SUPPLEMENTAL MATERIAL

1. Supplemental Methods

2. Supplemental Figures

- **Supplemental Figure 1**. Fluorescence-activated cell sorting plots illustrating gating approaches to efficiently isolate single cell

- **Supplemental Figure 2**. *EPOR* SNP assay amplification curves generate by Q-PCR for a single REH cell
- **Supplemental Figure 3**. Box and whisker plots to illustrate the range of raw C_T values generated by Q-PCR from wells that contained either one, two, three or four cells with a normal diploid karyotype
- **Supplemental Figure 4**. Example of Sankoff's algorithm

3. Supplemental Tables

- **Supplemental Table 1A.** QC metrics for Affymetrix® Cytogenetics Whole Genomic 2.7M arrays
- **Supplemental Table 1B.** QC metrics for Affymetrix® Genome-Wide Human SNP Array 6.0
- **Supplemental Table 2.** Custom and designed assays for each sample specific mutation used in this study
- **Supplemental Table 3.** Copy number assays for each sample specific alteration in this study
- **Supplemental Table 4.** Details and explanations of cell data removed from each case experiment
- **Supplemental Table 5.** FISH probes and scoring frequencies for each sample specific alteration in this study
- **Supplemental Table 6.** Character states graphs and corresponding matrices employed in this manuscript
- **Supplemental Table 7.** Jackknife analyses within the parsimonious trees of patient A and B employing three different percentage of character deletion

4. Supplemental Appendix.

- ***Basic terms and concepts in phylogenetics applied to cancer evolution***

5. Supplemental References

1. SUPPLEMENTAL METHODS

1.1 Bulk cell analysis

1.1.1 SNP array analysis for REH and the DS-ALL sample

To define copy number alterations and SNPs for the REH cell line and the DS-ALL sample we used the Affymetrix® Genome-Wide Human SNP Array 6.0 (Affymetrix®, Santa Clara, CA, USA). Remission DNA was not available for the DS-ALL case so controls for both samples were 20 HapMap Caucasian samples (hapmap.ncbi.nlm.nih.gov). Briefly, according to manufacturer's guidelines, 500ng of sample DNA was digested using restriction endonucleases *Nsp*I and *Sty*I, ligated to an adaptor, and then PCR amplified with adaptor-specific primers. The PCR products were digested using *DNase*I and labelled with a biotinylated nucleotide analogue. The resulting labelled DNA fragments were then hybridized to the microarray, stained using streptavidin-phycoerythrin conjugates and washed using the Affymetrix® Fluidics Station 450. The GeneChip® scanner 3000 7G was used to scan the arrays and the image was acquired using Affymetrix® GeneChip® Operating Software (GCOS version 1.4). Genotyping was performed in the Genotyping Console 4.0 software (Affymetrix®) using the Birdseed clustering algorithm. Contrast QCs and call rates are displayed in Supplemental Material Table 1B. Copy number analysis was completed using Partek® Genomics Suite™ Software (Partek Inc, Missouri, US) and the Hidden Markov Model default settings in the Copy Number Workflow.

1.1.2 Whole exome sequencing – additional information

The in house-variant caller CaVEMan (Cancer Variants through Expectation Maximisation) uses a naïve Bayesian classifier to estimate the posterior probability

of each possible genotype (wild-type, germline SNP, somatic SNV) at a given base, accounting for the effects of observables such as base quality (measuring signal:noise ratio), read position, sequencing lane, and read orientation. CaVEMan is configured to incorporate knowledge of copy number and normal cell contamination in the posterior probability calculations. Several post-processing filters were applied to the set of initial CaVEMan SNV calls in order to increase the specificity of the output. Initially, at least 1/3 of mutant alleles in tumour reads are of quality ≥ 25. At least 1 mutant allele in a tumour read must fall in the middle third of the read, unless the tumour read depth is less than 10, when a mutant allele the first third is acceptable. There is no more than 1 high quality (≥ 20) mutant allele in a normal read.

To call insertions and deletions, split-read mapping was implemented as a modification of the pindel algorithm [1]. The search for indels included read-pairs in which one or both ends map to the genome, but allow one of the pair to have mismatches, insertions or deletions. Pindel searches for reads where one end is anchored on the genome, and the other end can be mapped with high confidence in two (split) portions, spanning a putative indel. As completed for the CaVEMan output, we applied several post processing filters to the pindel output in order to improve specificity. Two classes of indel were identified: >4bp and <4bp. For both classes the following filters were applied to the raw output; >3 tumour reads must report putative indel variant; <5% of calls must occur in germline sequencing data and when no wild-type coverage in BAM; Pindel must not call an event in the wild-type. For small events the following filters were applied; tumour with BAM depth of <200 reads must have variant call in ≥8% of reads; tumour with BAM depth of ≥200 reads must have variant call in ≥4% of reads; germline BAM must have >5 reads

spanning the region; Pindel calls in germline reads must be ≤5% of the germline BAM depth; if the tumour BAM depth > wild-type BAM depth normalise the Pindel wild-type calls against this; discarding if new value is ≥5% reference; apply poly-nucleotide tract filter for events with repetitive region >9 repeats; germline BAM depth must be ≥8% of tumour BAM depth; tumour BAM must have <8% BWA reference calls vs BWA variant calls. Furthermore, for large events no germline sequencing reads should be called as part of an event by Pindel and exome data results must annotate to coding regions of the genome.

