 mutation treated with allogeneic hematopoietic stem cell transplantation.' *Leuk Res Rep*, 4(2) 2015/12/31, pp. 72-75.

Luskin, M. R., Lee, J. W., Fernandez, H. F., Abdel-Wahab, O., Bennett, J. M., Ketterling, R. P., Lazarus, H. M., Levine, R. L., Litzow, M. R., Paietta, E. M., Patel, J. P., Racevskis, J., Rowe, J. M., Tallman, M. S., Sun, Z. and Luger, S. M. (2016) 'Benefit of high-dose daunorubicin in AML induction extends across cytogenetic and molecular groups.' *Blood*, 127(12), Mar 24, 2016/01/13, pp. 1551-1558.

Luthra, R., Patel, K. P., Reddy, N. G., Haghshenas, V., Routbort, M. J., Harmon, M. A., Barkoh, B. A., Kanagal-Shamanna, R., Ravandi, F., Cortes, J. E., Kantarjian, H. M., Medeiros, L. J. and Singh, R. R. (2014) 'Next-generation sequencing-based multigene mutational screening for acute myeloid leukemia using MiSeq: applicability for diagnostics and disease monitoring.' *Haematologica*, 99(3), March 1, 2014, pp. 465-473.

Mace, E. M., Hsu, A. P., Monaco-Shawver, L., Makedonas, G., Rosen, J. B., Dropulic, L., Cohen, J. I., Frenkel, E. P., Bagwell, J. C., Sullivan, J. L., Biron, C. A., Spalding, C., Zerbe, C. S., Uzel, G., Holland, S. M. and Orange, J. S. (2013) 'Mutations in GATA2 cause human NK cell deficiency with specific loss of the CD56(bright) subset.' *Blood*, 121(14), Apr 04, 2013/02/01, pp. 2669-2677.

Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A. M. (2009) 'Transcriptome sequencing to detect gene fusions in cancer.' *Nature*, 458(7234) pp. 97-101.

Majewski, I. J. and Bernards, R. (2011) 'Taming the dragon: genomic biomarkers to individualize the treatment of cancer.' *Nature Medicine*, pp. 304-312.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J. and Turner, D. J. (2010) 'Target-enrichment strategies for next-generation sequencing.' *Nat Meth*, 7(2) pp. 111-118.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. and Visscher, P. M. (2009) 'Finding the missing heritability of complex diseases.' *Nature*, 461(7265), Oct 8, 2009/10/09, pp. 747-753.

Marcucci, G., Haferlach, T. and Döhner, H. (2011) 'Molecular Genetics of Adult Acute Myeloid Leukemia: Prognostic and Therapeutic Implications.' *Journal of Clinical Oncology*, 29(5), February 10, 2011, pp. 475-486.

Mardis, E. R. (2011) 'A decade's perspective on DNA sequencing technology.' *Nature*, 470(7333), Feb 10, 2011/02/11, pp. 198-203.

Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D., Fulton, L. A., Locke, D. P., Magrini, V. J., Abbott, R. M., Vickery, T. L., Reed, J. S., Robinson, J. S., Wylie, T., Smith, S. M., Carmichael, L., Eldred, J. M., Harris, C. C., Walker, J., Peck, J. B., Du, F., Dukes, A. F., Sanderson, G. E., Brummett, A. M., Clark, E., McMichael, J. F., Meyer, R. J., Schindler, J. K., Pohl, C. S., Wallis, J. W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M. E., Ivy, J. V., Kalicki, J., Elliott, G., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M. A., Baty, J., Heath, S., Shannon, W. D., Nagarajan, R., Link, D. C., Walter, M. J., Graubert, T. A., DiPersio, J. F., Wilson, R. K. and Ley, T. J. (2009) 'Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome.' *N Engl J Med*, 361(11), September 10, 2009, pp. 1058-1066.

Martelli, M. P., Gionfriddo, I., Mezzasoma, F., Milano, F., Pierangeli, S., Mulas, F., Pacini, R., Tabarrini, A., Pettirossi, V., Rossi, R., Vetro, C., Brunetti, L., Sportoletti, P., Tiacci, E., Di Raimondo, F. and Falini, B. (2015) 'Arsenic trioxide and all-trans retinoic acid target NPM1 mutant oncoprotein levels and induce apoptosis in NPM1-mutated AML cells.' *Blood*, 125(22), May 28, 2015/03/22, pp. 3455-3465.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads.' *2011*, 17(1)

Masetti, R., Pigazzi, M., Togni, M., Astolfi, A., Indio, V., Manara, E., Casadio, R., Pession, A., Basso, G. and Locatelli, F. (2013) 'CBFA2T3-GLIS2 fusion transcript is a novel common feature in pediatric, cytogenetically normal AML, not restricted to FAB M7 subtype.' *Blood*, 121(17), Apr 25, 2013/02/15, pp. 3469-3472.

Mattocks, C. J., Morris, M. A., Matthijs, G., Swinnen, E., Corveleyn, A., Dequeker, E., Muller, C. R., Pratt, V. and Wallace, A. (2010) 'A standardized framework for the validation and verification of clinical molecular genetic tests.' *European Journal of Human Genetics*, 18(12) pp. 1276-1288.

Mazumdar, C. and Majeti, R. (2017) 'The role of mutations in the cohesin complex in acute myeloid leukemia.' *Int J Hematol*, 105(1), Jan, 2016/11/01, pp. 31-36.

Mazzarella, L., Riva, L., Luzi, L., Ronchini, C. and Pelicci, P. G. (2014) 'The Genomic and Epigenomic Landscapes of AML.' *Seminars in Hematology*, 51(4), Oct, pp. 259-272.

McDermott, U. (2017) *Cancer diagnosis.* July 2017. Department of Health.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010) 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.' *Genome Res*, 20(9), Sep, 2010/07/21, pp. 1297-1303.

McKerrell, T., Moreno, T., Ponstingl, H., Bolli, N., Dias, J. M., Tischler, G., Colonna, V., Manasse, B., Bench, A., Bloxham, D., Herman, B., Fletcher, D., Park, N., Quail, M. A., Manes, N., Hodkinson, C., Baxter, J., Sierra, J., Foukaneli, T., Warren, A. J., Chi, J., Costeas, P., Rad, R., Huntly, B., Grove, C., Ning, Z., Tyler-Smith, C., Varela, I., Scott, M., Nomdedeu, J., Mustonen, V. and Vassiliou, G. S. (2016) 'Development and validation of a comprehensive genomic diagnostic tool for myeloid malignancies.' *Blood*, 128(1), Jul 07, 2016/04/29, pp. e1-9.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P. and Cunningham, F. (2016) 'The Ensembl Variant Effect Predictor.' *Genome Biology*, 17(1) p. 122.

Mead, A. J., Linch, D. C., Hills, R. K., Wheatley, K., Burnett, A. K. and Gale, R. E. (2007) 'FLT3 tyrosine kinase domain mutations are biologically distinct from and have a significantly more favorable prognosis than FLT3 internal tandem duplications in patients with acute myeloid leukemia.' *Blood*, 110(4), Aug, pp. 1262-1270.

Medeiros, B. C. (2012) 'Unveiling the complexity of CK+ AML.' *Blood*, 119(9), Mar 1, 2012/03/03, pp. 1958-1959.

Medical Research Council. (2013) *MRC Molecular Pathology Review.* Medical Research Council.

Meldrum, C., Doyle, M. A. and Tothill, R. W. (2011) 'Next-generation sequencing for cancer diagnostics: a practical perspective.' *Clin Biochem Rev*, 32(4), Nov, 2011/12/08, pp. 177-195.

Mendler, J. H., Maharry, K., Radmacher, M. D., Mrózek, K., Becker, H., Metzeler, K. H., Schwind, S., Whitman, S. P., Khalife, J., Kohlschmidt, J., Nicolet, D., Powell, B. L., Carter, T. H., Wetzler, M., Moore, J. O., Kolitz, J. E., Baer, M. R., Carroll, A. J., Larson, R. A., Caligiuri, M. A., Marcucci, G. and Bloomfield, C. D. (2012) 'RUNX1 mutations are associated with poor outcome in younger and older patients with cytogenetically normal acute myeloid leukemia and with distinct gene and MicroRNA expression signatures.' *J Clin Oncol*, 30(25), Sep 01, 2012/07/04, pp. 3109-3118.

Mertens, F., Johansson, B., Fioretos, T. and Mitelman, F. (2015) 'The emerging complexity of gene fusions in cancer.' *Nat Rev Cancer*, 15(6) pp. 371-381.

Mertes, F., ElSharawy, A., Sauer, S., van Helvoort, J. M. L. M., van der Zaag, P. J., Franke, A., Nilsson, M., Lehrach, H. and Brookes, A. J. (2011) 'Targeted enrichment of genomic DNA regions for next-generation sequencing.' *Briefings in Functional Genomics*, November 26, 2011,

Metzeler, K. H., Becker, H., Maharry, K., Radmacher, M. D., Kohlschmidt, J., Mrózek, K., Nicolet, D., Whitman, S. P., Wu, Y.-Z., Schwind, S., Powell, B. L., Carter, T. H., Wetzler, M., Moore, J. O., Kolitz, J. E., Baer, M. R., Carroll, A. J., Larson, R. A., Caligiuri, M. A., Marcucci, G. and Bloomfield, C. D. (2011a) 'ASXL1 mutations identify a high-risk subgroup of older patients with primary cytogenetically normal AML within the ELN Favorable genetic category.' *Blood*, 118(26), Dec 22, pp. 6920-6929.

Metzeler, K. H., Maharry, K., Radmacher, M. D., Mrózek, K., Margeson, D., Becker, H., Curfman, J., Holland, K. B., Schwind, S., Whitman, S. P., Wu, Y. Z., Blum, W., Powell, B. L., Carter, T. H., Wetzler, M., Moore, J. O., Kolitz, J. E., Baer, M. R., Carroll, A. J., Larson, R. A., Caligiuri, M. A., Marcucci, G. and Bloomfield, C. D. (2011b) 'TET2 Mutations Improve the New European LeukemiaNet Risk Classification of Acute Myeloid Leukemia: A Cancer and Leukemia Group B Study.' *Journal of Clinical Oncology*, 29(10), Apr, pp. 1373-1381.

Metzeler, K. H., Herold, T., Rothenberg-Thurley, M., Amler, S., Sauerland, M. C., Gorlich, D., Schneider, S., Konstandin, N. P., Dufour, A., Braundl, K., Ksienzyk, B., Zellmeier, E., Hartmann, L., Greif, P. A., Fiegl, M., Subklewe, M., Bohlander, S. K., Krug, U., Faldum, A., Berdel, W. E., Wormann, B., Buchner, T., Hiddemann, W., Braess, J. and Spiekermann, K. (2016) 'Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia.' *Blood*, 128(5), Aug 4, 2016/06/12, pp. 686-698.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation.' *Nat Rev Genet*, 11(1) pp. 31-46.

Meyer, S. C. and Levine, R. L. (2014) 'Translational implications of somatic genomics in acute myeloid leukaemia.' *The Lancet Oncology*, 15(9) pp. e382-e394.

Meyerson, M., Gabriel, S. and Getz, G. (2010) 'Advances in understanding cancer genomes through second-generation sequencing.' *Nat Rev Genet*, 11(10) pp. 685-696.

Meynert, A. M., Ansari, M., FitzPatrick, D. R. and Taylor, M. S. (2014) 'Variant detection sensitivity and biases in whole genome and exome sequencing.' *BMC Bioinformatics.*, 15

Micol, J. B. and Abdel-Wahab, O. (2014) 'Collaborating constitutive and somatic genetic events in myeloid malignancies: ASXL1 mutations in patients with germline GATA2 mutations.' *Haematologica*, 99(2), Feb, 2014/02/06, pp. 201-203.

Miles, G., Rae, J., Ramalingam, S. S. and Pfeifer, J. (2015) 'Genetic Testing and Tissue Banking for Personalized Oncology: Analytical and Institutional Factors.' *Seminars in Oncology*, 42(5), Oct, 2015/10/05, pp. 713-723.

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K., Krantz, I. D., Kuhn, R. M., Lee, C., Ostell, J. M., Rosenberg, C., Scherer, S. W., Spinner, N. B., Stavropoulos, D. J., Tepperberg, J. H., Thorland, E. C., Vermeesch, J. R., Waggoner, D. J., Watson, M. S., Martin, C. L. and Ledbetter, D. H. (2010) 'Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies.' *American Journal of Human Genetics*, 86(5), 02/15/received
04/12/revised
04/19/accepted, pp. 749-764.

Mitelman, F., Johansson, B., Mertens, F. and (Eds.). (2017) 'Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.' [Online]. [Accessed on 11/05/2017] http://cgap.nci.nih.gov/Chromosomes/Mitelman

Mohamedali, A. M., Alkhatabi, H., Kulasekararaj, A., Shinde, S., Mian, S., Malik, F., Smith, A. E., Gaken, J. and Mufti, G. J. (2013) 'Utility of peripheral blood for cytogenetic and mutation analysis in myelodysplastic syndrome.' *Blood*, 122(4), Jul 25, 2013/06/14, pp. 567-570.

Mohamedali, A. M., Gaken, J., Ahmed, M., Malik, F., Smith, A. E., Best, S., Mian, S., Gaymes, T., Ireland, R., Kulasekararaj, A. G. and Mufti, G. J. (2015) 'High concordance of genomic and cytogenetic aberrations between peripheral blood and bone marrow in myelodysplastic syndrome (MDS).' *Leukemia*, 29(9), Sep, 2015/05/07, pp. 1928-1938.

Monitor Deloitte. (2015) 'Genomics in the UK: An industry study for the Office of Life Sciences.' [Online]. [Accessed on 20/11/2016]

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq.' *Nat Methods*, 5(7), Jul, 2008/06/03, pp. 621-628.

Mrózek, K. (2008) 'Cytogenetic, molecular genetic, and clinical characteristics of acute myeloid leukemia with a complex karyotype.' *Seminars in Oncology*, 35(4), Aug, 2008/08/12, pp. 365-377.

Murray, R., Imison, C. and Jabbal, J. (2014) Financial failure in the NHS. (30/12/2016) *What causes it and how best to manage it*. The King's Fund.

Myllykangas, S. and Ji, H. P. (2010) 'Targeted deep resequencing of the human cancer genome using next-generation technologies.' *Biotechnology and Genetic Engineering Reviews*, 27 2010/01/01, pp. 135-158.

Nagy, E. and Maquat, L. E. (1998) 'A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance.' *Trends Biochem Sci*, 23(6), Jun, 1998/06/30, pp. 198-199.

Nakao, M., Yokota, S., Iwai, T., Kaneko, H., Horiike, S., Kashima, K., Sonoda, Y., Fujimoto, T. and Misawa, S. (1996) 'Internal tandem duplication of the FLT3 gene found in acute myeloid leukemia.' *Leukemia*, 10(12), Dec, 1996/12/01, pp. 1911-1918.

Naoe, T. and Kiyoi, H. (2014) 'Gene mutations of acute myeloid leukemia in the genome era.' *International Journal of Hematology*, 97(2) pp. 165-174.

National Cancer Institute. (2014) *Targeted Cancer Therapies; What targeted therapies have been approved for specific types of cancer?* : [Online] [Accessed on 21/05/2016] http://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheet

National Center for Biotechnology Information (NCBI). (2016) 'RefSeq: NCBI Reference Sequence Database.' [Online]. [Accessed on 17/092016] https://www.ncbi.nlm.nih.gov/refseq/

National Human Genome Research Institute. (2016) *DNA sequencing costs; data from the NHGRI Genome Sequencing Program (GSP)*. [Online] [Accessed on 14/04/2016] https://www.genome.gov/27541954/dna-sequencing-costs/

National Institute for Health and Care Excellence. (2016) 'Haematological cancers: improving outcomes.' [Online]. [Accessed on 20/11/2016] https://www.nice.org.uk/guidance/ng47

Ng, P., Tan, J. J., Ooi, H. S., Lee, Y. L., Chiu, K. P., Fullwood, M. J., Srinivasan, K. G., Perbost, C., Du, L., Sung, W. K., Wei, C. L. and Ruan, Y. (2006) 'Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes.' *Nucleic Acids Res*, 34(12), Jul 13, 2006/07/15, p. e84.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A. and Shendure, J. (2009) 'Targeted capture and massively parallel sequencing of 12 human exomes.' *Nature*, 461(7261), Sep, pp. 272-U153.

