Single cell mutational profiling and clonal phylogeny in cancer

Nicola E Potter, Luca Ermini, Elli Papaemmanuil, et al.

Genome Res. published online September 20, 2013
Access the most recent version at doi:10.1101/gr.159913.113

For *GENOME RESEARCH* (Methods)

Single cell mutational profiling and clonal phylogeny in cancer

Nicola E Potter – Institute of Cancer Research, UK

Luca Ermini – Institute of Cancer Research, UK

Elli Papaemmanuil – Wellcome Trust Sanger Institute, UK

Giovanni Cazzaniga - Clinica Pediatrica, Italy

Gowri Vijayaraghavan – Institute of Cancer Research, UK

Ian Titley – Institute of Cancer Research, UK

Anthony Ford – Institute of Cancer Research, UK

Peter Campbell - Wellcome Trust Sanger Institute, UK

Lyndal Kearney – Institute of Cancer Research, UK

Mel Greaves – Institute of Cancer Research, UK

Corresponding Author - Professor Mel Greaves FRS, Head, Haemato-Oncology Research Unit**,** Division of Molecular Pathology, Brookes Lawley Building, 15 Cotswold Road**,** Sutton, Surrey SM2 5NG
T +44 (0)20 8722 4073 (direct)
E mel.greaves@icr.ac.uk

Running title: Genetic analysis of single cancer cells

Keywords: Heterogeneity, Single cells, Cancer, Phylogeny

ABSTRACT

The development of cancer is a dynamic evolutionary process in which intra-clonal, genetic diversity provides a substrate for clonal selection and a source of therapeutic escape. The complexity and topography of intra-clonal genetic architecture has major implications for biopsy-based prognosis and for targeted therapy. High depth, next generation sequencing (NGS) efficiently captures the mutational load of individual tumours or biopsies. But, being a snapshot portrait of total DNA, it disguises the fundamental features of sub-clonal variegation of genetic lesions and of clonal phylogeny.

Single cell genetic profiling provides a potential resolution to this problem but methods developed to date all have limitations. We present a novel solution to this challenge using leukaemic cells with known mutational spectra as a tractable model. DNA from flow sorted single cells is screened using multiplex targeted Q-PCR within a micro-fluidic platform allowing unbiased single cell selection, high throughput and comprehensive analysis for all main varieties of genetic abnormalities: chimaeric gene fusions, copy number alterations and single nucleotide variants. We show, in this proof of principle study, that the method has a low error rate and can provide detailed sub-clonal genetic architectures and phylogenies.

INTRODUCTION

Cancer clones evolve by a dynamic Darwinian process of mutational diversification under selective pressures exerted by tissue ecosystems, the immune system and therapy (Nowell 1976; Merlo et al. 2006; Greaves and Maley 2012). Not only do cancers of a similar type differ in their genomic landscapes; indeed each is unique, but intra-clonal genetic and phenotypic diversity is an inherent feature of this disease (Marusyk et al. 2012). Progression of disease (Merlo et al. 2010; Park et al. 2010) and possibly the general intransigence of advanced disease may be attributable to the genetic diversity within the cells of a tumour and/or within the propagating or stem cells (Greaves 2013). The current vogue for personalised medicine and targeted therapy could well be thwarted or achieve only transient benefit if the targets are themselves sub-clonally segregated (Greaves and Maley 2012; Swanton 2012).

Second or next-generation sequencing (NGS) allows whole exome or whole genome sequencing of bulk cancer cells (Stephens et al. 2009; Yates and Campbell 2012) and with adequate depth of parallel sequence reads, it is apparent that many acquired mutations in the cancer genome are sub-clonally distributed (Nik-Zainal et al. 2012). Sub-clonal genetic diversity of genome wide copy number changes has also been demonstrated in a variety of cancers (Klein and Stoecklein 2009; Navin et al. 2010). Cancer cell genetic heterogeneity at the single cell level has long been recognised by chromosome karyotyping (Wolman 1986) and by fluorescence in situ hybridization (FISH) of tissue sections (Clark et al. 2008). The use of multiplex targeted FISH with three or four coloured probes to interrogate the copy number of specific DNA targets facilitates a deeper interrogation of clonal architecture in cancer cell populations from which evolutionary relationships of sub-clones at diagnosis and

relapse can be derived (Anderson et al. 2011). Whole genome amplification of single cells allowing both sequence and copy number analysis at the single cell level is now providing even more sub-clonal information (Navin et al. 2011; Baslan et al. 2012; Zong et al. 2012). These methods combined provide a striking portrait of cancer cell diversity and clonal evolution through the construction of phylogenetic trees characterised by a non-linear, branching architecture (Anderson et al. 2011; Navin et al. 2011; Gerlinger et al. 2012; Yates and Campbell 2012). But, the type of mutation that can be interrogated restricts this approach and the complexity of clonal architectures is underestimated.

Ideally, what is required for a comprehensive interrogation of the complex genomics of cancer cells is a methodology for single cell analysis that has the following attributes: (i) an unbiased cell sample from the cancer; (ii) highly efficient single cell sorting; (iii) a relatively high throughput analysis of at least a few 100's of cells; (iv) a method that allows the simultaneous detection of multiple genetic alterations of different types, e.g. chimaeric fusion genes, copy number alterations (CNA) and single nucleotide variants (SNV) in a single cell. We here provide proof of principle data using single acute lymphoblastic leukaemic (ALL) cells to demonstrate that this is feasible using multiplex targeted DNA amplification from flow sorted single cells followed by high throughput Q-PCR using the BioMark™ HD microfluidic platform (Fluidigm®, San Francisco, CA, USA). The versatility of Q-PCR allows simultaneous analysis of a variety of DNA alterations and coupling this technology with a microfluidic platform capitalises on a high-throughput system with small reaction volumes that lends itself to single cell analysis. Q-PCR analysis combined with the BioMark™ HD has been utilised for single cell gene expression analysis (Citri et al. 2012; Pina et al. 2012) and single nucleotide polymorphism (SNP) allelic

discrimination analysis with bulk DNA (Wang et al. 2009) but not to date for single cell mutational screening.

RESULTS

Development of the method using the REH leukaemia-derived cell line

The established B-cell precursor cell line, REH, was first used to develop and refine this single cell analysis method. Previous FISH (Horsley et al. 2008) and more recent SNP array analysis (unpublished data from our laboratory) has characterised this cell line as *ETV6-RUNX1* fusion gene positive with multiple CNAs including *CDKN2A* and *MX1* and single nucleotide polymorphisms. In this study the *EPOR* SNP (SNP rs318720) was adopted as a surrogate acquired heterozygous SNV anticipated to be present in every cell.

Briefly, single carboxyfluorescein diacetate succinimidyl ester (CFSE) labelled REH and cord blood cells (normal diploid control) were sorted into individual wells of a 96 well plate, lysed and DNA target amplification completed for regions encompassing the clonotypic *ETV6-RUNX1* fusion genomic sequence, *CDKN2A*, *MX1* and the SNP rs318720. The *B2M* locus, located in a diploid region of the genome, was used as a control. The resulting reaction mix was then diluted and Q-PCR completed using the 96.96 dynamic array and the BioMark™ HD. The workflow for this method is shown in Figure 1. A cell was deemed to be positive for a SNP (or SNV) if the Q-PCR cycle threshold (C_T) value was below 28. The presence or absence of the signal from the probe complementary to the wild-type sequence determined heterozygous or homozygous mutations respectively but the copy number of each allele cannot be inferred (Figure 2A). The $\Delta\Delta C_T$ method (Applied Biosystems®, Life Technologies Ltd, Paisley, UK) with modifications to incorporate

the results from multiple Taqman® assays targeting the same region was used to determine the relative copy number for each locus; the use of multiple assays to target one region increased the accuracy of attributed DNA CNAs (Figure 2B).

Several approaches were adopted during this experiment to optimise and confirm the presence of a single cell and ensure all assays performed efficiently under these experimental conditions (details can be found in Methods and f Material). Efficient FACS sorting of single cells was initially confirmed by microscopy. The fluorescent cells were sorted onto a glass slide with the aim of collecting 48 independent single cells; this established the efficiency of the BDFACSAria I SORP instrument (BD®, Franklin Lakes, NJ, USA). In our hands the failure rate is 2-4% (1-2 occasions per 48 attempts), the majority of which are a failure to sort a cell compared to two cells found in 0.002% of occasions (1 in every 528 attempts to sort a single cell). As it is not possible to visualize single cells in a 96 well plate, we sought to quantify the DNA in each well to identify those with high amounts (of the *B2M* gene) indicating that two or more cells may have inadvertently been collected. These data were consequently removed from the phylogenetic analysis but only constituted a maximum of two wells per plate.

