The Genomics of Acute Myeloid Leukaemia

An Investigation into the Molecular Pathogenesis of Acute Myeloid Leukaemia with t(8;21)

A Submission for the Degree of Doctorate of Philosophy

Deepak Mannari

Department of Medical Oncology
Institute of Cancer
Barts and The London School of Medicine
University of London

Dedicated to Mum & Dad for all their help and sacrifice for my education. Thank you!

Table of Contents

Contributions

The work presented here, unless stated, is the sole work of the author, Deepak Mannari and submitted from Queen Mary University, University of London.

Tracy Chaplin and Gael Molloy performed some of the exon arrays.

Jenny Dunne helped design the primers used in this project, some of the Western Blots and many of the ChIP experiments. She also supervised and introduced to me most of the techniques described in this project.

Duncan Gascoyne performed the retroviral transduction/transformation assay and performed the initial methylcellulose re-plating assay.

Probhir Chakraverty performed the GeneGO analysis.

Finlay Macdougall provided analysis and the Kaplain-Meier curves of the clinical data from Barts.

Jacques Merzeck and Sabha Khalid performed Bowtie analysis of the sequencing data.

Acknowledgements

I would like to thank all the patients at Barts who have contributed their samples to research.

I also thank all the staff, past and present, in the Department of Medical Oncology and in particular the Molecular Genomics Group who were always on hand to provide advice and support. I also like to thank the department of Genomics for their help in the sequencing project.

Specifically, I like to thank:

My supervisors Bryan Young and Andrew Lister
Tracy Chaplin for all my early supervision in the laboratory and introducing me to genomic techniques.
Sabha Khalid and Jacques Merzeck– the bioinformatics team for helping with the exon arrays and sequencing project.
Probhir Chakraverty for help with functional annotation analysis.
Duncan Gascoyne who gave so much of his time in performing and teaching me various techniques.
Jude Fitzgibbon and Helen Hurst for reading and providing constructive comments on my upgrade and thesis drafts.
Importantly, most of this entire project would have not remotely materialised without Jenny Dunne. I am so grateful to her for the time she gave up to supervise, educate and support me and whom I am now fortunate to consider as a friend.

Finally, a big thanks to my wife, Lisa for all her support, understanding and encouragement during these last three plus years. It is much appreciated even though it does not appear so at times!

Abstract

Acute myeloid leukaemia is a clonal disorder characterised by recurrent chromosomal translocations. One of the commonest, is the t(8;21) which results in part of the *AML1* gene being juxtaposed to most of the *ETO* gene with the resultant chimeric protein, AML1-ETO, acting predominantly as a transcriptional repressor. Despite the extensive literature available, the exact mechanism by which the chimeric protein results in AML has not been fully elucidated. By using exon arrays and high throughput sequencing as tools it was hoped to gain further insights into the molecular basis of this disease.

Gene expression profiling using the exon arrays highlighted molecular pathways and specific genes that play a key role in the pathogenesis in t(8;21). Exon arrays were also used to profile individual exon expression of the *ETO* gene. This demonstrated that the genomic breakpoint of *ETO* in the t(8;21) is variable between different patients. This technique also resulted in the discovery of a new exon in the *ETO* gene. This novel exon results in formation of alternative transcripts of AML1-ETO and was shown in mouse models to play a key role in leukaemogenesis. Chromatin immunoprecipitation followed by high throughput sequencing revealed novel aspects of AML1-ETO binding. A number of novel genes that bind AML1-ETO were recognized and in conjunction with the expression data, a number of hypothesis on how AML1-ETO binding effects gene expression are made.

Chapter 1

Introduction

Acute Myeloid Leukaemia

The first published description of leukaemia in medical literature was in 1827 when the French surgeon Velpeau described a febrile illness in a 63-year-old man with hepatosplenomegaly reporting, “A florist and seller of lemonade who had abandoned himself to the abuse of spirituous liquor and of women, without, however, becoming syphilitic….”. At autopsy the patient’s blood was noted to have a consistency “like gruel” and he postulated that this might be due to an excess of white blood corpuscles.

The accolade of the first person to describe leukaemia has been controversial. In his paper “two cases of disease and enlargement of the spleen in which death takes place from the presence of purulent matter in the blood” published in 1845 in the Edinburgh Medical and Surgical Journal, John Hughes Bennett termed the phrase leucocythaemia to describe this observation. However, four weeks after this publication Rudolph Virchow the German pathologist published a similar case but stated that the excess of white cells was not purulent matter as stated by Bennett, but originated from the blood. He also later termed the word leukaemia from the Greek “white blood” in his 1856 publication, which was based on his pioneering work with the light microscope noting the excess of white cells. The controversy remains although in 1995 the Leukaemia Research Fund commemorated Bennett’s work as the first to describe leukaemia.

In fact, in 1844 a full year before the publications of these two protagonists, Donne reported in his book *“Complimentary course on microscopy for medical studies”* a series of cases and his theories. *“Several cases exist with a great excess of white blood cells (...) Blood of such patients contains so many white blood cells that at first glance I thought they contained purulent matter. In fact, I believe that the excess of white blood cells is due to an arrest of maturation of blood. From my theory on the origin of blood cells, the overabundance of white blood cells should be the result of an arrest of development of intermediate cells.”* Although he gave no name for this description it was nevertheless a remarkable and often overlooked observation (Degos, 2001).

Over the course of the century the term acute myeloid leukaemia (AML) came into use, as through the work of investigators such as Ehrlich, Ebstein, Neumann and Naegeli, the underlying pathology was unravelled allowing the leukaemias to be differentiated. Today, work to further understand this heterogeneous disease at the cytogenetic and molecular level continues.

Pathogenesis & Clinical Features

AML is a malignant heterogeneous clonal disorder that is invariably fatal without treatment. It is characterised by an increase in myeloid precursor cells in the bone marrow and an arrest in their normal development. This results over time from an accumulation of acquired mutations in the early haemopoietic progenitor cells. The initial mutation results in a differential block of the early precursors and subsequent mutations result in proliferation of this clone. The underlying causes are unknown but carcinogens such as benzene found in chemicals as well as ionising radiation have been implicated in causing these mutations and giving rise to occasional cases of leukaemia. There is also a genetic predisposition with increase incidence of leukaemia in family members and an association with certain syndromes as well as a predisposition to a particular form of leukaemia seen in Down's syndrome. However, these cases only account for a minority and the cause for the majority of de-novo AML remains obscure. AML can also arise from a background of underlying pre-malignant haematological conditions such as myeloproliferative disease (MPD) or myelodysplasia (MDS). Together with AML arising from a background of previous chemotherapy exposure, which is usually associated with specific chromosomal abnormalities, they are termed secondary AML. These types of AML will not be the subject of further discussion here, which will focus only on primary AML.

The overall incidence of AML in the population is 3.4 per 100000 (Milligan et al., 2006). It occurs in all age groups but the incidence, as expected from its pathogenesis, is more common in the elderly with over two-thirds of cases occurring in those aged over sixty. It forms only a minor fraction of childhood leukaemia. There is variability in clinical presentation, response to treatment and overall prognosis. It frequently results in bone marrow failure with associated symptoms of bleeding, infection and those related to anaemia. Curative treatment of patients is based on intensive chemotherapy with an initial course of combination chemotherapy followed by further courses of consolidation chemotherapy or a stem cell transplant procedure in an appropriately selected subset of patients. The prognosis of AML, despite the recent improvements in treatment, has a 5-year survival of approximately 40%. In many patients, particularly the elderly, curative aim is not appropriate and prognosis is more guarded with life expectancy less than 1 year (Latagliata et al., 2006). It is hoped that newer treatments, targeting specific underlying molecular mechanisms causing leukaemia, would lead to improvements in these survival figures. The advances seen in the treatment of acute promyelocytic leukaemia (APML) and chronic myelocytic leukaemia (CML) targeting the specific mutations remain a paradigm for other disease subtypes.

Classification of AML for over 20 years was based on the French-British-American (FAB) system. This is a morphology-based system describing the differentiation of the myeloid blasts and results in eight categories ranging from M0 to M7 (Bennett et al., 1976) (Table 1). More recently, this has been superseded by the WHO classification, which incorporates and highlights the importance of cytogenetic abnormalities in addition to morphology (Vardiman et al., 2009) (Table 2).

Table 1 Summary of FAB Classification

FAB subtype	Name
M0	Undifferentiated acute myeloblastic leukaemia
M1	Acute myeloblastic leukaemia with minimal maturation
M2	Acute myeloblastic leukaemia with maturation
M3	Acute promyelocytic leukaemia
M4	Acute myelomonocytic leukaemia
M5	Acute monocytic leukaemia
M6	Acute erythroid leukaemia
M7	Acute megakaryocytic leukaemia

Table 2 WHO classification of AML and related neoplasms

AML WITH RECURRENT GENETIC ABNORMALITIES
• AML with t(8;21)(q22;q22); RUNX1-RUNX1T1
• AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); CBEB-MYH11
• Acute promyelocytic leukemia (APL) with t(15;17)(q22;q12); PML-RARA
• AML with t(9;11)(p22;q23); MLLT3-MLL
• AML with t(6;9)(p23;q34); DEK-NUP214
• AML with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); RPN1-EVI1
• AML (megakaryoblastic) with t(1;22)(p13;q13); RBM15-MKL1
• Provisional entity: AML with mutated NPM1

• Provisional entity: AML with mutated CEBPA
AML with myelodysplasia-related change
Therapy-related myeloid neoplasms
AML, not otherwise specified:
• Undifferentiated AML (M0)
• AML with minimal differentiation (M1)
• AML without maturation (M2)
• AML with maturation (M2)
• Acute myelomonocytic leukemia (M3)
• Acute monoblastic/monocytic leukemia (M4)
• Acute erythroid leukemia (M5)
• Pure erythroid leukemia (M6)
• Erythroleukemia, erythroid/myeloid (M6)
• Acute megakaryoblastic leukemia (M7)
• Acute basophilic leukemia
Acute panmyelosis with myelofibrosis
Myeloid sarcoma
Myeloid proliferations related to Down syndrome:
• Transient abnormal myelopoiesis
• Myeloid leukemia associated with Down syndrome
Blastic plasmacytoid dendritic cell neoplasm

Indeed, AML is associated with repeated frequent chromosomal abnormalities and a series of non-random translocations. These abnormalities are associated with different outcomes and three distinctive subsets have been recognised; good risk comprises the translocations t(15;17)(q22;q21), t(8;21)(q22;q22) and inversion 16(p13q22); poor risk comprises complex karyotype (>4 abnormalities) and certain specific abnormalities some of which include abn(3q), -5, -7, add(7q)/del(7q), t(6;11)(q27;q23), t(11q23) abnormalities, t(9;22)(q34;q11), -17, and abn(17p) ; standard risk comprises chromosome changes not part of the other groups and the normal karyotype (for full details see Grimwade et al., 2010). Despite this helpful and prognostically important categorisation almost half the patients have no identifiable cytogenetic change. In a bid to further stratify these patients and eventually have individualised risk stratification there has been a drive to identify further prognostic factors at the molecular level.

A mutli-step model for leukaemogenesis has been proposed and is based on collaboration of two types of mutations (Gilliland, 2002). Class I mutations activate signal transduction pathways conferring a proliferative advantage, often through activating point mutations in tyrosine kinases (such as *FLT3* and *KIT)*. Class II mutations frequently target transcription factors impairing differentiation and subsequently apoptosis (such as mutations of *AML1* and *MLL*). A whole host of Class I and Class II molecular prognostic factors have been identified and validated in clinical trials in large cohorts of patients (Tables 3 & 4). It is hoped that array profiling and specific expression signature patterns of individual patients may be of prognostic value in trying to unify and evaluate the prognostic implications of these newer molecular discoveries.

Table 3 Summary of Class I and II molecular prognostic factors (Takahashi, 2011).

CLASS I	CLASS II
Flt3 ITD	*AML1 (RUNX1)*
Flt3 TKD	*C/EBPα*
CBL	*MLL* re-arrangements
NRas	*PML/RARα*
KIT	*AML1-ETO*
TET2	*CBFβ/MYH11*
ASXL1	*WT1*
IDH 1/2	*NPM1*
	Dnmt3a

Table 4 Summary of the prevalence and prognostic implications of prognostic factors.

GENE	Prevalance	Prognostic Implications	Reference
Flt3-iTD	27%	Unfavourable	(Kottaridis et al., 2001)
Flt3-TKD	11%	Not significant	(Mead et al., 2007)
CBL	<2%	Unknown	(Bacher et al., 2010)
NRAS	10%	Not significant	(Bacher et al., 2006)
KIT	>20% (in CBF leukaemias)	Unfavourable	(Paschka et al., 2006)
RUNX1	6%	Unfavourable	(Gaidzik et al., 2011)
IDH1	7%	Unfavourable	(Schnittger et al., 2010)
IDH2	10%	Favourable (R140 mutation) Unfavourable (R172 mutation)	(Green et al., 2011)
Dnmt3a	22%	Unfavourable	(Ley et al., 2010) (Markova et al., 2011)
TET2	13%	Poorer prognosis	(Chou et al., 2011)
NPM1	35%	Favourable	(Falini et al., 2005)
ASXL1	16% (in>60 yrs)	Favourable	(Metzeler et al., 2011)
WT1	10%	Unfavourable	(Virappane et al., 2008)
CEBPα	12%	Favourable	(Taskesen et al.)

t(8;21) & CBF Leukaemia

The t(8;21)(q22;q22) is one of the commonest translocations in AML and was initially identified in 1973 (Rowley, 1973) although the molecular characterisation was only revealed in the early 1990's by several groups. The t(8;21) is seen in approximately 8-15% of AML cases and classically is associated with the FAB M2 phenotype (AML with granulocytic differentiation) and reported in up to 40% of these cases (Look, 1997). However, it is also observed in 6% of AML M1 (AML without maturation) cases and has also been described in AML M0 (AML undifferentiated), M4 (AML with myelomonocytic differentiation) and M5 (AML with monocytic differentiation), some myeloproliferative diseases and myelodysplasia (MDS) ("GroupeFrancais", 1990, Kojima et al., 1998). This translocation is also common in cases of extramedullary AML (granulocytic leukaemia).

As they share many clinical and molecular similarities, AML associated with t(8;21) and inv(16) are often grouped together and termed as core binding factor (CBF) leukaemias. CBF is a family of heterodimeric transcription factors containing one of three CBFalpha and a CBFbeta subunit. The *CBFA2* gene is *AML1* or *RUNX1* located on chromosome 21 and involved in the t(8;21) (Wang et al., 1993). The *CBFB* gene is involved in the inv(16) mutation (Liu et al., 1993). Either abnormality results in the formation of an abnormal CBF complex leading to disturbed haematopoiesis.

In the t(8;21) AML1's partner is the *ETO* gene on chromosome 8 (Erickson et al., 1992). The AML1 protein has a DNA binding domain and acts as a master regulator of haematopoiesis. The ETO protein contains repressor domains. The fusion product AML1-ETO acts as a transcriptional repressor and affects the normal transcriptional activity of AML1 in a dominant negative manner (Licht, 2001). However, despite being an early and critical mutation the resultant AML1-ETO fusion protein is not sufficient to cause leukaemia on its own, suggesting co-operating events are necessary. This is in keeping with the multistep model of leukaemogenesis where AML1-ETO provides the type 2 mutation impairing haemopoietic differentiation whilst a second acquired mutation, such as in a tyrosine kinase gene provides the type 1 defect conferring a proliferative advantage.

Clinical Aspects of CBF Leukaemia

CBF leukaemia is associated with a relatively good prognostic disease showing increased sensitivity to the chemotherapy agent cytarabine. The complete remission (CR) rates are approximately 90% for these leukaemias after standard induction treatment (Marcucci et al., 2005, Schlenk et al., 2008). Furthermore, use of high dose cytarabine as post induction treatment has improved the outcome in these patients (Bloomfield et al., 1998) with sequential doses of cytarabine being superior to a single high dose of cytarabine. (Byrd et al., 1999).

However, many clinical and pathological differences have been shown to exist between the two groups suggesting that these two types of leukaemia should be regarded as separate clinical entities (Table 5) (Appelbaum et al., 2006).

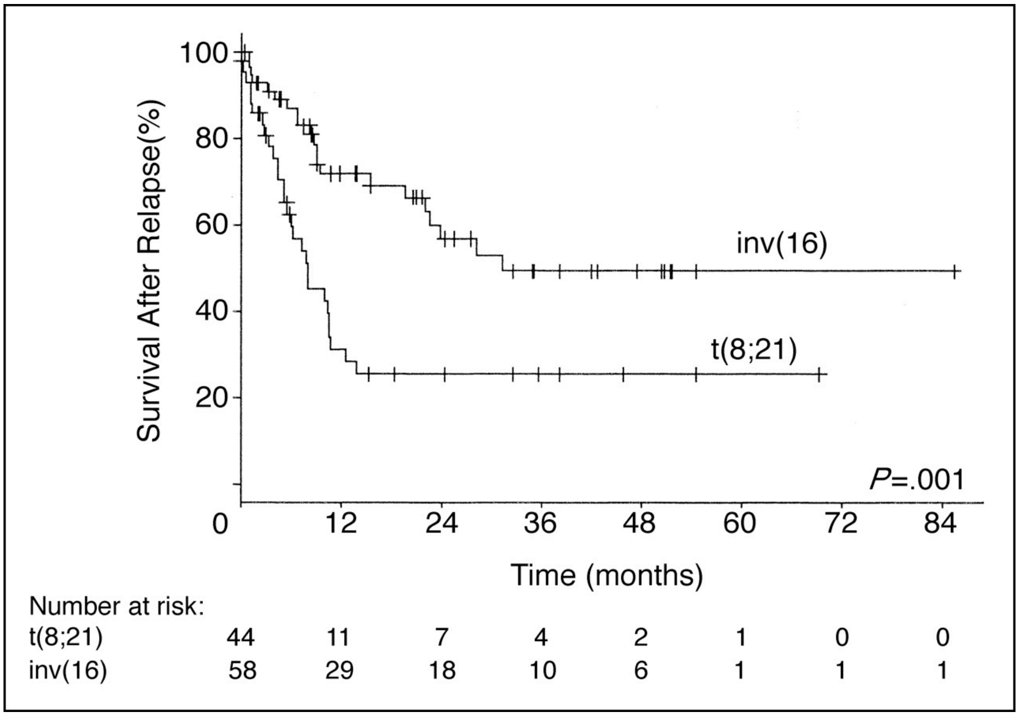
Table 5 Differences between t(8;21) and inv(16)

Cytogenetic Group	t(8;21)	Inv(16)
Race	More in blacks	More in whites
White Cell Count	Lower	Higher
FAB	M2	M4eo
Other chromosomal abnormalities	More frequent, X/Y losses, del (9q)	Less frequent Tri+22, +8, +21
Response post relapse	Poorer	Good
Gene expression profile	Different	
Survival	Shorter O/S	Longer O/S

Clinically, patients with t(8;21) have significantly shorter overall survival than inv(16) patients (Marcucci et al., 2005, Schlenk et al., 2004, Appelbaum et al., 2006) although this outcome has not been supported by recent data from the MRC (Grimwade et al., 2010, Harrison et al., 2010). In work by Marucci, 139 t(8;21) patients and 164 inv(16) patients received cytarabine based induction and consolidation therapy with 6 years follow-up. After adjusting for covariates, patients with t(8;21) compared to inv (16) had an overall survival of 4.4 years compared to 7.1 years and survival after first relapse of 14% compared to 34%. Similar findings were reported by Schlenk in a study comparing 191 and 201 patients with t(8;21) and inv(16) over 36 months. The results showed no difference in CR rates but poorer overall survival for the former. Whilst the relapse rates were similar the t(8;21) cohort were more difficult to salvage with only 33% entering a CR2 compared to 78%. (Figure 1).