NHS England. (2014) 'Five Year Forward View.' [Online]. [Accessed on 20/11/2016] https://www.england.nhs.uk/ourwork/futurenhs/

NHS England. (2015) 'Genomic Laboratory Service Re-design Service Specification.' [Online] E01/Sb. [Accessed on 18/11/2016] https://www.engage.england.nhs.uk/consultation/genomic-laboratories

Nowell, P. and Hungerford, D. (1960) 'A minute chromosome in human granulocytic leukaemia.' *Science*, 132(1497)

Ofran, Y. and Rowe, J. M. (2013) 'Genetic profiling in acute myeloid leukaemia - where are we and what is its role in patient management.' *British Journal of Haematology*, 160(3), Feb, pp. 303-320.

Ohgami, R. S. and Arber, D. A. (2015) 'The diagnostic and clinical impact of genetics and epigenetics in acute myeloid leukemia.' *International Journal of Laboratory Hematology*, 37 pp. 122-132.

Ommen, H. B. (2016) 'Monitoring minimal residual disease in acute myeloid leukaemia: a review of the current evolving strategies.' *Ther Adv Hematol*, 7(1), Feb, 2016/02/03, pp. 3-16.

Open Targets. (2017) *Open Targets.*: [Online] [Accessed on 23/07/2017] https://www.opentargets.org/

Oran, B. and Weisdorf, D. J. (2012) 'Survival for older patients with acute myeloid leukemia: a population-based study.' *Haematologica*, 97(12), 03/20/received
05/09/revised
06/22/accepted, pp. 1916-1924.

Ossenkoppele, G. and Löwenberg, B. (2015) 'How I treat the older patient with acute myeloid leukemia.' *Blood*, 125(5), 2015-01-29 00:00:00, pp. 767-774.

Ostergaard, P., Simpson, M. A., Connell, F. C., Steward, C. G., Brice, G., Woollard, W. J., Dafou, D., Kilo, T., Smithson, S., Lunt, P., Murday, V. A., Hodgson, S., Keenan, R., Pilz, D. T., Martinez-Corral, I., Makinen, T., Mortimer, P. S., Jeffery, S., Trembath, R. C. and Mansour, S. (2011) 'Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome).' *Nat Genet*, 43(10), Sep 04, 2011/09/06, pp. 929-931.

Ostronoff, F., Othus, M., Lazenby, M., Estey, E., Appelbaum, F. R., Evans, A., Godwin, J., Gilkes, A., Kopecky, K. J., Burnett, A., List, A. F., Fang, M., Oehler, V. G., Petersdorf, S. H., Pogosova-Agadjanyan, E. L., Radich, J. P., Willman, C. L., Meshinchi, S. and Stirewalt, D. L. (2015) 'Prognostic Significance of NPM1 Mutations in the Absence of FLT3 Internal Tandem Duplication in Older Patients With Acute Myeloid Leukemia: A SWOG and UK National Cancer Research Institute/Medical Research Council Report.' *Journal of Clinical Oncology*, 33(10), April 1, 2015, pp. 1157-1164.

Ottaviani, D., LeCain, M. and Sheer, D. (2014) 'The role of microhomology in genomic structural variation.' *Trends Genet*, 30(3), Mar, 2014/02/08, pp. 85-94.

Ozsolak, F. and Milos, P. M. (2011) 'RNA sequencing: advances, challenges and opportunities.' *Nat Rev Genet*, 12(2), Feb, 2010/12/31, pp. 87-98.

Pabst, T., Mueller, B. U., Zhang, P., Radomska, H. S., Narravula, S., Schnittger, S., Behre, G., Hiddemann, W. and Tenen, D. G. (2001) 'Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia.' *Nature Genetics*, 27(3), Mar, 2001/03/10, pp. 263-270.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D. R., Teague, J. W., Butler, A. P., Greaves, M. F., Ganser, A., Döhner, K., Schlenk, R. F., Döhner, H. and Campbell, P. J. (2016) 'Genomic Classification and Prognosis in Acute Myeloid Leukemia.' *New England Journal of Medicine*, 374(23) pp. 2209-2221.

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J. S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., Godfrey, A. L., Rapado, I., Cvejic, A., Rance, R., McGee, C., Ellis, P., Mudie, L. J., Stephens, P. J., McLaren, S., Massie, C. E., Tarpey, P. S., Varela, I., Nik-Zainal, S., Davies, H. R., Shlien, A., Jones, D., Raine, K., Hinton, J., Butler, A. P., Teague, J. W., Baxter, E. J., Score, J., Galli, A., Della Porta, M. G., Travaglino, E., Groves, M., Tauro, S., Munshi, N. C., Anderson, K. C., El-Naggar, A., Fischer, A., Mustonen, V., Warren, A. J., Cross, N. C. P., Green, A. R., Futreal, P. A., Stratton, M. R. and Campbell, P. J. (2011) 'Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts.' *New England Journal of Medicine*, 365(15) pp. 1384-1395.

Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011) 'Sequencing technologies and genome sequencing.' *Journal of Applied Genetics*, 52(4), Nov, pp. 413-435.

Park, S. H., Chi, H. S., Min, S. K., Park, B. G., Jang, S. and Park, C. J. (2011) 'Prognostic impact of c-KIT mutations in core binding factor acute myeloid leukemia.' *Leukemia Research*, 35(10), Oct, pp. 1376-1383.

Parkin, B., Ouillette, P., Yildiz, M., Saiya-Cork, K., Shedden, K. and Malek, S. N. (2015) 'Integrated Genomic Profiling, Therapy Response, and Survival in Adult Acute Myelogenous Leukemia.' *Clinical Cancer Research*, 21(9), May 1, pp. 2045-2056.

Parkin, B., Erba, H., Ouillette, P., Roulston, D., Purkayastha, A., Karp, J., Talpaz, M., Kujawski, L., Shakhan, S., Li, C., Shedden, K. and Malek, S. N. (2010) 'Acquired genomic copy number aberrations and survival in adult acute myelogenous leukemia.' *Blood*, 116(23), Dec 2, pp. 4958-4967.

Paschka, P., Marcucci, G., Ruppert, A. S., Mrózek, K., Chen, H., Kittles, R. A., Vukosavljevic, T., Perrotti, D., Vardiman, J. W., Carroll, A. J., Kolitz, J. E., Larson, R. A. and Bloomfield, C. D. (2006) 'Adverse prognostic significance of KIT mutations in adult acute myeloid leukemia with inv(16) and t(8;21): a Cancer and Leukemia Group B Study.' *Journal of Clinical Oncology*, 24(24), Aug 20, 2006/08/22, pp. 3904-3911.

Paschka, P., Schlenk, R. F., Gaidzik, V. I., Habdank, M., Kronke, J., Bullinger, L., Spath, D., Kayser, S., Zucknick, M., Gotze, K., Horst, H. A., Germing, U., Döhner, H. and Döhner, K. (2010) 'IDH1 and IDH2 Mutations Are Frequent Genetic Alterations in Acute Myeloid Leukemia and Confer Adverse Prognosis in Cytogenetically Normal Acute Myeloid Leukemia With NPM1 Mutation Without FLT3 Internal Tandem Duplication.' *Journal of Clinical Oncology*, 28(22), Aug, pp. 3636-3643.

Paschka, P., Du, J., Schlenk, R. F., Gaidzik, V. I., Bullinger, L., Corbacioglu, A., Spath, D., Kayser, S., Schlegelberger, B., Krauter, J., Ganser, A., Kohne, C. H., Held, G., von Lilienfeld-Toal, M., Kirchen, H., Rummel, M., Gotze, K., Horst, H. A., Ringhoffer, M., Lubbert, M., Wattad, M., Salih, H. R., Kundgen, A., Döhner, H. and Döhner, K. (2013) 'Secondary genetic lesions in acute myeloid leukemia with inv(16) or t(16;16): a study of the German-Austrian AML Study Group (AMLSG).' *Blood*, 121(1), Jan 03, 2012/11/02, pp. 170-177.

Pasquet, M., Bellanne-Chantelot, C., Tavitian, S., Prade, N., Beaupain, B., Larochelle, O., Petit, A., Rohrlich, P., Ferrand, C., Van Den Neste, E., Poirel, H. A., Lamy, T., Ouachee-Chardin, M., Mansat-De Mas, V., Corre, J., Recher, C., Plat, G., Bachelerie, F., Donadieu, J. and Delabesse, E. (2013) 'High frequency of GATA2 mutations in patients with mild chronic neutropenia evolving to MonoMac syndrome, myelodysplasia, and acute myeloid leukemia.' *Blood*, 121(5), Jan 31, 2012/12/12, pp. 822-829.

Patel, J. P., Gönen, M., Figueroa, M. E., Fernandez, H., Sun, Z., Racevskis, J., Van Vlierberghe, P., Dolgalev, I., Thomas, S., Aminova, O., Huberman, K., Cheng, J., Viale, A., Socci, N. D., Heguy, A., Cherry, A., Vance, G., Higgins, R. R., Ketterling, R. P., Gallagher, R. E., Litzow, M., van den Brink, M. R. M., Lazarus, H. M., Rowe, J. M., Luger, S., Ferrando, A., Paietta, E., Tallman, M. S., Melnick, A., Abdel-Wahab, O. and Levine, R. L. (2012) 'Prognostic Relevance of Integrated Genetic Profiling in Acute Myeloid Leukemia.' *New England Journal of Medicine*, 366(12) pp. 1079-1089.

Perez, B., Kosmider, O., Cassinat, B., Renneville, A., Lachenaud, J., Kaltenbach, S., Bertrand, Y., Baruchel, A., Chomienne, C., Fontenay, M., Preudhomme, C. and Cave, H. (2010) 'Genetic typing of CBL, ASXL1, RUNX1, TET2 and JAK2 in juvenile myelomonocytic leukaemia reveals a genetic profile distinct from chronic myelomonocytic leukaemia.' *Br J Haematol*, 151(5), Dec, 2010/10/20, pp. 460-468.

Pfeiffer, T., Schleuning, M., Mayer, J., Haude, K. H., Tischer, J., Buchholz, S., Bunjes, D., Bug, G., Holler, E., Meyer, R. G., Greinix, H., Scheid, C., Christopeit, M., Schnittger, S., Braess, J., Schlimok, G., Spiekermann, K., Ganser, A., Kolb, H. J. and Schmid, C. (2013) 'Influence of molecular subgroups on outcome of acute myeloid leukemia with normal karyotype in 141 patients undergoing salvage allogeneic stem cell transplantation in primary induction failure or beyond first relapse.' *Haematologica*, 98(4), Apr, pp. 518-525.

PHG Foundation. (2011) Next steps in the sequence; The implications of whole genome sequencing for health in the UK. PHG Foundation, Cambridge.

Placke, T., Faber, K., Nonami, A., Putwain, S. L., Salih, H. R., Heidel, F. H., Kramer, A., Root, D. E., Barbie, D. A., Krivtsov, A. V., Armstrong, S. A., Hahn, W. C., Huntly, B. J., Sykes, S. M., Milsom, M. D., Scholl, C. and Frohling, S. (2014) 'Requirement for CDK6 in MLL-rearranged acute myeloid leukemia.' *Blood*, 124(1), Jul 03, 2014/04/26, pp. 13-23.

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M. L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordonez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A. and Campbell, P. J. (2010a) 'A small-cell lung cancer genome with complex signatures of tobacco exposure.' *Nature*, 463(7278), Jan 14, 2009/12/18, pp. 184-190.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordonez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A. and Stratton, M. R. (2010b) 'A comprehensive catalogue of somatic mutations from a human cancer genome.' *Nature*, 463(7278), Jan 14, 2009/12/18, pp. 191-196.

Pratcorona, M., Abbas, S., Sanders, M. A., Koenders, J. E., Kavelaars, F. G., Erpelinck-Verschueren, C. A. J., Zeilemakers, A., Löwenberg, B. and Valk, P. J. M. (2012) 'Acquired mutations in ASXL1 in acute myeloid leukemia: prevalence and prognostic value.' *Haematologica-the Hematology Journal*, 97(3), Mar, pp. 388-392.

Qin, Y. Z., Zhu, H. H., Jiang, Q., Jiang, H., Zhang, L. P., Xu, L. P., Wang, Y., Liu, Y. R., Lai, Y. Y., Shi, H. X., Jiang, B. and Huang, X. J. (2014) 'Prevalence and prognostic significance of c-KIT mutations in core binding factor acute myeloid leukemia: A comprehensive large-scale study from a single Chinese center.' *Leukemia Research*, 38(12), Dec, pp. 1435-1440.

Quinlan, A. R. and Hall, I. M. (2012) 'Characterizing complex structural variation in germline and somatic genomes.' *Trends in Genetics*, 28(1) pp. 43-53.

Rabbani, B., Tekin, M. and Mahdieh, N. (2014) 'The promise of whole-exome sequencing in medical genetics.' *J Hum Genet*, 59(1), Jan, 2013/11/08, pp. 5-15.

Rampal, R. and Figueroa, M. E. (2016) 'Wilms tumor 1 mutations in the pathogenesis of acute myeloid leukemia.' *Haematologica*, 101(6), Jun, 2016/06/03, pp. 672-679.

Raphael, B. J. (2003) 'Reconstructing tumor genome architectures.' *Bioinformatics*, 19(Suppl. 2) pp. ii162-ii172.

Raphael, B. J. (2012) 'Chapter 6: Structural variation and medical genomics.' *PLoS Comput Biol*, 8(12) 2013/01/10, p. e1002821.

Rehm, H. L. (2013) 'Disease-targeted sequencing: a cornerstone in the clinic.' *Nat Rev Genet*, 14(4) pp. 295-300.

Rehm, H. L., Bale, S. J., Bayrak-Toydemir, P., Berg, J. S., Brown, K. K., Deignan, J. L., Friez, M. J., Funke, B. H., Hegde, M. R., Lyon, E., Amer Coll Med, G. and Genomics Lab Quality, A. (2013) 'ACMG clinical laboratory standards for next-generation sequencing.' *Genetics in Medicine*, 15(9), Sep, pp. 733-747.

Ribeiro, A. F. T., Pratcorona, M., Erpelinck-Verschueren, C., Rockova, V., Sanders, M., Abbas, S., Figueroa, M. E., Zeilemaker, A., Melnick, A., Löwenberg, B., Valk, P. J. M. and Delwel, R. (2012) 'Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia.' *Blood*, 119(24), June 14, 2012, pp. 5824-5831.

Rinke, J., Schafer, V., Schmidt, M., Ziermann, J., Kohlmann, A., Hochhaus, A. and Ernst, T. (2013) 'Genotyping of 25 Leukemia-Associated Genes in a Single Work Flow by Next-Generation Sequencing Technology with Low Amounts of Input Template DNA.' *Clinical Chemistry*, 59(8), Aug, pp. 1238-1250.

Rose, D., Haferlach, T., Schnittger, S., Perglerova, K., Kern, W. and Haferlach, C. (2017) 'Subtype-specific patterns of molecular mutations in acute myeloid leukemia.' *Leukemia*, 31(1), Jan, 2016/06/11, pp. 11-17.

Ross, J. S. and Cronin, M. (2011) 'Whole Cancer Genome Sequencing by Next-Generation Methods.' *American Journal of Clinical Pathology*, 136(4), October 1, 2011, pp. 527-539.

Roug, A. S., Hansen, M. C., Nederby, L. and Hokland, P. (2014) 'Diagnosing and following adult patients with acute myeloid leukaemia in the genomic age.' *British Journal of Haematology*, 167(2) pp. 162-176.

Rowley, J. D. (1973a) 'Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.' *Nature*, 243(5405), Jun 1, 1973/06/01, pp. 290-293.

Rowley, J. D. (1973b) 'Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia.' *Annales de Genetique*, 16(2), Jun, 1973/06/01, pp. 109-112.