To estimate the error rate of each assay (gene fusion, CNA and SNV assays) a control experiment consisting of 48 cord blood cells was completed for each patient-specific multiplex experiment. Gene fusion and SNV assay false-positive error rates can be found in Supplemental Table 2. Only two assays (SNVs in *EZH2* and *BAZ2A*) generated false-positive results in 2% (1/48) of cord blood cells. CNA assays had an error rate ranging from 4.3-7.1% (Supplemental Table 3). False-negative errors rates could not be directly determined as each assay is patient-specific and no positive single cell control sample is available. However, when using

allelic discrimination assays to detect SNVs each assays generates a signal either for the wild-type sequence, the mutant sequence or both (most common here). Each single cord blood control cell produced a signal for the wild-type sequence confirming the assay efficiency in the multiplex system. The SNP rs318720 analysed in the REH cell line produced a signal for the wild-type and polymorphic sequence in each cell interrogated. The assay error rates were used to define a bone-fide minor sub-clonal population (Methods - *Defining sub-clonal populations at low frequencies*). The authors were also careful to compare the mutation allele burdens generated by each approach utilised in this study, including exome sequencing, 454 pyrosequencing, allelic discrimination by digital PCR using bulk DNA and genotyping of each single cell (Table 1).

Data from single cells that were removed from the Q-PCR analysis included those wells that showed no data (no cell), those wells in which all *B2M* assays did not have a strong signal (<28 C_T) and wells in which all CNA assays for a target region of interest did not produce C_T results within one C_T. Data from suggested minor sub-clonal populations that did not exceed assay error rates was also removed. On average 75% of interrogated single cells generated complete comprehensive results. A detailed breakdown of the single cell experiment data with explanations as to why data were removed can be found in Supplemental Table 4.

All REH cells (n = 126 – single cells considered for phylogenetic analysis) scored positive for the *ETV6-RUNX1* fusion gene and the *EPOR* SNP rs318720. Major sub-clonal populations had varying copies of the *MX1* gene; 49% of cells retained two copies of *MX1* compared to 42% of cells that had gained either one or two copies of this gene (Figure 2C). The majority of cells showed loss of both *CDKN2A* copies, but minor sub-clones were characterised by either loss of *MX1*

and/or loss of only one copy of *CDKN2A*. These DNA copy number data were independently confirmed by FISH using BAC probes complementary to the *ETV6-RUNX1* fusion gene, *CDKN2A* and *MX1*. The sub-clonal structure and percentages obtained by FISH correlated well with the results obtained by single cell genetic analysis using multiplex targeted Q-PCR and the BioMark™ HD (Figure 2C).

Single cell genetic analysis of a Down's syndrome ALL case

We replicated this approach using leukaemic cells from a child with Down's syndrome and acute lymphoblastic leukaemia (DS-ALL) for which we had previously characterised the genomic alterations using SNP analysis and targeted Sanger Sequencing. This case was expected to have a simple but informative clonal architecture involving a *P2RY8-CRLF2* fusion gene, loss of *CDKN2A* and an *IL7R* mutation involving the deletion of eight base-pairs and the insertion of ten non-consensus base-pairs in exon 6.

The sub-clonal genetic architecture of this DS-ALL case was more complex than suggested by SNP analysis. Using the single cell multiplex targeted Q-PCR approach 115 single cells were analysed and compared to a further 100 cells using FISH (Figure 2D). Cells that did not harbour any alterations were present at 4% and 2% by FISH and multiplex targeted Q-PCR experiments respectively, reflecting the 90% blast count/low non-leukaemic cell mix in this diagnostic sample. The *P2RY8-CRLF2* fusion and the *IL7R* mutation were present in 98% of cells by multiplex targeted Q-PCR. A minor sub-clone (2%) had lost the *IL7R* wild-type allele but retained the mutant allele. Loss of *CDKN2A* was sub-clonal to the *IL7R* mutation and proceeded to homozygous loss in 11% of cells. FISH for the *P2RY8-CRLF2* fusion and *CDKN2A* copy number confirmed these results with respect to these loci. The

loss of the wild-type *IL7R* allele in a minor sub-clone was not anticipated. Sub-clonal segregation of a homozygous mutation would be very difficult to clarify, except using single cell analysis.

Single cell genetic analysis of *ETV6-RUNX1* positive ALL cases with exome sequencing data

To expand our approach and confirm that it could be applied to more complex genetic data sets we selected two *ETV6-RUNX1* positive ALL diagnostic bone marrow samples which had been subjected to both whole exome sequencing to identify SNVs (Table 1) and SNP arrays for CNA (confirmed by exome analysis) (Table 2). Variations in mutation allelic burden suggested the presence of sub-clonal populations in both cases and was confirmed using Digital PCR (Table 1). Each case also harboured varied chromosomal alterations in both size and location and known recurrent secondary CNAs in ALL (Mullighan et al. 2007) were identified including loss of *PAX5* and *CDKN2A*.

Genomic targets for Case A included SNVs in *BCHE* (exome SNV read depth 71×, variant reads 30), *EZH2* (exome SNV read depth 139×, variant reads 53), *PIK3R1* (exome SNV read depth 67×, variant reads 9), *DAXX* (exome SNV read depth 131×, variant reads 14) and *BAZ2A* (exome SNV read depth 113×, variant reads 42) and CNAs in *CCNC* and *TBL1X.* Targets for Case B included SNVs in *KRAS* (exome SNV read depth 65×, variant reads 28), *RB1* (exome SNV read depth 129×, variant reads 6), and *SRSF11* (exome SNV read depth 302×, variant reads 16) and CNAs in *VPREB1*, *PAX5*, *CDKN2A* and *DPF3* (Tables 1 and 2). SNVs were chosen based on allelic burden encompassing both high, low and intermediate frequencies; not all SNVs were included. However, allelic discrimination assays

designed for *DNAH17*, *MAN21* and *SLC7A14* did not show sufficient target specificity to wild-type and mutant sequence in the multiplex Q-PCR reaction and resulting data was inconclusive. Only one attempt was made to design assays for each chosen SNV. The patient-specific *ETV6-RUNX1* genomic fusion gene sequence was obtained by long distance PCR (Wiemels and Greaves 1999) guided by fusion break-point co-ordinates from whole exome sequencing. Comprehensive data was collected from 261 single cells for Case A and 254 from Case B.

Derivation of phylogenetic trees using Maximum Parsimony

To uncover the clonal phylogeny from these single cell data we used the Maximum Parsimony (MP) method (Page and Holmes 1998). The most parsimonious phylogenetic tree is the tree describing the best estimated phylogenetic relationships given the included taxa (group of related cells or clones). Tree branch lengths are directly proportional to the number of evolutionary changes inferred and the points at which the branches diverge (nodes) represent the ancestor state of a clonal clade (a monophyletic group which includes all descendants of the ancestor). A phylogeny inferred using this criterion shows how the clonal expansion has evolved from a common ancestor towards the observed states. In the event that two (or more) trees hold equal parsimony all trees are accepted (Page and Holmes 1998). Detailed method explanations of this approach can be found in Methods and Supplemental Material.

Maximum parsimony searches for Case A resulted in one maximum parsimonious tree (Figure 3A). The phylogenetic architecture of the tree shows a quasi-linear structure with a direction towards sub-clone A4 representing 87.4% of the clonal population; the sub-clonal heterogeneity in this case is therefore modest,

given the number of genomic alterations investigated. The most recent common ancestor (MRCA) of this tree, which is the most recent ancestor from which all clones of the group directly descend, is sub-clone A3. This clone harbours the *ETV6-RUNX1* fusion in addition to *BCHE* and *EZH2* mutations suggesting that these alterations were relatively early mutational changes in the pathogenesis of this individual ALL. The major clone A4 may have a selective fitness advantage over all other sub-clones associated with the acquisition of *CCNC* deletions and a *BAZ2A* SNV. But, as this is a single time point snapshot of a dynamic process, we cannot exclude the possibility that sub-clone A4 was spawned before sub-clones A5 and A2, which are characterised by heterozygous mutations in *PIK3R1* and *DAXX* respectively. The maximum parsimony algorithm shows the inferred MRCA of the A4-A5 clonal clade (Figure 3A - grey box), a group of cells which has died out or been outcompeted or if still present, exists at to a low frequency to detect.

Maximum parsimony searches for Case B produced two equally parsimonious trees with identical topological structures (Figure 3B). Both trees show complex branching structures with a marked sub-clonal heterogeneity of seven clones that differ only for the position of sub-clones B4 and B5. The MRCA of the clonal expansion is represented by sub-clone B2, which shows a homozygous deletion of *VPREB1* in addition to the *ETV6-RUNX1* fusion. This sub-clone is the biggest clonal group after clone B3 (at diagnosis) despite the genetic diversity of the progressive sub-clonal populations, suggesting that this sub-clone may be quiescent or reside in a niche environment. The acquisition of the *KRAS* mutation presumably provides selective advantage as a secondary driver event. A minor sub-clone was also identified in which the wild-type *KRAS* allele was lost suggesting the acquisition of homozygousity for the *KRAS* mutation. The number of mutant gene copies (one or

two) in each single cell cannot be determined using the Taqman® *KRAS* SNV assay used here but the increased allele burden characterised by whole exome sequencing, 454 pyrosequencing and digital PCR validates this conclusion.