The reason for the inferior response to salvage therapy has not been elicited but this resistance to salvage treatment is further supported by Japanese transplant data. In this retrospective study patients with t(8;21) had poorer outcomes compared to patients with inv(16) when having transplantation procedures beyond first CR, although there was no difference when compared in CR1 (Kuwatsuka et al., 2009). Generally, the use of transplantation in CBF leukaemias in CR1 is not recommended although this is not universally agreed. The Medical Research Council (MRC) showed minimal survival differences between transplantation and chemotherapy for the good risk cohort of patients (Grimwade et al., 1998). This is in contrast to the

South-West Oncology Group (SWOG) results suggesting higher survival rates with transplant procedures (Slovak et al., 2000). More recently the German AML group published data suggesting that transplantation is worse than chemotherapy for t(8;21), at least for those with no other risk factors (Schlenk et al., 2008).



Number at risk:								
t(8;21)	44	11	7	4	2	1	0	0
inv(16)	58	29	18	10	6	1	1	1

Figure 1 Survival for relapsed CBF Leukaemias (Schlenk et al., 2004)

In summary, these clinical findings suggest that patients with t(8;21) behave differently to patients with inv(16). Although t(8;21) patients have a good response to induction chemotherapy, a subset of patients will relapse and be resistant to further chemotherapy or transplant procedures. The ability to define this cohort of patients is required to firstly alter their initial management strategy and secondly to attempt to detect the specific defects which may enable development of targeted treatments.

AM1

AML1 is also known as CBFA2, PEBP2aB and RUNX1 and acts as a transcription factor (Table 6). The slightly confusing nomenclature derives from its discovery by several independent groups. It was originally purified and characterised due to its ability to bind with a sequence motif located within enhancer core sites of the murine polyomavirus. This DNA binding factor was called polyoma enhancer binding protein 2 (PEBP2) and found to consist of two sets of polypeptide subunits, PEBP2α and PEBP2β (Kamachi et al., 1990). Independently, several polypeptides that bound the core enhancer site in another murine virus, namely the Moloney leukaemia virus were also identified and referred to as core binding factors (CBFs) (Wang and Speck, 1992). Molecular cloning of these components came with identification of the fusion gene in the t(8;21). The gene was named *AML1* (Miyoshi et al., 1991) with its partner gene *ETO* (Erickson et al., 1992) or *MTG8* (Miyoshi et al., 1993). Subsequently, a series of reports showed that these genes were the human homologues of *PEBP2/CBF* genes. The AML1 gene product was shown to be the human analogue of the PEBP2α subunit (Bae et al., 1993) and the CBFα subunit (Wang et al., 1993). Similarly, the non-DNA binding subunit of PEBP2 was also cloned and it was shown that PEBP2β and CFBβ were identical (Ogawa et al., 1993, Wang et al., 1993). These proteins were shown to have a high degree of homology to the *drosophila melanogaster* gene, a transcriptional regulator known as *runt* (Kania et al., 1990). The HUGO nomenclature committee redesignated *AML1* to be officially called *RUNX1* based on this homology. Other members of this small family of transcription factors found both in human and mouse include *RUNX2* and *RUNX3*. *RUNX2* is located on human chromosome 6p21 and appears to be a key regulator of osteogenesis in the developing embryo. Mutation of this gene results in a human autosomal disorder cleidocranial dystosis, which results in multiple bony abnormalities. *RUNX3*, found on human chromosome 1p36, appears to be important for immunoglobulin class switching although no disease has been associated with this gene (Licht, 2001).

Table 6: Nomenclature used for AML genes and gene products

HUGO	Older Name	Polyomavius	Retroviral
RUNX1	AML1	PEBP2αB	CBFα2
RUNX3	AML2	PEBP2αC	CBFα3
RUNX2	AML3	PEBP2αA	CBFα1
CBFB	CBFB	PEBP2β	CBFβ
CBFA2T1 or RUNX1T1	ETO or MTG8		
RUNX1/RUNX1T1	AML1-ETO		

Structure and Function

The *AML1* gene is located at 21q22.3, spans 260Kb and contains 12 exons of which 8 are coding. There is a complex pattern of transcription and translation as a result of distinct 5'UTRs, differential use of 3' polyadenylation sites and differential exon usage. This results in a range of different alternative splice AML1 mRNAs generating a repertoire of proteins ranging from 20-52kDa (Levanon et al., 2001). The *AML1* gene has two promoter regions, the distal P1 and the proximal P2, which yield different sized primary transcripts containing different length 5' UTRs. The different roles of P1 and P2 allow subtle regulation of the AML1 protein. P1 utilises cap-dependent translation increasing the efficiency of translation while the P2 transcript allows tighter control as its translation is tightly controlled through IRES. In the mouse model, recent work has shown that there is sequential activation of P2 promoter followed by the P1 promoter. The P2 transcript appears essential for the generation of haemopoietic cells from haemogenic endothelium while the P1 transcript appears to mark the onset of definitive haematopoiesis but is dispensable in haemopoietic maintenance (Sroczynska et al., 2009).

AML1 is expressed in all tissues except the heart and brain (Miyoshi et al., 1995). It is highly expressed in lymphoid tissue as well as B and T lymphoid lines. It is expressed in various haemopoietic cell lines including HL60, K562 and U937 and foetal liver cells (Licht, 2001).

The AML1 protein contains various domains including transactivation and inhibitor domains, a nuclear matrix attachment signal (NMTS) and importantly for its role as a transcription factor the runt homology DNA binding domain (RHD). (Figure 2)

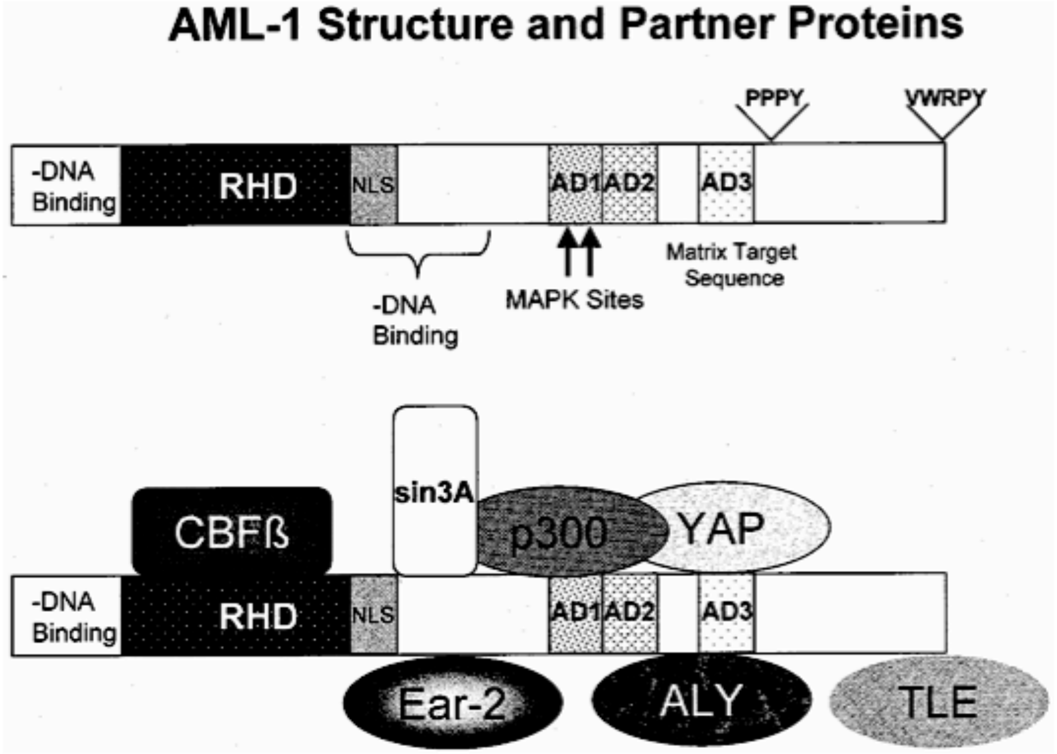


Figure 2 Structure of AML1 and binding proteins. (Licht, 2001) The activation domains can interact with activators p300/CBP, YAP and ALY. Sites for co-repressors Ear2 and Sin3 are noted. At the extreme C-terminal end the conserved VWRPY sequence (recognition motif of the Groucho, or its human homologue TLE, family) is shown and also acts as a co-repressor.

The 128 amino acid DNA binding domain is located at the N-terminal end of the AML1 protein and shows over 90% homology to two drosophila transcriptional regulators: runt involved in segmentation, sex determination and neurogenesis and lozenge involved in eye development and haemopoiesis (Lutterbach and Hiebert, 2000). This domain is responsible for both DNA binding and heterodimerization with CBFβ. In drosophila this region binds the DNA sequence PuACCPuCA (Kamachi et al., 1990) while the human homologue AML1 protein binds the sequence TG(T/C)GGT (Meyers et al., 1993). 3-D structures show that the runt homology domain (RHD) forms a 12-stranded barrel that adopts an s-type immunoglobulin fold and recognizes specific bases in both major and minor grooves of the DNA (Asou, 2003). Mutations in the RHD affecting the DNA binding face of the protein have been associated with familial platelet disorder associated with predisposition to leukaemia (FPD) and in sporadic cases of AML and MDS without chromosomal translocations (Owen et al., 2008).

The RHD is also involved in binding its heterodimer partner CBFβ, which is crucial for the role of AML1. The gene for *CBFB* is located at 16q and is ubiquitously expressed. It has homology to Drosophila big brother and brother proteins (Golling et al., 1996). CBFβ does not bind DNA directly or interact with any other co-factors but rather has an allosteric effect on AML1 DNA binding. CBFβ binds on the RHD domain of AML1 close to the DNA binding region and results in the conformational change of critical DNA binding residues allowing AML1 to overcome the adjacent segments that inhibits its binding to its sequence (Warren et al., 2000). The dimerisation with CBFβ also protects AML1 from ubiquitin-mediated degradation (Huang et al., 2001). In fact, the dimerisation of AML1 with CBFβ does not readily occur suggesting this might be a rate-limiting step for the functions of AML1. In contrast to these observations, the interaction of CBFβ with AML1-ETO does not appear essential in inducing transformation (Kwok et al., 2009) although this has been challenged (Roudaia et al., 2009).

The products of *AML1* are localised to the nucleus unlike those of the *CBFB*, which remain in the cytoplasm unless heterodimerised with AML1 (Tanaka et al., 1997). Localisation occurs through two mechanisms: in part by sequences in the RHD domain (Lu et al., 1995) but also through a specific targeting sequence of 31aa within the C-terminus of the protein (NMTS) (Zeng et al., 1997).

At the C-terminal end of the AML1 protein, activation and repression domains have been discovered and will be described in the following section.

Target Genes of AML 1

Recently, expression array profiling has identified a large number of genes relevant to haemopoietic differentiation whose transcription is regulated by AML1 and AML1-ETO proteins. In contrast, traditional techniques had previously highlighted the complex nature of the interactions required between AML1 and regulation of its target genes. It appears that AML1 alone has only weak transactivator properties and therefore requires the presence of other co-operating factors binding to adjacent promoter sites for haemopoietic expression of these genes. Some of these such as *TCR, IL-3, GM-CSF, MPO, M-CSF* and *granzyme B* are described in detail highlighting the complex nature of some of these interactions. These examples suggests a mechanism for cell specific transcription where AML1 acts as the master transcriptional organiser recruiting co-operating factors to a complex that stimulates lineage-restricted transcription.

AML1 requires the co-expression of c-myb to transactivate the MPO gene (Britos-Bray and Friedman, 1997). AML1 binds to MEF causing activation of the IL-3 promoter (Mao et al., 1999). AML1 synergistically activates transcription of TCRα and TCRβ enhancers. Activation of the TCRα receptor is complex requiring LEF and CREB as well as recruiting Ets-1 and AML1 into a ternary DNA-protein complex (Giese et al., 1995). It appears that direct binding by Ets-1 to AML1 augments DNA binding by AML1 to the TCRβ promoter and in a reciprocal manner the binding of AML1 to Ets activates DNA binding by Ets (Goetz et al., 2000, Gu et al., 2000, Kim et al., 1999). On the M-CSF receptor promoter AML1 synergistically activates in combination with both C/EBPα and PU.1 (Zhang et al., 1996). The mechanisms are different. C/EBPα binds to AML1 leading to a co-operative binding of DNA. In contrast PU.1 has weak binding to AML1 and does not bind DNA but appears to recruit the activator CBP/p300 to the promoter (Petrovick et al., 1998). There is a physical interaction between AML1 and SMAD protein and this can mediate the TGF-β responsiveness of the IgA1 promoter (Pardali et al., 2000).

Furthermore, immunoprecipitation experiments have highlighted partner proteins that act as co-activators. The p300/CBP proteins directly bind with the C-terminal activation domain of AML1 (Kitabayashi et al., 1998a). These co-activators have intrinsic histone acetylation (HAT) activity as well as binding to yet another HAT, P/CAF. This activity results in acetylation of lysine residues in chromatin associated histones, causing an opening of the chromatin structure and leading to enhanced transcription. p300/CBP also serve as integrator proteins between AML1 and other co-activators such as c-MYB, C/EBPα and PU.1. Other transactivators discovered to bind AML1 include ALY and YAP. ALY is ubiquitously expressed but lacks direct activating properties. Rather, by binding to AML1, it acts as a bridge between AML1 and other transcription factors such as TCR/LEF (Bruhn et al., 1997). YAP binds to the PPPY motif of AML1 and contains an activation domain making AML1 a stronger activator (Yagi et al., 1999) (Figure 2).

Although AML1 is seen as a transcriptional activator it also has domains associated with transcriptional repression. AML1 co-precipitates with the Sin3 co-repressor and this interaction appears critical for AML1 to repress a wide number of promoters including the cyclin-dependent kinase inhibitor *p21 waf* (Lutterbach et al., 2000). The VWRPY motif recruits the Groucho/TLE family of transcriptional repressors and limits transactivational activity of AML1 (Aronson et al., 1997, Levanon et al., 1998). The interaction of AML1 and TLE can lead to decreased transcription of the neutrophil elastase and M-CSF genes (Imai et al., 1998).

AML1 in haemopoiesis

The development of the haemopoietic system is tightly regulated by a series of transcription factors carefully co-ordinating the generation of haemopoietic stem cells and the differentiation of the progenitor cells. Primative haemopoiesis is a transient phase and is followed by the appearance of definitive haemopoiesis (Keller et al., 1999).

AML1 is essential for development of the definitive haemopoietic system acting as a key regulator of the formation of the haemopoietic stem cell. AML1 null mice die in-utero, incur severe CNS haemorrhages and lack foetal liver haematopoiesis (Okuda et al., 1996). A similar phenotype results with loss of the CBFβ gene and the finding that CBFβ protects AML1 from degradation may explain this finding (Wang et al., 1996b). These defects are intrinsic to the haemopoietic cells rather than the microenvironment and the addition of AML1 restores haemopoiesis (Okuda et al., 2000). AML1 null mice have a defect in forming haemopoietic cell clusters (North et al., 1999). Haemopoietic stem cells play a key role in angiogenesis during embryogenesis (Takakura et al., 2000). This suggests that the haemorrhages seen in AML1 null mice are secondary effects from a lack of definitive haemopoiesis. These reports suggest a widespread role for AML1 in induction of cell differentiation and proliferation (Licht, 2001).

AML1 in leukaemia

The *AML1* gene fuses with many different translocation partners, which result in leukaemia. (Table 7)

Table 7: AML1 Partner Genes involved in Leukaemia

TRANSLOCATION PARTNERS	GENE	DISEASE	REFERENCE
8q22	*ETO*	AML (FAB M2)	
12p13	*TEL*	B-ALL	(Golub et al., 1995)
3q26	*EVI1,* *MDS1,* *EAP*	Therapy Related (TRL) & CML transformation	(Yin et al., 2006, Nucifora et al., 1993)
16q24	*MTG16*	Secondary AML / TRL	(Salomon-Nguyen et al., 2000)
19q13	*AMP19*	Radiation associated TRL	(Hromas et al., 2001)
8q24	*TRPS1*	Relapse AML	(Asou et al., 2007)

Point mutations particularly in the RHD of AML1 are also involved in causing leukaemia and found in cases of AML-M0, AML-MDS and therapy-related AML/MDS (Asou, 2003). RHD mutations are also a common feature of the familial platelet disorder with a predisposition to AML (Dowton et al., 1985).

ETO

Structure and Function

ETO (eight twenty-one) was initially identified as the fusion partner of *AML1* in t(8;21) (Erickson et al., 1992). In this translocation, almost the entire open reading frame of ETO is retained. Consequently, much of the work conducted on *ETO* has looked at its role in leukaemia and conducted with regard to the t(8;21) and its product.

The *ETO* gene is highly conserved and represents the mammalian homologue of the drosophila nervy gene. The human and mouse proteins have 90% homology. The role of *ETO* in humans is still unclear although the widely accepted function is that *ETO* acts as a transcriptional co-repressor (Hiebert et al., 2001). Its expression is found in a variety of human tissues including the brain, heart, lung and testis (Wolford and Prochazka, 1998). *ETO*, although down regulated in haemopoietic cells, is expressed in CD34+ progenitor cells (Erickson et al., 1996).

ETO was named and identified in 1992 (Erickson et al., 1992) and also by another group where it was referred to as *MTG* (myeloid translocation gene) (Miyoshi et al., 1993). Other members of the ETO family identified in humans are MTGR1 and MTG16 whose mouse homologue is ETO2 (Davis et al., 2003). MTG16 has been identified as a fusion partner to AML1 in cases of MDS (Gamou et al., 1998, Salomon-Nguyen et al., 2000). MTGR1 has been shown to heterodimerise with AML1-ETO (Kitabayashi et al., 1998a).

ETO consists of 14 exons covering 87Kb (Figure 3). It has alternative first exons 1a and 1b each with its own promoter. This yields two transcripts giving rise to two proteins of 577aa and 604aa. The breakpoint region in the t(8;21) is between the first two exons; thus almost the entire reading frame of ETO is translocated to the AML1 gene. Alternatively spliced transcripts incorporating alternative exons resulting in the introduction of premature stop codons have been identified (Kozu et al., 2005, Wolford and Prochazka, 1998). The generation of alternative transcripts and in particular the exon 9a transcript has a key role to play in the leukaemic potential of the AML1-ETO product (Yan et al., 2006).

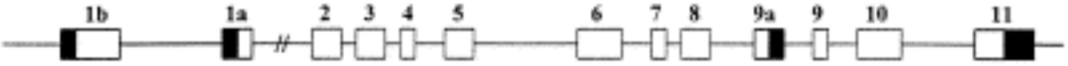


Figure 3 Structure of ETO gene (Wolford and Prochazka, 1998)

ETO has four domains with extensive homology to the drosophila nervy protein, involved in axon guidance. These are known as nervy homology regions 1-4 (NHR) NHR1 is homologous to several TATA binding protein-associated factors. This region has been shown to be important in the interaction of AML1-ETO with the activation domain of E-proteins, which play a role in cell cycle, causing their inactivation (Zhang et al., 2004). It also has a role in nuclear sub- localization of ETO (Odaka et al., 2000). The NHR2 is the hydrophobic heptad repeat region and has a α-helical tetramer structure that is important for its role in acting as an oligomerisation domain. Through this domain it interacts with AML1-ETO proteins and with other ETO family members including ETO and MTGR1. It also has a binding site for the co-repressor Sin3. The NHR2 appears essential for the block of haematopoietic differentiation, repression of basal transcription and for the dimerisation. The NHR3 has mainly an α-helical structure and may assist the NHR4 interaction with the SMRT repressor (Zhang et al., 2001). The NHR4 (MYND) has two zinc fingers but does not bind DNA. Instead, it acts as a protein–protein interaction motif playing a critical role in regulating activity of AML1 promoters by recruiting the transcriptional co-repressors NCOR and SMRT. Other regions outside these areas also have a role to play in recruiting HDACs and assisting interactions with the Nervy domains (Amann et al., 2001). In between the Nervy domains the PST rich regions, phosphorylated at these residues, could play a role in protein stability (Figure 4).

The ETO protein is found in the cell nucleus. A nuclear localisation sequence (NLS) located between NHR1 and NHR2 regions is required for ETO localisation (Davis et al., 2003). ETO and AML1-ETO are localised into distinct subnuclear speckles, which is distinct from AML1 (Odaka et al., 2000).