Rowley, J. D., Golomb, H. M. and Dougherty, C. (1977) '15/17 translocation, a consistent chromosomal change in acute promyelocytic leukaemia.' *Lancet*, 1(8010), Mar 5, 1977/03/05, pp. 549-550.

Royal College of Pathologists. (2010) 'The future provision of molecular diagnostic services for acquired disease in the UK.' 13/10/2010,

Royal College of Physicians. (2016) 'Underfunded. Underdoctored. Overstretched. The NHS in 2016.' [Online]. [Accessed on 30/12/2016] file://christie/dfsroot$/homedrives/NTelford/Downloads/Underfunded,%20underdoctored,%20overstretched_0_0.pdf

Ruan, Y., Ooi, H., Choo, S., Chiu, K., Zhao, X., Srinivasan, K., Yao, F., Choo, C., Liu, J., Ariyaratne, P., Bin, W., Kuznetsov, V., Shahab, A., Sung, W., Bourque, G., Palanisamy, N. and Wei, C. (2007) 'Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs).' *Genome Res*, 17 pp. 828 - 838.

Rubin, E. H. and Gilliland, D. G. (2012) 'Drug development and clinical trials--the path to an approved cancer drug.' *Nat Rev Clin Oncol*, 9(4), Feb 28, 2012/03/01, pp. 215-222.

Rücker, F. G., Schlenk, R. F., Bullinger, L., Kayser, S., Teleanu, V., Kett, H., Habdank, M., Kugler, C.-M., Holzmann, K., Gaidzik, V. I., Paschka, P., Held, G., von Lilienfeld-Toal, M., Lübbert, M., Fröhling, S., Zenz, T., Krauter, J., Schlegelberger, B., Ganser, A., Lichter, P., Döhner, K. and Döhner, H. (2012) 'TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome.' *Blood*, 119(9), Mar 1, pp. 2114-2121.

Rustagi, N., Hampton, O. A., Li, J., Xi, L., Gibbs, R. A., Plon, S. E., Kimmel, M. and Wheeler, D. A. (2016) 'ITD assembler: an algorithm for internal tandem duplication discovery from short-read sequencing data.' *Bmc Bioinformatics*, 17, Apr,

Rykalina, V. N., Shadrin, A. A., Amstislavskiy, V. S., Rogaev, E. I., Lehrach, H. and Borodina, T. A. (2014) 'Exome Sequencing from Nanogram Amounts of Starting DNA: Comparing Three Approaches.' *Plos One*, 9(7), Jul, p. 13.

Salipante, S. J., Fromm, J. R., Shendure, J., Wood, B. L. and Wu, D. (2014) 'Detection of minimal residual disease in NPM1-mutated acute myeloid leukemia by next-generation sequencing.' *Mod Pathol*, 27(11), Nov, 2014/04/20, pp. 1438-1446.

Samorodnitsky, E., Datta, J., Jewell, B. M., Hagopian, R., Miya, J., Wing, M. R., Damodaran, S., Lippus, J. M., Reeser, J. W., Bhatt, D., Timmers, C. D. and Roychowdhury, S. (2015) 'Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing.' *Journal of Molecular Diagnostics*, 17(1), Jan, pp. 64-75.

Samtools. (2015) 'The Variant Call Format (VCF) Version 4.2 Specification.' [Online]. [Accessed on 25/04/2017] https://samtools.github.io/hts-specs/VCFv4.2.pdf

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors.' *Proc Natl Acad Sci U S A*, 74(12), Dec, 1977/12/01, pp. 5463-5467.

Sato-Otsubo, A., Sanada, M. and Ogawa, S. (2012) 'Single-nucleotide polymorphism array karyotyping in clinical practice: where, when, and how?' *Semin Oncol*, 39(1), Feb, 2012/02/01, pp. 13-25.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C. (2015) 'Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform.' *Nucleic Acids Research*, 43(6) pp. e37-e37.

Schlenk, R. F., Döhner, K., Krauter, J., Fröhling, S., Corbacioglu, A., Bullinger, L., Habdank, M., Spath, D., Morgan, M., Benner, A., Schlegelberger, B., Heil, G., Ganser, A. and Döhner, H. (2008) 'Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia.' *New England Journal of Medicine*, 358(18), May 1, 2008/05/03, pp. 1909-1918.

Schnepp, R. W., Bosse, K. R. and Maris, J. M. (2015) 'Improving Patient Outcomes With Cancer Genomics: Unique Opportunities and Challenges in Pediatric Oncology.' *JAMA*, 314(9), Sep 01, 2015/09/02, pp. 881-883.

Schnittger, S., Haferlach, C., Kern, W. and Haferlach, T. (2014) 'Analysis for Loss of Heterozygosity on Chromosome Arm 13q by STR Analysis or SNP Sequencing Can Replace Analysis of FLT3-ITD to Detect Patients with Prognostically Adverse AML.' *Genes Chromosomes & Cancer*, 53(12), Dec, pp. 1008-1017.

Schnittger, S., Bacher, U., Kern, W., Alpermann, T., Haferlach, C. and Haferlach, T. (2011a) 'Prognostic impact of FLT3-ITD load in NPM1 mutated acute myeloid leukemia.' *Leukemia*, 25(8), Aug, pp. 1297-1304.

Schnittger, S., Dicker, F., Kern, W., Wendland, N., Sundermann, J., Alpermann, T., Haferlach, C. and Haferlach, T. (2011b) 'RUNX1 mutations are frequent in de novo AML with noncomplex karyotype and confer an unfavorable prognosis.' *Blood*, 117(8), Feb 24, 2010/12/15, pp. 2348-2357.

Schnittger, S., Schoch, C., Kern, W., Mecucci, C., Tschulik, C., Martelli, M. F., Haferlach, T., Hiddemann, W. and Falini, B. (2005) 'Nucleophosmin gene mutations are predictors of favorable prognosis in acute myelogenous leukemia with a normal karyotype.' *Blood*, 106(12), Dec 1, 2005/08/04, pp. 3733-3739.

Schnittger, S., Kinkelin, U., Schoch, C., Heinecke, A., Haase, D., Haferlach, T., Buchner, T., Wormann, B., Hiddemann, W. and Griesinger, F. (2000) 'Screening for MLL tandem duplication in 387 unselected patients with AML identify a prognostically unfavorable subset of AML.' *Leukemia*, 14(5), May, 2000/05/10, pp. 796-804.

Schoch, C., Kern, W., Kohlmann, A., Hiddemann, W., Schnittger, S. and Haferlach, T. (2005) 'Acute myeloid leukemia with a complex aberrant karyotype is a distinct biological entity characterized by genomic imbalances and a specific gene expression profile.' *"Genes, Chromosomes and Cancer"*, 43(3), Jul, 2005/04/23, pp. 227-238.

Schork, N. J. (2015) 'Personalized medicine: Time for one-person trials.' *Nature*, 520(7549), Apr 30, 2015/05/01, pp. 609-611.

Seifert, H., Mohr, B., Thiede, C., Oelschlagel, U., Schakel, U., Illmer, T., Soucek, S., Ehninger, G. and Schaich, M. (2009) 'The prognostic impact of 17p (p53) deletion in 2272 adults with acute myeloid leukemia.' *Leukemia*, 23(4) pp. 656-663.

Shaffer, L. G., Beaudet, A. L., Brothman, A. R., Hirsch, B., Levy, B., Martin, C. L., Mascarello, J. T. and Rao, K. W. (2007) 'Microarray analysis for constitutional cytogenetic abnormalities.' *Genet Med*, 9(9), 09//print, pp. 654-662.

Shah, S. P., Kobel, M., Senz, J., Morin, R. D., Clarke, B. A., Wiegand, K. C., Leung, G., Zayed, A., Mehl, E., Kalloger, S. E., Sun, M., Giuliany, R., Yorida, E., Jones, S., Varhol, R., Swenerton, K. D., Miller, D., Clement, P. B., Crane, C., Madore, J., Provencher, D., Leung, P., DeFazio, A., Khattra, J., Turashvili, G., Zhao, Y., Zeng, T., Glover, J. N. M., Vanderhyden, B., Zhao, C., Parkinson, C. A., Jimenez-Linan, M., Bowtell, D. D. L., Mes-Masson, A.-M., Brenton, J. D., Aparicio, S. A., Boyd, N., Hirst, M., Gilks, C. B., Marra, M. and Huntsman, D. G. (2009) 'Mutation of FOXL2 in Granulosa-Cell Tumors of the Ovary.' *N Engl J Med*, 360(26), June 25, 2009, pp. 2719-2729.

Shendure, J. and Stewart, C. J. (2009) 'Cancer Genomes on a Shoestring Budget.' *N Engl J Med*, 360(26), June 25, 2009, pp. 2781-2783.

Shiba, N., Ichikawa, H., Taki, T., Park, M. J., Jo, A., Mitani, S., Kobayashi, T., Shimada, A., Sotomatsu, M., Arakawa, H., Adachi, S., Tawa, A., Horibe, K., Tsuchida, M., Hanada, R., Tsukimoto, I. and Hayashi, Y. (2013) 'NUP98-NSD1 gene fusion and its related gene expression signature are strongly associated with a poor prognosis in pediatric acute myeloid leukemia.' *Genes Chromosomes Cancer*, 52(7), Jul, 2013/05/01, pp. 683-693.

Shivarov, V. and Bullinger, L. (2014) 'Expression profiling of leukemia patients: key lessons and future directions.' *Exp Hematol*, 42(8), Aug, 2014/04/22, pp. 651-660.

Shlush, L. I., Zandi, S., Mitchell, A., Chen, W. C., Brandwein, J. M., Gupta, V., Kennedy, J. A., Schimmer, A. D., Schuh, A. C., Yee, K. W., McLeod, J. L., Doedens, M., Medeiros, J. J., Marke, R., Kim, H. J., Lee, K., McPherson, J. D., Hudson, T. J., Brown, A. M., Yousif, F., Trinh, Q. M., Stein, L. D., Minden, M. D., Wang, J. C. and Dick, J. E. (2014) 'Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia.' *Nature*, 506(7488), Feb 20, 2014/02/14, pp. 328-333.

Sindi, S. S., Onal, S., Peng, L. C., Wu, H. T. and Raphael, B. J. (2012) 'An integrative probabilistic model for identification of structural variation in sequencing data.' *Genome Biol*, 13(3) 2012/03/29, p. R22.

Singh, R. R., Patel, K. P., Routbort, M. J., Reddy, N. G., Barkoh, B. A., Handal, B., Kanagal-Shamanna, R., Greaves, W. O., Medeiros, L. J., Aldape, K. D. and Luthra, R. (2013) 'Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes.' *J Mol Diagn*, 15(5), Sep, 2013/07/03, pp. 607-622.

Sinha, C., Cunningham, L. C. and Liu, P. P. (2015) 'Core Binding Factor Acute Myeloid Leukemia: New Prognostic Categories and Therapeutic Opportunities.' *Seminars in Hematology*, 52(3), Jul, 2015/06/27, pp. 215-222.

Siravegna, G., Marsoni, S., Siena, S. and Bardelli, A. (2017) 'Integrating liquid biopsies into the management of cancer.' *Nat Rev Clin Oncol*, Mar 02, 2017/03/03,

Solh, M., Yohe, S., Weisdorf, D. and Ustun, C. (2014) 'Core-binding factor acute myeloid leukemia: Heterogeneity, monitoring, and therapy.' *Am J Hematol*, 89(12), Dec, 2014/08/05, pp. 1121-1131.

Spencer, D. H., Tyagi, M., Vallania, F., Bredemeyer, A. J., Pfeifer, J. D., Mitra, R. D. and Duncavage, E. J. (2014) 'Performance of Common Analysis Methods for Detecting Low-Frequency Single Nucleotide Variants in Targeted Next-Generation Sequence Data.' *The Journal of Molecular Diagnostics*, 16(1) pp. 75-88.

Spencer, D. H., Abel, H. J., Lockwood, C. M., Payton, J. E., Szankasi, P., Kelley, T. W., Kulkarni, S., Pfeifer, J. D. and Duncavage, E. J. (2013) 'Detection of FLT3 Internal Tandem Duplication in Targeted, Short-Read-Length, Next-Generation Sequencing Data.' *Journal of Molecular Diagnostics*, 15(1), Jan, pp. 81-93.

Spinner, M. A., Sanchez, L. A., Hsu, A. P., Shaw, P. A., Zerbe, C. S., Calvo, K. R., Arthur, D. C., Gu, W., Gould, C. M., Brewer, C. C., Cowen, E. W., Freeman, A. F., Olivier, K. N., Uzel, G., Zelazny, A. M., Daub, J. R., Spalding, C. D., Claypool, R. J., Giri, N. K., Alter, B. P., Mace, E. M., Orange, J. S., Cuellar-Rodriguez, J., Hickstein, D. D. and Holland, S. M. (2014) 'GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics, and immunity.' *Blood*, 123(6), 2014-02-06 00:00:00, pp. 809-821.

Stein, B. L., Williams, D. M., O'Keefe, C., Rogers, O., Ingersoll, R. G., Spivak, J. L., Verma, A., Maciejewski, J. P., McDevitt, M. A. and Moliterno, A. R. (2011) 'Disruption of the ASXL1 gene is frequent in primary, post-essential thrombocytosis and post-polycythemia vera myelofibrosis, but not essential thrombocytosis or polycythemia vera: analysis of molecular genetics and clinical phenotypes.' *Haematologica*, 96(10), Oct, 2011/06/30, pp. 1462-1469.

Stein, E. M. and Tallman, M. S. (2016) 'Emerging therapeutic drugs for AML.' *Blood*, 127(1), Jan 7, 2015/12/15, pp. 71-78.

Stephens, P. J., McBride, D. J., Lin, M. L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., Greenman, C. D., Jia, M. M., Latimer, C., Teague, J. W., Lau, K. W., Burton, J., Quail, M. A., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A. M., Martens, J. W. M., Silver, D. P., Langerod, A., Russnes, H. E. G., Foekens, J. A., Reis-Filho, J. S., van't Veer, L., Richardson, A. L., Borresen-Dale, A. L., Campbell, P. J., Futreal, P. A. and Stratton, M. R. (2009) 'Complex landscapes of somatic rearrangement in human breast cancer genomes.' *Nature*, 462(7276), Dec, pp. 1005-U1060.

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009) 'The cancer genome.' *Nature*, 458(7239) pp. 719-724.

Strom, S. P. (2016) 'Current practices and guidelines for clinical next-generation sequencing oncology testing.' *Cancer Biology & Medicine*, 13(1), 01/06/received
02/02/accepted, pp. 3-11.

Sukhai, M. A., Craddock, K. J., Thomas, M., Hansen, A. R., Zhang, T., Siu, L., Bedard, P., Stockley, T. L. and Kamel-Reid, S. (2015) 'A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer.' *Genet Med*, Apr 16, 2015/04/17,

Suzuki, T., Kiyoi, H., Ozeki, K., Tomita, A., Yamaji, S., Suzuki, R., Kodera, Y., Miyawaki, S., Asou, N., Kuriyama, K., Yagasaki, F., Shimazaki, C., Akiyama, H., Nishimura, M., Motoji, T., Shinagawa, K., Takeshita, A., Ueda, R., Kinoshita, T., Emi, N. and Naoe, T. (2005) 'Clinical characteristics and prognostic implications of NPM1 mutations in acute myeloid leukemia.' *Blood*, 106(8), Oct 15, 2005/07/05, pp. 2854-2861.

Swerdlow, S. H., Campo, E., Pileri, S. A., Harris, N. L., Stein, H., Siebert, R., Advani, R., Ghielmini, M., Salles, G. A., Zelenetz, A. D. and Jaffe, E. S. (2016) 'The 2016 revision of the World Health Organization classification of lymphoid neoplasms.' *Blood*, 127(20), May 19, 2016/03/17, pp. 2375-2390.

Swerdlow, S. H., Campo, E., Harris, N. L., Jaffe, E. S., Pilieri, S. A., Stein, H., Thiele, J. and Vardiman, J. W. (eds.) (2008) *WHO Classification of tumours of haematopoietic and lymphoid tissues*.*World Health Organization Classification of Tumours,* 4th ed., Lyon: IARC.