The position of sub-clones B4 or B5 indicates that these populations are equally parsimonious ancestors of the clonal clade constituted by sub-clones B3, B7 and B8 representing ~65% of all examined cells. This represents the reiteration of either *PAX5* or *CDKN2A* loss independently in two different branches. Reiterated CNA of the same gene is a feature of the sub-clonal genetic architecture in paediatric ALL (Anderson et al. 2011; Waanders et al. 2012) suggesting that these deletions may not only provide selective advantage but may be a target for DNA level breakage via, for example, RAGS (Kitagawa et al. 2002; Waanders et al. 2012) or AID (Longerich et al. 2006; Klemm et al. 2009).

DISCUSSION

Cancers have complex patterns of acquired mutations and patients with closely related subtypes of cancer have unique or clone-specific mutation patterns. Additional complexity clearly exists within each individual cancer, as mutations are acquired serially and with a variegated pattern of distribution in sub-clones. As mutational signatures are increasingly used for differential diagnosis, prognostication and therapeutic targeting, this multi-dimensional complexity is of some consequence.

Systematically interrogating sub-clonal genetic complexity poses a considerable technical challenge. A clonal architecture or phylogeny can be inferred bioinformatically by analysis of high depth NGS data (Nik-Zainal et al. 2012; Yates and Campbell 2012). Nik-Zainal et al. reconstructed clonal phylogenies through the development of novel algorithms that rely on whole genome sequencing data (200×

coverage) to phase mutations with germline polymorphisms and define clonal and sub-clonal phylogenies. This is in marked contrast to the exome sequencing data used in this study, which do not offer the opportunity of continuous genomic information that would allow one to reconstruct such phylogenies with confidence. Additionally, exome data in leukaemia's show low total mutation burden (on average 10) which renders them further limited to differentiate clonal phylogenies. It is in fact this limitation of exome sequencing data and to a lesser extent whole genome sequencing that the present study is precisely addressing where, as shown in the example of Case B, two SNVs of close allele burden estimates *RB1* and *SRSF11* (4.65 and 5.30 respectively) could be in the same (or separate) sub-clone with equal probability. Without single cell data this level of phylogenetic resolution is not possible. We conclude that single cell genetic analysis is required for a robust and definitive designation of the segregation pattern of mutations within cancer clones. Theoretically, the best approach might be to sequence the genomes of individual cells. Significant progress has recently been made in this regard (Baslan et al. 2012; Zong et al. 2012) but error rates and low cell throughput are, currently, significant limitations. We adopted the alternative strategy of analysing single cells from cancers whose genomic, mutational profile at the cell population level was already established by high-resolution SNP arrays and whole exome sequencing.

Using the strategy outlined here, we were able to detect all three categories of mutational change – fusion gene, CNA and SNV in the one assay. Our method is amenable to high throughput analysis: in each case we were able to analyse 200-300 leukaemic cells. We assume that the profiles that emerge from these number of cells are representative of the patients' leukaemia. This would be more demanding with adult carcinomas where the topographical segregation of distinct sub-clones

could result in selective sampling but clearly multiple small biopsies could be assessed (Park et al. 2010; Gerlinger et al. 2012). The screening of many thousands of single cells is possible by the method we report but is restrained at present by cost.

The phylogenetic trees inferred from the cellular distribution of SNVs in larger numbers of cells provide additional evidence that cancers evolve within a branching architecture of sub-clones. These detailed and complex sub-clonal architectures would not be detected by other genetic techniques. Prior studies on ALL suggested that the latter is generated and sustained by genetically diverse cancer propagating or stem cells (Anderson et al. 2011; Notta et al. 2011) and it will be important to confirm this using the current microfluidic platform.

This proof of principle study using leukaemic cells illustrates that it is possible to assess single cells simultaneously for different types of genetic lesions – fusion genes, copy number alterations and single nucleotide variants, and from these data to construct a clonal, evolutionary phylogeny. Leukaemias are likely to be significantly less complex in this respect than carcinomas (Vogelstein et al. 2013) but nevertheless, the sub-clonal architectures we illustrate here for ALL are likely to be a significant under-estimation of complexity. Higher definition of clonal structure, including minor clones at <1%, is however achievable by both increasing the depth of initial sequencing and by screening larger numbers of single cells. Also, informative though these clonal analyses are, they remain single time point snapshots of a very dynamic evolutionary process. Ideally, single cell genetics and phylogenetic tree construction should be applied to serial samples from individual patients, i.e. diagnosis versus relapse, primary versus metastases, pre- versus post-chemotherapy. It is known that these major transitions can involve selective sweeps

or stringent clonal selection (Mullighan et al. 2008; Liu et al. 2009; Anderson et al. 2011; Diaz Jr et al. 2012).

Dissecting the detailed clonal architecture of cancer has significant clinical implications. The extent of intra-clonal genetic diversity may be predictive of progression of disease (Maley et al. 2006) or clinical outcome (Mroz et al. 2013). Cancer genomics holds the promise of personalised medicine with therapy targeted at products of recurrent 'driver' mutations (Chin et al. 2011). Sub-clonal segregation would not be a desirable credential of any candidate target. Targeting even a major branch of the phylogenetic tree rather than the founder lesion would be predicted to have only transient benefit and, moreover, provide selective pressure for the emergence of previously minor clones (Greaves and Maley 2012; Swanton 2012).

METHODS

Samples

The precursor B-cell leukaemic cell line, REH, was purchased from American Type Culture Collection® (ATCC®, Virginia, USA) and cultured in RPMI-1640 medium 10% FCS.

The patient samples studied in this investigation were collected from Italian or UK hospitals, with local ethical review committee approval (CCR 2285, Royal Marsden Hospital NHS Foundation Trust). Bone marrow aspirates or peripheral blood samples underwent lymphoprep separation and were viably frozen in 10% DMSO, 90% FCS and stored in liquid nitrogen (10% DMSO, 90% FCS).

Bulk sample analysis

Single nucleotide polymorphism and DNA copy number array analysis for Case A and Case B

To define DNA copy number alterations and SNVs for these cases we employed the Affymetrix® Cytogenetics Whole Genome 2.7M Array (Affymetrix®, Santa Clara, CA, USA). Briefly, 100ng of genomic DNA from both diagnostic and remission samples was whole-genome amplified (WGA) using the Affymetrix® Cytogenetics Reagent Kit and the Affymetrix® assay protocol according to the manufacturer's instructions. Samples were then fragmented to generate small products (<300 bp), which were subsequently biotin-labelled, denatured and loaded into the arrays. After hybridization the chips were washed, stained (streptavidin-PE) and scanned using the GeneChip® Scanner 3000. CEL files were generated using Affymetrix® GeneChip® Command Console™ (AGCC) v3.1 and analyzed by

Chromosome Analysis Suite (Affymetrix®) software, version 1.2.2. Quality control metrics for each case can be found in Supplemental Material Table 1A. SNP and DNA copy number analysis for REH and the DS-ALL sample was completed using the Affymetrix® Genome-Wide Human SNP Array 6.0. Method details can be found in Supplemental Material.

ETV6-RUNX1 fusion breakpoints

Genomic sequences for the *ETV6-RUNX1* positive samples (REH cell line and patients Cases A and B) were cloned using long distance inverse PCR using DNA from bulk cells as described before (Wiemels and Greaves 1999). *P2RY8-CRLF2* fusions were sequenced according to Mullighan et al. (Mullighan et al. 2009).

Whole exome sequencing

Matched genomic DNA (3-5μg) from leukemic and complete remission samples from the two cases with childhood acute lymphoblastic leukaemia was prepared for Illumina® paired end sequencing (Illumina® Inc, San Diego, CA, USA). Exome enrichment was performed using the Agilent SureSelect^{XT} Human All Exon 50Mb kit (Agilent Technologies Ltd, Berkshire, UK) as per manufacturer's guidelines but without the pre-enrichment PCR amplification_ENREF_1. Solid Phase Reversible Immobilization (SPRI) bead clean up was used to purify products in preparation for sequencing (Agencourt® AMPure® XP beads, Beckman Coulter, UK). Flow-cell preparation, cluster generation and paired end sequencing (75 bp reads) was performed according to the Illumina® protocol guidelines on an Illumina® GAII Genome Analyzer. The target coverage per sample was for 70% of the captured regions at a minimum depth of 30× sequencing coverage.