ETO Interactions

The actions of AML1-ETO directly oppose the activity of AML1. This is thought to occur through ETO's interactions with HDACs. Deacetylation of histone tails influences interactions between key regulatory proteins (Iizuka and Smith, 2003). HDAC inhibitors can block the effects of AML1-ETO on the cell cycle suggesting that at least some of these HDAC interactions are functionally significant (Klisovic et al., 2003, Liu et al., 2005, Wang et al., 1999). ETO has been shown to directly interact with HDACs 1,2,3, (Amann et al., 2001, Lutterbach et al., 1998). Direct interactions of HDACs require contact with the NHR2 and NHR4 domains (Hug and Lazar, 2004). Indirect interactions with HDACs occur through ETO's interactions with co-repressors and was initially shown using yeast two-hybrid screens. NCoR was cloned from a screen for leukaemic proteins interacting with ETO whilst independently ETO was cloned when screening for co-repressors interacting with SMRT, a homologue of NCoR (Gelmetti et al., 1998, Wang et al., 1998). The NHR4 domain of ETO is required for this interaction. The NHR2 domain enhances this reaction although the NHR3 domain also assists in the binding of the SMRT repressor (Hug and Lazar, 2004). The NCoR/SMRT complex also includes other proteins such as TBL1, GPS, IR10 and importantly HDAC3, the enzymatic subunit required for the deacetylation of histones. Using a candidate approach another repressor Sin3 was detected (Lutterbach et al., 1998). This interaction was dependent on the NHR2 domain of ETO although additional contacts were required. Sin3 has a diverse role, being recruited by transcription factors such as MAD1 as well as causing global deacetylation. Sin3 requires co-operating components and has been found to complex with HDAC1 and 2 as well as SAP and RbA.

In addition to these interactions with co-repressors ETO is able to interact with itself and other family members forming high molecular weight oligomers via the NHR2 domain (Davis et al., 1999). Homodimerisation and formation of high molecular weight complexes may be a common mechanism for leukaemogenesis in fusion proteins with deletion of its oligodimerisation domains resulting in loss of transformation potential.

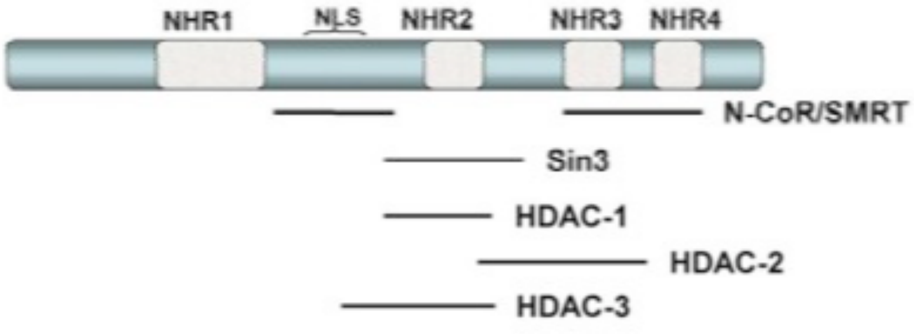


Figure 4 Structure of ETO and its interactions adapted from (Davis et al., 2003). The four NHR domains are highlighted as well as the nuclear localisation signal (NLS) on the ETO molecule. Below the location of the interactions between ETO and the repressor molecules are depicted. (see text for further details)

AML1-ETO

The t(8;21) results in the formation of one *AML1-ETO* fusion gene the remaining allele being unaffected. The reciprocal product *ETO-AML1* has not been identified suggesting that this transcript is either not expressed or that it is unstable and degraded rapidly.

The breakpoint regions for the *AML1-ETO* translocation are consistent. The AML breakpoint is in intron 5 and has 3 breakpoint cluster regions (BCR). The breakpoint of *ETO* occurs in two regions, one in intron 1a containing one BCR and the other in intron1b that has three BCR. As exon 1b of *ETO* does not have any splice acceptor sites the chimeric protein produced by either truncation in *ETO* is the same (Peterson and Zhang, 2004). (Figure 5)

This translocation leads to a consistent AML1 product of 177 aa containing the RHD domain. However, *ETO* is alternatively spliced and gives rise to co-existence of different *AML1-ETO* products. The main product though involves almost the whole of *ETO* (exon 2 through to 11) and gives rise to an ETO product of 575 amino acids. The resultant AML1-ETO protein is 752 aa and is detected as a band of approximately 85kDa. AML1-ETO is under the control of the 2 AML1 promoters with the published AML1-ETO transcript derived from the P2 promoter.

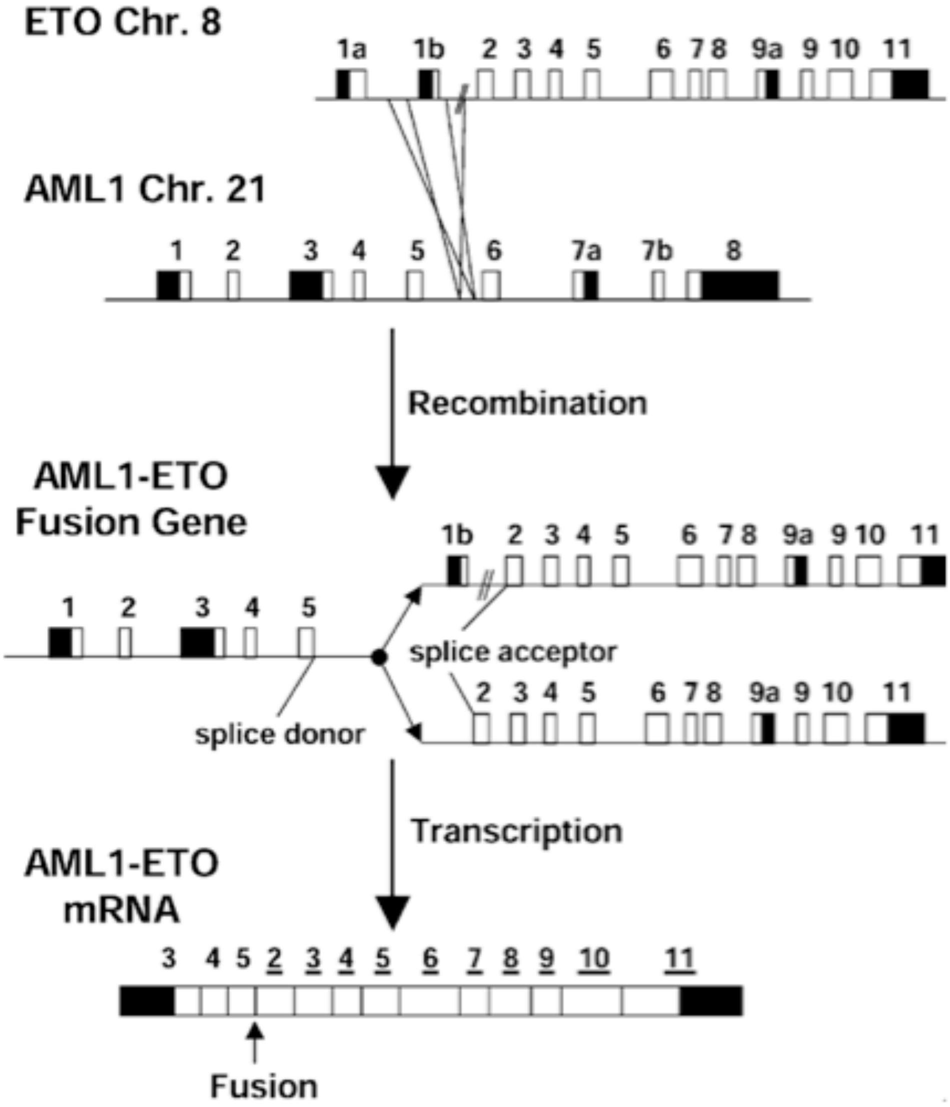


Figure 5 Genomic Structure of t(8;21) (Peterson and Zhang, 2004). Breakpoints in *AML1* and *ETO* and their recombinations as well the final AML1-ETO mRNA produced are shown.

Pathogenesis of AML1-ETO

Through the use of transcription assay models, haemopoietic cell lines and mouse models extensive literature is available and has led to a broad understanding of the mechanisms by which AML1-ETO induces leukaemia. However, the understanding of the exact mechanisms by which this chimeric transcription factor causes leukaemia still remains incomplete.

In the t(8;21) there is loss of one *AML1* and one *ETO* but a gain of *AML1-ETO* gene. The majority of studies, which are discussed in detail below, suggest a general mechanism where transcriptional activation of *AML1* is replaced by *AML1-ETO* transcriptional repression. This is mediated through the loss of the activation properties of AML1, the dominant inhibition of wild type AML1 and the interaction with the nuclear co-repressor domains of ETO. Repression of genes occurs through ETO's interaction with HDACs leading to a closed chromatin structure and hence repression of transcription. However, while these earlier studies strongly favour transcriptional repression as a basis for t(8;21) induced leukaemia the more recent evidence challenges this simplistic assumption. It has been noted that AML1 can act as a transcriptional repressor in addition to an activator. Evidence described below shows that AML1-ETO is also able to act as a transcriptional activator, in addition to its repressor activities. This and other evidence such as the role of alternative transcripts and the requirement for a "second hit" complicates the understanding of t(8;21) induced leukaemogenesis further.

Transcription Assay Models

Using transfection reporter gene studies it was initially demonstrated that AML1-ETO blocks the ability of AML1 to activate promoters of its target genes such as *MDR-1, GM-CSF* receptor,*IL-3, fos, IgA* and the *TCRβ* receptor. These reports also highlighted the complex interactions between genes and the transcription factor. However, evidence also suggests that AML1-ETO has positive regulator activities and negatively affects transcriptional repressors contrary to its "repressor role".

The first promoter shown to be negatively regulated by AML1-ETO was on *TCRβ.* A very low amount of AML1-ETO was required to block AML1 activation of *TCRβ* and AML1-ETO only showed negative effects in co-transfection with AML1 and not alone suggesting a dominant block of AML1 function (Meyers et al., 1995). AML1-ETO still retains the ability to bind DNA and heterodimerise with CFBβ (Meyers et al., 1993) as would be predicted as it retains the RHD. In fact, the fusion protein binds DNA and CBFB more efficiently than wild type AML1 (Tanaka et al., 1998). This sequesters CBFB away from the wild type AML1. This competitive advantage results in even very low levels of AML1-ETO dominating AML1.

Similar reports on the *GM-CSF* gene promoter were also seen. However, in this case AML1-ETO did show negative effects on its own suggesting that AML1-ETO could also play an active role in repression of promoters (Frank et al., 1995). The mechanism of active repression was shown to be due to the repressor activity of the

ETO part of the protein and its interaction with HDACs. This interaction occurs directly or indirectly through its interaction with NCoR/SMRT or Sin3 leading to deacetylation of histones and inhibition of transcription. AML1-ETO has also been shown to recruit DNMT1, leading to DNA methylation, which can further cause active repression (Liu et al., 2005).

AML1-ETO also binds to and negatively regulates the transcriptional activity of other transcription factors independent of AML1. These include C/EBPα, MEF, PU.1 and Jun. AML1-ETO binds PU.1 displacing the co-activator Jun thus down regulating *PU.1* transcription (Vangala et al., 2003). C/EBPα is essential for myeloid differentiation and is down regulated by AML1-ETO and has been shown to block C/EBPα activation of the *NP3* promoter (Westendorf et al., 1998). MEF is an Ets family member and can regulate cell cycle of haemopoietic stem cells. AML1-ETO can bind and inhibits its activity and transcription and also suppress activating properties of MEF on the *IL-3* promoter (Mao et al., 1999). TGFβ activation of the *IgA* promoter mediated by SMAD can be repressed by AML1-ETO (Jakubowiak et al., 2000).

AML1-ETO is also able to repress tumour suppressor genes directly. *p14ARF*, a mediator of the p53 oncogene checkpoint, is a direct transcription target of AML1-ETO and its down regulation leads to an increase in MDM2 and the subsequent inactivation of p53. Thus down-regulation of p14ARF impairs p53-mediated growth arrest and apoptosis in response to activated oncogenes and renders cells immortal

(Hiebert et al., 2003, Linggi et al., 2002). AML1-ETO has been shown to inhibit the function of retinoic acid (RA). RA binds to it receptor RARα and is then able to regulate its target gene *RARβ2* which has strong promoter functions and also acts a tumour suppressor gene. AML1-ETO by recruiting HDACs into this complex can silence the RA signalling pathway (Fazi 2007).

In contrast to its repressor properties AML1-ETO was also shown to have positive regulator activities. The M-CSF receptor contains an AML1 binding site and AML1 is able to enhance its activity through collaboration with PU.1 or C/EBPα. AML1-ETO actually enhances AML1 transactivation of M-CSF receptor (Rhoades et al., 1996). It is postulated that AML1-ETO may remove Sin3 from bound AML1 making the AML1 more active. BCL-2 and the G-CSF receptor can also be up regulated by AML1-ETO (Ahn et al., 1998, Klampfer et al., 1996, Shimizu et al., 2000).

AML1-ETO has effects on transcriptional repressors such as PLZF and GFI proteins. PLFZ acts as a repressive transcription factor blocking myeloid cell growth. AML1-ETO interacts with PLZF and inhibits the repressor activity of PLZF thus activating gene expression. It can be postulated that AML1-ETO may block factors required to repress genes involved in cell cycle (Yeyati et al., 1999).

The pathogenesis caused by AML1-ETO also takes into account the loss of the functions of AML1 due to the loss of the carboxyl domain as well as the gains of function of ETO. The important domains of the *AML1* gene, which are lost, include the activation, repressor and NMTS domains. The loss of the activation domains results in a loss of p300/CBF and repression of target gene transactivation. Despite the loss of the NMTS domain the fusion product can still localise into the nucleus as ETO sequences take over this role. ETO contains one nuclear localisation signal and 2 NMTS. However, the subnuclear localisation of these signals does not overlap with AML1 (Davis et al., 2003). Therefore this relocalisation of AML1 binding domain by AML1-ETO may contribute to differential regulation of target genes.

Haemopoietic Cell Models

AML1-ETO has been stably transfected into many haemopoietic cell lines including 32Dcl3 and U937. The transcription control mediated by t(8;21) has been addressed in several different ways in these studies and include its effects on differentiation, proliferation and apoptosis. Generally these studies report that AML1-ETO inhibits differentiation but, contradicting its function in promoting leukaemogenesis, appear to inhibit proliferation and promote apoptosis of haemopoietic cell lines.

The forced expression of AML1-ETO in cell lines results in a block in both myeloid and erythroid differentiation. 32Dcl3, a murine myeloid progenitor cell line and L-G a mouse myeloid cell line, require IL-3 for growth but undergo granulocytic differentiation when IL-3 is replaced by G-CSF. Expression of AML1-ETO blocks the granulocytic differentiation in response to G-CSF and instead increases proliferation (Kitabayashi et al., 1998b). AML1-ETO expression also up-regulates the G-CSF receptor in 32Dcl3 cells, explaining the increased sensitivity of these cells to G-CSF. This effect is by an indirect mechanism inducing the expression of C/EBPα. However, overexpression of C/EBPα or G-CSF receptor in myeloid cell lines does result in differentiation unlike AML1-ETO expressing cells where differentiation is blocked. This suggests that overexpression of the G-CSF receptor is not sufficient to cause differentiation and does not fully explain the effects of AML1-ETO (Shimizu et al., 2000). BCL2 is also up regulated in 32Dcl3 cell lines in response to G-CSF. The BCL2 promoter contains AML1 binding sites and AML1-ETO can activate the BCL2 promoter in cell lines (Klampfer et al., 1996). However, overexpression of BCL2 does not block differentiation suggesting that stimulation of BCL2 is not the primary target of AML1-ETO (Kohzaki et al., 1999).

The increase in cell apoptosis appears to involve the JNK pathway. Experiments in the myeloid U937 cell line confirmed that AML1-ETO expression induces *Jun*. Inhibition of the JNK pathway reduced the transactivation capacity of AML1-ETO on the *Jun* promoter and consequently decreased AML1-ETO induced apoptosis (Elsasser et al., 2003).

In addition to cell lines which stably express exogenous AML1-ETO there are two cell models of the t(8;21): Kasumi-1 (Asou et al., 1991) and SKNO-1 (Matozaki et al., 1995). It is noted that both these cell lines have mutated *p53* and *KIT* genes. Knock down of AML1-ETO in Kasumi cell lines by si-RNA can give rise to a distinctive gene expression signature and in particular demonstrates up regulation of *p21waf* (Dunne et al., 2006). Using Kasumi cell lines it was shown that AML1-ETO and AML1 could bind to ribosomal RNA at nuclear-organising regions and result in epigenetic regulation of cell growth through up regulation of ribosomal gene transcription (Bakshi et al., 2008).

Haemopoietic Stem Cell Models

The expression of AML1-ETO in stem cells is more in keeping with the leukaemic effects of t(8;21) and favours long-term expansion of both mouse and human stem cells.

Murine HSC transduced with an AML1-ETO expressing retrovirus promoted self-renewal, did not cause AML but developed a MDS phenotype (de Guzman et al., 2002, Schwieger et al., 2002). In primary human CD34+ PBSC cells the introduction of an AML1-ETO vector results in a more complex pattern of growth. Initially there is a block in colony formation followed by proliferation on both methylcellulose plates and liquid cultures i.e. there is inhibition of growth of the committed progenitor cells but growth of more primitive cells (Mulloy et al., 2002). Furthermore, introduction of AML1-ETO into normal CD34+ cells has also been shown to give rise to long-term (8 months) growth of cells that retain the ability to renew and differentiate. These long-term in-vitro cultures of CD34+ cells can engraft but do not cause leukaemia in NOD/SCID mice (Mulloy et al., 2003). This observation is also apparent in vitro assays of mouse bone marrow cells (Hug et al., 2002, Odaka et al., 2000). A similar complex pattern was seen in CD34+/CD13- erythroid cells. Colony formation was abrogated with the introduction of AML1-ETO. However, in further liquid culture studies AML1-ETO inhibited proliferation during early EPO independent erythropoiesis but then gave way to cell expansion, confirming that AML1-ETO has the capacity to promote stem cell self renewal (Tonks et al., 2003).

Studies of stem cell expansion have focused on the inhibition of function of lineage promoting transcription factors by AML1-ETO and the ability of AML1-ETO to promote self-renewal through the WNT and Notch signalling pathways.

AML1-ETO directly targets the transcription factors PU.1, C/EBPα and GATA-1 and through these interactions represses lineage commitment (Elagib and Goldfarb, 2007a). PU.1 is an ETS transcription factor involved in early haemopoiesis and crucial for generating all haemopoietic lineages. Decreased PU.1 can lead to AML in mouse models (Rosenbauer et al., 2004). However, mutations in PU.1 although rarely present in AML patients are not associated with t(8;21). C/EBPα acts as an activator. Mice with a truncated form of C/EBPα develop leukaemia (Rosenbauer et al., 2005). However, although mutants are found in 9% of AML patients they are not associated with t(8;21) and in fact C/EBPα levels are lower in t(8;21) as AML1-ETO negatively interacts with C/EBPα. GATA1 is a zinc finger transcription factor and has an important role in haemopoietic development. A 5% expression of GATA1 in mice causes leukaemia (Shimizu et al., 2004). *GATA1* mutations are found in AML –FAB-M7 associated with Down’s syndrome. AML1-ETO is a potent inhibitor of GATA-1 transcription and blocks erythroid differentiation by inhibiting p300 acetylation of GATA1 through its NHR4 domain. The GATA1 transcriptional programme consists of down regulating KIT expression (Munugalavadla et al., 2005). Abnormally sustained KIT expression in GATA-1-deficient cells may then predispose these genetically unstable cells to positive selection for receptor-activating mutations, accounting for the high frequency of KIT mutations seen in t(8;21) AML patients (Elagib and Goldfarb, 2007b).