Szpurka, H., Jankowska, A. M., Makishima, H., Bodo, J., Bejanyan, N., Hsi, E. D., Sekeres, M. A. and Maciejewski, J. P. (2010) 'Spectrum of mutations in RARS-T patients includes TET2 and ASXL1 mutations.' *Leuk Res*, 34(8), Aug, 2010/03/26, pp. 969-973.

Tang, J. L., Hou, H. A., Chen, C. Y., Liu, C. Y., Chou, W. C., Tseng, M. H., Huang, C. F., Lee, F. Y., Liu, M. C., Yao, M., Huang, S. Y., Ko, B. S., Hsu, S. C., Wu, S. J., Tsay, W., Chen, Y. C., Lin, L. I. and Tien, H. F. (2009) 'AML1/RUNX1 mutations in 470 adult patients with de novo acute myeloid leukemia: prognostic implication and interaction with other gene alterations.' *Blood*, 114(26), Dec 17, 2009/10/08, pp. 5352-5361.

Taskesen, E., Bullinger, L., Corbacioglu, A., Sanders, M. A., Erpelinck, C. A., Wouters, B. J., van der Poel-van de Luytgaarde, S. C., Damm, F., Krauter, J., Ganser, A., Schlenk, R. F., Löwenberg, B., Delwel, R., Döhner, H., Valk, P. J. and Döhner, K. (2011) 'Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity.' *Blood*, 117(8), Feb 24, 2010/12/24, pp. 2469-2475.

Technology Strategy Board. (2011) 'Stratified Medicine in the UK; Vision and Roadmap.' [Online]. [Accessed on 29/12/2016] https://connect.innovateuk.org/documents/2843120/3724280/Stratified+Medicines+Roadmap.pdf/fbb39848-282e-4619-a960-51e3a16ab893

ten Bosch, J. R. and Grody, W. W. (2008) 'Keeping Up With the Next Generation Massively Parallel Sequencing in Clinical Diagnostics.' *Journal of Molecular Diagnostics*, 10(6), Nov, pp. 484-492.

The Cancer Genome Atlas Research Network. (2013) 'Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia.' *New England Journal of Medicine*, 368(22) pp. 2059-2074.

The Genome Institute at Washington University School of Medicine. (2014) *Pindel User Manual*. Genome Modeling Tools. [Online] [Accessed on 24/03/2017] http://gmt.genome.wustl.edu/packages/pindel/user-manual.html

Theilgaard-Monch, K., Boultwood, J., Ferrari, S., Giannopoulos, K., Hernandez-Rivas, J. M., Kohlmann, A., Morgan, M., Porse, B., Tagliafico, E., Zwaan, C. M., Wainscoat, J., Van den Heuvel-Eibrink, M. M., Mills, K. and Bullinger, L. (2011) 'Gene expression profiling in MDS and AML: potential and future avenues.' *Leukemia*, 25(6), Jun, pp. 909-920.

Thiede, C., Koch, S., Creutzig, E., Steudel, C., Illmer, T., Schaich, M. and Ehninger, G. (2006) 'Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML).' *Blood*, 107(10), May 15, 2006/02/04, pp. 4011-4020.

Thiede, C., Steudel, C., Mohr, B., Schaich, M., Schakel, U., Platzbecker, U., Wermke, M., Bornhauser, M., Ritter, M., Neubauer, A., Ehninger, G. and Illmer, T. (2002) 'Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis.' *Blood*, 99(12), Jun 15, 2002/05/31, pp. 4326-4335.

Thol, F., Damm, F., LÃ¼deking, A., Winschel, C., Wagner, K., Morgan, M., Yun, H., GÃ¶hring, G., Schlegelberger, B., Hoelzer, D., LÃ¼bbert, M., Kanz, L., Fiedler, W., Kirchner, H., Heil, G., Krauter, J. r., Ganser, A. and Heuser, M. (2011) 'Incidence and Prognostic Influence of DNMT3A Mutations in Acute Myeloid Leukemia.' *Journal of Clinical Oncology*, 29(21), July 20, 2011, pp. 2889-2896.

Thorvaldsdottir, H., Robinson, J. T. and Mesirov, J. P. (2013) 'Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.' *Brief Bioinform*, 14(2), Mar, 2012/04/21, pp. 178-192.

Thota, S., Viny, A. D., Makishima, H., Spitzer, B., Radivoyevitch, T., Przychodzen, B., Sekeres, M. A., Levine, R. L. and Maciejewski, J. P. (2014) 'Genetic alterations of the cohesin complex genes in myeloid malignancies.' *Blood*, 124(11), Sep 11, 2014/07/10, pp. 1790-1798.

Tian, Q., Price, N. D. and Hood, L. (2012) 'Systems Cancer Medicine: Towards Realization of Predictive, Preventive, Personalized, and Participatory (P4) Medicine.' *Journal of internal medicine*, 271(2) pp. 111-121.

Tipping, A. J., Pina, C., Castor, A., Hong, D., Rodrigues, N. P., Lazzari, L., May, G. E., Jacobsen, S. E. and Enver, T. (2009) 'High GATA-2 expression inhibits human hematopoietic stem and progenitor cell function by effects on cell cycle.' *Blood*, 113(12), Mar 19, 2009/01/27, pp. 2661-2672.

Tiu, R. V., Gondek, L. P., O'Keefe, C. L., Elson, P., Huh, J., Mohamedali, A., Kulasekararaj, A., Advani, A. S., Paquette, R., List, A. F., Sekeres, M. A., McDevitt, M. A., Mufti, G. J. and Maciejewski, J. P. (2011) 'Prognostic impact of SNP array karyotyping in myelodysplastic syndromes and related myeloid malignancies.' *Blood*, 117(17), 2011-04-28 00:00:00, pp. 4552-4560.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) 'The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.' *Contemporary Oncology*, 19(1A), 01/20, pp. A68-A77.

Tosi, S., Harbott, J., Teigler-Schlegel, A., Haas, O. A., Pirc-Danoewinata, H., Harrison, C. J., Biondi, A., Cazzaniga, G., Kempski, H., Scherer, S. W. and Kearney, L. (2000) 't(7;12)(q36;p13), a new recurrent translocation involving ETV6 in infant leukemia.' *Genes Chromosomes Cancer*, 29(4), Dec, 2000/11/07, pp. 325-332.

Tsai, F. Y., Keller, G., Kuo, F. C., Weiss, M., Chen, J., Rosenblatt, M., Alt, F. W. and Orkin, S. H. (1994) 'An early haematopoietic defect in mice lacking the transcription factor GATA-2.' *Nature*, 371(6494), Sep 15, 1994/09/15, pp. 221-226.

Tucker, T., Marra, M. and Friedman, J. M. (2009) 'Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine.' *American Journal of Human Genetics*, 85(2), Aug, pp. 142-154.

Tutt, B. (2012) *New Treatments for Acute Myelogenous Leukemia May Improve Patient Outcomes*. OncoLog. [Online] [Accessed on 16/04/2016] https://www2.mdanderson.org/depts/oncolog/articles/12/4-apr/4-12-2.html

UKNEQAS. (2017) *Pre Pilot Acute Myeloid Leukaemia Gene Panels (Not Accredited).* . 24/03/2017. Sheffield Teaching Hospital NHS Foundation Trust,. (AML GP 161701)

University of Birmingham. (2015) *MyeChild 01 - International Randomised Phase III Clinical Trial in Children with Acute Myeloid Leukaemia.* 04/08/2015.

Valk, P. J., Verhaak, R. G., Beijen, M. A., Erpelinck, C. A., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J. M., Beverloo, H. B., Moorhouse, M. J., van der Spek, P. J., Löwenberg, B. and Delwel, R. (2004) 'Prognostically useful gene-expression profiles in acute myeloid leukemia.' *N Engl J Med*, 350(16), Apr 15, 2004/04/16, pp. 1617-1628.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. and DePristo, M. A. (2013) 'From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.' *Curr Protoc Bioinformatics*, 43 2014/11/29, pp. 11 10 11-33.

Vardiman, J. W., Arber, D. A., Brunning, R. D., Larson, R. A., Matutes, E., Baumann, I. and Thiele, J. (2008) 'Therapy-related myeloid neoplasms.' *In* Swerdlow, S. H., Campo, E., Harris, N. L., Jaffe, E. S., Pilieri, S. A., Stein, H., Thiele, J. and Vardiman, J. W. (eds.) *WHO Classification of tumours of haematopoietic and lymphoid tissues*. 4th ed., Lyon: IARC, pp. pg 127-129.

Verhaak, R. G., Wouters, B. J., Erpelinck, C. A., Abbas, S., Beverloo, H. B., Lugthart, S., Löwenberg, B., Delwel, R. and Valk, P. J. (2009) 'Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling.' *Haematologica*, 94(1), Jan, 2008/10/08, pp. 131-134.

Vermeesch, J. R., Brady, P. D., Sanlaville, D., Kok, K. and Hastings, R. J. (2012) 'Genome-Wide Arrays: Quality Criteria and Platforms to be Used in Routine Diagnostics.' *Human Mutation*, 33(6), Jun, pp. 906-915.

Vermeulen, N. (2009) Big science; characterising transformations in science. *Supersizing science; on building large scale research projects in Biology.* Florida, USA: Dissertation.com.

Vermeulen, N. (2016) 'Big Biology.'
NTM Journal of the History of Science, Technology and Medicine, 24(2) pp. 195-223.

Vertitas Genetics. (2015) *Veritas Genetics Breaks $1,000 Whole Genome Barrier*. [Online] [Accessed on 14/04/2016] http://www.prnewswire.com/news-releases/veritas-genetics-breaks-1000-whole-genome-barrier-300150585.html

Vinh, D. C., Patel, S. Y., Uzel, G., Anderson, V. L., Freeman, A. F., Olivier, K. N., Spalding, C., Hughes, S., Pittaluga, S., Raffeld, M., Sorbara, L. R., Elloumi, H. Z., Kuhns, D. B., Turner, M. L., Cowen, E. W., Fink, D., Long-Priel, D., Hsu, A. P., Ding, L., Paulson, M. L., Whitney, A. R., Sampaio, E. P., Frucht, D. M., DeLeo, F. R. and Holland, S. M. (2010) 'Autosomal dominant and sporadic monocytopenia with susceptibility to mycobacteria, fungi, papillomaviruses, and myelodysplasia.' *Blood*, 115(8), Feb 25, 2009/12/31, pp. 1519-1529.

Voelkerding, K. V., Dames, S. A. and Durtschi, J. D. (2009) 'Next-Generation Sequencing: From Basic Research to Diagnostics.' *Clin Chem*, 55(4), April 1, 2009, pp. 641-658.

Volik, S., Zhao, S., Chin, K., Brebner, J. H., Herndon, D. R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W. L., Magrane, G., De Jong, P., Gray, J. W. and Collins, C. (2003) 'End-sequence profiling: sequence-based analysis of aberrant genomes.' *Proc Natl Acad Sci U S A*, 100(13), Jun 24, 2003/06/06, pp. 7696-7701.

von Bergh, A. R., van Drunen, E., van Wering, E. R., van Zutven, L. J., Hainmann, I., Lonnerholm, G., Meijerink, J. P., Pieters, R. and Beverloo, H. B. (2006) 'High incidence of t(7;12)(q36;p13) in infant AML but not in infant ALL, with a dismal outcome and ectopic expression of HLXB9.' *Genes Chromosomes Cancer*, 45(8), Aug, 2006/04/29, pp. 731-739.

von Neuhoff, C., Reinhardt, D., Sander, A., Zimmermann, M., Bradtke, J., Betts, D. R., Zemanova, Z., Stary, J., Bourquin, J. P., Haas, O. A., Dworzak, M. N. and Creutzig, U. (2010) 'Prognostic impact of specific chromosomal aberrations in a large group of pediatric patients with acute myeloid leukemia treated uniformly according to trial AML-BFM 98.' *J Clin Oncol*, 28(16), Jun 01, 2010/05/05, pp. 2682-2689.

Wade, N. (2010) 'A decade later, genetic map yields few new cures.' *New York Times.* 12th June 2010.

Wadhwa, V. (2014) 'The triumph of genomic medicine is just beginning.' *The Washington Post.* Innovations. 13th March 2014.

Wall, J. D., Tang, L. F., Zerbe, B., Kvale, M. N., Kwok, P. Y. and Schaefer, C. (2014) 'Estimating genotype error rates from high-coverage next-generation sequence data.' *Genome Research*, 24

Wan, J. C., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R. and Rosenfeld, N. (2017) 'Liquid biopsies come of age: towards implementation of circulating tumour DNA.' *Nat Rev Cancer*, Feb 24, 2017/02/25,

Wang, F., Travins, J., DeLaBarre, B., Penard-Lacronique, V., Schalm, S., Hansen, E., Straley, K., Kernytsky, A., Liu, W., Gliser, C., Yang, H., Gross, S., Artin, E., Saada, V., Mylonas, E., Quivoron, C., Popovici-Muller, J., Saunders, J. O., Salituro, F. G., Yan, S., Murray, S., Wei, W., Gao, Y., Dang, L., Dorsch, M., Agresta, S., Schenkein, D. P., Biller, S. A., Su, S. M., de Botton, S. and Yen, K. E. (2013) 'Targeted inhibition of mutant IDH2 in leukemia cells induces cellular differentiation.' *Science*, 340(6132), May 03, 2013/04/06, pp. 622-626.

Wang, X., Muramatsu, H., Okuno, Y., Sakaguchi, H., Yoshida, K., Kawashima, N., Xu, Y., Shiraishi, Y., Chiba, K., Tanaka, H., Saito, S., Nakazawa, Y., Masunari, T., Hirose, T., Elmahdi, S., Narita, A., Doisaki, S., Ismael, O., Makishima, H., Hama, A., Miyano, S., Takahashi, Y., Ogawa, S. and Kojima, S. (2015) 'GATA2 and secondary mutations in familial myelodysplastic syndromes and pediatric myeloid malignancies.' *Haematologica*, May 28, 2015/05/30,

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics.' *Nat Rev Genet*, 10(1), Jan, 2008/11/19, pp. 57-63.

Watson, J. D. and Crick, F. H. (1953) 'Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.' *Nature*, 171(4356), Apr 25, 1953/04/25, pp. 737-738.

Weissmann, S., Alpermann, T., Grossmann, V., Kowarsch, A., Nadarajah, N., Eder, C., Dicker, F., Fasan, A., Haferlach, C., Haferlach, T., Kern, W., Schnittger, S. and Kohlmann, A. (2012) 'Landscape of TET2 mutations in acute myeloid leukemia.' *Leukemia*, 26(5), May, 2011/11/26, pp. 934-942.

Welch, J. S. (2014) 'Mutation position within evolutionary subclonal architecture in AML.' *Seminars in Hematology*, 51(4), Oct, 2014/10/15, pp. 273-281.

Welch, J. S. and Link, D. C. (2011) 'Genomics of AML: Clinical Applications of Next-Generation Sequencing.' *ASH Education Program Book*, 2011(1), December 10, 2011, pp. 30-35.

Welch, J. S., Westervelt, P., Ding, L., Larson, D. E., Klco, J. M., Kulkarni, S., Wallis, J., Chen, K., Payton, J. E., Fulton, R. S., Veizer, J., Schmidt, H., Vickery, T. L., Heath, S., Watson, M. A., Tomasson, M. H., Link, D. C., Graubert, T. A., DiPersio, J. F., Mardis, E. R., Ley, T. J. and Wilson, R. K. (2011a) 'Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene.' *JAMA: The Journal of the American Medical Association*, 305(15), April 20, 2011, pp. 1577-1584.

Welch, J. S., Larson, D., Ding, L., McLellan, M. D., Lamprecht, T., Kandoth, C., Payton, J. E., Baty, J., Harris, C. C., Lichti, C. F., Fulton, R. S., Dooling, D. J., Koboldt, D. C., Schmidt, H., Zhang, Q. Y., Osborne, J. R., Lin, L., O'Laughlin, M., McMichael, J. F., Delehaunty, K. D., McGrath, S. D., Fulton, L. A., Magrini, V. J., Vickery, T. L., Wylie, T., Walker, J., Westervelt, P., Tomasson, M. H., Watson, M. A., Heath, S., Shannon, W. D., Nagarajan, R., Link, D. C., Graubert, T., DiPersio, J. F., Mardis, E. R., Wilson, R. K. and Ley, T. J. (2011b) 'Complete Sequencing and Comparison of 12 Normal Karyotype M1 AML Genomes with 12 t(15;17) Positive M3-APL Genomes.' *Blood*, 118(21), Nov, pp. 185-185.