Sequencing reads were aligned to the human genome (NCBI build 37) using the BWA algorithm on default settings (Li and Durbin 2010). Duplicate reads derived by PCR were removed using Picard and reads mapping outside the targeted region of the genome were excluded from the analysis. Standard internal quality control evaluation of sequencing data including % of unique mapped reads, % of target region covered, % of unmapped reads, sequence quality metrics and total sequencing output (in GB) was performed for all samples. The remaining uniquely mapping reads (~60%) provided 60–80% coverage over the targeted exons at a minimum depth of ×30. Leukemic sample identity relative to the matched remission sample was controlled by digital genotyping of 100 genomewide SNP markers prior to variant calling.

In house-variant caller CaVEMan (Cancer Variants through Expectation Maximisation) was used to call single nucleotide substitutions (Varela et al. 2011). To call insertions and deletions, split-read mapping was implemented as a modification of the pindel algorithm (Ye et al. 2009). Copy number and LOH analysis was performed using ASCAT (Van Loo et al. 2010). Further details of these steps can be found in Supplemental Material. For validation all putative somatic indels were confirmed by capillary sequencing via Roche pyrosequencing of both tumour and remission samples from each patient. Mutant allele burden estimates were derived from the fraction of reads reporting the mutant allele over the total read depth at each genomic location and confidence intervals were derived using the binomial distribution.

Commercial and custom primers for Q-PCR and digital PCR

Primer Express Software (Applied Biosystems®, Warrington, UK) was used to design custom genotyping Taqman® Q-PCR assays for an SNV that could distinguish the mutant allele from its wild-type counterpart. Each SNV assay contained allele-specific minor-groove binder (MGB) probes for the wild-type allele (FAM-labelled) and the mutant allele (VIC-labelled) (Supplemental Material Table 2). These assays were tested to ensure specificity and reliability (refer to Assay validation in Supplemental Material). Assays to detect a fusion breakpoint were designed using a similar approach with a FAM labelled MGB probe straddling the fusion break point. DNA copy number Taqman® assays were purchased from Applied Biosystems® as these have been design for uniform amplification efficiency and have been commercially validated. Three CNA assays were chosen within each DNA target region of interest and the diploid reference region encompassing *B2M*.

Digital PCR

Digital PCR was used to quantify target sequences and estimate mutant allele burdens within bulk DNA from each patient at diagnosis and in cord blood. This was completed using the 12.765 digital array (Fluidigm®) and the BioMark™ HD. This digital array contains 12 panels each with 765 individual microfluidic chambers (6nl volume per chamber). Six targets were simultaneously interrogated according to the manufacturer's instructions; 5ng DNA per panel. The number of target molecules per panel was determined using BioMark™ HD Digital PCR software; SNV frequencies were calculated by dividing the number of mutant allele copies by the total number of copies for the wild-type and mutant alleles.

Single cell analysis

Single cell labelling and flow sorting

Single cell sorting was performed on a BDFACSAria I SORP instrument (BD®, Franklin Lakes, NJ, USA) equipped with an automated cell deposition unit using the following settings: 100micron nozzle, 1.4bar sheath pressure, 32.6KHz head drive and a flow rate that gave 1-200 events per second. Details of viable cell thawing, single cell CFSE staining (according to manufacturer's instructions), cell sorting parameter explanations and the assessment of single cell sorting efficiencies can be found in Supplemental Material. Two 96 well plates of single cells were collected for REH and the DS-ALL Case and four plates were collected for the two ALL Cases. The plates were composed of a no template control (NTC), 11 control cord blood cells and 84 target cells (REH or patient cells).

Single cell multiplex targeted pre-amplification and Q-PCR

Labelled single cells were sorted into 2.5 µl lysis buffer composed of 1 mg/ml proteinase K (Qiagen Ltd, Manchester, UK) and 0.5% Tween20 in HEPES buffered saline (Sigma-Aldrich®, Gillingham, UK). Lysis was carried out for 50 min at 60^{o}C followed by 10 min at 98^{o}C. Specific (DNA) targeted amplification (STA) was then performed prior to Q-PCR. This multiplex STA reaction was composed of 5µl pre-amplification master mix (Life Technologies Ltd) and 2.5µl 1:40 primer mix (containing all primers for the gene targets of interest). Denaturation was completed at 95^{o}C for 15 min, followed by 24 cycles of amplification at 95^{o}C for 15 s and 60^{o}C for 4 min. The STA product was then diluted 1:6 using DNA suspension buffer (Teknova®, Surrey, UK). Finally, 2.7 µl of the single cell target amplified DNA was interrogated by Q-PCR for each DNA target of interest using the 96.96 dynamic

microfluidic array and the BioMark™ HD as recommended by the manufacturer; thermal phase 70°C for 1800 secs, 25°C for 60°C secs followed by a hot start phase of 95°C for 60 secs. This was followed by 35 cycles of 96°C for 5 secs and 60°C for 20 secs. CNA assays were completed in quadruplicates and SNV or fusion assays were completed in duplicates.

Single cell Q-PCR analysis

The BioMark™ HD generates a C_T value for each reaction. A heterozygous mutation was considered to be present if the signals from the mutant and wild-type sequence probes (FAM and VIC respectively) had a C_T value <28 in a single cell. A homozygous mutation was considered to be present if there was no wild-type sequence signal.

To ensure robust DNA CNA data from a system that can be influenced by assay efficiency and experimental variation we employed the $\Delta\Delta C_T$ method (Applied Biosystems®) to determine a copy number for each locus with modifications to incorporate data from three distinct assays targeting the control region (*B2M)* and the region of interest. The $\Delta\Delta C_T$ value was calculated for every target gene assay using each of the three reference gene C_T values generating nine estimated DNA copy number results for a region of interest. A confidence metric was assigned to the estimated copy number inferring the confidence with which an estimated copy number could be deemed true (according to Applied Biosystems CopyCaller® Software v2) (details of this approach can be found in Supplemental Material). The weighted mean of the nine estimated DNA copy numbers (for a region of interest) was used as the final DNA copy number taking into consideration the confidence metric attributed to each. This reduced the contribution of less reliable estimated

DNA copy numbers to the final DNA copy number. Estimated copy number results were not considered if the confidence value was less than 50% or the estimated copy number was greater than four (with only quadruplicates per assay the results are not robust enough to accurately detect DNA copy numbers greater than four (Weaver et al. 2010)). At least two of the nine estimated copy numbers must have a confidence value above 50% to calculate the final copy number for a region of interest.

Defining sub-clonal populations at low frequencies

Each assay type (gene fusion, CNA or SNV) varied in error rate demanding careful consideration when defining sub-clonal populations at low frequencies. Gene fusion assays and SNV assays did not yield any false-positive results in the control experiments, except *EZH2* and *BCHE* for which we saw one cell (Supplemental Table 2). However, CNA assays had an average error rate of 5.4% (Supplemental Table 3). Consequently the following criteria was used to define a bona fide sub-clonal population: an observed signal pattern (character state) must be attributed to four or more cells (in this experiment the equivalent of approximately 1%): a population present at less than the error rate for a given CNA assay cannot be defined by that single CNA alone but if two CNA alterations define a sub-clonal population it was deemed to be true. A single fusion event or SNV can distinguish a minor sub-clonal population. For example in Case B sub-clonal B6 is defined from the ancestral sub-clone B2 by a *KRAS* mutation.

Interphase fluorescence in situ hybridization (FISH)

Methanol-acetic acid fixed cells, prepared by standard cytogenetic techniques, were used for all FISH studies. Interphase FISH for the *ETV6–RUNX1* fusion gene in combination with probes for regions of copy number alteration was performed using a commercial LSI *ETV6-RUNX1* extra signal (ES) probe (Vysis, Abbott Laboratories Ltd, Illinois, USA) as previously described (Horsley et al. 2008; Bateman et al. 2010). The *P2RY8-CRLF2* gene fusion was identified using a dual colour breakapart probe consisting of two bacterial artificial chromosome (BAC) probes flanking the *CRLF2* locus (Supplemental Table 5). BAC and fosmid probes for this and other regions of interest were obtained from the BACPAC Resource Centre (Children's Hospital, Oakland Research Institute) (http://bacpac.chori.org). Probes were labelled by nick translation with either biotin-16-dUTP (Roche Ltd), SpectrumRed™ or SpectrumOrange™ (Vysis) and hybridized in combination. FISH was performed by standard protocols (Horsley et al. 2008; Bateman et al. 2010) with a single detection layer of streptavidin-Cy5 for biotinylated probes. Fluorescent signals were viewed using an Olympus AX2 fluorescence microscope equipped with narrow bandpass filters for FITC, SpectrumOrange™, TexasRed™, and Cy5. Images were captured and analysed using a charge-coupled device (Photometrics) and SmartCapture 3 software version 3.0.4 (Digital Scientific UK Ltd, Cambridge, UK). In each case, 100 nuclei were scored for the presence of the fusion gene and other targets of interest. Nuclei from karyotypically normal cells (peripheral blood samples from normal individuals) were used to assess probe hybridization efficiency. The percentage of cells with the expected normal signal pattern was 97–100% (mean = 98%) for each probe (Supplemental Table 5).