Copy number analysis was performed using ASCAT (version 2.2) [2] taking into account non-neoplastic cell infiltration and tumour aneuploidy, and resulted in integral allele-specific copy number profiles for the tumour cells. Allele-specific copy number estimates for point mutations and indels were obtained by integrating copy number and sequencing data. In a sample containing only tumour cells, the number of reads, r, with a mutation can be expressed as:

Equation (1.1)

$$r = \frac{n_{mut} R}{n_{locus}}$$

In equation (1.1), n_{locus} is the copy number of the locus, n_{mut} is the number of mutated copies and R is the total number of reads from that locus. In case of a tumour sample consisting of a fraction of tumour cells ρ, infiltrated with a fraction of normal cells 1 − ρ (assumed to have two copies), equation (1.1) becomes equation (1.2):

Equation 1.2

$$r = \frac{n_{mut} R \rho}{\rho n_{locus} + 2(1 - \rho)}$$

Hence, allele-specific copy number estimates for point mutations and indels can be described as:

Equation 2

$$n_{mut} = f_s \frac{1}{\rho}\left(\rho n_{locus} + 2(1-\rho)\right)$$

In equation (2), fs = r/R is the frequency of mutated reads observed in the sequencing data, and ρ and n locus can be obtained from the ASCAT copy number analysis. These copy number estimates of mutations were used to determine which mutations are likely sub-clonal: if nmut ≥ 0.8, the mutation is called likely clonal and if nmut < 0.8, the mutation is called likely sub-clonal.

In the case of indels, reads with an insertion or deletion may not map as well as reads without insertions and deletions. Therefore, a procedure was followed to estimate fs for indels that was independent of ease of mapping. Reads were obtained by matching flanking sequence (10 bp on each side) around the indel, further filtered to exclude spurious matches. The mutated read frequency was subsequently calculated, accounting for the difference in sequence lengths with and without the indel:

Equation 3

$$f_s = \frac{\frac{r_{indels}}{(l_s - l_{indels} + 1)}}{\frac{r_{indels}}{(l_s - l_{indels} + 1)} + \frac{r_{normal}}{(l_s - l_{normal} + 1)}}$$

$$f_s = \frac{r_{indels}/(l_s - l_{indels} + 1)}{r_{indels}/(l_s - l_{indels} + 1) + r_{normal}/(l_s - l_{normal} + 1)}$$

In equation (3), r_{indel} and r_{normal} are the respective numbers of reads with and without the indel, l_s is the read length (76 bp), and l_{indel} and l_{normal} are the respective lengths of the matching fragment in sequences with and without the indel.

For validation all putative somatic indels were confirmed by capillary sequencing on the tumour and germline DNA from that patient. In <10% of calls, the capillary sequencing gave noisy traces, and we report variants where there was convincing evidence for the mutation on exome data (high coverage; good quality sequencing and mapping; high fraction of reads reporting the variant in the tumour; no reads reporting variant in matched germline sequencing. Validation of putative somatic substitutions was performed via Roche pyrosequencing of both tumour and remission samples. Primers were designed to generate 275-425 bp fragments suitable for Roche 454 pyrosequencing. Pyrosequencing data were evaluated for the presence of the mutant allele in the tumour sample. SNVs were annotated as somatic only when mutant alleles were present in the tumour sample. 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.

1.2 Single cell analysis

1.2.1 Single cell labeling and flow sorting

Patient samples were thawed from liquid nitrogen stored cryovials by gentle warming in lukewarm water followed by resuspension in 9ml RPMI-1640 medium 10% FCS. Cells were pelleted and washed in PBS ($1x10^6$ cultured cells were pelleted prior to washing). The cells were resuspened in 5ml PBS containing 10µM carboxyfluorescin diacetate, succinimidyl ester (CFSE) and incubated at 37°C for 15 minutes. The cells

were then pelleted, resuspended in RPMI-1640 medium 10% FCS and incubated at 37°C for 30 minutes. Finally, the cells were pelleted and resuspended in 100μl PBS prior to sorting. CFSE is an *in vivo* cell viability tracer that passively diffuses into cells and only fluoresces once intracellular esterases cleave the acetyl groups from the compound.