Animal Models

Studies on mouse models have been vitally important in the understanding of cancer development and leukaemia. However, this has proved more difficult for studies on the t(8;21) animal model. Heterozygous AML1-ETO knock in mice die at 12.5 days (Okuda et al., 1998, Yergeau et al., 1997). Foetal mice do not establish definitive liver haematopoiesis and develop CNS haemorrhage. This is identical to the phenotypes seen in both AML-1 deficient mice and CBFβ knock out mice (Okuda et al., 1996, Wang et al., 1996a). This same phenotype is also seen in mice heterozygous for the knock-in of the CBFβ-MYH11 fusion gene seen in inv(16). This implies that AML1-ETO and the CBFβ fusion product dominantly block AML-1 function.

Culture of yolk sac cells from the AML1- ETO knock in mice gives rise to dysplastic colonies that have a high self renewal capacity (Okuda et al., 1998) and macrophages (Yergeau et al., 1997). This is not seen in the AML1 or CBFβ negative mutants, which lack any detectable haemopoietic progenitors. This further implies that AML1-ETO has other roles besides blocking wild type AML1.

To circumvent the lethality of the AML1-ETO allele transgenic mice are made in which the expression of AML1-ETO is under the control of regulatory systems. Many different mouse models have been established and include tetracycline regulatable (Tet) (Rhoades et al., 2000), myeloid lineage specific MRP8 promoter directed (mrp8) (Yuan et al., 2001), haemopoietic stem cell Sca-1 locus (sca1) (Fenske et al., 2004) and Cre recombinase-mediated (cre) transgenic mice (Buchholz et al., 2000). With the exception of the Sca1 mouse these mice remained healthy with normal haemopoiesis. The sca model developed a myeloproliferative disorder (MPD) after about 6 months (Fenske et al., 2004). In the tet and Cre mice myeloid progenitor cells were shown to increase self renewal and decrease differentiation in vitro. In the mrp8 mice treatment with an alkylating agent resulted in the transgene mice developing AML or in some cases T-ALL (Yuan et al., 2001).

These findings suggest that while the AML1-ETO mutation has the potential to cause leukaemia it requires additional mutations to transform from the preleukaemic state to the leukaemic state.

Secondary events

AML1-ETO mRNA has been detected in patients with long-term remission (Guerrasio et al., 1995) and in blood spots taken from identical twins (Wiemels et al., 2002). Furthermore, the AML1-ETO transcript has been demonstrated in a fraction of colony-forming cells of erythroid, granulocyte and megakaryocyte lineage in both leukaemic and remission marrow. AML1-ETO is detected in stem cells with both myeloid and B-lymphoid capabilities (Miyamoto et al., 1996, Miyamoto et al., 2000). This suggests that the translocation occurs at an early stage of stem cell development and at an early stage of leukaemogenesis. Along with the evidence from AML1-ETO mouse models it appears that the t(8;21) translocation on its own is not sufficient to cause leukaemia but requires an additional mutation. This would concur with the Gilliland hypothesis proposing a two-hit model. AML1-ETO provides one hit conferring a class II type mutation. Many potential second hits have been investigated to discover the class I mutation.

Secondary non-random chromosomal changes are more common in t(8;21) patients. The commonest changes are the loss of sex chromosomes and deletion in 9q. The loss of sex chromosomes in haemopoietic cells naturally occurs with increasing age. However, the incidence in t(8;21) cohort is higher and occurs at a significantly younger age compared to normal populations and to AML with other chromosome translocations. The loss of the Y chromosome in male patients seemed to lead to a weak but significant poor prognostic factor for overall survival

(Schlenk et al., 2004). The deletion of 9q resulting in loss of the tumour suppressor genes TLE1 and TLE4 was shown to have a more favourable prognosis in non-white t(8;21) patients but were not reproduced in further studies. Other chromosome abnormalities associated with t(8;21) include trisomy 8 and trisomy 4 and more rarely t(5;12) TEL-PDGFR and t(9;22) BCR-ABL (Paschka, 2008).

Translocations in AML may also occur with molecular changes and particularly together with mutations of tyrosine kinases (TK) which may provide the class I mutation. Kit is a tyrosine kinase receptor and although the *KIT* mutation is relatively rare in AML it is increased in AML with CBF mutations occurring in 59% of AML t(8;21) patients (Wang et al., 2005). In another series, *KIT* and *RAS* mutations were increased in 70% of paediatric and 48% of adult cases of CBF leukaemia (Goemans et al., 2005). Furthermore, these reports also found *KIT* to be over-expressed in over 90% of t(8;21) cases regardless of mutational status. The main mutations have been located in exon 8, which encodes for an extracellular part of the receptor and exon 17, which encodes the activation loop in the kinase domain. Mutations in the exon 17 are associated with poorer prognosis. KIT mutations pose a potential target for imatinib, a tyrosine kinase inhibitor active against BCR-ABL, Kit and PDGFR receptor kinases. Although active against mutations in exon 8 and exon 17 involving the codon N822 it is inactive against the mutants of exon 17 involved in codon D816, the most common form. However, newer tyrosine kinases inhibitors such as dasatinib are active against this common mutant.

Other possible candidates for a type I collaborating mutation include *FLT3*, *JAK2* and *RAS* present in 2-9%, 6% and 8-11% of t(8;21) patients respectively. In human AML samples the internal tandem duplication of the FLT3 gene can occur in 20-30%. Although this mutation most commonly occurs in APML it is also seen in 9% of t(8;21) (Kottaridis et al., 2001). The prognostic significance for most of these reported mutations is unavailable. However, a recent report implicates *FLT3* mutations as conferring poorer prognosis in this cohort of patients. In a 146 patients with t(8;21) AML, the presence of *FLT3* mutations occurred in 13% and were associated with more frequent relapse and shorter survival (Paschka et al., 2009).

Mutations in M-CSF, TRKA and TRKC receptors have also been recorded in cases of t(8;21) (Abu-Duhier et al., 2003, Reuther et al., 2000). Although an activated form of the IL-3Rβ chain can promote the development of leukaemia in mouse, mutations of IL-3 receptors have not been found in humans (Phan et al., 2003). TP53 mutations causing inactivation has a key role in development of many malignancies. However, this mutation is rarely seen in AML with t(8;21). Our own laboratory has shown loss of *EZH2*, through a microdeletion in one allele and a mutation in the other allele, in a case of t(8;21) primary sample (personal communication).

Mice models have confirmed that a constitutive TK with a CBF mutation can lead to leukaemia. Thus, AML1-ETO with a *FLT3* mutation was shown to cause leukaemia in mice (Grisolano et al., 2003). Similarly, the TEL-PDGFRβ fusion TK can also co-operate with AML1-ETO in mouse models to cause leukaemia (Schessl et al., 2005).

Other mouse models give further clues to secondary hits. Transgenic mice expressing WT1 can also develop leukaemia when exposed to AML1-ETO (Nishida et al., 2006). Mice lacking the interferon regulatory factor ICSBP, a tumour suppressor, develop myeloblastic transformation when exposed to AML1-ETO (Schwieger et al., 2002). AML1-ETO in p21 waf deficient mouse stem cells induced leukaemia although paradoxically AML1-ETO expressing cells have been shown to increase the cell cycle inhibitor p21 waf (Peterson et al., 2007b, Yan et al., 2006). It has been shown that there is up regulation in p53 in response to DNA damage affected through the t(8;21) which may explain the good response to treatment. Loss of the p53 pathway may be associated with disease progression and indirect mechanisms, such as repression of CDKN2, involved with p53 pathway, have been suggested (Krejci et al., 2008).

Gene Expression Profiling & Newer Models of Leukaemogenesis

Gene expression profiling (GEP) has been shown to sub-classify leukaemias based on their chromosomal abnormalities This includes being able to distinguish between the inv(16) and t(8;21) leukaemias (Debernardi et al., 2003). More recently GEP has been used as a prognostic indicator. GEP on 93 CBF leukaemias reported that these patients could be separated into 2 groups that differ in overall survival. Group 1, which had a poorer prognosis, was characterised by overexpression of genes involved in JNK and MAPK pathways and aberrant chemo-resistance. Group II was characterised by genes involved in mTOR signalling and anti-apoptosis (Bullinger et al., 2007). Further GEP studies in t(8;21) patients lacking KIT mutations was able to separate patients into prognostic groups based on their expression profile (Paschka et al., 2007). These expression studies are also able to highlight potential target genes as therapeutic targets. Although GEP has generally not been shown to classify different molecular mutations recent evidence has highlighted a specific signature for CBF AML with *KIT* mutations. (Luck et al., 2010).

GEP has also shown that AML1-ETO, in addition to regulating a number of genes that are not normally regulated by AML1, can activate as many genes as it represses, in contrast to its perceived role as a transcriptional repressor (Shimada et al., 2000). This includes many of the genes involved in DNA repair and stem cell maintenance and renewal. For example, AML1-ETO up regulated Jagged 1 and drives stem cell proliferation via the Notch pathway (Alcalay et al., 2003)). AML1-

ETO directly induced the expression of plakoglobin, a mediator of Wnt signalling by binding to TCF/LEF and activating its target genes such as *c-myc* and *cyclin D1*. In functional studies, plakoglobin was shown to cause clonal growth when transfected in myeloid 32D cell lines and enhanced self-renewal when expressed in murine haemopoietic progenitor cells (Muller-Tidow et al., 2004). AML1-ETO also up regulates TRKA in CD34+ stem cells and causes stem cell proliferation in response to NGF and IL-3 (Mulloy et al., 2005). These results suggest that AML1-ETO may drive expansion of early progenitors through up regulation of target genes rather than repression and appears to complement the differentiation blockade seen in cell lines.

Further evidence against transcriptional repression as the sole mechanism for t(8;21) leukaemogenesis comes from studies of alternative transcripts of AML1-ETO. Although the role of many of these transcripts has not been elucidated, an alternative transcript utilising an alternative exon 9a, resulted in a truncated AML1-ETO protein, which was shown in mouse models to cause leukaemia on its own (Yan et al., 2006). Furthermore, cells with the AML1-ETO 9a variant escapes mitotic arrest and may affect normal mitotic checkpoint promoting secondary mutagenic events (Boyapati et al., 2007). Along with gene expression data showing that AML1-ETO leads to repression of genes involved in base excision repair (Alcalay et al., 2003) it suggests that AML1-ETO itself could promote additional mutations in the preleukaemic clone. This suggestion is further supported in studies where AML1-

ETO expression in haemopoietic stem cells results in an increase in DNA double-strand breaks and consequently in genetic mutations (Krejci et al., 2008).

In summary, whilst on-going work at the molecular level continues to gives us clues as to mechanisms of leukaemogenesis and provide potential targets for treatments, further work is still required to elucidate the pathogenesis of leukaemia in the t(8;21) and enable targeted treatments for this disease.

In this thesis I have set out to further explore the pathogenic mechanisms of the t(8;21). Our laboratory has a specific expertise in AML with t(8;21), a great deal of experience working with high throughput technologies and access to many primary AML samples. These factors allied with the clinical identification of a need to explore this particular translocation lead to the development of our aims. I have worked on primary leukaemia samples using the latest molecular techniques of expression array profiling and ChIP followed by high throughput sequencing. The aims were to discover new key molecules that are regulated by AML1-ETO giving potential targets for treatment. By using exon arrays the aim was to discover new insights into the role of alternative transcripts. Using ChIP it was hoped to provide a detailed analysis of which genes the transcription factor AML1-ETO binds and thus give further insights into possible mechanisms of this form of leukaemia.

Chapter 2

Materials and Methods

I. Common Methods

Samples and Ethics

Presentation bone marrow and blood samples were obtained from patients of St. Bartholomew’s Hospital diagnosed with AML. Written informed consent, for storage of samples for research purposes, was obtained prior to obtaining samples. Samples were stored and maintained at the Tissue Bank by the Medical Oncology Department of Bart’s Hospital according to the Human Tissue Act 2004.

Cell Lines

All cell lines were obtained from the cell production department at CRUK, Lincoln’s Inn Fields. Kasumi-1 cell line is an established cell line which harbours the t(8;21) translocation. RNA was obtained from this cell line and used as a positive control for PCR and cloning experiments.

Thawing of Cryopreserved Cells

Sample vials were removed from liquid nitrogen and immediately thawed in 37°C water bath. The contents were transferred into a 15ml FALCON tube and 1xPBS medium was added one drop at a time to make solutions of 10mls. These were centrifuged at 1300g for 5 minutes. The supernatants were discarded and 2-5 mls of fresh PBS was added according to pellet sizes. A 10μl sample was kept aside to perform cell count and to assess viability of cells. The samples were topped up with PBS to 10mls and centrifuged at 1300g for 5 minutes. Once again the supernatants were discarded and the pellets resuspended and transferred to a 2ml Eppendorf tubes and spun at 3000g to remove all PBS.

RNA Extraction using TRIzol (Invitrogen) Method

0.1 ml of TRIzol (Invitrogen) per million cells was added (1.0 ml for less than 10 million cells), homogenised with a pipette and incubated at room temperature for 5 minutes. 0.2 ml of chloroform per 1.0ml of TRIzol was added and the mixtures shaken vigorously for 15 seconds before incubating for 2 minutes at room temperature. Samples were centrifuged for 15 minutes at 4°C and 12,000g, which separated the mixture into 3 phases. The colourless upper phase containing the RNA was recovered into separate tubes. RNA was precipitated by mixing with 0.5ml isopropanol per 1ml of TRIzol used and incubated at room temperature for 10 minutes. Samples were centrifuged at 12,000g for 10mins at 4°C to obtain a RNA pellet on the bottom of tubes. The supernatants were removed and the pellets washed with 75% ethanol. Samples were air dried and resuspended with DEPC water and concentrations measured using a spectrophotometer (Agilent Bioanalyzer). (Figure 6)

Ethanol precipitation was then performed to improve the purity of RNA. To the RNA solutions, a 1/10 of 3M sodium acetate pH5.2 and 2.5x cold 100% ethanol were added, mixed and incubated at -20°C overnight. Samples were centrifuged at 12,000g for 20mins and the obtained pellets were washed twice with 80% ethanol. Pellets were air dried and mixed with DEPC water to obtain an appropriate solution.

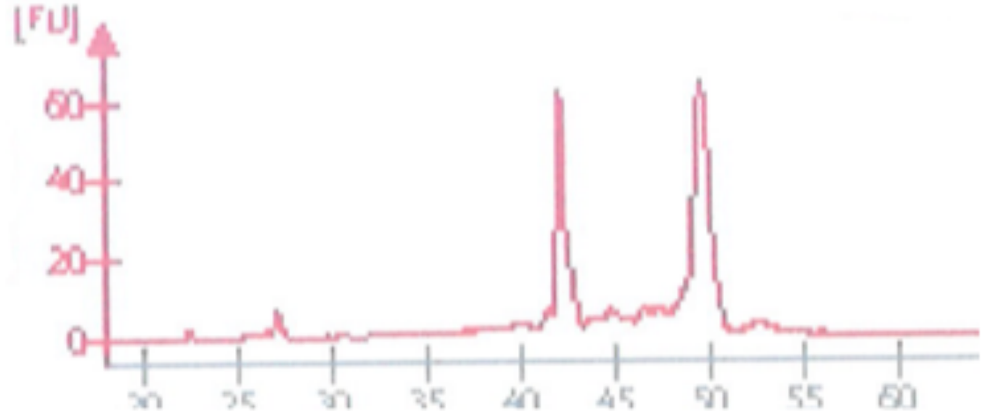


Figure 6 An example of RNA from a primary sample run on an Agilent Bioanalyser to assess quality. The two large peaks correspond to ribosomal RNA.

cDNA Synthesis

Two protocols were used for RQ analysis: (Details of Mastermix in Appendix D)

A mastermix consisting of distilled water, 5x synthesis buffer, 2.5 μM dNTPs, 50u μ M random hexamers 1.5μl and RT was made and added to 1μg/1μl of sample RNA. The mixture was put on a thermal cycler with the following protocol; 25°C for 10 minutes, 42°C for 60 minutes and 95°C for 5 minutes.

100ng of RNA was mixed with 1.5μl of random hexamer (50μM) and the solution made up to 12μl with water. This mixture was put at 70°C for 5 minutes and cooled quickly on ice. A mastermix was made consisting of 5x synthesis buffer, 2.5uM dNTPs and M-MLV Reverse Transcriptase RNase H minus point mutant (Promega). This was added to the RNA/hexamer mixture above. The 30ul solution was placed in a thermal cycler at 42°C for 1hr and 95°C for 5mins.

RT-PCR

Master mix was made of 10x PCR Buffer (5μl), 2.5μm dNTPs (4μl), forward and reverse primers 2μl each, Taq polymerase (0.3μl), distilled water (35.7μl) and cDNA (1μl). The mixture was run on a thermal cycler using various appropriate protocols. (Details in Appendix D)

RQ-PCR

Primers were constructed using the software Primer Express® (Applied Biosystems).

Primers used for the expression assays were obtained from Applied Biosystems and had been validated by them. Primers used for the RQ in the ChIP experiments were subjected to melting curve analysis to check a single product was being produced.

Real time assays were conducted in triplicate and included non-test controls (NTC) as negative controls where distilled water replaced the cDNA. Samples were mixed with Universal Master Mix or SYBR Green (both Applied Biosystems) as required and along with the appropriate probes and primers (Details in Appendix D). Samples were placed into wells, a film applied and the sample spun briefly before being placed in the PCR machine. The assay was performed using the ABI PRISM 7900 Detection System (PE Applied Biosystems). The amplification protocol followed was as follows; 2 minute at 50°C, 10 minutes at 95°C and then 50 cycles of 15 seconds at 95°C and 1 minute at 58°C. 18S (Applied Biosystems) was used as control and values based on cycle threshold (C_T) values for each gene and the controls were generated performed on Excel. A ΔC_T value was derived by subtracting the mean C_T

value of the reference gene from the mean C_T of the target gene. This normalises for variability of amounts of RNA added to RT reactions. A $\Delta\Delta C_T$ is then calculated by subtracting the ΔC_T of the positive control from the C_T of each sample. This allows comparison between plates. Fold change was calculated by 2 C_T. To correlate gene expression log base 2 of $\Delta\Delta C_T$ was plotted against log base 2 of the ratio of normalised microarray values for target gene 18S.

Agarose Gels

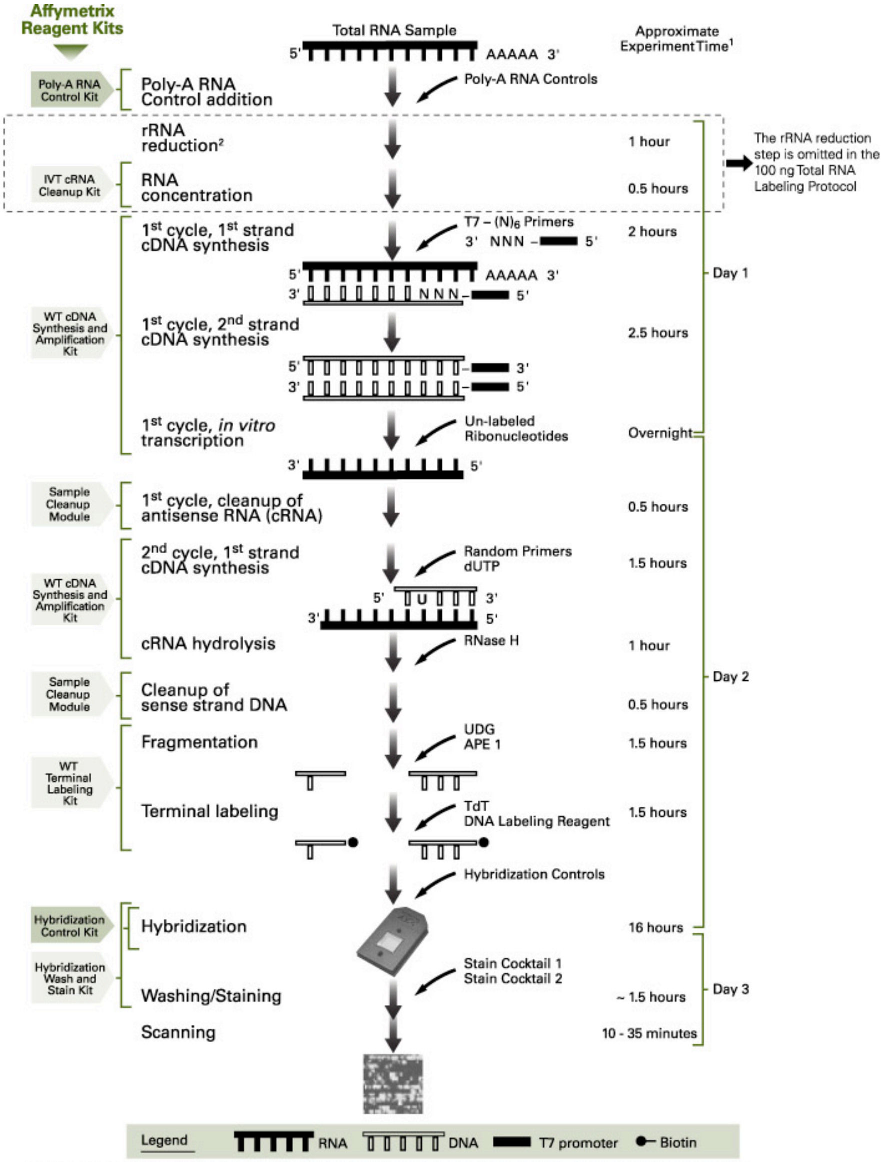
Agarose gels were made of 1 or 2% agarose with TBE. The appropriate volumes were mixed and heated in a microwave loosely covered at medium power. Once the agarose had fully dissolved the solution was cooled and a small amount ethidium bromide was added and mixed. The solution was poured into a mould, a comb added and the gel was left to set. The set gel was placed in a gel tank, the comb removed and the tank filled with TBE buffer to the level of the wells. Samples were mixed with an appropriate amount of loading dye and the mixture loaded into the wells. The gels were run at 80V for an appropriate time making sure that the current flowed in the right direction. The gels were photographed under UV light. TAE buffer was used if the resultant gel was used for cutting out DNA and further ligation experiments as this buffer lacks borate which can interfere with ligation.