Phylogenetic analysis and clonal evolution

Clonal groups

Copy numbers and genotypes for each interrogated cell were concatenated in a linear array of nine characters and then aligned to identify the same motif in the array. Cells sharing identical alterations were grouped together in the same clonal group. Clones were then aligned and used to infer the evolutionary history of each patient's leukaemia.

Maximum parsimony

Maximum parsimony searches were conducted using heuristic searches. Heuristic searches were performed with a series of 1,000,000 random additional clones and tree branch-swapping using a bisection-reconnection (TBR) algorithm. All characters except the initiating genetic lesion *ETV6-RUNX1* fusion gene were weighted equally and state graphs and step matrices were used to assign equal costs to each character state transition in different genes. The cost assigned for each transition is linear and results from the equation $y_i=2x_i$. We assigned a transition cost equal to 1 ($y_0=x_0$) only to the transition 0 (no fusion) to 1 (one fusion) for the *ETV6-RUNX1* fusion. Character state graphs and corresponding matrices employed are shown in Supplemental Table 6. The normal clone (according to the alterations interrogated) found in each case was assumed to be the ancestral clone and included in the analysis as the root of the tree/trees. All parsimony analyses were performed using the computer software PAUP* version 4.0b10 for Linux (Swofford 2005). Trees were visualized using Dendroscope Software version 3 (Huson and Scornavacca 2012).

Node support: jackknife

Support for the internal branches was assessed in PAUP* by jackknife with 1000 pseudo-replicates. Heuristic searches with randomly added taxa followed by TBR were used for each jackknifing iteration deleting 12.5% of the characters in each pseudo-replica. A jackknife 50% majority-rule consensus tree (Margush and McMorris 1981) was used to support the node of phylogenetic trees inferred.

Mimicking the bootstrap procedure

We wrote an R in house-script to mimic the bootstrap re-sampling method. The script samples without replacement from the aligned clones and generates replicas with columns shuffled in different orders.

The script was run on fifty separate occasions for Cases A and B, using R software for statistical computing version 2.15 (RCoreTeam 2013). Each resulting replica was used as input for a maximum parsimony search employing the same settings as described in Supplemental Material.

DATA ACCESS

Data accession numbers

Whole exome sequencing - The European Genome-phenome Archive (EGA) (www.ebi.ac.uk/ega/submission): accession number EGAD00001000636.

Single nucleotide polymorphism and copy number analysis by SNP arrays - Gene Expression Omnibus (GEO) database (www.ncbi.nlm.nih.gov/geo/): accession number GSE49215.

ACKNOWLEDGMENTS

This work was supported by Leukaemia & Lymphoma Research United Kingdom and the Kay Kendall Leukaemia Fund United Kingdom.

DISCLOSURE DECLARATION

We have nothing to disclose.

FIGURE LEGENDS

Figure 1. Schematic workflow of the multiplex targeted Q-PCR approach for the simultaneous detection of gene fusions, DNA copy number alterations and mutations in single cells. Initially, CFSE labelled single cells are sorted into each well of a 96 well plate and then lysed. Multiplex DNA specific target amplification is then completed to amplify a DNA region of interest (fusion gene, copy number alteration or SNV) from the two copies found in a single cell to an amount that can be detected by Q-PCR using the BioMark™ HD. Amplified samples and assays are then loaded into a 96.96 dynamic array which utilizes a valve controlled capillary network to bring these two mixes together at nanolitre volumes (completed in the IFC controller) for the Q-PCR reaction to take place. This final Q-PCR step determines the gene fusion, mutation or copy number alteration status for each single cell.

Figure 2. Single cell genetic analysis of leukaemic cells using the multiplex targeted Q-PCR approach and the BioMark™ HD platform. (A) Heatmap depicting an example of raw Q-PCR data from the BioMark® HD. The rows represent single cells including six cord blood cells and seven REH cells. The columns represent assays, each completed in quadruplicate including *B2M* (one of three assays), *MX1* (two of three assays), *CDKN2A* (two of three assays), *ETV6-RUNX1* fusion gene assay and the *EPOR* SNP (rs318720) assay containing two Taqman® probes, one complementary to the wild-type sequence (labelled with FAM) the other complementary to the SNP sequence (labelled with VIC). The coloured boxes at the junction of a row and column indicate the raw C_T value (according to the key on the right) obtained for a Q-PCR reaction involving the indicated cell and assay. Assays targeting a mutation or the fusion gene provide a definitive positive or negative result indicating the

presence or absence respectively of an alteration. The DNA copy number assays provide a raw C_T value which requires further analysis (standard $\Delta\Delta C_T$ method Applied Biosystems®) to attribute a DNA copy number to the target gene of interest for a single cell; an example can be found in (B) (refer to Single cell analysis in Methods and Supplemental Material). The Q-PCR amplification curves generated from each copy number assay for a single cord blood cell are indicated by the green arrow. A white cross in a coloured box indicates an inadequate amplification curve. (B) Graph depicting the estimated DNA copy number of *MX1* attributed reliably to 89 single cells given the assay results from *B2M* (assay 3) and *MX1 (*assay 3); one of the nine estimated copy number results used to confidently attribute a DNA copy number to the gene of interest for a single cell. The height of the bar indicates the estimated DNA copy number and the colour of the bar the integer; light blue–one copy, dark blue–two copies, green–three copies and yellow-four copies. CB - cord blood. (C) Sub-clonal genetic architecture of the REH cell line inferred by multiplex targeted Q-PCR and confirmed by FISH analysis (126 and 100 cells respectively); percentages in parenthesis are those obtained by FISH analysis. All cells harboured the *ETV6-RUNX1* fusion (F) and the *EPOR* SNP compared to cord blood cells. DNA copy number is indicated for each gene and sub-clonal population. Representative FISH images are shown next to their respective sub-clone. (D) Sub-clonal genetic architecture of leukaemic cells from a child with Down's syndrome and acute lymphoblastic leukaemia (DS-ALL) generated by multiplex targeted Q-PCR and FISH analysis (115 and 100 cells respectively); 98% of cells harboured the *P2RY8-CRLF2* fusion and the *IL7R* mutation (*IL7R*^{m}) by multiplex targeted Q-PCR. Of these, the majority were heterozygous mutations (*IL7R*^{m} hete). A minor sub-clone (2%) had a homozygous *IL7R* mutation (*IL7R*^{m} homo). Loss of *CDKN2A* was sub-clonal to the

IL7R mutation and proceeded to homozygous loss in 11% of cells. FISH for the *P2RY8-CRLF2* fusion and *CDKN2A* copy number confirmed these results (percentages in parentheses); the *IL7R* mutation cannot be detected by FISH.

Figure 3. Phylogenetic analysis results for Cases A and B. In each case the observed clone is indicated by a circle, the yellow circles indicate tumour clones, and black the normal cell population. The alterations are listed below each sub-clone; excluding *ETV6-RUNX1* those without a number indicate the presence of a mutation and those with a number indicate DNA copies accordingly. The boxed sub-clone in grey is inferred; a group of cells which has died out or been outcompeted or if still present, exists at a low frequency that cannot be reliably detected by this approach. The number in italics at each node indicates the jackknifing value. The distance unit is indicated. (A) One parsimonious tree was found for Case A consisting of four sub-clones with modest heterogeneity. The major clone (A4) represents 87.4% of the population. The size of each circle is proportional to the number of single cells included in the sub-clone except A4, which has been reduced by a third in this tree (B) In Case B there are two equally parsimonious trees composed of seven sub-clones. These two trees differ by the position of sub-clones B4 and B5, which are equal parsimonious ancestors to sub-clones B3, B7 and B8. This case shows increased heterogeneity with the major clone representing 54.9% of the population (B3). The major clone B3 is reduced by half in this tree.