Wellcome Trust Sanger Institute. (2016) 'Cancer Genome Project; Catalogue Of Somatic Mutations In Cancer (COSMIC).' [Online] Vol. v78. [Accessed on 10/10/2016] http://cancer.sanger.ac.uk/cosmic

Wen, H., Li, Y., Malek, S. N., Kim, Y. C., Xu, J., Chen, P., Xiao, F., Huang, X., Zhou, X., Xuan, Z., Mankala, S., Hou, G., Rowley, J. D., Zhang, M. Q. and Wang, S. M. (2012) 'New fusion transcripts identified in normal karyotype acute myeloid leukemia.' *Plos One*, 7(12) 2012/12/20, p. e51203.

West, A. H., Godley, L. A. and Churpek, J. E. (2014a) 'Familial myelodysplastic syndrome/acute leukemia syndromes: a review and utility for translational investigations.' *Ann N Y Acad Sci*, 1310, Mar, 2014/01/29, pp. 111-118.

West, R. R., Hsu, A. P., Holland, S. M., Cuellar-Rodriguez, J. and Hickstein, D. D. (2014b) 'Acquired ASXL1 mutations are common in patients with inherited GATA2 mutations and correlate with myeloid transformation.' *Haematologica*, 99(2), Feb, pp. 276-281.

Whitman, S. P., Archer, K. J., Feng, L., Baldus, C., Becknell, B., Carlson, B. D., Carroll, A. J., Mrózek, K., Vardiman, J. W., George, S. L., Kolitz, J. E., Larson, R. A., Bloomfield, C. D. and Caligiuri, M. A. (2001) 'Absence of the wild-type allele predicts poor prognosis in adult de novo acute myeloid leukemia with normal cytogenetics and the internal tandem duplication of FLT3: a cancer and leukemia group B study.' *Cancer Research*, 61(19), Oct 1, 2001/10/05, pp. 7233-7239.

Whitman, S. P., Ruppert, A. S., Marcucci, G., Mrózek, K., Paschka, P., Langer, C., Baldus, C. D., Wen, J., Vukosavljevic, T., Powell, B. L., Carroll, A. J., Kolitz, J. E., Larson, R. A., Caligiuri, M. A. and Bloomfield, C. D. (2007) 'Long-term disease-free survivors with cytogenetically normal acute myeloid leukemia and MLL partial tandem duplication: a Cancer and Leukemia Group B study.' *Blood*, 109(12), Jun 15, 2007/03/08, pp. 5164-5167.

Whitman, S. P., Ruppert, A. S., Radmacher, M. D., Mrózek, K., Paschka, P., Langer, C., Baldus, C. D., Wen, J., Racke, F., Powell, B. L., Kolitz, J. E., Larson, R. A., Caligiuri, M. A., Marcucci, G. and Bloomfield, C. D. (2008) 'FLT3 D835/I836 mutations are associated with poor disease-free survival and a distinct gene-expression signature among younger adults with de novo cytogenetically normal acute myeloid leukemia lacking FLT3 internal tandem duplications.' *Blood*, 111(3), Feb, pp. 1552-1559.

Whitman, S. P., Caligiuri, M. A., Maharry, K., Radmacher, M. D., Kohlschmidt, J., Becker, H., Mrózek, K., Wu, Y. Z., Schwind, S., Metzeler, K. H., Mendler, J. H., Wen, J., Baer, M. R., Powell, B. L., Carter, T. H., Kolitz, J. E., Wetzler, M., Carroll, A. J., Larson, R. A., Marcucci, G. and Bloomfield, C. D. (2012) 'The MLL partial tandem duplication in adults aged 60 years and older with de novo cytogenetically normal acute myeloid leukemia.' *Leukemia*, 26(7), Jul, 2012/03/03, pp. 1713-1717.

Will, C. L. and Lührmann, R. (2011) 'Spliceosome Structure and Function.' *Cold Spring Harbor Perspectives in Biology*, 3(7) p. a003707.

Wlodarski, M. W., Collin, M. and Horwitz, M. S. (2017) 'GATA2 deficiency and related myeloid neoplasms.' *Seminars in Hematology*, 54(2) pp. 81-86.

Wlodarski, M. W., Hirabayashi, S., Pastor, V., Starý, J., Hasle, H., Masetti, R., Dworzak, M., Schmugge, M., van den Heuvel-Eibrink, M., Ussowicz, M., De Moerloose, B., Catala, A., Smith, O. P., Sedlacek, P., Lankester, A. C., Zecca, M., Bordon, V., Matthes-Martin, S., Abrahamsson, J., Kühl, J. S., Sykora, K.-W., Albert, M. H., Przychodzien, B., Maciejewski, J., Schwarz, S., Göhring, G., Schlegelberger, B., Cseh, A. m., Noellke, P., Yoshimi, A., Locatelli, F., Baumann, I., Strahm, B. and Niemeyer, C. M. (2016) 'Prevalence, clinical characteristics and prognosis of GATA2-related myelodysplastic syndromes (MDS) in children and adolescents.' *Blood*, 2015-01-01 00:00:00, World Health Organization. (2017) *Human Genomics in Global Health; WHO definitions of genetics and genomics*. [Online] [Accessed on 19/06/2017] http://www.who.int/genomics/geneticsVSgenomics/en/

Wouters, B. J., Löwenberg, B., Erpelinck-Verschueren, C. A., van Putten, W. L., Valk, P. J. and Delwel, R. (2009) 'Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome.' *Blood*, 113(13), Mar 26, 2009/01/28, pp. 3088-3091.

Yan, X. J., Xu, J., Gu, Z. H., Pan, C. M., Lu, G., Shen, Y., Shi, J. Y., Zhu, Y. M., Tang, L., Zhang, X. W., Liang, W. X., Mi, J. Q., Song, H. D., Li, K. Q., Chen, Z. and Chen, S. J. (2011) 'Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia.' *Nature Genetics*, 43(4), Apr, pp. 309-U351.

Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C. H., Zhang, C., Ren, X., Protopopov, A., Chin, L., Kucherlapati, R., Lee, C. and Park, P. J. (2013) 'Diverse mechanisms of somatic structural variations in human cancer genomes.' *Cell*, 153(4), May 09, 2013/05/15, pp. 919-929.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R. and Ning, Z. (2009) 'Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.' *Bioinformatics*, 25(21) pp. 2865-2871.

Zelent, A., Guidez, F., Melnick, A., Waxman, S. and Licht, J. D. (2001) 'Translocations of the RARalpha gene in acute promyelocytic leukemia.' *Oncogene*, 20(49), Oct 29, 2001/11/13, pp. 7186-7203.

7.0 Appendices

7.1. Appendix 1. Concentration of DNA from samples used in the project

Table 7.1 Concentration of DNA from 36 samples use in the project

Sample ID	Concentration Nanodrop (ng/µl)	Concentration Qubit (ng/µl)	Absorbance $A_{260/280}$	DIN	Comments
13.0816	21.2	110	1.85	8.8	
13.0881	26.4	71.6	1.85	8.2	
13.0883	34.9	34.6	1.95	7.1	
13.0886	38.3	79.4	1.95	7.3	
13.0895	767.0	1060	1.84	7.9	
13.0970 (tubeA)	9.6	13.1	1.74	8.2	2 tubes. Low yield
13.0970 (tubeB)		12.8		8.8	
13.1004	15.9	17.9	2.02	7.8	
13.1027	14.7	24.2	1.76	7.4	
13.1050	84.7	63.6	1.98	8.3	
13.1141 (tubeA)	7.1	3.48	2.12	8.3	2 tubes. Low yield
13.1141 (tubeB)		12		8.3	
13.1216	130.0	129.2	1.92	8.1	
13.1249	107.0	124.4	1.90	8.8	
13.1420 (tubeA)	8.0	2.92	2.06	OK	2 tubes. Low yield 1.5µg
13.1420 (tubeB)		14.1		8.6	
13.1485	59.7	60.8	1.82	8.4	
13.1568	13.4	30.8	1.95	7.2	
13.1760 (tubeA)	12.8	16.2	1.96	8	2 tubes
13.1760 (tube B)		5.36		OK	
13.1772	33.5	63.4	1.99	6.7	
13.1931	178.1	58.2	1.87	8.4	
13.2006	63.8	20.6	2.04	6.5	
13.2017	40.3	114	1.85	8.5	
13.2231	55.1	72.8	1.85	8.1	
13.2246	5.7	32.3	2.16	8.3	Low yield 2µg
13.2326	31.9	41	1.87	8.4	
13.2412	74.9	66.6	1.86	8.8	
13.2419	31.4	28.6	1.91	7.4	
13.2540		14.5		8.6	Nanodrop n/a
13.2541	37.9	27.8	1.93	8.5	
13.2679	49.8	17.7	2.03	8.1	
13.2680	97.9	79	1.93	8.9	

13.2758	243.4	207	1.87	9.6	
13.2878	42.0	40.4	1.84	9	
13.2957	77.4	17.8	2.00	6.4	
13.3005	68.9	75	1.84	8.3	
13.3120	45.1	7.8	1.96	OK	
13.3157	57.6	77.6	1.85	8.1	
13.3334	47.9	74.6	1.86	8.3	

DNA concentration was recorded in house using a Nanodrop lite at the time of extraction. Concentration was reassessed at the time of sequencing using a Qubit, which was considered to be more reliable. The Absorbance ratio ($A_{260/280}$) was estimated by the Nanodrop and is used to assess the purity of nucleic acids, with pure DNA being ~1.8. DNA Integrity Number (DIN) is produced by the Agilent 2200 TapeStation instrument to score gDNA samples. The DNA integrity number (DIN) ranges from 1 to 10 and is a reliable and reproducible objective measure of gDNA degradation.

7.2. Appendix 2. Read Metrics from Main Sequencing Run on Illumina Nextseq

Table 7.2 showing run statistics broken down into lanes for each of paired read 1 and 2. The number of tiles per lane (Tiles) and the density of clusters (in thousands per mm^2) detected by image analysis, +/- 1 standard deviation (Density) and the percentage of clusters passing filtering, +/- 1 standard deviation (Cluster PF) are shown. Phase and Prephase are the percentage of molecules in a cluster for which sequencing falls behind (phasing) or jumps ahead (prephasing) cycles within a read, for the NextSeq, estimated empirically for every cycle and this figure is an aggregate value determined from the first 25 cycles. The number of reads (clusters in millions), the number of clusters (in millions) passing filtering (Reads PF), and the percentage of bases with a quality score of 30 or higher (%Q ≥ 30) and the number of bases sequenced which passed filter (Yield) are displayed. 150 cycles were error-rated using PhiX and the alignment rate for each read is shown (Aligned) and error ate calculated from the PhiX alignment is shown (Error Rate %) as well as error rate at specific cycles (35, 75 and 100). The intensity at Cycle 20 (as a % of the intensity at Cycle 1 (Intensity at cycle 20)/(Intensity at cycle 1) x100) is shown.

Table 7.2. Read Tables Metrics

Lane	Read	Tiles	Density (k/mm^2)	Cluster PF (%)	Phase/ Prephase (%)	Reads	Reads PF	%≥Q30	Yield (Gbp)	Aligned (%)	Error Rate (%)	Error Rate 35 Cycles (%)	Error Rate 75 Cycles (%)	Error Rate 100 Cycles (%)	Intensity Cycle 20/1*100
1	1	216	203 ±6	91.59 ±1.68	0.148 / 0.094	131,999,848	120,908,800	88.2	18.14	0.96 ±0.09	0.89 ±0.25	0.23 ±0.19	0.33 ±0.21	0.44 ±0.20	5,244 ±524
	2	216	203 ±6	91.59 ±1.68	0.187 / 0.147	131,999,848	120,908,800	82.03	18.12	0.93 ±0.09	1.31 ±0.43	0.31 ±0.13	0.58 ±0.24	0.76 ±0.28	5,789 ±534
2	1	216	200 ±10	91.77 ±1.53	0.141 / 0.086	130,027,752	119,339,056	88.86	17.90	0.93 ±0.09	0.86 ±0.16	0.23 ±0.08	0.30 ±0.08	0.41 ±0.09	5,407 ±558
	2	216	200 ±10	91.77 ±1.53	0.194 / 0.150	130,027,752	119,339,056	82.44	17.88	0.90 ±0.09	1.25 ±0.28	0.29 ±0.06	0.55 ±0.13	0.73 ±0.17	5,538 ±567
3	1	216	195 ±5	89.21 ±1.86	0.135 / 0.091	126,349,864	112,689,376	86.23	16.90	0.83 ±0.04	1.25 ±0.22	0.37 ±0.22	0.50 ±0.20	0.65 ±0.19	3,413 ±267
	2	216	195 ±5	89.21 ±1.86	0.189 / 0.152	126,349,864	112,689,376	78.16	16.90	0.80 ±0.03	1.96 ±0.26	0.46 ±0.11	0.88 ±0.15	1.16 ±0.18	4,194 ±622
4	1	216	192 ±5	88.58 ±2.04	0.127 / 0.090	124,502,872	110,321,568	86.22	16.55	0.83 ±0.03	1.34 ±0.28	0.34 ±0.07	0.49 ±0.12	0.68 ±0.16	3,102 ±196
	2	216	192 ±5	88.58 ±2.04	0.177 / 0.152	124,502,872	110,321,568	77.04	16.55	0.78 ±0.03	2.24 ±0.51	0.52 ±0.12	1.01 ±0.23	1.35 ±0.31	3,866 ±497

7.3. Appendix 3. Life Technologies Ion Torrent PGM sequencing

Six Ion Torrent PGM chips were used to provide sequencing proton… sequencing for the 36 AML samples. Sample DNA density varied significantly depending on the yield from the original sample, mean 114 ng/ul (range 3.7 – 786). Volumes added to the Chip were adjusted accordingly to ensure the correct amount of DNA. All samples used 10ng with the exception of 3 (mean 10.15, range 9.42 – 12.9).

Table 7.3 Ion Torrent sequencing - Chip overview and metrics

Mean no. mapped reads per sample	832258 (397752 - 1147952)
Mean On Target % (range)	89.68 (75.18 - 97.04)
Mean Depth (range)	3431 (1701-4889)
Mean Uniformity %	96.3
End to end reads %	85.9
Amplicons with at least 500 reads %	97.93 (94.5 - 99.2)
Mean Total reads per chip	5139307 (4533640 - 5583212)

7.4. Appendix 4. Published work

7.4.1 Myeloproliferative neoplasm with eosinophilia and T-lymphoblastic lymphoma with ETV6-LYN gene fusion.

SNP microarray testing was performed in a limited number of cases during the course of the project, to gain experience in handling and analysing genomic data.

The following article was published regarding the detection of an unusual chromosomal translocation in myeloproliferative neoplasm with eosinophilia and T-lymphoblastic lymphoma. The methodology was not the same as that used for the NGS project and therefore is not referred to directly. This early work does not necessarily conform to the Manchester Metropolitan University 'Institutional Codes of Practice and Research Degrees Regulations'.

Blood Cancer J. 2016 Apr 8;6:e412. doi: 10.1038/bcj.2016.11.

Myeloproliferative neoplasm with eosinophilia and T-lymphoblastic lymphoma with ETV6-LYN gene fusion.

Telford N(1), Alexander S(2), McGinn OJ(2), Williams M(3), Wood KM(4), Bloor A(5), Saha V(2)(6).

Author information:

(1) Oncology Cytogenetics, The Christie Pathology Partnership, The Christie NHS Foundation Trust, Manchester, UK. (2) Children's Cancer Group, Centre for Paediatric, Teenage and Young Adult Cancer, Institute of Cancer, University of Manchester, Manchester, UK. (3) Leukaemia Biology Group, Institute of Cancer, University of Manchester, Manchester, UK. (4)Department of Cellular Pathology, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, UK. (5)Haematology and Transplant Unit, The Christie NHS Foundation Trust, Manchester, UK. (6)Tata Translational Cancer Research Centre, Tata Medical Center, Kolkata, India.