II. Exon Array Expression Profile Analysis

RNA Processing

Primary samples were obtained from storage, thawed and RNA extracted. RNA was processed using a whole transcript (WT) assay for to obtain appropriate targets for array analysis (Figure 7). This incorporates ribosomal RNA removal, a random priming strategy, in vitro transcription and a novel fragmentation and labelling method. Single stranded DNA targets are generated in the sense orientation from the entire length of the transcripts.

Although the WT assay may be attractive for profiling partially degraded samples, good quality total RNA is recommended as tested on the bioanalyser (Figure 6). The protocol uses only 1.0μg of RNA due to the incorporation of in vitro transcription amplification step.



[1] Assume that only one sample is carried through the assay. The time may be longer if multiple samples are processed simultaneously.

[2] RiboMinus Human/Mouse Transcriptome Isolation Kit is required for this step, and needs to be ordered directly from Invitrogen.

Figure 7 Summary of RNA processing for Exon Array (from Affymetrix Assay Manual)

Ribosomal RNA removal

As over 90% of all eukaryotic total RNA is accounted by ribosomal RNA with only approximately 2% comprising mRNA, an rRNA removal step is initially used. Four biotinylated RiboMinus probes are used to specifically bind to the 18S and 28S rRNA. Subsequent addition of RiboMinus Magnetic Beads coated with streptavidin result in about 60-80% of rRNA removal. (Figure 8)

Step A- Preparation of Dilutions of poly-A RNA controls-

Two μl of poly-A RNA control stock was added and mixed to 38ul of control dilution buffer to make first dilution. Two μl of this first dilution was added and mixed to 98μl of control dilution buffer to make second dilution. Two μl of this second dilution was added and mixed to 98μl of control dilution buffer to make third dilution. 2μl of this third dilution was added to 1μg of total RNA (RNA concentration greater than 0.31μg/μl) to make the total RNA/Poly-A RNA controls mix.

Step B- Preparation of Hybridisation buffer with betaine

Buffer was prepared with 5M betaine (54μl) and Invitrogen Hybridization buffer (126μl) to make a total of 180μl for each reaction.

Step C- RiboMinus Probe Hybridization

Total RNA/Poly-A mixture from step A and hybridisation buffer with betaine from step B and RiboMinus probe was mixed and incubated at 70°C for 5 minutes. The reaction was quenched by placing on ice. (Details in Appendix D)

Step D- Preparation of Beads

50μl of magnetic beads in suspension were placed in tubes and incubated in a magnetic stand for 1 minute. The supernatant was discarded and the beads washed with 50μl of RNA-ase free water before placing on magnetic stand and discarding the supernatant. This wash procedure was repeated with water and then again for the third time with 50μl of hybridisation buffer (from step B). The beads were then re-suspended in 30μl of hybridisation buffer and incubated for 2 minutes at 37°C.

Step E- rRNA reduction

The ice-cooled sample from step C was mixed with the beads and incubated at 37°C for 10 minutes. The sample was then placed on magnetic stand for 2 minutes and the supernatant containing the rRNA-reduced total RNA/Poly-a control mix was transferred to fresh tube and left on ice. The beads were resuspended with a further 50μl hybridisation buffer and incubated for 5 minutes at 50°C. This was placed on a magnetic tube and the supernatant transferred to the previously acquired tube to give a volume of approximately 100μl.

Step F- Concentration

This was performed with IVT cRNA Cleanup kit. 350μl of binding buffer was added and mixed to sample from step E. 250μl of 100% ethanol was then added and mixed. The mixture was applied to the spin column and centrifuged for 15 seconds at 8000g. The flow through was discarded and a new collection tube applied to spin column. 500μl of wash buffer (with 20ml of 100% ethanol added previously) was added and centrifuged at 8000g for 15 seconds. The spin column was washed again with 500μl of 80% ethanol. The flow through was discarded. The column cap was left open and centrifuged at maximum speed for 5 minutes. The IVT column was transferred to a new collection tube and 11μl RNAase free water was added directly to the membrane and spun at maximum speed for 1 minute. Approximately 10μl of eluate was obtained.

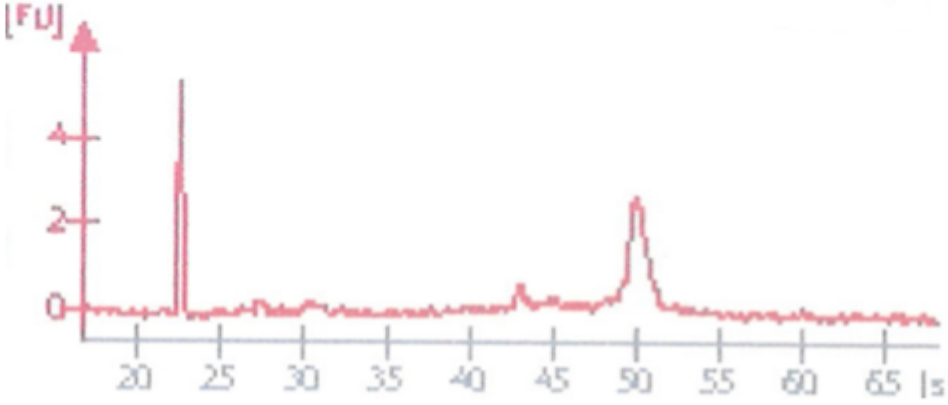


Figure 8 An example of RNA from a primary sample run on an Agilent sphectophotometer after ribosomal RNA removal step of exon array protocol. The two large peaks corresponding to ribosomal RNA are reduced (compared to Figure 6 p 60).

Random Priming

A random-priming strategy, using random primers incorporating a T7 promoter sequence, is used to generate cDNA from all RNA transcripts. (Details of the master mixes used found in Appendix D)

Step A - T7-(N)6 Primers mix

A Whole Transcript cDNA synthesis kit was used. 500ng/μl working solution was made from the stock. 1μl of the primer solution was mixed with 4μl of RNA/Poly-A RNA obtained from step F above. The mixture was incubated at 70°C for 5 minutes and then 4°C for at least 2 minutes.

Step B –First cycle, First-strand cDNA synthesis

A first strand master mix was made and 5μl of this was added to 5μl of the RNA/Primer mixture from step A. The mixture was incubated at 25°C for 10 minutes, 42°C for 60 minutes and 70°C for 10 minutes and cooled at 4°C for at least 2 minutes.

Step C – First cycle, second strand cDNA synthesis

10ul of second strand master-mix was prepared and mixed with mixture from step B. The total 20μl mixture was incubated 16°C for 120 minutes without heated lid and 75°C for 10 minutes with heated lid. The sample was cooled for at least 2 minutes.

Step D –IVT

A 30ul IVT master mix was prepared and mixed with the 20μl solution from step C and incubated for 16hrs at 37°C. The sample was then concentrated using the same methodology as above. The cRNA yield was quantified by spectrophotometer and checked on Agilent (Figure 9).

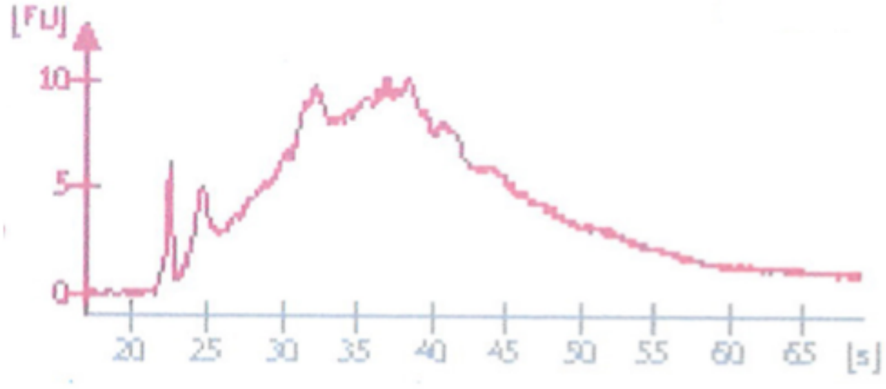


Figure 9 An example of RNA from a primary sample run on an Agilent after IVT step of exon array protocol.

Step E –Second cycle, first strand cDNA synthesis

10μg of cRNA was mixed with 1.5μl of random primers and the mixture was made up to 8μl total volume and incubated at 70°C for 5 minutes and 25°C for 5 minutes. In a separate tube the second cycle master-mix was made and 12ul transferred to the 8μl sample from step D. The mixture was incubated at 25°C for 10 minutes, 42°C for 90 minutes and 70°C for 10 minutes. The sample was cooled for at least 2 minutes.

Step F –Hydrolysis and Concentration

1ul RNase was added to the samples and incubated at 37°C for 45 minutes and 95°C for 5 minutes. The sample was cooled for at least 2 minutes. The mixture was then cleaned and a total volume of 28μl was obtained of single stranded cDNA.

DNA Fragmentation, Labelling and Hybridisation

Fragmentation involves incorporation of a modified dUTP instead of dTTP in the reverse transcription reaction of the second cycle cDNA synthesis. The sense single-stranded DNA strand now has the unnatural uracil base incorporated at predefined intervals and this is treated with a combination of enzymes. UDG specifically removes the uracil residue from single strand DNA molecules and APE 1 then cleaves the phosphodiester backbone where the base is missing, leaving a 3'-hydroxyl and a 5'-deoxyribose phosphate terminus. Thus DNA is cut in specific rather than random locations providing consistency and reproducibility of the fragmentation product.

Step G –Fragmentation

A fragmentation master-mix was made and 48μl was added to the sample from step F and incubated at 37°C for 60 minutes and 93°C for 2 minutes. The sample was cooled for at least 2 minutes.

Step H –Labelling

A labelling reaction was set up using 45μl from the sample from step G. The mixture was incubated at 37°C for 60 minutes and 70°C for 10 minutes. The sample was cooled for at least 2 minutes.

Hybridisation

The labelled DNA target was mixed into a hybridisation cocktail (Appendix D) and given to laboratory for completion of hybridisation and loading onto the chips.

Background Correction

A common set of probes is used for background correction for all probes on the exon array. In contrast the expression array uses a probe specific background extraction method where a specific mismatch for each perfect match probe is selected i.e. on average 11 mismatched probes in each probeset.

There are two collections of background probes used for the exon array with each probe consisting of 25-mer sequences. Antigenomic background probes are collections that are not represented on the human genome (or seven other genomes including mouse and rat) and should not cross hybridise to human sequences. Genomic background probes are mismatch probes whose perfect match counterparts do match the genome although they are from regions in which there are less likely to be expressed. Both collections are organized into 26 bins of varying GC content, from all 25 bases being Gs or Cs to none of the 25-mer sequence containing Gs or Cs. There are approximately 1000 probes for each bin.

These collections of probes are used to estimate the probe-specific background by comparing perfect match probe intensity to the median intensity of all background probes with a matching GC count.

In addition to these background probes other additional critical controls have been represented on the exon array. These include Affymetrix controls, intron-exon controls and unmapped human mRNAs.

* Intron controls — for approximately 100 genes with relatively constitutive expression, both exon-based and intron-based probe sets were tiled. The intron/exon normalization control probe sets can be used to monitor contamination from genomic DNA, RNA as well as to provide a baseline for experiment quality control.

* Hybridization controls — bioB, vbioC, bioD and cre

* Poly-A RNA controls — lys, dap, phe, and thr

The standard Affymetrix expression control sets (i.e., bacterial spikes) including both the hybridization control spikes and poly-A RNA control spikes are present on the exon array.

In addition, the platform has a series of control probes, which hybridise to a series of RNA “spike ins”. The degree of hybridization between the control probes and the “spike ins” is used to control for the hybridisation of target probes.

Exon Array Analysis

Gene expression is estimated by measuring the fluorescence emitted by a labelled mRNA once hybridisation to a probe occurs. The fluorescence detected is compared to the fluorescence emitted by a series of background probes contained on the platform.

Software provided by Affymetrix. GCOS v1.3 is provided to control the 7G Scanner and generate .dat (raw image file relating to pixel values) and .cel files (image file based on intensity calculations). DABG stands for "detection above background" and is a detection metric generated by comparing perfect match probes to a distribution of background probes. This comparison yields a p-value that is then combined into a probe set level p-value. (Figure 10)

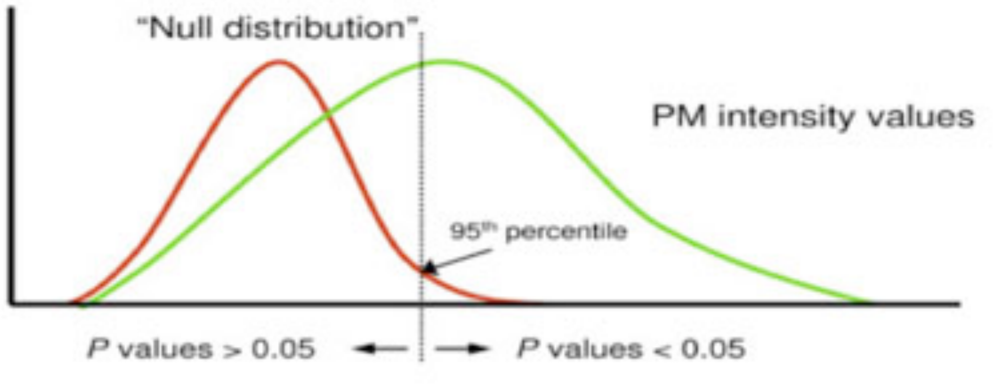


Figure 10 A comparison of background signal to perfect match (PM) signal. PM signal intensities at 95% of background probes with the same GC content are given a p-value of 0.05.

Normalisation is a mathematical algorithm, which is designed to correct for experimental error and account for outliers thus removing systemic bias. A number of algorithms such as PLIER (Probe Logarithmic Intensity ERror) and RMA (Robust Multi-array Average) are available. Applying these algorithms to both background and test values, results in an expression value for each individual exon.

Expression data generated can be retrieved in a number of ways: at the exon level, generating a value for each individual probeset (probeset ID) or at the gene level generating an expression value for each gene based on the measurements of all its exons (transcript cluster ID). Data can also be retrieved using any one of three different metaprobeset files, which are based on confidence levels of the probe sets.

Subsequent analysis requires third party software to assess and interpret the signals. Affymetrix provides a visualisation software tool, the Integrated Genome Browser (IGB) to inspect signal intensities. Our in-house GOLF software was used to visualise probeset intensities. Most of the analysis was performed using a commercial software programme Partek™. This programme requires the input of the .cel files and uses an RMA normalisation algorithm. For detecting alternative transcripts Partek™ uses an ANOVA alternative transcript algorithm based on the equation:

$$y = \mu + E + T + T*E + S(T) + \varepsilon$$

y= exon expression, μ = exon expression, E = exon expression, T = tissue, S = array, ε = error factor

III. Novel Transcript Identification

Cloning & Transformation

The Ta Topo Kit C (Invitrogen) was used according to the manufacturer's instructions. The salt solution and vector provided in the kit were defrosted slowly on ice. 1ul of vector and 1ul salt solution were mixed with 4ul of the PCR product to be cloned. This was incubated for 5 minutes at room temperature and 2μl of this solution was added to a vial of chemically component E.Coli cells (Invitrogen) that had been slowly thawed on ice. This was mixed gently and kept on ice for 30 minutes. The cells were subjected to heat shock by placing the vials in a water bath at 37°C for 30 seconds and then cooled on ice for 2 minutes. 250μl of SOC medium (Invitrogen) was added and shaken at room temperature for 60 minutes. 100ul solution was plated out and incubated at 37°C overnight. White colonies that indicate the incorporation of the vector were selected for growing up and DNA purification.

Mini-Preps

Selected colonies from the cloning process were added to 2ml of LB medium (CRUK) containing ampicillin and incubated overnight at 37°C. The DNA extraction was performed using the Wizard Plus SV Minipreps DNA purification system (Promega) (See Appendix D for details of buffers). The solutions were centrifuged for 5mins at 10,000g and the supernatants discarded. Pellets were resuspended in 250μl of cell resuspension solution. 250μl of cell lysis solution was added and mixed by inversion until cell suspension cleared. Vortexing the mixtures were avoided to minimise chromosome shearing. Alkaline protease solution (10μl) was added to inactivate endonucleases and other proteins released during cell lysis. This was incubated for no longer than 5 minutes to prevent nicking of plasmid DNA. 350μl of Neutralization solution was added and mixed to stop lysis. The bacterial lysate was then centrifuged at maximum speed for 10 minutes. The cleared lysate was transferred to the spin column and centrifuged for 1 minute at room temperature. The follow through was discarded, the column washed twice with wash solution (previously diluted with 95% ethanol) and the spin column transferred to a new sterile tube. Plasmid DNA was eluted by adding 100μl nuclease free water and centrifuging at maximum speed for 1 minute. The DNA was checked by digesting with appropriate enzymes and running on an agarose gel. The DNA was stored at -20°C. Glycerol stocks were made from the positive clones for long-term storage at -80°C by mixing 850μL of culture with 150μL of glycerol (CRUK).

Restriction Enzyme Digestion

The mixture set up for digestion included distilled water, appropriate buffer, DNA and the relevant enzyme. Digestion was performed in a water bath at 37°C for 2 hours. Products were run on a 1% gel at 80V and made with TAE buffer rather than TBE buffer if fragments were going to be used for ligation assays. (Details in Appendix D)

Gel Extraction

This was performed using the MinElute Gel Extraction Kit (Qiagen). The DNA fragment was excised from the agarose under UV light. The piece was weighed and x3 volume of Buffer QG (solubilisation and binding buffer) added. This was incubated at 50°C for 10 minutes until it had solubilised completely. 10μl of 3M sodium acetate, pH 5.0 was added to ensure acidification and efficient DNA adsorption onto membrane. One gel volume of isopropanol was added, mixed and applied to the MinElute column provided. This was centrifuged for 1 minute, the flow-through discarded and Buffer QG added to the spin column and centrifuged for 1 minute. The flow through was discarded and the column washed with 750μl of Buffer PE (wash buffer). After discarding the flow-through the column was centrifuged for an additional minute. The column was placed in a clean 1.5ml tube and the DNA was eluted by adding 10μl of Buffer EB (10mM Tris.Cl, pH 8.5).