Table 1. Whole exome sequencing results for two *ETV6-RUNX*1 positive acute lymphoblastic leukaemia cases

Case	Chr.	Gene[a]	p.Change	Wt Allele	Mut Allele	Effect	Allele Burden estimates by NGS (%) (CI)	Allele Burden estimates by 454 pyroseq (%) (CI)	Allele Burden estimates by Digital PCR (%)	Allele burden estimates for single cells (%) (CI)
Case A	5	***PIK3R1***	p.E518Q	G	C	missense	13.43 (6.7-24.5)	4 (2.1-8.1)	4.30	1.33 (0.6-2.8)
	6	***DAXX***	p.Q565*	G	A	nonsense	10.69 (6.2-17.6)	4 (1.6-10.3)	1.70	0.76 (0.2-2.1)
	7	***EZH2***	p.P577L	G	A	missense	38.13 (30.1-46.8)	56 (48.3-63.3)	36.70	46.58 (42.3-50.9)
	2	*FSIP2*	p.N4564Y	A	T	missense	11.11	-	-	-
	17	*DNAH17*	p.A559G	G	C	missense	33.82	-	-	-
	7	*PON3*	p.R32Q	C	T	missense	85.71	-	-	-
	5	*DNAH5*	p.G3653E	C	T	missense	45.31	-	-	-
	1	*TNR*	p.D1000D	G	A	silent	29.47	-	-	-
	3	***BCHE***	p.I141T	A	G	missense	42.25 (30.8-54.5)	48 (43.1-53.1)	37.10	46.58 (42.3-50.9)
	12	***BAZ2A***	p.P676L	G	A	missense	37.17 (28.4-46.8)	34 (25.8-42.3)	33.40	43.35 (39.1-47.7)
	10	*C10orf112*	p.R545C	C	T	missense	9.48	-	-	-
	8	*DKK4*	p.D95H	C	G	missense	52.80	-	-	-
	13	*SMAD9*	p.P62P	C	T	silent	34.72	-	-	-
Case B	13	***RB1***	p.K810N	G	C	missense	4.65 (1.9-10.3)	4 (1.3-10.5)	1.60	2.39 (1.3-4.3)
	12	***KRAS***	p.G12C	C	A	missense	43.08 (31.1-55.9)	43 (38.2-47.1)	51.0	43.43 (39.1-47.9)
	5	*MAN2A1*	p.T554M	C	T	missense	34.57	-	-	-
	3	*SLC7A14*	p.G150A	C	G	missense	47.37	-	-	-
	3	*KALRN*	p.T1215M	C	T	missense	35.21	-	-	-
	1	***SRSF11***	p.Q22E	C	G	missense	5.30 (3.2-8.6)	4.9 (2.4-9.4)	2.40	2.39 (1.3-4.3)
	3	*COL6A5*	p.V32M	G	A	missense	40.00	-	-	-
	8	*TRHR*	p.I131M	C	G	missense	43.14	-	-	-
	12	*WDR66*	T529FS>NS	A	TCCCC	nonsense	0.17	-	-	-

[a] Mutation targets selected for single cell interrogation are shown in bold text. Confidence intervals (CI) were established using binomial prop. test for each approach. An allele burden of 50% confers either a heterozygous mutation in every cell or a homozygous mutation in 25% of cells.

Table 2. DNA copy number data for two *ETV6-RUNX*1 positive acute lymphoblastic leukaemia cases

Case	CNA	Type	Chr.	Cytoband start	Cytoband end	Size (bp)	Genes[a]
Case A	1	Loss	3	p21.31	p21.31	736.86	*ELP6, CSPG5, SMARCC1, DHX30, MIR1226, MAP4, CDC25A, CAMP*
	1	Loss	3	p21.31	p21.31	222.80	*PFKFB4, UCN2, COL7A1, MIR711, UQCRC1, TMEM89, SLC26A6, CELSR3, NCKIPSD, IP6K2, PRKAR2A*
	1	Loss	3	p21.31	p21.31	383.99	*RHOA, TCTA, AMT, NICN1, DAG1, BSN, APEH, MST1, RNF123, AMIGO3, GMPPB, IP6K1*
	1	Loss	3	p21.31	p21.2	634.70	*RBM5, SEMA3F, GNAT1……..CACNA2D2, C3orf18, HEMK1, CISH, MAPKAPK3, DOCK3*
	1	Loss	6	q14.1	q27	91165.16	many genes including ***CCNC***
	0	Loss	7	p14.1	p14.1	49.85	*TRG*
	1	Loss	8	p23.1	p23.1	217.17	*SgK223*
	1	Loss	9	p24.3	p24.3	65.91	*SMARCA2*
	1	Loss	9	p24.3	p24.3	126.76	*DMRT1, DMRT3, DMRT2*
	1	Loss	21	q22.11	q22.11	255.24	*MRPS6, SLC5A3, LINC00310*
	1	Loss	X	p22.33	q25	121029.37	many genes including ***TBL1X***
Case B	1	Loss	4	q31.3	q31.3	283.60	*FBXW7*
	1	Loss	4	q28.3	q28.3	260.02	
	1	Loss	4	q33	q33	95.85	*MFAP3L, AADAT*
	1	Loss	7	p14.1	p14.1	116.23	*TRG*
	1	Loss	7	q34	q34	140.04	*TRB*
	1	Loss	8	q24.11	q24.11	58.36	*EXT1*
	1	Loss	9	p21.3	p21.3	511.07	*LOC554202, IFNE, MIR31, MTAP, C9orf53,* ***CDKN2A***
	1	Loss	9	p13.3	p13.2	1572.02	*CREB3, GBA2, RGP1, MSMP, NPR2, SPAG8……RNF38, MELK,* ***PAX5****, ZCCHC7*
	1	Loss	10	q21.1	q21.1	53.17	*BICC1*
	1	Loss	11	q13.4	q13.4	55.52	*CHRDL2*
	1	Loss	14	q24.2	q24.2	123.19	***DPF3***
	1	Loss	22	q11.22	q11.22	818.48	***VPREB1****, LOC96610, ZNF280B, ZNF280A, PRAME, LOC648691, POM121L1P, GGTLC2, MIR650*

[a] Target genes selected for single cell analysis located in regions of loss are shown in bold text.

REFERENCES

Anderson K, Lutz C, van Delft FW, Bateman CM, Guo Y, Colman SM, Kempski H, Moorman AV, Titley I, Swansbury J et al. 2011. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**: 356-361.

Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B et al. 2012. Genome-wide copy number analysis of single cells. *Nat Protoc* **7**(6): 1024-1041.

Bateman CM, Colman SM, Chaplin T, Young BD, Eden TO, Bhakta M, Gratias EJ, van Wering ER, Cazzaniga G, Harrison CJ et al. 2010. Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood* **115**(17): 3553-3558.

Chin L, Andersen JN, Futreal PA. 2011. Cancer genomics: from discovery science to personalized medicine. *Nat Med* **17**(3): 297-303.

Citri A, Pang ZP, Sudhof TC, Wernig M, Malenka RC. 2012. Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat Protoc* **7**(1): 118-127.

Clark J, Attard G, Jhavar S, Flohr P, Reid A, De-Bono J, Eeles R, Scardino P, Cuzick J, Fisher G et al. 2008. Complex patterns of *ETS* gene alteration arise during cancer development in the human prostate. *Oncogene* **27**: 1993-2003.

Diaz Jr LA, Williams RT, Wu J, Kinde I, Hecht JR, Berlin J, Allen B, Bozic I, Reiter JG, Nowak MA et al. 2012. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**: 537-540.

Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpet P et al. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine* **366**(10): 883-392.

Greaves M. 2013. Cancer stem cells as 'units of selection'. *Evol Appl* **6**(1): 102-108.

Greaves M, Maley CC. 2012. Clonal evolution in cancer. *Nature* **481**: 306-313.

Horsley SW, Colman S, McKinley M, Bateman CM, Jenney M, Chaplin T, Young BD, Greaves M, Kearney L. 2008. Genetic lesions in a preleukemic aplasia phase in a child with acute lymphoblastic leukemia. *Genes, chromosomes & cancer* **47**(4): 333-340.

Huson DH, Scornavacca C. 2012. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Syst Biol*.

Kitagawa Y, Inoue K, Sasaki S, Hayashi Y, Matsuo Y, Lieber MR, Mizoguchi H, Yokota J, Kohno T. 2002. Prevalent involvement of illegitimate V(D)J recombination in chromosome 9p21 deletions in lymphoid leukemia. *J Biol Chem* **277**: 46289-46297.

Klein CA, Stoecklein NH. 2009. Lessons from an aggressive cancer: evolutionary dynamics in esophageal carcinoma. *Cancer Res* **69**(13): 5285-5288.

Klemm L, Duy C, Iacobucci I, Kuchen S, Von Levetzow G, Feldhahn N, Henke N, Li Z, Hoffmann TK, Kim Y-m et al. 2009. The B cell mutator AID promotes B lymphoid blast crises and drug resistance in chronic myeloid leukemia. *Cancer Cell* **16**: 232-245.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5): 589-595.

Liu W, Laitinen S, Khan S, Vihinen M, Kowalski J, Yu G, Chen L, Ewing CM, Eisenberger MA, Carducci MA et al. 2009. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* **15**(5): 559-565.