DOI: 10.1038/bcj.2016.11. PMCID: PMC4855251.PMID: 27058227 [Indexed for MEDLINE]

7.4.2 Germline GATA2 mutation in three siblings with familial Myelodysplastic syndrome

This is an unpublished, draft report of a family with *GATA2* germline mutation, whose MDS and AML samples were tested by the methodology described in the project. My analysis confirmed a frameshift deletion of CTTCTGGCGGC in *GATA2* at exon 5 c.1021_1031delGCCGCCAGAAG p.(Ala341SerfsTer39). A WT1 (32417910 G>T) missense substitution was also detected in the sibling with AML.

Clinical and cytogenetic features of myeloid malignancy in three siblings with familial *GATA2* frameshift mutation without syndromic features

Nick Telford [1], Jamie Ellingford[2], Jill Urquard[2], Marion Schneider[3], Rob Wynn[4], Andrew Will[4], Denise Bonney[2], Bronwyn Kerr[5], Graeme Black[5], Stefan Meyer[3,4,5]

Introduction

Familial myelodysplastic syndrome (MDS) and acute myeloid leukaemia (AML) are rare and their origins were obscure until recent advances in genomic testing revealed the genetic basis of these disorders. An increasing number of myeloid neoplasms with hereditary predisposition are now recognised and included as distinct entities in the latest diagnostic classification (Arber *et al.*, 2016). In particular, heterozygous germline mutations of the GATA binding protein 2 (*GATA2)* gene were first identified in familial myeloid neoplasia (Hahn *et al.*, 2011) and in several autosomal dominant disorders with an increased risk of MDS and AML; Emberger (Ostergaard *et al.*, 2011) and MonoMAC syndromes (Hsu *et al.*, 2011), and Dendritic cell, Monocyte and Lymphoid deficiency (DCML) (Dickinson *et al.*, 2011). Patients with congenital neutropaenia were later detected (Pasquet *et al.*, 2013; Ganapathi *et al.*, 2015). A complex assortment of diseases are now recognised, considered to be part of the same autosomal dominant genetic disorder, collectively known as '*GATA2* deficiency syndrome,' which can also include different neoplastic diseases, and a broad range of disease characteristics, including non-haematological and non-infectious complications (reviewed in

Crispino & Horwitz, 2017; Wlodarski *et al.*, 2017). There is considerable variation in the presenting phenotypes of these disorders, with overlap of clinical features, different ages of onset and penetrance of features (Dickinson *et al.*, 2014; Spinner *et al.*, 2014; Collin *et al.*, 2015). The precise incidence of *GATA2* deficiency is unknown, however, evidence suggests that disease penetrance is high and that the vast majority of mutation carriers will develop neoplastic disease or immunological problems during their lifetime (Spinner *et al.*, 2014; Collin *et al.*, 2015).

The *GATA2* gene, located at 3q21.3, is one of a family of fifteen 'GATA zinc finger domain containing' genes (HUGO Gene Nomenclature Committee (HGCN) Database). *GATA2* encodes a transcription factor, with two zinc finger DNA binding regions (ZF1 and ZF2) involved in binding to the promoters and enhancers of numerous target genes to form functional heterodimers (Hahn *et al.*, 2011). Furthermore, *GATA2* (and *GATA1*) has been shown to be self-activating and regulators of their own transcription (Cortés-Lavaud *et al.*, 2015). *GATA2* is a critical regulator of gene expression in haematopoietic cells and is necessary for the production and function of haematopoietic stem cells (HSC) and for the maintenance of the HSC reservoir (Tsai *et al.*, 1994; de Pater *et al.*, 2013). *GATA2* controls the production of specific progenitors of monocytes, B and NK lymphocytes and dendritic cells. More than one hundred different germline *GATA2* mutations are now known, approximately a third of which are inherited. Three main types have been described; approximately 60% are nonsense and frameshift mutations located at different positions within the gene, upstream of or within ZF2. 30% are missense single nucleotide substitutions, which cluster in ZF2 and the adjacent C-terminal domains, unlike somatic missense mutations, which are predominantly found in ZF1 (Vinh *et al.*, 2010; Hahn *et al.*, 2011; Hsu *et al.*, 2011). Thirdly, recurrent noncoding variants are found within intron 4 (NM_032638.4), affecting the -9.5kb regulatory site containing E-box and GATA and ETS elements and are detected in up to 10% of patients (Johnson *et al.*, 2012; Hsu *et al.*, 2013). A few cases of *GATA2* deficiency syndrome result from large, whole gene deletions, leading to complete hemizygosity and absent transcription of one allele. These cases could also involve deletion of contiguous genes at 3q21.3, and are associated with wider spectrum of features including developmental delay and mental impairment.

All heterozygous germline mutations are predicted to reduce or nullify *GATA2* transcriptional activity, resulting in *GATA2* deficiency from haploinsufficiency. This creates an imbalance in the concentration-dependent equilibrium of *GATA2*, relative to other transcription factors, and leads to disordered expression of downstream target genes (Collin *et al.*, 2015). Maintenance of HSC is acutely sensitive to the levels of *GATA2* and haploinsufficiency leads to aberrant haematopoiesis and progressive depletion of the HSC pool (Crispino & Horwitz, 2017; Wlodarski *et al.*, 2017). The main clinical features of certain subtypes of the syndrome are a distinct pattern of infections, which are the consequences of immunodeficiency brought about by the disturbance in production of specific cell lineages and resulting peripheral cytopenias (Hsu *et al.*, 2011; Dickinson *et al.*, 2011; Dickinson *et al.*, 2014; Spinner *et al.*, 2014; Collin *et al.*, 2015).

Haploinsufficiency is common to all *GATA2* deficiency and so a simple model is unable to explain the heterogeneity of disease types. Similar mutations have been observed amongst families with different clinical syndromes and it has been suggested that there are no clear associations between types of mutation and clinical features (Hyde & Liu, 2011; Kazenwadel *et al.*, 2012; Spinner *et al.*, 2014). The study of genotype–phenotype correlations is confounded by this discrepancy, as well as the variable onset of features and differences in penetrance. However, the molecular pathogenesis is becoming better understood; different studies have remarked that there is phenotype clustering in families with the same mutation (Spinner *et al.*, 2014) and in a study of specific, recurrent germline missense mutations, clinical features were found to be remarkably consistent (Chong *et al.*, 2017). Evidence of the variable effect of different types of *GATA2* mutation influencing phenotypic features is slowly becoming apparent. Differences may be partly explained that levels of *GATA2* activity can be modified by residual mutant *GATA2* expression and the DNA binding affinity of the mutant protein, affecting the dominant-negative transactivation of the wild-type allele and the degree of haploinsufficiency (Cortés-Lavaud *et al.*, 2015; Chong *et al.*, 2017). The importance of autoregulation of expression is demonstrated by the observation of *GATA2* mutations within the autoregulatory binding sites of its own promoter and enhancer and therefore lack the regulatory element responsible for maintaining the self-activation, but otherwise have intact coding sequences (Johnson *et al.*, 2012; Hsu *et al.*, 2013). It has been proposed that there are two classes of germline *GATA2* mutation with 'reduced' or 'greatly reduced' DNA binding affinity (Chong *et al.*, 2017). Furthermore, the differential binding of cofactors to *GATA2* in haematopoietic differentiation, such as PU.1 (Chong *et al.*, 2017) and impairment of interactions in a complex network of transcription factors, have the potential to affect downstream expression and changes in cellular responses, which could be critical to the development of disease and influence the disease phenotype. It appears that the mode of activity of *GATA2* mutation is complex involving interplay with multiple factors, making geneotype-phenotype correlations difficult to interpret. However, the subtly different effects of specific mutation will eventually be revealed with growing evidence from new findings.

The *GATA2* gene is also critical for the development of the lymphatic vasculature (Kazenwadel *et al.*, 2012) and when defective, can promote primary lymphedema, a key characteristic of Emberger syndrome (Collin *et al.*, 2015). The majority of these patients tend toward an earlier age of onset; have frameshift or nonsense mutations which are predicted to eliminate *GATA2* activity. Complete or nearly complete haploinsufficiency may be more disruptive to lymphatic development than missense mutations (Kazenwadel *et al.*, 2012; Dickinson *et al.*, 2014; Chong *et al.*, 2017). Mutations, such as amino acid missense substitution, that have more likely to have reduced transcriptional activity but with residual function exclude Emberger and therefore might confer a resistance to lymphedema.

GATA2 deficiency promotes clonal haematopoiesis, which is likely to contribute to pathogenesis in these disorders (Dickinson *et al.*, 2014; Collin *et al.*, 2015). However, it is unclear how germline loss-of-function mutations result in myeloid neoplasms (Wlodarski *et al.*, 2017). *GATA2* deficiency imparts a high risk of a familial myeloid neoplasia (Hahn *et al.*, 2011; Bodor *et al.*, 2014; Holme *et al.*, 2012; Wlodarski *et al.*, 2016); overall prevalence is estimated to be approximately 75% amongst mutation carriers with a median age of onset of approximately 20 years, but with a wide range (<1~78 years) (Wlodarski *et al.*, 2017). The risk of developing MDS or AML appears to be universal and not to be confined to distinct germline *GATA2* mutations (Collin *et al.*, 2015), however, an exception may be that missense ZF2 mutations are more marginally associated with the familial MDS/AML phenotype (Hahn *et al.*, 2015) or that the frequency of accessory features is reduced. An earlier age of onset has been reported in patients with antecedent accessory features (Pasquet *et al.*, 2013). The non-neoplastic complications resulting from mononuclear cytopaenia has been described as an ‘accessory phenotype,’ and a significant proportion of patients present with a primary MDS or AML, apparently lacking other non-neoplastic haematological features, although large pedigrees demonstrating ‘pure’ familial MDS as the sole feature are unusual (Hahn *et al.*, 2011; Bodor *et al.*, 2014; Holme *et al.*, 2012; Wlodarski *et al.*, 2016). However, it is possible that cellular deficiencies could be overlooked if not thoroughly investigated or may not have been recorded in family histories (Collin *et al.*, 2015). Development of MDS is often associated with acquisition cytogenetic abnormalities, especially monosomy 7 in 41% of patients and trisomy 8 in 14% (Wlodarski *et al.*, 2017). Other studies have reported acquired *ASXL1* mutations in about 30% of patients, often coinciding with the progression from hypoplastic MDS to a more proliferative disease and particularly CMML (Micol & Abdel-Wahab, 2014; Bodor *et al.*, 2014; West *et al.*, 2014b; Churpek *et al.*, 2015; Wang *et al.*, 2015). We report a family of three siblings presenting in childhood with MDS or AML at different ages, who were found to have a novel frameshift *GATA2* germline mutation inherited from the unaffected father. There were no apparent accessory features of germline *GATA2* deficiency syndrome and so myeloid neoplasias were the only presenting feature. The clinical and laboratory details of this unusual family are presented and the implications of the genetic changes on disease phenotype and clinical course are discussed

Case studies and Family history

The findings and routine clinical investigations of three siblings are described.

Clinical description

The family comprised three children, 1 boy, two girls (see Figure 1).

Sibling 1 (boy, 11 years and 8 months) investigated for pancytopenia and diagnosed with MDS. Sibling haematopoietic stem cell transplant age using HLA compatible sister aged as stem cell donor (Sibling 3). Developed severe chronic GVHD and died of pulmonary complications.

Sibling 2 (girl, 12 years and 1). Acute presentation with high WCC (49 x 10^9/l). Morphologically and immunologically AML. Commenced treatment according to the AML 10 protocol and had unrelated HSCT with FLU-ATG conditioning with minimal post-transplant complications. FU 12 year post HSCT minimal long term effects.

Sibling 3 (16 years, donor for sibling 1 age 6?) presented with thrombocytopenia. Bone marrow morphology was consistent with MDS. Had unrelated HSCT with Flu ATG without significant complications. Well 4 years post HSCT.

Materials and methods

Germline Sequencing

Germline whole exome sequencing (WES) was performed to investigate the family and first detect the GAT2 germline mutation.

Cytogenetics and microarray of leukaemic cells

Cytogenetic analysis was performed to local adaptions of standard techniques (Czepulkowski, 2001) and reported using standard nomenclature (*ISCN 2016; An International System for Human Cytogenomic Nomenclature* 2016).

SNP Microarray. Normal and leukaemia DNA was tested by SNP microarray to look for evidence of germline and acquired abnormalities, with particular attention to chromosome 7 due to monosomy in the tumour.

Next Generation sequencing of DNA from leukaemic cells from diagnostic bone marrows

DNA was extracted from previously cultured cells surplus to cytogenetic investigations, which had been stored in a 3:1 mixture of methanol and glacial acetic acid at -20°C. Stored cells were available from the time of diagnosis of the myeloid neoplasms in the three siblings. Post-treatment remission samples were available from Sibs 1 and 2 and a pre-symptomatic screening bone marrow, 7 years before diagnosis of MDS, was available from Sib 3, which were used as normal controls.

Sure Select solution phase cRNA biotinylated oligonucleotide baits (Agilent Technologies, Santa Clara, CA, USA) were designed to capture relevant sequences from 45 target genes, including all exons from the set of 30 genes known to be frequently recurrently mutated in acute myeloid leukaemia for the detection of single nucleotide variants. The exonic and intronic positions of typical junctions of *FLT3* internal tandem duplications, *KMT2A* partial tandem duplications and typical breakpoints in genes of recurrent gene fusions in AML were also sequenced (see Table 2).

GATA2 RT-PCR and sequencing

RNA was extracted from viable bone marrow cells stored from Sibling 3. cDNA was made using oligo dT and hexamer primers and *GATA2* was assessed by PCR using forward primer FW4 (GACAAGGACGGCGTCAAGTA) and reverse primer RV1 (CGCCCCTTTCTTGCTCTTCT). Unrelated cells from leukaemic cell line (OCI-AML5) with wild type *GATA2* were used as normal control. The PCR was cleaned up using the Qiagen PCR clean up kit and then Sanger sequenced using either the FW4 or RV1 primers in separate reactions.

Results

GATA2 mutation analysis

Whole exome sequencing revealed a deletion of CTTCTGGCGGC in *GATA2* at exon 5 c.1021_1031delGCCGCCAGAAG p.(Ala341SerfsTer39) (Annotated against Ensembl 68). This was confirmed by Sanger sequencing. The same germline deletion was detected in peripheral blood DNA of all three siblings and their father.

Cytogenetics and microarray analysis of bone marrow aspirate specimens

Conventional cytogenetic analysis of bone marrow aspirate specimens at the time of diagnosis of myeloid neoplasia showed abnormal karyotypes, with monosomy 7 in all three siblings and with trisomy 8 in a separate apparently unrelated cell line in sibling 3 (see Table 1). Using SNP microarray, applying a cut-off of 50 kb, the tumour samples confirmed loss of the whole of chromosome 7. There was a 32kb deletion of chromosome 7 in the normal sample (pos 124,236,837-124,268,655), which was not reported in the Database of Genomic

Variants but there were no known genes in this region (http://dgv.tcag.ca/dgv/app/home). There were no other copy number variants in either the tumour or normal samples, which were not reported in DGV and so there was no indication of tumour predisposition genes. The deletions and duplications called in the tumour and not seen in the normal DNA were difficult to distinguish from background noise.

Next Generation sequencing of DNA from diagnostic bone marrows compared to normal controls

Targeted next generation sequencing of DNA from dysplastic marrow or leukaemia cells from diagnostic bone marrows also showed germline *GATA2* replacement of TCTTCTGGCGGC with T at chr3:128,200,774-128,200,784 (GRCh37/hg19 reference genome), confirming the findings of exome sequencing. A point mutation in exon 7 of WT1 (32417910 G>T), a nonsense mutation introducing a stop codon at S152* was identified in Sibling 2. This was detected in 168 variant reads at an overall read depth of 441 reads (VAF 0.38). Of note, somatic mutations of *ASXL1* were not detected. Germline missense *ASXL1* variants were found in our cohort; L815P as a homozygous variant in each of our siblings and E1102D as a heterozygous change in Sibling 2. Following realignment and further investigation for low frequency variants, two low VAF variants were observed in Sibling 3; homozygous, non-homologous missense variant *U2AF1* (chr21:44513243 G>A, p.S158L/S231L) was detected at VAF 0.052 (14/272 reads) and homozygous, non-homologous missense variant in *IDH1* (chr2:209116179 G>A, p.P33S) at VAF 0.048 (12/252 reads).