Agar Preparation

Solutions of agar were heated in the microwave at low power for 20minutes. When the agar had cooled to approximately 50°C, 400μl ampicillin and 400μl Xgal was added. The agar was poured onto plates, cooled and labelled.

Ligation

This protocol used the T4 DNA Ligase (New England Biolabs). The DNA products were mixed with enzyme and enzyme buffer and incubated according to manufacturer's instructions. (Details in Appendix D)

Sequencing

Samples (400ηg of plasmid DNA or 200ηg of PCR product) and primers (10ρmol/μl) were provided to Department of Genomics at Institute of Cancer for sequencing.

Western Blot

This was performed with direct supervision from J. Dunne. (Details of buffers and antibodies in Appendix D) (D. Gascoyne provided the lysates)

Protein Electrophoresis

Samples were mixed with x4 loading buffer and x10 reducing agent. The samples were then incubated at 70°C for 10 minutes to denature protein and then placed on ice for 5 minutes. An electrophoresis tank was set up and the samples were loaded into the wells of the NuPage gel (Invitrogen) alongside a rainbow marker (Amersham). The tank was filled with Running Buffer and 500μl antioxidant (Invitrogen) added to the buffer. The gel was run at 200V for 50 minutes.

Protein Transfer

The protein was transferred to nitrocellulose. A sandwich was constructed using the following layers noting the importance of getting the order of the layers right. On saran paper add foam, then x2 Whatman paper, the gel (face up), the nitrocellulose membrane then Whatman paper and then foam. These components have all been soaked in running buffer and after each layer added the sandwich rolled to remove air bubbles. In addition the nitrocellulose is labelled to allow easier orientation. The sandwich is placed in a tank and filled with running buffer. It is important to face the gel as the cathode and the nitrocellulose as the anode. The transfer was run for 90 minutes at 25V.

Probing and Detecting

Initially, the membrane was incubated in 8μl of blocking buffer for 30 minutes. The membrane was then incubated overnight with antibody in 8ml BSA. The antibody was washed off with three 5-minute washes of TBS-T. The membrane was then incubated with a second antibody mixed in binding buffer for 30 minutes. The antibody was washed x3 with TBS-T. All incubations and washes were performed on a rocking shaker at room temperature. Detection of protein was performed using the SuperSignal West Dura Substrate kit (Pierce Biotechnology). On a saran film 500μl of substrate1 was mixed with 500μl of substrate 2. The membrane was incubated for 5 minutes. The excess was blotted off and the cellulose exposed to film for varying length of times and the film developed using an automatic SRX-101A (Konica Minolta).

Retroviral Transduction/Transformation Assay

This was wholly performed by Duncan Gascoyne and is subsequently only briefly described to outline the method.

Plasmid DNA was transfected into the LinXE ecotropic retrovirus packaging cell line18. The supernatant, containing newly formed retrovirus was harvested after 48 hours. c-Kit+Ter-119– hematopoietic progenitor cells (HPCs) were purified from murine foetal liver by fluorescence-activated cell sorting (FACS). HPCs were infected with retrovirus and subjected to methylcellulose re-plating assays.

Methylcellulose Re-plating assay

D. Gascoyne performed the initial assays but latter ones were performed by the author. Colonies were counted and morphology assessed. Cells were then harvested with 2mls of warmed (to 37°C) FCS medium. The sample was spun at 1200g for 5 minutes. The supernatant was drawn off and the pellet resuspended in 1ml of medium. Cell counting was performed using a counter. 30,000 cells were re-plated by making upto 300μl of FCS and adding to methylcellulose containing 3.3μl GM-CSF. The mixture was shaken and left to settle before transferring 2.2ml into 2 plates and grown for 6-7 days before counting and harvesting.

FACS

Cells not used for re-plating were spun down at 12000g for 5 minutes. The supernatant was removed and $5x10^{6}$ cells were resuspended in 10% blocking buffer. 100ul was aliquoted and incubated for 10 minutes at room temperature. The plate was spun at 1000g for 5 minutes and the supernatant removed. The cells were resuspended in the first antibody, anti-human CD2 biotin at 1 in 100 dilution, and incubated for 30 minutes at 4°C. The antibody was washed with 2% FCS/PBS and centrifuged for 5 minutes at 1000g and supernatant removed. The cells were resuspended in 30μl of the 2nd antibody (strep-apc) at 1 in 1000 dilution and incubated at 4°C for 30 minutes. The sample was washed in FCS, resuspended in 100ul of FCS and transferred to a FACs tube. The sample is made upto 400μl with FCS and analysed on LSR cytometer (BD). Analysis was performed on FlowJo (Tree Star Inc) software.

IV. ChIP-Seq

ChIP

In summary, cells were treated with formaldehyde to cross-link DNA-protein interactions. Cells were then lysed, sonicated and incubated with specific antibody overnight. The antibody was immobilised onto sepharose beads and after washing eluted. Samples were then heated to 65C to reverse cross-link the DNA-protein bonds. Protein was digested and the DNA was extracted with phenol/chloroform.

Cells were obtained from storage, defrosted as previously described, washed and counted. 5 x 10^7 cells were used per treatment. Cells were treated in 10 mls medium with 270ul 37% formaldehyde (fresh Sigma, F8775). Flasks were swirled and then incubated for 10 min at 37ºC. Cells were washed in ice-cold PBS in a 50ml Falcon and then washed a final time in PBSA containing protease inhibitors in 1.5ml tubes. Samples were centrifuged 1000g for 5 minutes to pellet cells and the supernatant removed. Cells were lysed in 500μl lysis buffer and the DNA was then sheared by sonication. Sonication was performed using the Vibra-Cell VCX-500 (Sonics) according to the manufacturer's instructions. This was performed in an ice bath using an amplitude of 30% and intervals of 10/30 seconds on and off respectively.

After sonication samples were centrifuged for 10 minutes at 13000g 4ºC and the supernatant removed to new cold Eppendorf tube. 50μl of this was removed as input sample and 200μl of IP dilution buffer was added. The mixture was stored at 4ºC until required. The remaining 450μl lysate was divided into 2 x 225μl samples in fresh screw-cap Eppendorfs and 900μl IP dilution buffer was added to each tube. To these samples the test antibody and a negative control IgG antibody were added respectively and incubated overnight at 4°C with rotation.

To immobilise the antibody 50μl protein G/sepharose (washed in IP buffer) was added to each tube for 1 hr at 4°C. The beads were spun down and subjected to a series of washes with 3 wash buffers and final wash with TE. 250μl elution buffer was added to each pellet of beads, vortexed briefly and incubated for 15min with rotation at room temperature. After spinning, the supernatant was stored and the beads were further incubated with elution buffer with a further 250μl for 15 minutes. The eluates were combined to make 500μl and 20μl of 5M NaCl was added.

All eluate samples and input samples were then heated at 65°C for 4 hrs or overnight to reverse the cross-links between DNA and protein. To digest all traces of proteins 20μl 1mg/ml Proteinase K to eluates were added to all the samples. 10ul 0.5M EDTA and 20ul 1M Tris pH6.8 was also added to non-input samples. Samples were then incubated at 45°C for 1 hr or overnight.

The input sample volume was increased to 500μl by adding 250μl of water. All samples were then extracted with 250μl phenol / 250μl chloroform following by a further 500μl chloroform extraction**.** An extra sonication using 3 cycles step was performed at this stage. To precipitate DNA 1 ml ice-cold ethanol and 1/10 volume Ammonium Acetate was added. To visualise the pellet of DNA upon precipitation

20μg glycogen (1μl, Roche) was also added to each tube and incubated at -20 for 30min. The mixture was spun for 10min at 13000g 4°C, supernatant removed and air dried. Input samples were resuspended in 75μl whilst ChIP samples were resuspended in 30μl of TE. (Details of buffer used found in Appendix D)

High Throughput Sequencing

Library Preparation

The DNA molecules are end repaired using, Klenow and DNA polymerase to digest 3' overhangs and fill in recesses as well as a polynucleotide kinase (PNK) to phosphorylate the 5' end. A poly "A" tail is added using ATP and Klenow exo minus, which lacks the 3' to 5' exonuclease activity. Specific adaptors are then ligated to the DNA. Importantly, the prepared DNA is then size selected before PCR amplification is performed, with the reaction limited to 18 cycles, well before the plateau phase, in an attempt to keep the DNA quantification in correct proportions.

End Repair

The following reaction mix was prepared: ChIP enriched DNA (30μl), Water (10μl), T4 DNA ligase buffer with 10mM ATP (5μl), dNTP mix (2μl), T4 DNA polymerase (1μl), Klenow DNA polymerase (1U/μl) (1μl) and T4 PNK (1μl)**.** The 50μl mixture was incubated in a thermal cycler for 30 minutes at 20ºC and purified on one QIAquick column, eluting in 34μl of EB according to the instructions in the QIAquick PCR Purification Kit.

Add “A” Tail

The following reaction mix was prepared: DNA sample (34μl), Klenow buffer (5μl), dATP (10μl) and Klenow exo (3’ to 5’ exo minus) (1μl). The 50ul mixture was incubated in a thermal cycler for 30 minutes at 37°C and purified on one MinElute column, eluting in 10μl of EB according to the instructions in the MinElute PCR Purification Kit.

Ligate adaptors

The adapter oligo mix was diluted 1:10 with water to adjust for the smaller quantity of DNA. The following reaction mix was prepared, DNA sample (10μl), DNA ligase buffer (15μl), diluted (1 in10) adapter oligo mix (1μl) and DNA ligase (4μl). The 30μl mixture was incubated at room temperature for 15 minutes and purified on one MinElute column, eluting in 10μl of EB according to the instructions in the MinElute PCR Purification Kit.

Size Selection

A 2% agarose gel with TAE buffer was prepared. The sample was loaded alongside a 100bp ladder. Only one sample was loaded per gel. At least one empty lane was left between the ladder and the sample. The gel was run at 120 V for 60 minutes. A region of gel between 300 bp +/- 25bp was excised with a clean scalpel. A QIAGEN Gel Extraction Kit (QIAGEN, part # 28704) was used to purify the DNA from the agarose slices and eluted DNA in 36μl.

PCR

The following PCR reaction mix was prepared: DNA (36μl), 5x Phusion* buffer (10μl), dNTP mix (1.5μl), PCR primer (1μl), PCR primer (1μl) and Phusion* polymerase (0.5μl). The following PCR protocol was used: **a.** 30 seconds at 98°C **b.** 18 cycles of: 10 seconds at 98°C 30 seconds at 65°C 30 seconds at 72°C **c.** 5 minutes at 72°C **d.** Hold at 4°C. The MinElute PCR Purification Kit was used according instructions to purify the sample, eluting in 15μl of EB.

Cluster Formation

This takes place on a specific Illumina provided cluster station. The DNA sample is applied to a flow cell. This is an eight-lane optically transparent glass surface, which contains on its surface a bed of adaptors and its complementary adaptors. Separate DNA libraries can be used for each lane. The DNA is applied to the lane, denatured and as it flows over the cell is immobilised to the surface of cell through its adaptor. DNA molecules can then form bridges, hybridising its free end to the complementary adaptor immobilised on the surface. Reagents are passed over to allow a PCR type amplification with the immobilised adaptors acting as primers. After several cycles random clusters containing 1000 copies of single stranded DNA are produced (Figure 11).

Libraries were initially checked for quality and size on Agilent bioanalyser. RQ analysis was performed by the Genome Centre to quantify the library. Appropriate volumes of samples were loaded onto the machine as per protocol (details available from Illumina). Accurate quantification allows for a high number of clusters to be formed whilst ensuring that the maximum cluster generation capability was not surpassed. This would otherwise, result in overlapping clusters, leading to poor image intensity of the sequenced reads.

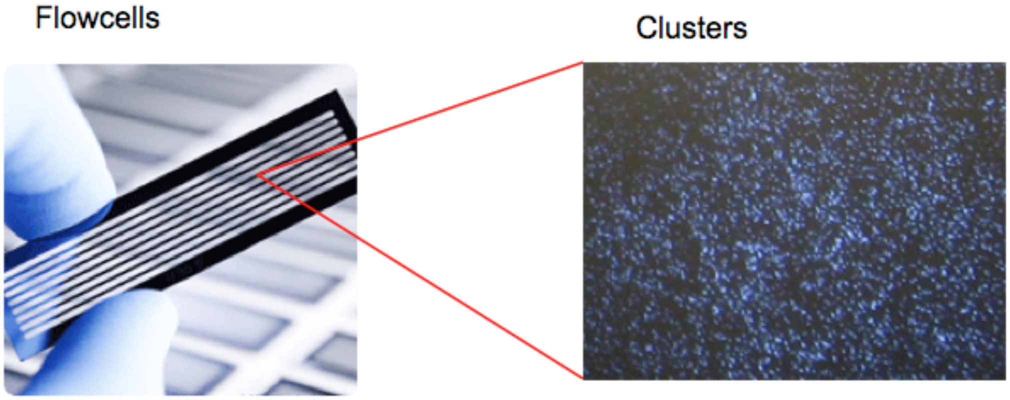


Figure 11 A flow cell demonstrating the eight lanes. And an example of cluster generation post bridge amplification. (Figure from www.illumina.com)

Sequencing

This uses a sequencing-by-synthesis chemistry. All four nucleotides, which are chemically blocked at the 3’ end and labelled with a different fluorescent dye, are added along with a special DNA polymerase. After each occasion a base is incorporated, the image is read. The blocking group is then chemically removed and sequencing reaction repeated to incorporate the next base. The reading sequences currently give up to 50 bases of sequence with plans for longer sequence runs being implemented. Images are converted to sequence files that are suitable for analysis. The sequences are filtered and then aligned to a reference genome. The aligned data can then be analysed for peak detection.

Paired End Reads

Paired end reads were used in this project allowing the opposite end of the same read to be sequenced. The use of paired ends allows easier and more complete assembly. It allows reads that go across repetitive regions to be sequenced as well as allowing structural re-arrangements to be mapped. Although more mappable the disadvantage of paired-end module is that overall there are less reads. The paired-end module directs the regeneration and amplification operations to prepare the templates for the second round of sequencing. First, the newly sequenced strands are stripped off and the complementary strands are bridge amplified to form clusters. Once the original templates are cleaved and removed, the reverse strands undergo sequencing-by-synthesis. The second round of sequencing occurs at the opposite end of the templates. (Figure 12)

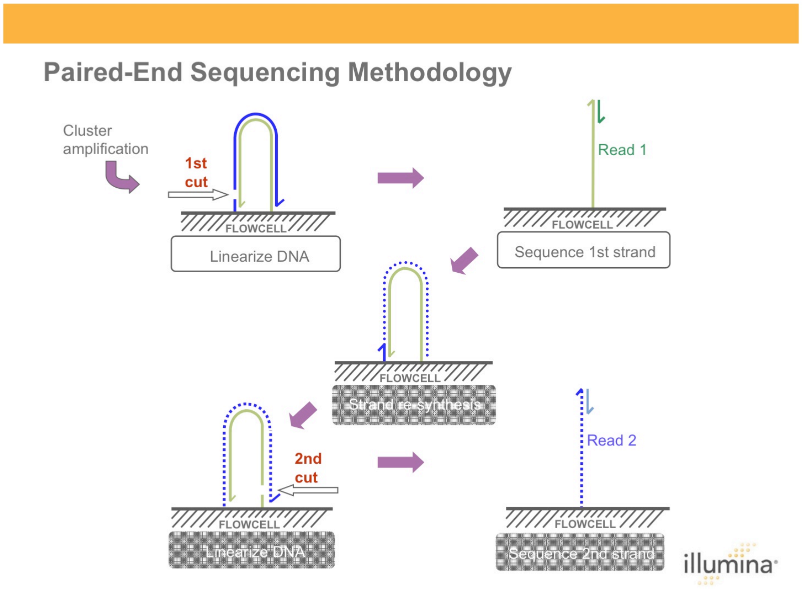


Figure 12 Paired end sequencing (Details in text) (figure from www.illumina.com)

Result Analysis

Illumina provides an Integrated Primary Analysis and Reporting system (IPAR) to perform image analysis. The Genome Analyzer Pipeline Software (Pipeline) version 1.4 is used to process the IPAR image files. The Pipeline is used to convert the raw image into base calls and intensity scores. It also provides a quality metric comparing it to a reference genome and results in the production of sequence files in the “fastq” format that are suitable for further downstream analysis. This initial data is generated by the sequence run by the Genome centre.

The freely available software Bowtie was used for alignment and filtering of the generated reads against a reference genome. Initial filtering involved removing polyA reads (>20). The reads were then mapped to the human genome (build 36.1, hg 18) allowing a maximum mismatch number of 2 (in the first 28 bases) to account for possible polymorphisms in the genome and any unmapped reads were filtered out. Reads mapping to more than 1 location on the genome, which suggests areas of repeat sequences, were also removed. The limited PCR amplification step in the formation of the library may introduce errors as reads may undergo amplification bias. To account for this potential PCR bias, duplicated reads were assumed to be artefacts and only one read was kept. The Illumina base quality >30 indicates that the probability of the base being called incorrectly is <0.001 and reads not passing this QC metric (<30) were also filtered out. (Figure 13) This Bowtie assessment was performed by bioinformatics team.

Aligned, filtered reads were then analysed using the commercial software Partek^{TM} to detect peaks and motifs. Aligned reads generated from Bowtie were converted to a text file and imported into the software. Peak detection is performed by the software utilising data from both the forward and reverse strands. A set window size is used to count peaks in each window. Numbers of overlapping reads in a region are counted. The software also generates two p-values; one the binomial p-value for comparing the test sample and the input; the other is a Mann-Whitney p=value which compares the degree of separation of the forward and reverse reads. The peaks detected are annotated enabling identification of genes. A visualisation interface can be used to inspect peaks and gene locations. The Partek analysis and subsequent interpretation and filtering was performed by myself.

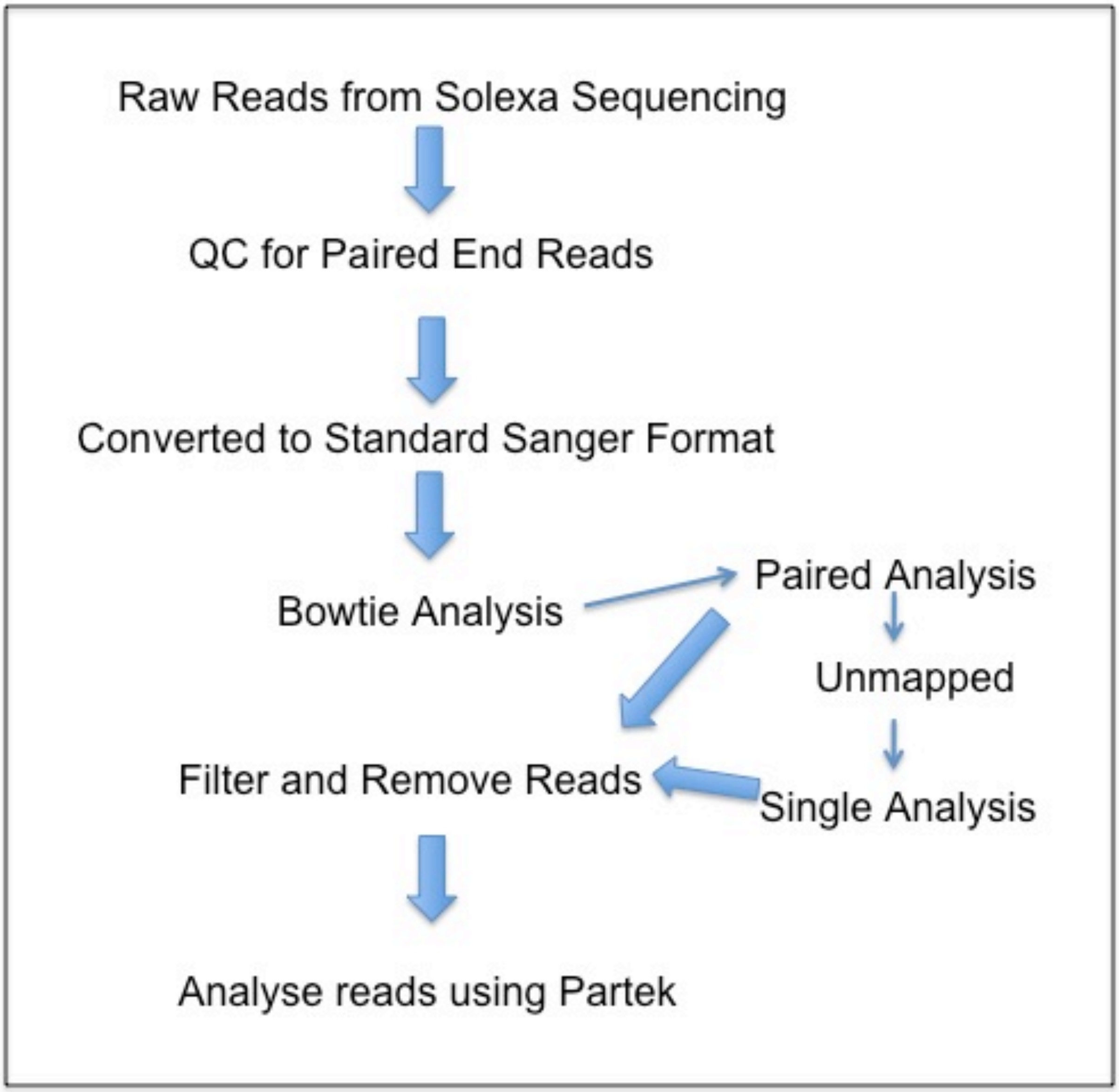


Figure 13 Flow diagram of Analysis Pathway (see text for details)

Chapter 3 Results

Gene Expression Profiling using Exon Arrays in Acute Myeloid Leukaemia with t(8;21)

Introduction

Array technology is used as a fast, high throughput method to explore genomes in various disease states. They serve as hypothesis-generating tools leading to a greater understanding of disease pathogenesis at the molecular level. The underlying basis for expression arrays is relatively simple and relies on hybridization of amplified and labelled RNA/cDNA from target genes, obtained from the test sample, to specific probes. These probes, which are either cDNA or oligonucleotides, are arrayed and immobilised onto a platform. Detection of fluorescence of hybridized targets and subsequent computational analysis provides a single estimate of gene expression for every gene.