Longerich S, Basu U, Alt F, Storb U. 2006. AID in somatic hypermutation and class switch recombination. *Curr Opin Immunol* **18**: 164-174.

Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, Paulson TG, Blount PL, Risques RA, Rabinovitch PS et al. 2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet* **38**(4): 468-473.

Margush T, McMorris FR. 1981. Consensus n-trees. *Bull Math Biol* **43**: 239-244.

Marusyk A, Almendro V, Polyak K. 2012. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer* **12**: 323-334.

Merlo LM, Shah NA, Li X, Blount PL, Vaughan TL, Reid BJ, Maley CC. 2010. A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prev Res (Phila)* **3**(11): 1388-1397.

Merlo LMF, Pepper JW, Reid BJ, Maley CC. 2006. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer* **6**: 924-935.

Mroz EA, Tward AD, Pickering CR, Myers JN, Ferris RL, Rocco JW. 2013. High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer*.

Mullighan CG, Collins-Underwood JR, Phillips LA, Loudin MG, Liu W, Zhang J, Ma J, Coustan-Smith E, Harvey RC, Willman CL et al. 2009. Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. *Nat Genet* **41**(11): 1243-1246.

Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB et al. 2007. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**: 758-764.

Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, Downing JR. 2008. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**: 1377-1380.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* **472**(7341): 90-94.

Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge J, Grubor V et al. 2010. Inferring tumor progression from genomic heterogeneity. *Genome Res* **20**: 68-80.

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M et al. 2012. The life history of 21 breast cancers. *Cell* **149**: 994-1007.

Notta F, Mullighan CG, Wang JC, Poeppl A, Doulatov S, Phillips LA, Ma J, Minden MD, Downing JR, Dick JE. 2011. Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. *Nature* **469**(7330): 362-367.

Nowell PC. 1976. The clonal evolution of tumor cell populations. *Science* **194**: 23-28.

Page RMD, Holmes EC. 1998. *Molecular evolution: a phylogenetic approach*. Wiley-Blackwell.

Park SY, Gönen M, Kim HJ, Michor F, Polyak K. 2010. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J Clin Invest* **120**(2): 636-644.

Pina C, Fugazza C, Tipping AJ, Brown J, Soneji S, Teles J, Peterson C, Enver T. 2012. Inferring rules of lineage commitment in haematopoiesis. *Nat Cell Biol* **14**(3): 287-294.

RCoreTeam. (2013). R: A language and environment for statistical computig. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Stephens PJ, McBride DJ, Lin M-L, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ et al. 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**: 1005-1010.

Swanton C. 2012. Intratumor heterogeneity: evolution through space and time. *Cancer Res* **72**(19): 4875-4882.

Swofford D. 2005. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) Version 4.0 Beta 10*. Sinauer Associates, Inc. Sunderland.

Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B et al. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**(39): 16910-16915.

Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J et al. 2011. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**(7331): 539-542.

Vogelstein B, Papadopoulos N, Velculescu V, Zhou S, Diaz L, Kinzler K. 2013. Cancer genome landscapes. *Science (New York, NY)* **339**(6127): 1546-1558.

Waanders E, Scheijen B, van der Meer LT, van Reijmersdal SV, van Emst L, Kroeze Y, Sonneveld E, Hoogerbrugge PM, van Kessel AG, Van Leeuwen FN et al. 2012. The origin and nature of tightly clustered *BTG1* deletions in precursor B-cell acute lymphoblastic leukemia support a model of multiclonal evolution. *PLoS Genet* **8**(2): e1002533.

Wang J, Lin M, Crenshaw A, Hutchinson A, Hicks B, Yeager M, Berndt S, Huang W-Y, Hayes R, Chanock S et al. 2009. High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. *BMC genomics* **10**: 561.

Weaver S, Dube S, Mir A, Qin J, Sun G, Ramakrishnan R, Jones RC, Livak KJ. 2010. Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods* **50**(4): 271-276.

Wiemels JL, Greaves M. 1999. Structure and possible mechanisms of TEL-AML1 gene fusions in childhood acute lymphoblastic leukemia. *Cancer Res* **59**(16): 4075-4082.

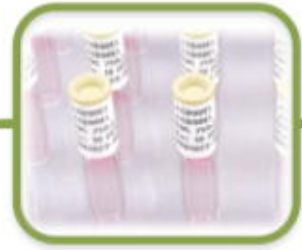
Wolman SR. 1986. Cytogenetic heterogeneity: its role in tumor evolution. *Cancer Genet Cytogenet* **19**: 129-140.

Yates LR, Campbell PJ. 2012. Evolution of the cancer genome. *Nat Rev Genet* **13**(11): 795-806.

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

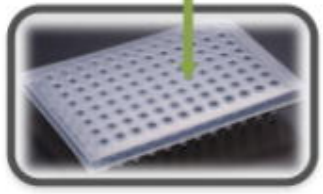
Zong C, Lu S, Chapman AR, Xie XS. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**(6114): 1622-1626.

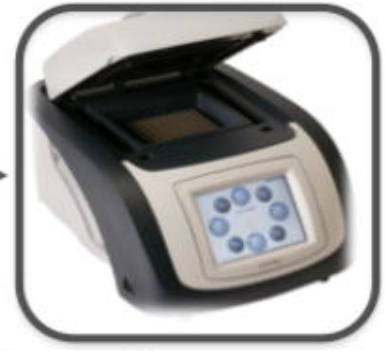
Single cell analysis workflow



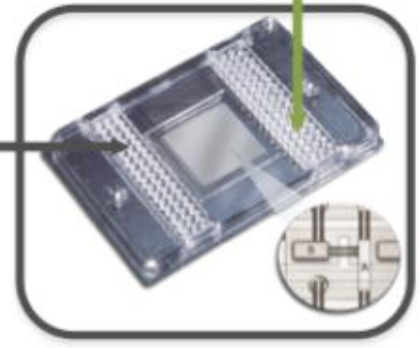
TaqMan assays for gene targets

Single cell sort





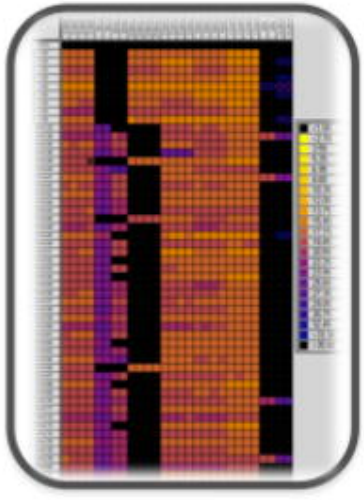
Cell lysis & DNA-specific target amplification