RNA expression

Test cDNA from the bone marrow specimen of Sibling 3 (from the forward sequence) showed expression of the mutated allele only; there was no evidence of the wild-type allele. There were two separate aberrant sequences, one corresponding to the expected *GATA2* sequence with the deletion (deletion of GCCGCCAGAAG). A second cDNA sequence with the deletion appeared to represent an aberrant splice form three bases upstream of the deletion, and corresponding with sequence from within intron 6, which suggests that aberrant splicing was occurring. This might be due to the deletion as it is close to an intron/exon boundary and whilst not predicted to alter splicing it may change the binding of exon splice enhancers Testing of the control cell line (OCI-AML5) detected the expected intact *GATA2* sequences.

Discussion

The current study presents a family with a *GATA2* germline mutation, inherited from the asymptomatic father by three offspring and presenting with MDS/AML without other features typical of subtypes of the *GATA2* deficiency syndrome. Myeloid neoplasia is the presenting feature in up to 50% of cases of *GATA2* deficiency (Collin *et al.*, 2015) and the overall prevalence is estimated to be approximately 75%, with a median age of onset of approximately 20 years but with a wide range (<1~78 years) (Wlodarski *et al.*, 2017). The lifetime risk of developing MDS is estimated to be near to 90% (Micol & Abdel-Wahab, 2014). However, the rate differs between families, and the risk is reduced in cohorts selected by a history of recurrent infections (Spinner *et al.*, 2014). Prevalence estimates, therefore, depend on the ascertainment of the proband, compounded by a possible phenotype clustering effect in individuals with the same mutation. The three affected children in the current study all developed myeloid neoplasms of less than the median age of onset, albeit at different ages; 11, 12 and 18 years, two with MDS and one with AML.

GATA2 mutations do not confer poor prognosis in childhood MDS. In GATA2 deficiency syndrome, the high risk of progression from cytpopaenia and MDS to AML informs the decision to proceed to timely haematopoietic stem cell transplantation (HSCT) (Wlodarski *et al.*, 2016) and outcomes are less favourable after development of myeloid neoplasm or after the onset of chronic infection. Recovery of normal phenotype and reversion to full immune reconstitution following transplant can take as long as several years but many patients derive many years free of disease symptoms (Crispino & Horwitz, 2017). Unfortunately, potential related donors may share a familial *GATA2* mutation, rendering them unsuitable and many patients may not be transplantation candidates. Transplantation from sibling carrier in our case 1 which now would probably be reconsidered with current understanding of the syndrome.

The absence of additional haematological features of *GATA2*-deficiency is not uncommon; half of *GATA2* mutation patients ascertained by presentation of a paediatric MDS lacked an accessory phenotype (Wlodarski *et al.*, 2016). However, there is a high rate of *de novo GATA2* mutation and the majority of *GATA2*-related MDS patients present irregularly with no family history of MDS. Therefore, a significant proportion of probands present with no prior features to indicate *GATA2* deficiency. Usually, within *GATA2*-related MDS pedigrees, all affected relatives were diagnosed with MDS/AML as children or young adults. Within paediatric *GATA2*-related MDS, the low number of silent carriers are consistent with supporting high penetrance and expressivity (Wlodarski *et al.*, 2016). The typical appearance of large hereditary MDS/AML kindreds, therefore, is with younger generations with clear DCML deficiency and MDS (Collin *et al.*, 2015) and pedigrees with multiple members with *GATA2*-related MDS without additional accessory phenotypes are uncommon.

The father is a clinically silent carrier of the germline *GATA2* mutation, which is unusual, whereas his offspring apparently have a highly penetrant and expressive disease (Wlodarski *et al.*, 2016). Such phenotypic discordance of MDS has been noted previously within kindreds (Spinner *et al.*, 2014); a father of children with MDS with trisomy 8 and a family history of *GATA2*-related malignancy was clinically asymptomatic (kindred 40) and in a separate family (kindred 4), a father of two children presenting with later onset MDS (at ages 26 and 19) was unaffected until the age of 78 when he developed CMML. Only 8 silent carriers were apparent in more than 200 *GATA2* individuals (Micol & Abdel-Wahab, 2014). Evidence suggests that the father is at risk of myeloid (or other) malignancy and that a series of somatic mutations and clonal expansion are required for overt neoplastic disease to develop; a cooperating ‘second hit’ mutation or chromosomal abnormality is necessary. There is growing evidence from chromosomal studies and mutational profiling in *GATA2* deficiency of the significance of secondary acquired abnormalities.

Neoplastic disease develops from the stepwise acquisition of additional mutations and clonal expansion, which is different from the other features of *GATA2* deficiency and is unlikely to show the same course and timing. It is accepted that additional somatic genetic abnormalities are acquired in addition to the germline defect leading to transformation to overt neoplastic disease. Cytogenetic abnormalities are well-documented in *GATA2*-related MDS and the most common are monosomy 7 (30%), trisomy 8 (25%) and der(1;7) and other abnormalities resulting in deletion of 7q (Wlodarski *et al.*, 2016). Monosomy 7 is found in all three affected individuals in this family and trisomy 8 in one (see Figure 2). The study acquisition of molecular somatic mutations as routes of neoplastic disease progression requires further study. However, *ASXL1* mutations have been identified in up to 30% of *GATA2*-related MDS, as a major route of leukaemia progression. Acquired ASXL1 mutation is strongly associated with the presence of monosomy 7, bone marrow hypercellularity and chronic myelomonocytic leukaemia (West *et al.*, 2014b; Micol & Abdel-Wahab, 2014; Bodor *et al.*, 2014; Lubking *et al.*, 2015; Churpek *et al.*, 2015). However, when corrected for ascertainment bias, large studies suggest that this is an overestimate and it is no larger than the general frequency in paediatric MDS without germline *GATA2* mutations (Wlodarski *et al.*, 2017). Acquired *ASXL1* mutations were not detected in our cohort but germline missense *ASXL1* variants were detected; L815P as a homozygous variant in each of our siblings and E1102D as a heterozygous change in Sibling 2. c.L815P has not been fully characterised and its effect on ASXL1 protein function is not known but it does not appear to lie within any known functional domains of the ASXL1 protein (UniProt.org). *ASXL1* missense variant p.E1102D has been previously reported as a pathological mutation, presumably due to it not appearing as a commonly reported SNP (dbSNP 147) (Szpurka *et al.*, 2010; West *et al.*, 2014b). This variant appears in the germline in our patient which supports its interpretation as a non-pathological polymorphism, as other studies of the gene (Clin Var) (Carbuccia *et al.*, 2010; Perez *et al.*, 2010; Stein *et al.*, 2011).

With the exception of *ASXL1*, there are few other investigations of acquired somatic mutations. Whole genome analysis from a single patient with trisomy 8 found mutations in *EZH2, HECW2* and *GATA1* from that may contribute to MDS/AML (Fujiwara *et al.*, 2014). In a study of three *GATA2* germline families, 4 out of 6 individuals showed mutations in *RUNX1, GATA2, GPRC5A, STAG2, WT1, NRAS, TP53, SETBP1* and *ASXL1* and others at low frequency (Wang *et al.*, 2015). Frameshift mutations in *PDS5B*, a cohesion family gene, was found in a *GATA2* patient and a germline *RUNX1* patient in a study of familial MDS/AML, as well as recurrent mutations in *BCOR, DNMT3A, ASXL1, PTPN11*, and *STAG2* genes (Churpek et al., 2015). After conservative filtering of our data from a targeted panel of common AML-associated genes matched against the paired normal samples by removal of common SNPs and variants unlikely to be pathogenic, only one mutation in WT1 was detected at a significant variant allele frequency (VAF). *HECW2, GATA1* and *PDS5B* genes were not tested in our panel. *WT1* mutations are common in paediatric malignancy and one was detected in a previous patient with *GATA2* mutation (Wang *et al.*, 2015). The *WT1* mutation in Sib II.2 with AML was found in exon 7 (32417910 G>T) 168 variant reads at overall read depth of 441 reads (VAF 0.38), a nonsense mutation introducing a stop codon at S152*. *WT1* exon7 stop mutations represent the vast majority of acquired *WT1* mutations in AML, resulting in a truncated protein with loss-of-DNA binding function by elimination of the zinc finger domain or nonsense-mediated RNA decay and complete loss of expression (Rampal & Figueroa, 2016). Homozygous, non-homologous missense variant *U2AF1* (chr21:44513243 G>A, p.S158L/S231L) was detected at VAF 0.052 (14/272 reads) and homozygous, non-homologous missense variant in *IDH1* (chr2:209116179 G>A, p.P33S) at VAF 0.048 (12/252 reads). These are included to be comparable to previous studies (Wang *et al.*, 2015) and although the quality of the reads was satisfactory, the small number of reads from previously fixed tissue casts doubt on the significance of these findings.

The spectrum of recurrently mutated genes in *GATA2* familial MDS/AML appears to be different from sporadic disease, albeit with considerable overlap in the genes and gene families that serve to enhance the disease progression. The common driver mutations and gene fusions found in sporadic, such as *NPM1, FLT3-ITD, RUNX1-RUNX1T1, CBFB-MYH11*, *KMT2A* fusions and *NUP98* fusions, are notoriously absent from familial disease. Furthermore, it has been reported that the mutational burden of familial disease is less. The relative paucity of mutations in these diseases and lack of additional recognised powerful driver events, may indicate that the most significant and prevalent genetic abnormalities have already been demonstrated; *GATA2* germline mutations, together with monosomy 7 and trisomy 8 and *ASXL1* mutations, particularly in CMML.

The current family was found to have a germline *GATA2* mutation involving an 11bp deletion (CTTCTGGCGGC/GCCGCCAGAAG) at position 128,200,774-128,200,784 (GRCh37/hg19 reference genome) of chromosome 3. This is a frameshift mutation in exon 5 of *GATA2* and is predicted to change the amino acid sequence of the translated protein from amino acid position 341, 71% along the amino acid sequence of the protein

(ENST00000341105), towards the C-terminus (see SIFT analysis Table 2). This has not been reported in the literature and is not on the ClinVar database (accessed 21/09/2017). A heterozygous 4-bp insertion (c.1025GCCG) in *GATA2* gene, at the same genomic position GATA2:c.1021_1024dupGCCG (p.Ala342Glyfs) has been reported in a patient with Dendritic cell, monocyte, B lymphocyte, and natural killer lymphocyte deficiency and later aplastic anaemia. This frameshift was predicted to result in premature termination (Ala342GlyfsTer41), which may explain the different phenotype (Mace *et al.*, 2013). Our family's *GATA2* mutation is novel although frameshifts are generally amongst the most common nonsense mutations causing the syndrome and have been found previously in MDS; e.g. 18 truncating mutations (including frameshifts) were identified in a large cohort with childhood MDS as the presenting feature (Wlodarski *et al.*, 2016). They occurred randomly in advance of C-terminal end of ZF2, almost all of which are unique because they are not derived from simple base pair changes. Amino acid 341, in our case, is towards the end of the first zinc finger domain and this is predicted to replace the subsequent 140 amino acid sequence with a 38 amino acids, including complete disruption of the ZF2. Consistent with the zinc finger domains being critical for effective *GATA2* function, unsurprisingly, the majority of nonsense mutations eliminate the ZF domains and non-synonymous missense substitutions of single amino acids are found within ZF2, affecting protein function. Gene deletions have a minimal region of deletion encompassing zinc fingers (Collin *et al.*, 2015).

It appears that this frameshift mutation, c.1021_1031delGCCGCCAGAAG p.(Ala341SerfsTer39), does not predispose accessory features of *GATA2* deficiency. The children present with MDS/AML without other features and the father is asymptomatic at age ?. However, patients with *GATA2* germline mutation can lack a disease phenotype as cellular deficiency tends to evolve gradually over several decades and could remain subclinical (Dickinson *et al.*, 2014). The age of onset of clinical symptoms was highly variable with 50% of individuals being without any symptoms of the disease by age 20 and peak age for haemato-immunologic disease manifestation is in the second and third decade of life (Spinner *et al.*, 2014). It was reported that frameshift mutations favour an earlier onset age of cellular deficiency, compared with missense mutations but this may depend on the underlying mechanism of *GATA2* deficiency (see below) (Dickinson *et al.*, 2014). The development of MDS and AML, however, appears to be an equal risk with all types of *GATA2* mutation and depend on the chance acquisition of secondary driver genetic abnormalities (Dickinson *et al.*, 2014; Spinner *et al.*, 2014; Collin *et al.*, 2015). These observations are consistent with the theory that the cytopaenia phenotype is progressive whereas that leukaemic onset is multimodal, reflecting different causative factors. Therefore, it is feasible that the earlier the malignancy presents, the less likely non-haematologic accessory features will develop as a presenting feature.

Despite random variation within families and an unpredictable influence of environmental factors, particularly infectious exposure, the accumulation of evidence suggests that the clustering of similar disease within family groups is influenced by the specific underlying genetic constitution. Each specific mutation is likely to have subtly different activity and exert different phenotypic affects (Hahn *et al.*, 2011; Chong *et al.*, 2017). This has been demonstrated previously with differences in their ability to inhibit HSC differentiation and maintain the progenitor phenotype, thereby disrupting the balance in the HSC pool from hyperstimulation, and depletion of HSCs in the cytopaenic subtypes of the syndrome (Tipping *et al.*, 2009; Cortés-Lavaud *et al.*, 2015). The activity of mutant *GATA2* expression, the length and precise structure of the residual *GATA2* mutated protein, retention of the ZF binding domains, and whether nonsense mediated decay (NMD) occurs would all determine the (non- or hypo-) functionality of *GATA2* protein. The findings of Cortés-Lavaud and colleagues (2015) support haploinsufficiency being not just a function of a quantitative reduction of intact *GATA2* but induced by impaired binding of mutant *GATA2* to the promoter 2.4kb upstream of wild type *GATA2*, leading to secondary loss of expression from the intact allele due to reduced *GATA2* occupancy at its own promoter. *GATA2* germline mutations could affect DNA binding other *GATA2* transactivation target genes, lead to putative dysregulation of specific *GATA2* targets and therefore exaggerate the variable degree of functional haploinsufficiency in different families.

Furthermore, nonsense mediated decay (NMD) is a cellular mechanism of mRNA surveillance to detect nonsense mutations and prevent the expression of truncated or erroneous proteins and may result in undetectable transcription of one allele (SIFT). SIFT has predicted from the absence of a premature STOP codon in the last exon or within 50 nucleotides from the end of the of the second to last exon, that NMD does not occur with this mutation (Nagy & Maquat, 1998). This is demonstrated by the detection of transcribed mutant mRNA in bone marrow cells of sibling II.3 in our study (as cDNA by reverse transcriptase-PCR). From our tentative evidence (?), it appears that the active expression of *GATA2* p.(Ala341SerfsTer39) but complete disruption of the DNA binding domain of mutant *GATA2* has resulted in no detectable expression of wild-type *GATA2*. This is surprising but would support the observations of Cortés-Lavaud et al (2015) that GATA2 protein transactivates its own transcription and in the present case, this process is suppressed in a dominant negative fashion by expressed mutant alleles with defects in the zinc finger domains, results in loss of DNA-binding ability and leading to the wild-type *GATA2* deficiency. Previous analyses that found differentially affected genes depending on the mutation (Hahn *et al.*, 2015; Chong *et al.*, 2017) explain in part the clinical heterogeneity amongst patients with GATA2 deficiency. Symptoms could differ from a pure haploinsufficiency phenotype. This effect is postulated to be different for various mutation types which might be subtle for simple single amino acid change missense mutation. This effect would be less subtle for the protein disruption resulting from frameshift mutation, as in our family. Frameshift or null mutations have been linked to a higher risk of lymphedema and severe viral infections and Emberger (Hyde & Liu, 2011; Spinner et al, 2014). However, frameshifts with residual *GATA2* mutant function, without NMD, may behave as missense mutations.