The use of gene expression profiling in AML has been widely studied. Gene expression profiling can be used to classify AML subtypes related to major cytogenetic classes (Debernardi et al., 2003). They have also been used to discover novel subclasses and used as prognostic classifiers (Bullinger et al., 2007).

However, standard expression arrays are prone to a number of limitations. The paradigm that one gene encodes one protein has proven to be an oversimplification. With the number of genes discovered in the human genome proving to be less than anticipated, alternative splicing is believed to play an important role in increasing the diversity of proteins produced from the genome, with studies suggesting that 73% of human genes are alternatively spliced (Lee and Roy, 2004).

To address this, modifications to expression arrays have resulted in the introduction of exon arrays, which aim to provide complementary analysis of both expression data and alternative splicing events.

There are many differences between exon arrays and expression arrays ranging from the platform, the probe design and the chemistry used to prepare samples for hybridization (Table 8). Exon arrays are exon focused rather than gene focused with each exon an individual target of expression analysis. Furthermore, many hypothetical exons, in addition to well-established exons, are represented on the exon array, allowing for novel exon discovery. These changes have been facilitated by an increase in the number of probes available to interrogate the target. The Human Exon 1.0 ST array has 1.4 million probesets compared to 54,000 probesets on the Affymetrix U133 plus 2.0 array.

I set out to investigate the t(8;21) using this novel platform. To understand the techniques used in the analysis a description of the design of the exon arrays is described in the next section.

Table 8 Differences between Exon arrays and Standard Expression Arrays

	Human Exon 1.0 ST Array	Human U133 Plus 2.0 Array
Probe sets	1.4 million	54,000
Supported by putative full-length mRNA	289,961 probe sets	N/A
# of perfect match probes for each probe selection	4	11
Number of probes for each RefSeq Sequence	Median 30 - 40	11
Probe selection region location	Along the entire length of the transcripts	Most 3' end
Probe selection region length	Median 123 bp	600 bp
Nucleic acid Hybridisation	cDNA	cRNA
Interrogated strand	Sense	Antisense
Priming	Random Priming	Priming from 3' end
Fragmentation	Specific	Random
Background subtraction strategy	Median intensity of up to 1,000 background probes with the same GC content	One mismatch probe for every PM probe
Detection Call	% above background	% present

Exon Arrays - Probe Design & Annotations

A Probe Selection Region (PSR) represents the smallest region of the genome that is predicted to act as an integral, coherent unit of transcriptional behaviour and in many cases usually represents a single exon. In some cases, several PSRs may form subsets of a true biological exon as may occur with potentially overlapping exon structures (Figure 14). PSRs act as the target sequence from which probes are designed. In the final design, there are approximately 1,400,000 PSRs.

The median size of a PSR is 123bp with a minimum of 25bp. Each PSR is represented by an individual probe set. 90% of the probe sets contain 4 probes with the remaining 10% split between 3, 2, and 1 probes for each probe set. Probe sets span the entire gene rather than simply the 3' end as in expression arrays and result in a single measure for each unique exon sequence. Most probe sets map to a unique location in the putative transcriptome. Probes within each probeset frequently overlap with each other. Probes are selected to hybridise to DNA targets and to targets in the sense orientation unlike the expression array. Therefore, samples prepared for expression arrays cannot be used for exon arrays. The median number of probes for each gene is 30-40 although many genes have more than a hundred probes on the exon array. This high-density coverage potentially offers a sensitive and a statistically robust measure of gene expression quantification.

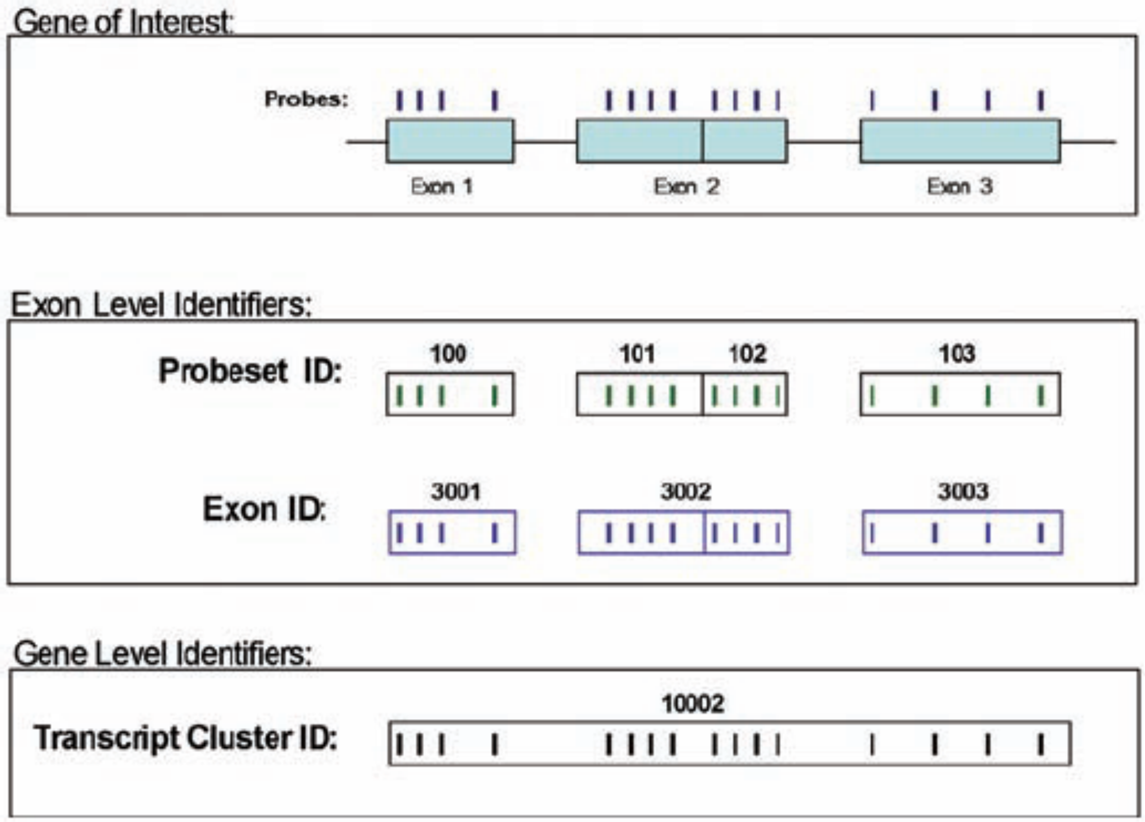


Figure 14 Probes in relation to Probesets/Exons/Transcripts. This sample gene has three exons with exon 2 containing a long and a short form. Affymetrix produce 4 probesets each with 4 probes. An expression value is generated for each probeset (exon level data). Probesets can be mapped to exons (Exon ID) or genes (Transcript Cluster ID). Expression across the transcript is known as gene-level expression. (Goncalves, 2007)

It is estimated that the genome database contains approximately only 10% of the known transcripts. Therefore, in the design of the PSRs, sequences identifying both characterized and putative exons have been selected, reflecting the exploratory nature of these arrays (Table 9).

Table 9 Summary of Annotations used to establish PSRs. Sequences from human genomes, animal libraries and predicted gene sets have been used. In addition mitochondrial RNAs (87 exons) based on mitomap and approximately 190 unprocessed human microRNA sequences from the Sanger MicroRNA registry are also represented on this array.

Established Annotations	-cDNA	human refseq mrnas, genbank mrnas est from dbest.
	-syntenic cDNAs	mouse and rat genomes mapped to human using genome synteny maps (UCSC Genome)
Predicted Annotations	Geneid, Genscan, Vega, Ensembl, Exoniphy, RNAgene, SgpGene TWINSCAN, Mitomap, microRNA	

To delineate relationships between exons and genes there is a post design process, which can map probe sets (exon level) into transcript clusters (gene level). Mapping of probesets are defined by metaprobeset lists, which are based on the confidence level of supporting evidence of the probesets. This gives rise to core, extended and full metaprobeset lists (Table 10). Consequently, looking for new splicing events may require use of the full metaprobeset file that includes predicted sequences whilst other queries may require established annotated exons needing the core metaprobeset file. Using the core metaprobeset file at gene level means that exon arrays can be used to accomplish the same analysis as expression arrays.

Table 10 Evidence used to define Metaprobeset Lists - Of the 1.4 million probesets comprising the full probe sets, approximately 290,000 are supported by full-length mRNAs (core set) and 800,000 are part of the extended set.

CORE	refseq full length genbank mRNAs	
EXTENDED	cDNA transcripts, syntenic rat and mouse RNA and Ensembl, micro RNA, Mitomap, Vegagene and VegaPseudogene annotations	
FULL	ab-initio predications from	Geneid, Genscan, GENSCAN Suboptimal, Exoniphy, RNAgene, SgpGene, TWINSCAN

Summary

The clinical outcome of the t(8;21) AML suggests a comparatively favourable outlook. However, this is an oversimplification with long term survival of around 50%, which is only slightly higher compared to standard risk AML (Marcucci et al., 2005). Furthermore, it is noted there is a particularly poor response to treatment following relapsed disease.

Initially, these clinical findings were confirmed by analysing a sample from our local population. The findings validated our approach to further investigate the subset of t(8;21) AML by gene expression profiling.

Using the Human Exon 1.0 ST array for gene expression profiling insights into the pathobiology of the t(8;21) were provided by identifying critical pathways and candidate oncogenes. In addition, by using this newer platform, novel exons and alternative transcripts of specific genes related to t(8;21) were identified.

The data was analysed by two methodologies identifying two separate aspects of how exon arrays may be analysed. One analysis inspected the *ETO* gene specially and showed that exon arrays can be used to identify genomic breakpoint of translocations.

The second analysis involved global gene expression profiling and several observations were made, generating a number of hypotheses as to the underlying mechanisms of this translocation. The results validated the exon array as a suitable platform for generating expression data. It also highlighted novel genes whose expression is differentially expressed in t(8;21). Using functional gene annotation, pathways and processes implicated in the pathogenesis of the t(8;21) were described.

Furthermore, the use of the exon array platform has further advantages compared to standard expression arrays. This enabled us to explore the data at an individual exon level to look for alternative transcripts as well as potentially to look for novel genes and in particular miRs. Our observations implicate genes whose splicing may be controlled by the fusion gene product AML1-ETO.

The results of these findings are intriguing and as well as implying potential novel mechanisms of leukaemogenesis and suggest approaches to treatment strategies.

Clinical Results

A retrospective analysis of clinical data from St. Bartholomew's Hospital from the last ten years regarding the outcomes for patients with t(8;21) and inv(16) AML was performed.

The total numbers in each group were small with 10 and 15 patients for AML with inv(16) and with t(8;21) respectively. The 5-year overall survival was approximately 50% for both groups (Figure 15).

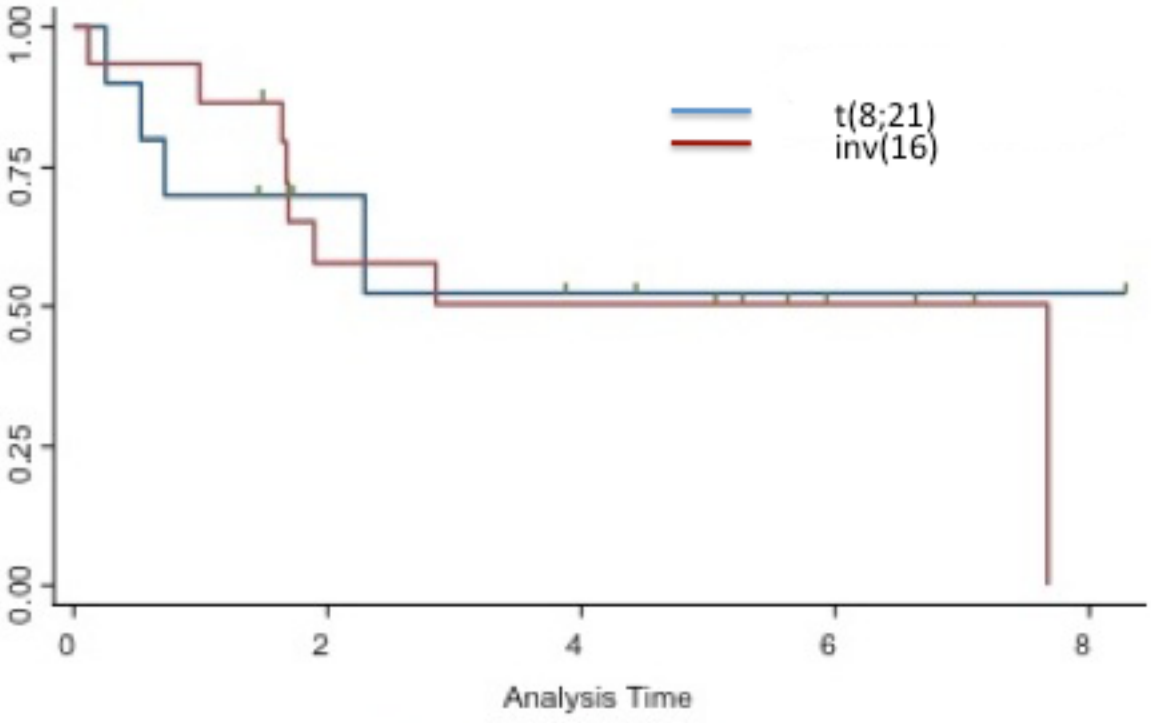


Figure 15 Overall survival for t(8;21) and inv(16) patients from Barts Hospital (n=15 & 10 respectively) with y axis depicting percentage of patients and x axis depicting time in years. (Produced by Finlay Macdougall)

However, the overall survival for relapsed CBF patients is poor. Within each subgroup only three patients with inv(16) and four with t(8;21) had disease relapse. Two of three inv(16) patients had further treatment, both achieving a CR and one long term remission. Three of the four t(8;21) patients had further treatment with two achieving remission but both having a further relapse. With such low numbers it is difficult for any comparison of the two groups to achieve statistical significance.

Overall, these findings are in keeping with current outcomes. We conclude there is still a critical need to investigate the molecular basis of t(8;21) and in particular regarding disease relapse.

Patients and Methods used for Exon Array Analysis

Patient samples were initially, selected based on their FAB subtypes and karyotype. (Eleven with FAB M2 and t(8;21) and nine controls with FAB M2 and normal karyotype) (Table 11) Further selection of patients was based on amount of material stored such that there was a bias to high white cell counts.

Table 11 Patient Characteristics – Age measured in years, disease free survival (DFS) defined as time from diagnosis to time to relapse, overall survival (OS) defined as time from diagnosis to death from any cause both in years

SEX	AGE	TISSUE	BLAST %	FAB	KARYOTYPE	DFS	OS
F	24	PB	98	M2	46,XX,t(1;6)(p36;p23),t(8;21)(q22;q22)/46,idem,der(22)t(1;22)(q23;p11.2)/47,idem,+8	0	0.49
M	51	BM	51	M2	46,XY,del(7)(q32q36),t(8;21)(q22;q22)	0.65	1
M	20	PB	N/A	M2	45,X,Y,t(8;21),add(14)(q32)	0.2	0.7
M	67	PB	30	M2	46,XY,t(8;21)(q22;q22)	0	0.2
M	67	BM	60	M2	46,XY,t(8;21)(q22;q22)	19	19.2
F	49	N/A	N/A	M2	46,XX,t(8;21)(q22;q22)	N/A	N/A
F	38	PB	70	M2	46,XX,t(8;21)(q22;q22)	0.82	12.8
M	34	PB	N/A	M2	45,X,-Y, t(8;21)(q22;q22)	N/A	N/A
F	27	PB	88	M2	45,X,-Y,t(8;21)(q22;q22)	9	9
M	68	BM	50	M2	45,X,-Y, t(8;21)(q22;q22)	3.3	3.3
M	18	PB	50	M2	45,X,-Y, t(8;21)(q22;q22)	11.6	11.7
M	56	PB	53	M2	46,XY	0.4	1.8
M	53	N/A	N/A	M2	46,XY	N/A	N/A
M	18	BM	80	M2	46,XY	2.2	2.3
F	72	PB	75	M2	46,XY	0.3	1.2
M	60	BM	62	M2	46,XY	0	1
F	44	PB	56	M2	46,XX	0.4	1
F	57	PB	75	M2	46,XX	1	1.8
M	51	PB	35	M2	46,XY	0	0.2
F	50	N/A	N/A	M2	46,XX	N/A	N/A

RNA was extracted using the TriZol method (see Chapter 2 p62) and the quality tested on the Agilent bioanalyser. Samples were tested for the t(8;21) by RT-PCR to confirm the translocation was expressed (Figure 16). RNA was then processed and hybridised to the array (Chapter 2 p65). The array results were analysed as described and assessed in two main ways.