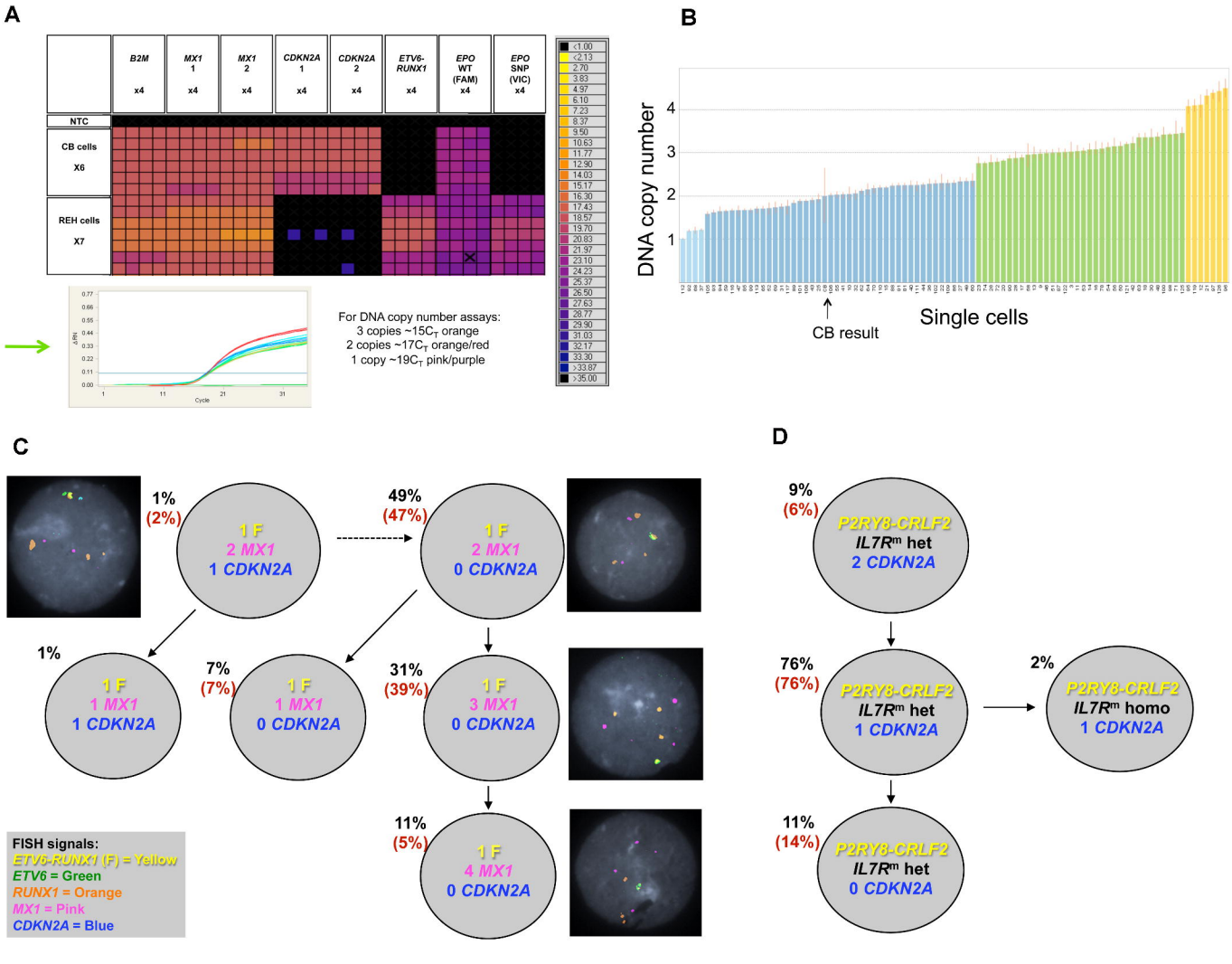
Microfluidic dynamic array

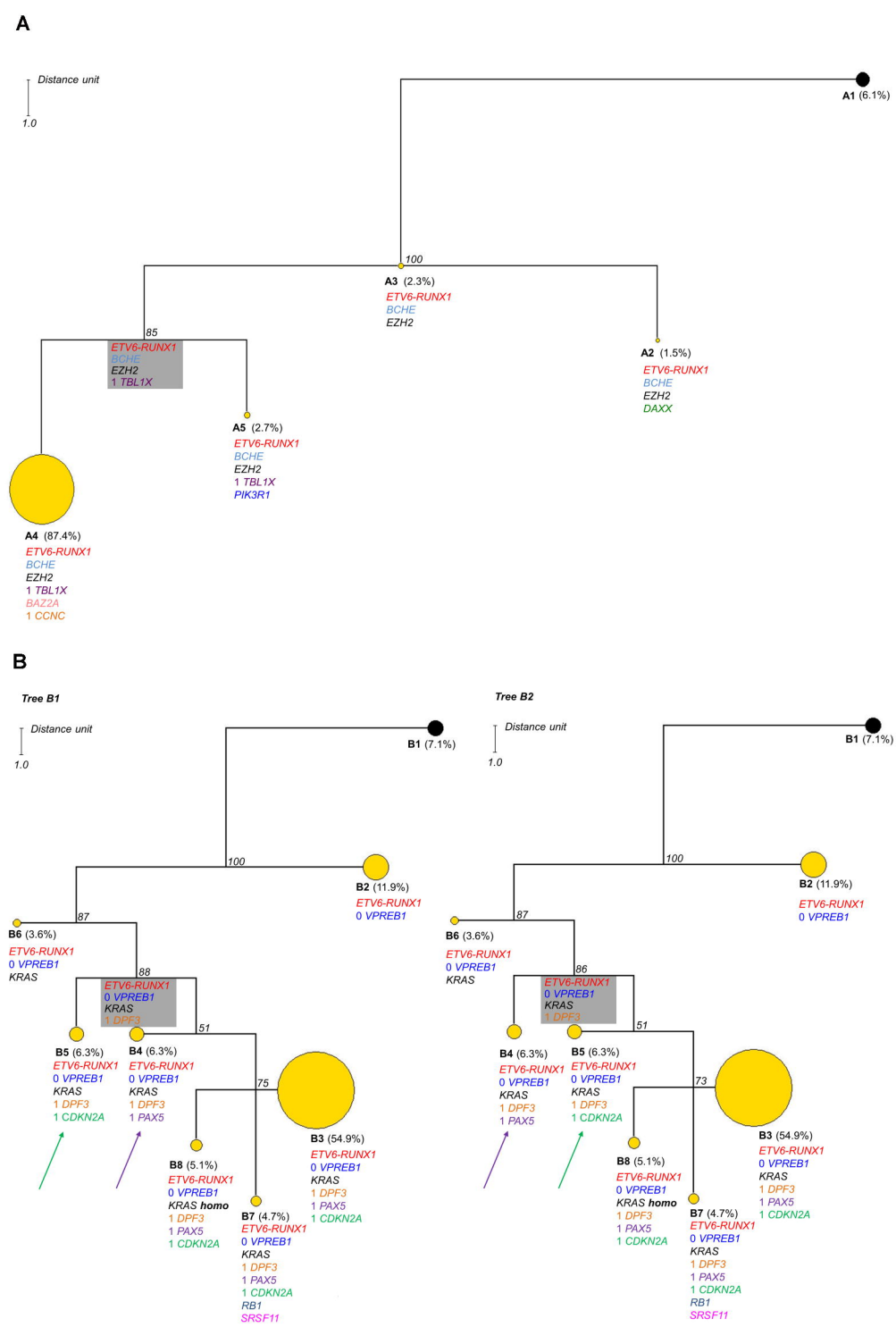
Integrated fluidic circuit (IFC) controller for array loading

Biomark HD for Q-PCR



Q-PCR analysis


A
B2M x4
MX1 1 x4
MX1 2 x4
CDKN2A 1 x4
CDKN2A 2 x4
ETV6-RUNX1 x4
EPO WT (FAM) x4
EPO SNP (VIC) x4
NTC
CB cells X6
REH cells X7
For DNA copy number assays:
3 copies ~15CT orange
2 copies ~17CT orange/red
1 copy ~19CT pink/purple
B
DNA copy number
Single cells
CB result
C
1% (2%)
1 F
2 MX1
1 CDKN2A
49% (47%)
1 F
2 MX1
0 CDKN2A
1%
1 F
1 MX1
1 CDKN2A
7% (7%)
1 F
1 MX1
0 CDKN2A
31% (39%)
1 F
3 MX1
0 CDKN2A
11% (5%)
1 F
4 MX1
0 CDKN2A
FISH signals:
ETV6-RUNX1 (F) = Yellow
ETV6 = Green
RUNX1 = Orange
MX1 = Pink
CDKN2A = Blue
D
9% (6%)
P2RY8-CRLF2
IL7Rm het
2 CDKN2A
76% (76%)
P2RY8-CRLF2
IL7Rm het
1 CDKN2A
2%
P2RY8-CRLF2
IL7Rm homo
1 CDKN2A
11% (14%)
P2RY8-CRLF2
IL7Rm het
0 CDKN2A


A
Distance unit
1.0
A1 (6.1%)
100
A3 (2.3%)
ETV6-RUNX1
BCHE
EZH2
85
ETV6-RUNX1
BCHE
EZH2
1 TBL1X
A2 (1.5%)
ETV6-RUNX1
BCHE
EZH2
DAXX
A5 (2.7%)
ETV6-RUNX1
BCHE
EZH2
1 TBL1X
PIK3R1
A4 (87.4%)
ETV6-RUNX1
BCHE
EZH2
1 TBL1X
BAZ2A
1 CCNC
B
Tree B1
Distance unit
1.0
B1 (7.1%)
100
B2 (11.9%)
ETV6-RUNX1
0 VPREB1
87
B6 (3.6%)
ETV6-RUNX1
0 VPREB1
KRAS
88
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
51
B5 (6.3%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 CDKN2A
B4 (6.3%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 PAX5
75
B3 (54.9%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 PAX5
1 CDKN2A
B8 (5.1%)
ETV6-RUNX1
0 VPREB1
KRAS homo
1 DPF3
1 PAX5
1 CDKN2A
B7 (4.7%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 PAX5
1 CDKN2A
RB1
SRSF11
Tree B2
Distance unit
1.0
B1 (7.1%)
100
B2 (11.9%)
ETV6-RUNX1
0 VPREB1
87
B6 (3.6%)
ETV6-RUNX1
0 VPREB1
KRAS
86
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
51
B4 (6.3%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 PAX5
B5 (6.3%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 CDKN2A
73
B3 (54.9%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 PAX5
1 CDKN2A
B8 (5.1%)
ETV6-RUNX1
0 VPREB1
KRAS homo
1 DPF3
1 PAX5
1 CDKN2A
B7 (4.7%)
ETV6-RUNX1
0 VPREB1
KRAS
1 DPF3
1 PAX5
1 CDKN2A
RB1
SRSF11