Table 1 – presenting features and acquired genetic abnormalities detected in the kindred

	Age at diagnosis	**Diagnosis**	**Karyotype**	**CGH**	**Targeted NGS**
Sibling 1		MDS	45,XY,-7[10]/46,XY[1]		
Sibling 2		AML	45,XX,-7[10]		WT1 (32417910 G>T),
Sibling 3		MDS	45,XX,-7[7]/47,XX,+8[3]/46,XX[10]		

Figure 1

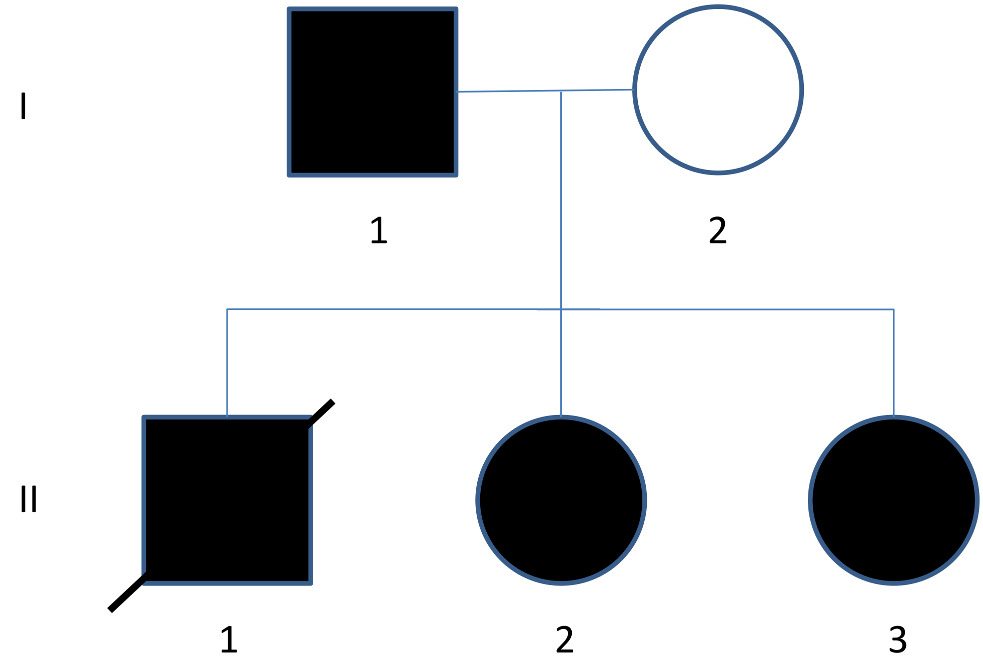


Table 2. Results of SIFT analysis to predict the effect of germline *GATA2* mutation c.1021_1031delGCCGCCAGAAG on protein function (http://sift.jcvi.org/ accessed 02/02/2017)

nput Coordinates	3,128200773,128200784,-1,-/,TCTTCTGGCGGC/------------
Gene ID	ENSG00000179348
Transcript ID	ENST00000341105
Substitution Type	FRAMESHIFT
Region	CDS
Amino acid position change	340-480
ndel Location	71% (towards C-terminus of the protein)
Nucleotide change	CGGCT-cttctggcggc-CGACT
Amino acid change	KRRLSaarragtccancqtt*->KRRLSsrhllcklsddnhhl*
Causes NMD	NO
Transcript visualisation	3'<---3'UTR[e5][*.e4]--[e3]--[e2]--[e1]5'UTR—5'UTR-----|5'
Gene Name	GATA2
Gene Desc	Endothelial transcription factor GATA-2 (GATA-binding protein 2) [Source:UniProtKB/Swiss-Prot;Acc:P23769]
Protein Family ID	ENSFM00500000269911
Protein Family Desc	RECNAME: FULL=GATA BINDING
Transcript Status	KNOWN
Protein Family Size	5
Protein Sequence Change	ENST00000341105 change mismatch for this transcript in red
Original Protein Sequence	>ENST00000341105; MISMATCH = 340-480
	MEVAPEQPRWMAHPAVLNAQHPDSHHPGLAHNYMEPAQLLPPDEVDVFFNHLDSQGNPYYANPAHARARV SYSPAHARLTGGQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHNPWTVSPFSKTPLHPSAAGGPGGPLSVYP GAGGGSGGGSGSSVASLTPTAAHSGSHLFGFPPTPPKEVSPDPSTTGAASPASSSAGGSAARGEDKDGVKYQVS LTESMKMESGSPLRPGLATMGTQPATHHPIPTYPSYVPAAAHDYSSGLFHPGGFLGGPASSFTPKQRSKARSCSE GRECVNCGATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSaarragtccancqtttttlwrrnangdpvc nacglyyklhnvnrpltmkkegiqtrnrkmsnkskkskkgaecfeelskcmqeksspfsaaalaghmapvghlppfshsghilptptpih

	pssslsfghphpssmvtamg
Modified Protein Sequence	>ENST00000341105; MISMATCH = 340-378
	MEVAPEQPRWMAHPAVLNAQHPDSHHPGLAHNYMEPAQLLPPDEVDVFFNHLDSQGNPYYANPAHARARVSYSPAHARLTGGQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHNPWTVSPFSKTPLHPSAAGGPGGPLSVYPGAGGGSGGGSGSSVASLTPTAAHSGSHLFGFPPTPPKEVSPDPSTTGAASPASSSAGGSAARGEDKDGVKYQVSLTESMKMESGSPLRPGLATMGTQPATHHPIPTYPSYVPAAAHDYSSGLFHPGGFLGGPASSFTPKQRSKARSCSEGRECVNCGATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSsrhllcklsddnhhlmapkrqrgpclqrlwpllqaaqc

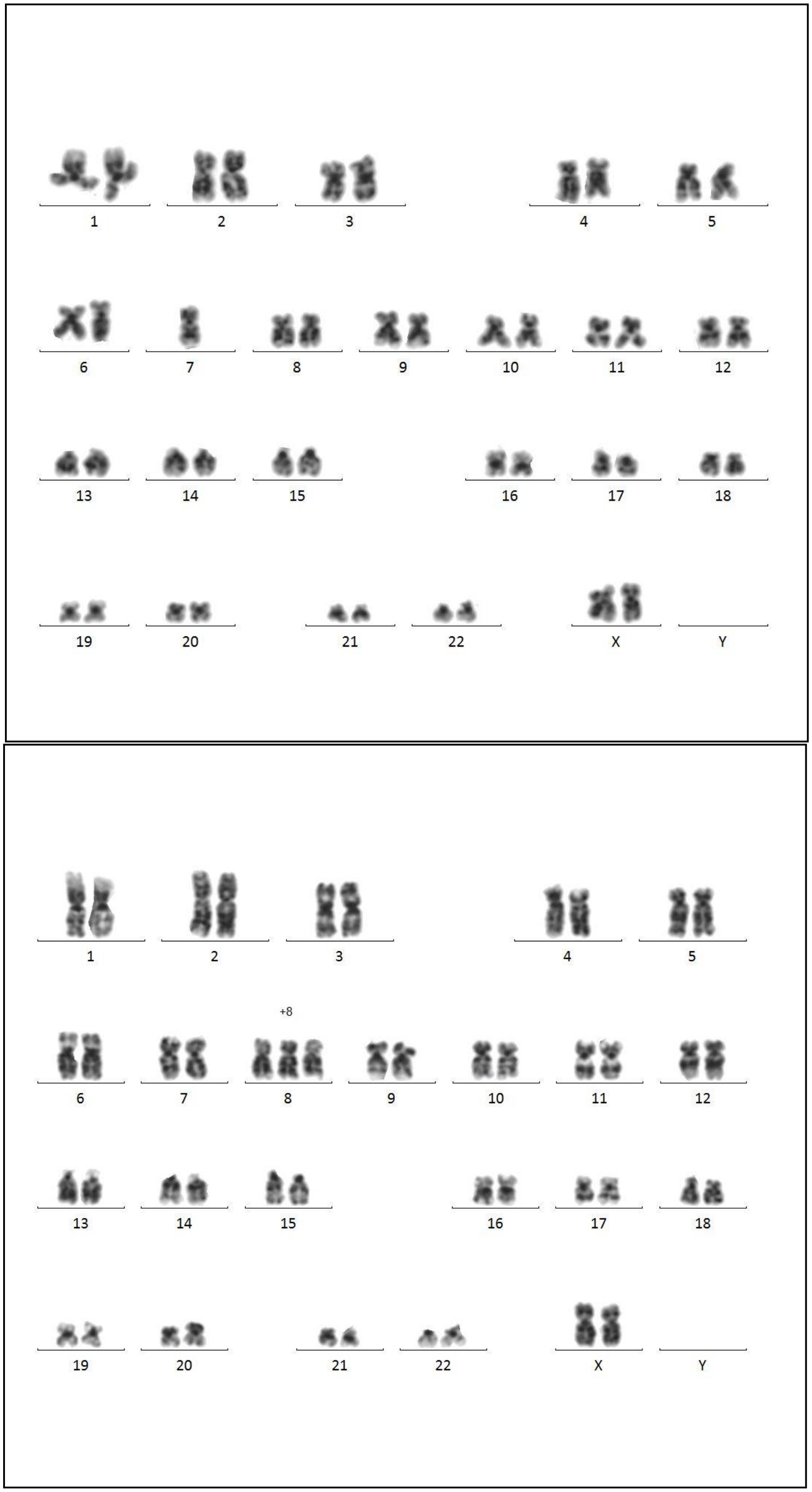
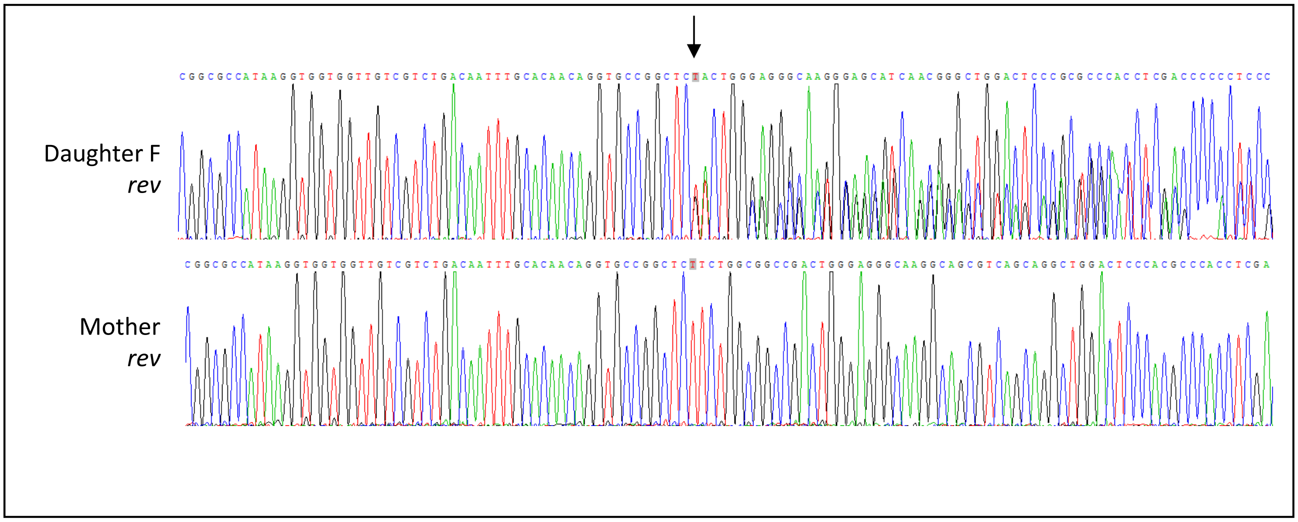


Figure 2; Karyotypes from sibling 2.III showing 1. monosomy 7 and 2. trisomy 8



Daughter (F). The same germline sequence is present in the other two siblings

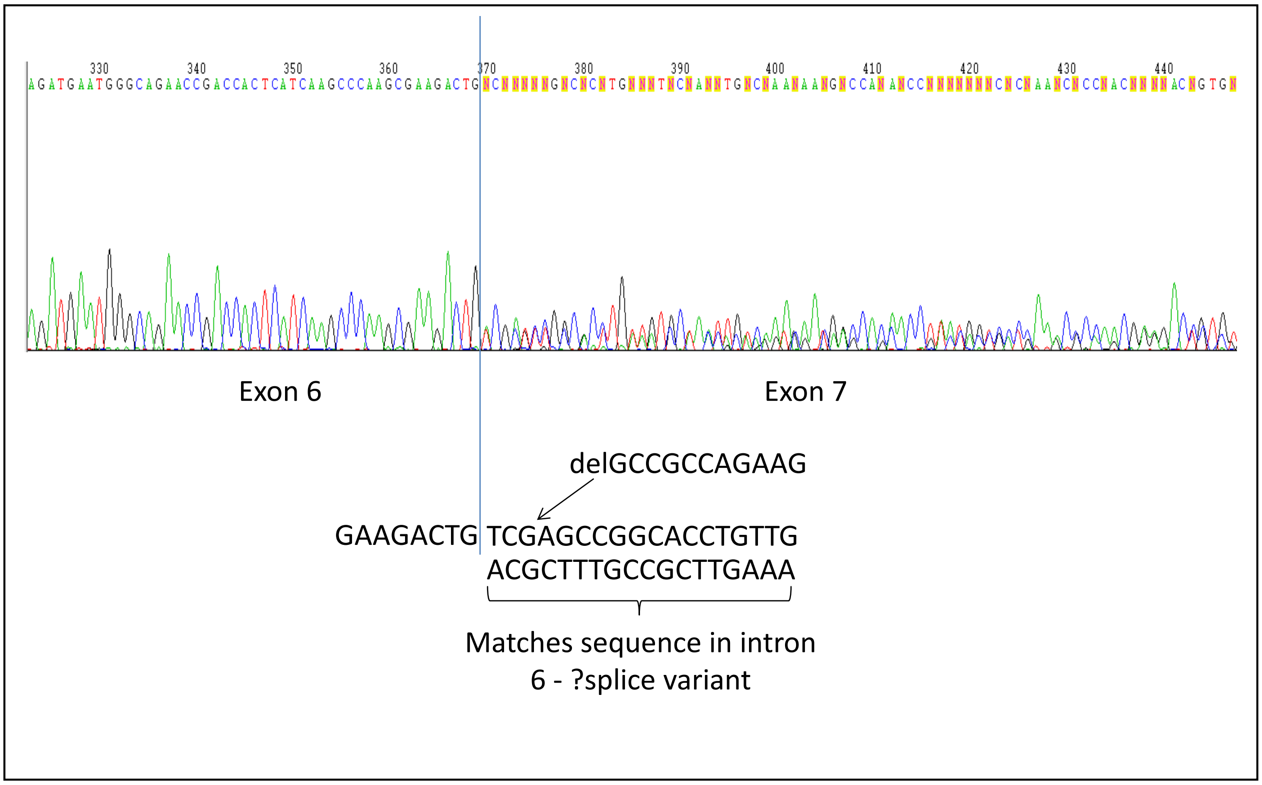


Figure 3; Sequence of cDNA from expressed mutant allele from leukaemia cell line of sib 2.III

Table 3. QC statistics of NGS run for acquired mutations in bone marrow cells from MDS/AML samples

Sib	Reads in BAM passing quality filters	Reads discarded	% duplicate reads	Reads in Covered regions	% in covered regions	Average read length	Average read depth	Bases with ≥50 reads	Bases with ≥100 reads	Regions with ≥50 reads	Regions with ≥100 reads
1	3813753	22	2.61	1657325	43.46	142	400	95.44	91.1	96.41	93.35
2	3127903	16	2.28	1353498	43.27	143	328	94.97	89.79	96.26	93.53
3	3285266	20	2.52	1452149	44.2	142	349	95.09	90.48	96.41	93.82