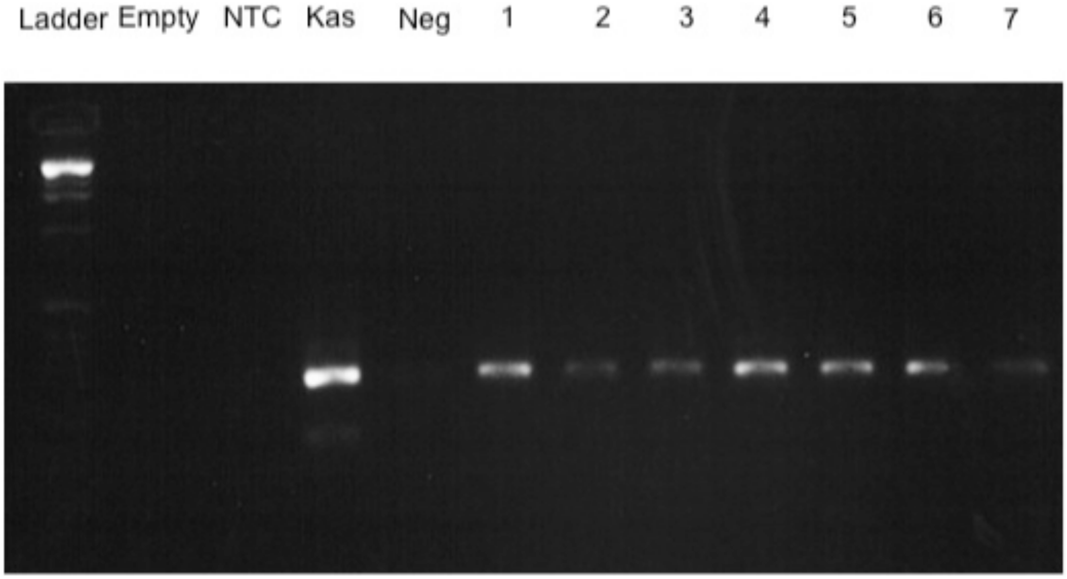


Figure 16 RT-PCR showing bands at appropriate size to confirm the presence of t(8;21) product. Seven of ten primary samples with the translocation, as well as three NK karyotype, were performed by the author with the remaining samples being performed by collaborators. The seven t(8;21) patient samples used by the author are are shown. Primers to AML exon 4 and ETO exon 3 were used to amplify a PCR product of 395bp. Kasumi cell line was used as a positive control and a no test control (NTC) and Molt4 cell line as negative control. (Details of primers used in Appendix D)

Individual Gene Assessment

Firstly, at the individual gene level, exon arrays were used to profile the expression of individual exons for the *ETO* gene specifically. For this analysis the software provided by Affymetrix was used. GCOS scanner was used to generate .dat and .cel files. Probe level analysis of the .cel files was carried out by the ExACT software, which uses a PLIER algorithm to normalise the data. This generates signal estimates and detection p-values at the probeset level for either exon-level or gene-level analysis. A visualisation software tool, the Integrated Genome Browser (IGB), was used to inspect signal intensities. The signal outputs on the IGB provided a number of observations that were explored

Global Gene Expression

Secondly, at a global genome level, analysis was aimed to detect differential gene expression signatures for the two sample groups. Partek^{TM} software was used to generate of lists of differentially expressed genes. The .cel files generated were imported into this software. Normalisation was performed using a RMA algorithm. The software provides a Principal Component Analysis (PCA) method for quality control (Figure 17).

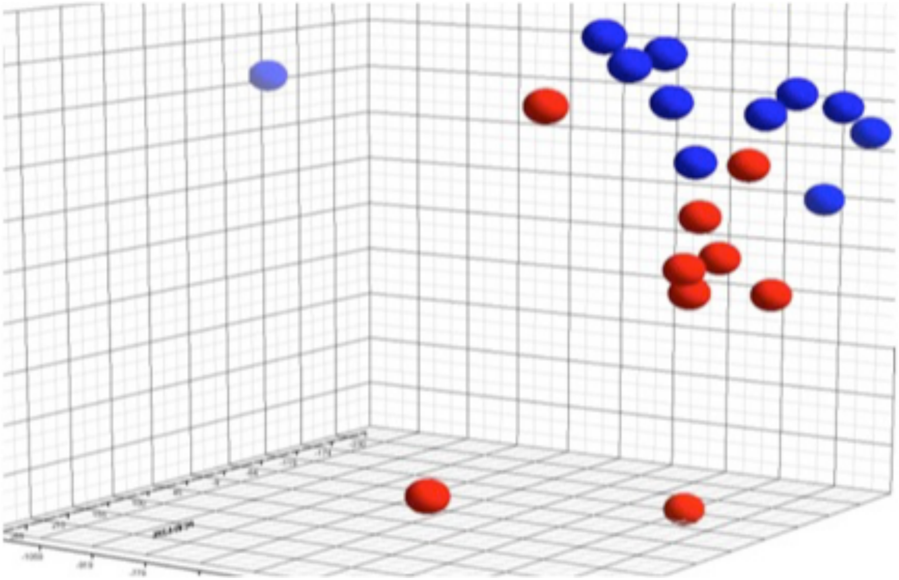


Figure 17 PCA of exon array data. Blue represents t(8;21) samples and red represents NK samples. There are 3 outliers identified.

Three outliers were detected and assessed as to whether this was due to poor hybridisation of these arrays or as a consequence of important biological differences. A pictorial representation of the .dat and .cel files demonstrated by the Partek software indicated poor hybridisation and therefore the data was inspected using the GOLF software. This suggested that the likely cause of the outliers was due to poor hybridisation of the chips (Table 12). Subsequently, these three samples were omitted, resulting in a total of 10 t(8;21) samples and 7 controls for downstream data analysis.

Table 12 Table of data generated from GOLF representing expression values for three potential outliers and all other samples. This demonstrates the low readings for maximum and median values as well as increased numbers of probesets with no signal value, for the three outliers compared to all other samples.

Patient	Median value	Maximum value	Numbers of Probesets <0
1	5	3658	207171
2	5	8269	213182
3	8	7136	168074
Others	23-93	12342-32584	15888-67466

For gene-level analysis the normalised array data was summarised into gene-level data based on mean values by Partek. An ANOVA one-way test on the data between the two groups was then applied. From the ANOVA analysis a list of significant probesets was created using an appropriate multiple testing correction metric such as the false discovery rate (FDR) and determining appropriate fold change (FC) values. Annotation files supplied by Affymetrix were then used to generate gene lists from the table of probesets.

For standard gene expression the convention has been to use fold change values of +/- 2. The significance, and particularly the clinical significance of this cut-off have been debated. For the exon arrays a standard convention has not been developed and therefore for this newer platform a more inclusive policy using a FC cut-off of +/- 1 was adopted to generate gene lists. However, for direct comparisons with gene lists derived from standard expression data, gene lists generated with FC +/-2 from both the exon and standard expression array were used.

For studying gene expression data the most robust metaprobeset file is the core probeset file, which contains only established exons. For alternative splicing and microRNA discovery the extended and full metaprobeset files were used.

For alternative splicing events, Partek uses an alternative splicing ANOVA model on the exon level data sets. A p-value is generated to assess tissue dependent alternative splicing. Small p-values suggest alternative splicing occurrence but these have to be visually inspected to confirm this.

Results

Analysis of the *ETO* gene

The data on the 20 samples are summarised in figures 20, 21 & 22 that depict signal intensities as seen on the IGB. The figures represent values using the full metaprobeset file. There were 3 major findings, which both validated the approach and identified areas for further investigation.

ETO is not expressed in haemopoietic cells

Figure 18 shows the exon expression in 2 patients (patient 1: control t(8;21) –ve and patient 2 t(8;21) +ve). *ETO* has been shown to be transcriptionally silent in haemopoietic cells (Erickson et al., 1996). These findings confirm that wild type *ETO* is not expressed in the absence of the translocation.

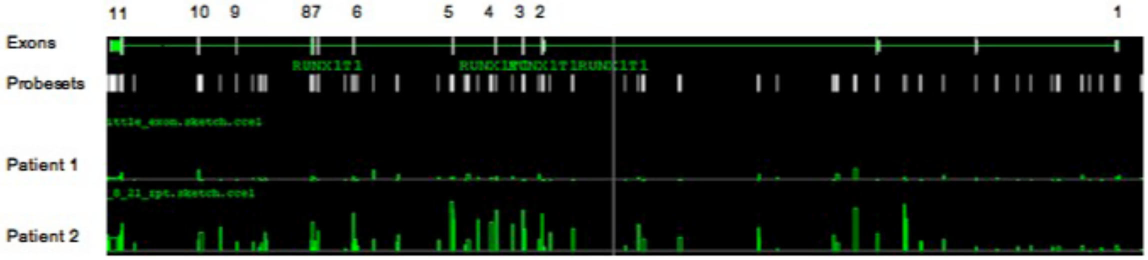


Figure 18 The reverse strand of the *ETO* gene with the 5' end on the right is illustrated on the IGB. The individual exons of *ETO* and Affymetrix probesets are shown at the top of the figure. Patient 1 is a control and Patient 2 has t(8;21). Signal values for each individual probeset are shown, illustrating that ETO is transcribed in t(8;21) but not in the control.

Genomic breakpoints of *ETO* vary

Figure 19 shows the exon expression for eleven patients with t(8;21) and two controls. The signal values, in the intronic region between exons 1 and 2, show a marked variation between each t(8;21) sample; with signals detected at varying distances from the 5' end (figures 19 & 20). The break point for the ETO gene in the t(8;21) are located within this intronic region.

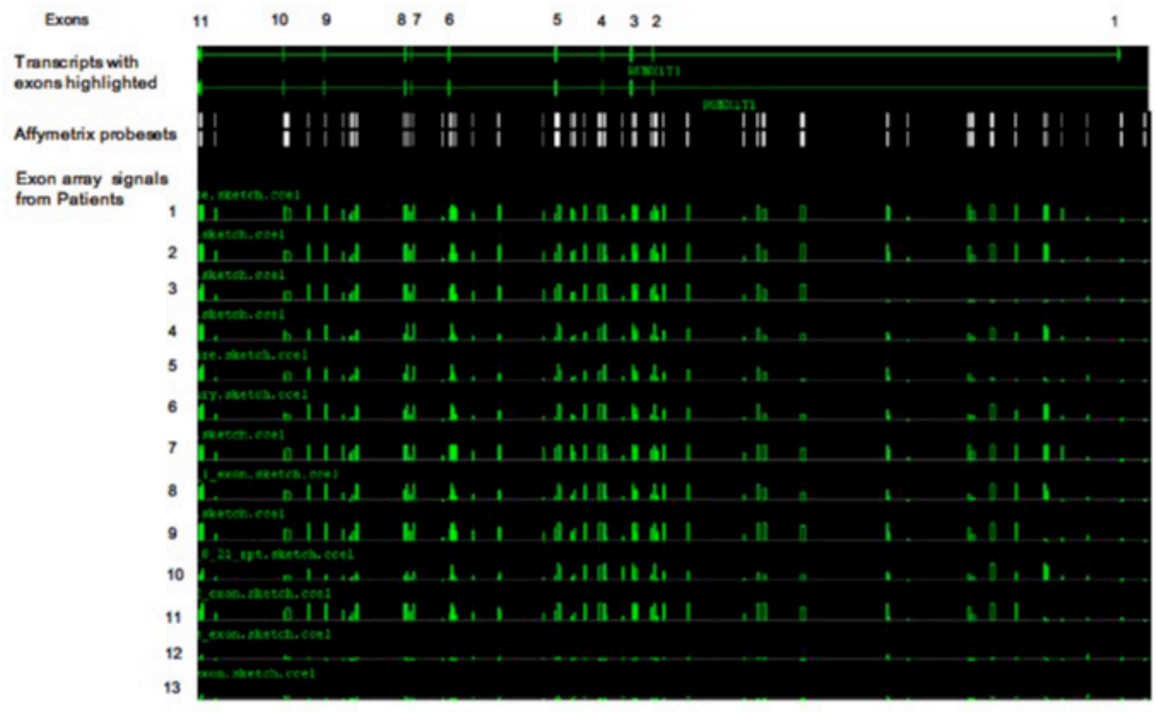


Figure 19 Exon Arrays represented on IGB showing the ETO gene. Layout is the same as figure 20. Samples 1-11 are from patients with AML t(8;21). Samples 12-13 are controls. Signals are seen in the t(8;21) across both the confirmed and putative exons. The breakpoint of ETO in the t(8;21) is) between exons 1 and 2. The signals detected from different patients vary in this region.

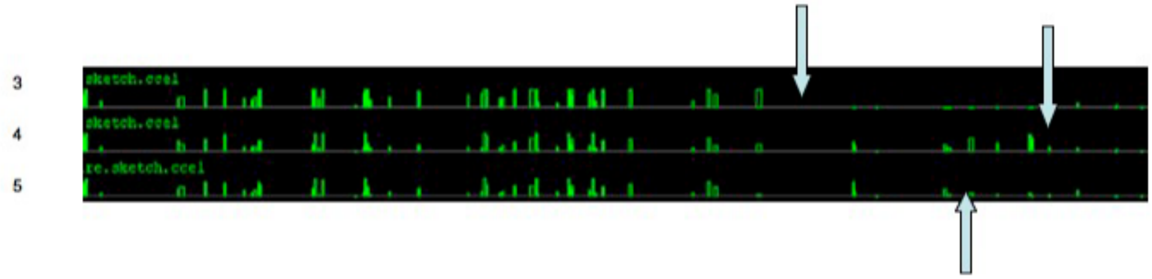


Figure 20 focuses on samples 3-5 from figure 21. Signals start to be detected at different distances from the 5' end clearly demonstrating that the actual genomic breakpoint varied between patients. Arrows mark the potential breakpoint regions.

This demonstrates that the actual genomic breakpoint of ETO is variable between patients. From this small number of samples analysed there appear to be four breakpoint regions detected consistent with current understanding.

Exon arrays may detect novel exons

In the t(8;21) samples the arrays, in addition to showing expression of all 11 reported exons in ETO, identified other signals (Figure 19). These correspond to probesets representing putative exons. As the full metaprobeset was used for this analysis it implies that some of these probesets may actually represent true novel exons not previously reported.

To confirm that exon arrays detected additional novel exons, an RT-PCR approach was used to determine the authenticity of the putative novel exons. Primers specific to 3 novel exons located within intron 1 (the site of the ETO breakpoints) termed A, B and C were designed and used in combination with an AML1 exon 4 specific primer to amplify putative products (Figure 21).

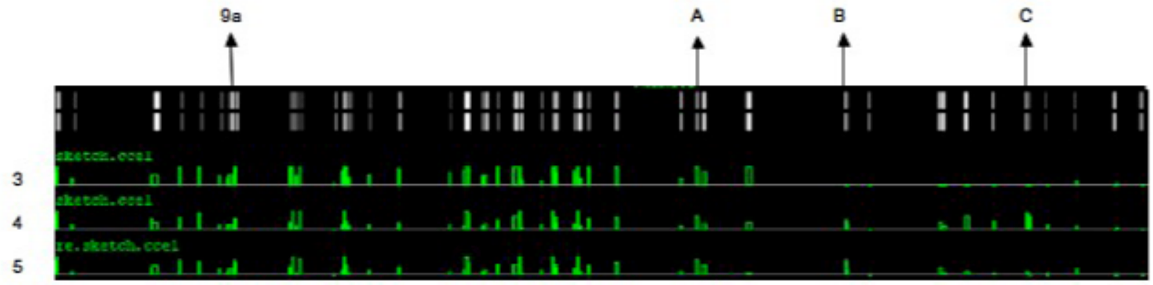


Figure 21 The reverse strand of ETO, illustrated on the IGB. Affymetrix probesets are shown at the top, corresponding to known and putative exons. Samples 3-5 are shown, from patients with t(8;21). Primers A, B and C were designed against 3 probesets, chosen as they correspond to areas with varying signals from different patients. The probeset for exon 9a is also highlighted.

No evidence for expression of these novel sequences was found in either Kasumi cell lines or in 3 patient samples, despite attempts to optimise thermal cycler conditions. PCR amplification of AML1 exon 4 and a confirmed alternative exon of ETO exon 9a was used as a positive control. This amplification did produce appropriate sized bands of 1400bp. Cloning and sequencing of these bands revealed 2 different products, one of which was the expected exon 9a transcript but also a second similar sized transcript, containing a novel ETO sequence referred to as ETO 6a, was also found. This unexpected result was further investigated and is described in the chapter 4.

Global Gene Expression Profiling

Expression profiling of the 17 samples remaining after quality control assessment was performed using the Partek software as described. This analysis generated a gene list of 448 probesets that are differentially expressed, most of which represent annotated genes (Appendix A). Initially, inspection of this list highlighted individual genes well known to be associated with t(8;21) giving confidence to the validity of the technique. However, to both further strengthen this validity and to highlight novel genes associated with t(8;21) a series of comparisons with previously published expression data from t(8;21) samples and cell lines was performed. Furthermore, experimental validation using RQ on a number of genes was also performed. The gene lists obtained were interrogated both at an individual level, using Pubmed searches and at a global level, using functional annotation software and lead to a number of observations.

AML1-ETO causes up regulation as well as gene repression

Approximately half of the genes regulated by AML1-ETO are up regulated confirming previous evidence that AML1-ETO does not only cause gene repression but also up regulation (Nimer and Moore, 2004).

Many of these genes were noted to be involved with the t(8;21) subclass in previous expression studies (Debernardi et al., 2003). *RUNX1T1* (*ETO*) was the most statistically significant gene to be up regulated. *POU4F1* has been shown to be highly up regulated in primary samples although not significantly in the Kasumi cell line (Dunne et al., 2006). *KIT* has been shown to be consistently over-expressed in t(8;21) samples (Wang et al., 2005). Both these known observations were confirmed on the exon array platform (Figure 22 & 23). 12 *HOX* genes were noted as differentially expressed confirming previous findings that down regulation of *HOX* genes is associated with t(8;21) (Debernardi et al., 2003). These findings provisionally validated our approach for using the exon array platform for gene expression analysis.

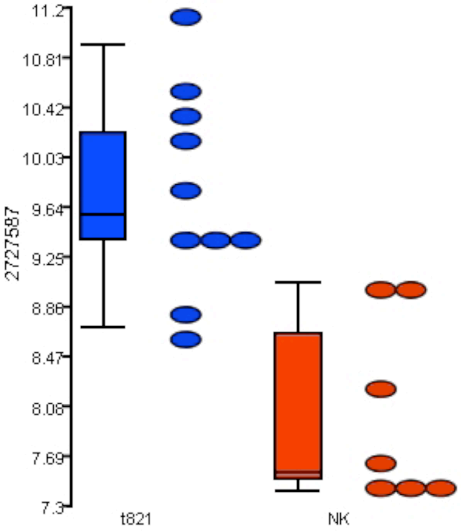


Figure 22 Box plot of expression signals for KIT from t(8;21) patients in blue and NK patients in red.

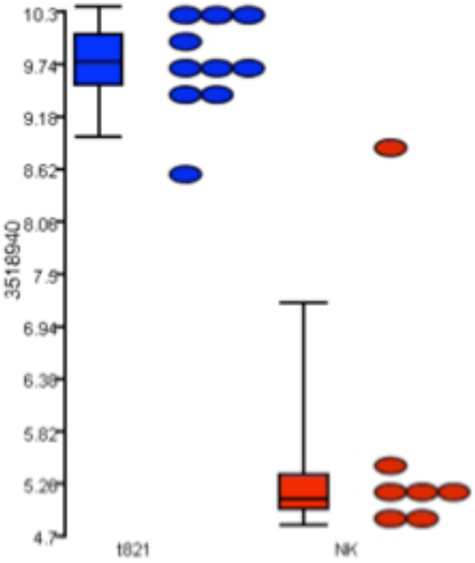


Figure 23 Box plot of expression signals for POU4F1 from t(8;21) patients in blue and NK patients in red

However, the major aim was to discover novel genes associated with t(8;21). Previous expression studies have compared t(8;21) samples to all other AML subclasses. Aiming to be more specific for the translocation, this experiment compares expression data between FAB M2 t(8;21) and FAB M2 (NK) samples. Comparison of our data with previously published expression data was performed to select a subset of the 448 genes that identified novel candidate genes associated with t(8;21).

Gene expression profiling on a series of 100 AML patients using the ABI platform has been previously conducted in this laboratory (Debernardi et al., 2003). 9 t(8;21) primary samples were identified and compared to the remaining 91 primary samples, which compromised all other subgroups including the NK. For ABI data text files of the expression data were imported into Partek. Gene lists were generated using FC +/-2 and FDR p<0.05 as the multiple testing correction method. This resulted in identifying 991 probesets that passed the threshold criteria. This list was compared to the 229 probesets (223 annotated) from the exon array that was generated using FDR p<0.05 and FC +/- 2. Of these probesets, 130 genes (58%) were found to be in both datasets with a further 93 only in the exon array set. Annotation files were compared to exclude probesets from the exon array that were not present in the ABI annotation files. Of these 93 probesets, 66 probesets were in both annotation files. Thus, 66 genes were identified that were unique to the exon array analysis and it is likely that expression changes in many of these genes occur due to the specific effects of the t(8;21) (Table 13).