Single cell sorting was performed on a BDFACSAria1–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. These flow settings allow an average time between events (at worst) just less than 100 times larger than the window used to select the event; therefore with a monodispersed sample the chances of selecting two events in one sort window is small. Cell selection by forward-scattered light (FSC) and side-scattered light (SSC) accounted for cell size and internal complexity allowing accurate selection of single cells avoiding doublets and clumps (Supplemental Figure 1).

To further assess the efficiency of single cell sorting Beckman Coulter Flow-check beads (Beckman Coulter Inc®, CA, USA) were sorted singly onto an Ampligrid slide (Beckman Coulter Inc®) and counted using a Nikon Eclipse 50i fluorescence microscope (Nikon Corporation®, Japan) to confirm one particle per sorted droplet. Once labelled with CFSE as previously described, cell suspensions were sorted and assessed in the same way to confirm single cells by microscopy. If success was achieved on 98% of occasions we proceeded with single cell sorting. This was completed prior to each experiment.

1.2.2 Single Cell Q-PCR analysis

The BioMark™ HD generates a C_T value for each reaction (this is the PCR cycle at which the concentration of free emitter dye fluorescence is detected by the instrument). The C_T value therefore indicates the amount of DNA after the amplification phase and confers the DNA copy number or the presence of a SNV or fusion. DNA copy number (*CDKN2A* and *MX1*) and *EPOR* SNP assay amplification curves in a single REH cell can be found in Figure 2a and Supplemental Figure 2 respectively. 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.

The $\Delta\Delta C_T$ method (Applied Biosystems®) was employed to determine a copy number for each locus with modifications to incorporate data from multiple Taqman® assays targeting the same genome region. This method determines the mean ΔC_T value of quadruplicates from an endogenous reference gene, in this case *B2M* (diploid), and a target gene of interest for both a calibrator cell (diploid) and the cell of interest (REH/ALL - unknown ploidy). Normalising the result of each target gene by a reference gene corrects for experimental variations. The corrected mean ΔC_T value for each target gene of the calibrator cell is then subtracted from that of the cell of interest generating a ratio referred to as the $\Delta\Delta C_T$. This ratio is an estimated copy number and represented by the following equation:

$$\text{Copy number} = cn_c 2^{-\Delta\Delta C_T}$$

where cn_c is the copy number of the target gene in the calibrator cell and $\Delta\Delta C_T$ is the difference between ΔC_T of the cell of interest and the calibrator cell.

To ensure robust data from a system that can be influenced by assay efficiency and experimental variability we used three distinct assays to target *B2M* and the region of interest and calculated the DNA copy number estimates as described above. Normalising every target gene assay by each reference gene assay generated nine estimated copy number results for a region of interest. A confidence metric was assigned to the estimated copy number. For example a confidence value of 90% indicates that there is a 10% chance that the true copy number differs from the estimated copy number (according to ABI CopyCaller® Software v2). The inferred confidence is a function of the estimated copy number and replicate mean and is calculated as:

$$\text{Confidence}\left(\mu_r, cn_{estimated}\right) = \left[1 + \sum_{cn \neq cn_a} \frac{\Pi_{cn}}{\Pi_{cn_a}} e^{-\Omega}\right]^{-1}$$

where *a* (subscripted) = estimated, μ_r = replicate mean for the sample, $cn_{estimated}$ = copy number given to the sample, Π_{cn} = probability of copy number *cn*, Ω is calculated as:

$$\Omega = \frac{1}{\sigma^2 \log(1+E)} \log\left(\frac{cn}{cn_a}\right)\left((\hat{\mu}_r - K) + \frac{\log(cn_a cn)}{2\log(1+E)}\right)$$

where σ^2 = standard deviation of the sub-distributions, *E* = PCR efficiency of the target assay, *K* = constant in the function relating the sub-distribution mean (μ_{cn}) to copy number (*cn*) calculated as:

$$\mu_{cn} = K - \frac{1}{log(1+E)} log(cn)$$

To calculate the actual DNA copy number for the region of interest taking into consideration all nine estimates (where they were deemed to be reliable), we then

calculated the weighted mean of the estimated copy numbers according to the confidence metric attributed to each. This reduced the contribution of less reliable estimated DNA copy number results to the final DNA copy number result. 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 determine DNA copy numbers greater than four [3] or only one of the nine DNA copy number results for a given region was deemed reliable.

The weighted standard error was also calculated in conjunction with the weighted mean and the weighted mean only finally accepted for a cell of interest if the attributed weighted standard error did not exceed the maximum weighted standard error value generated from the control plate of 48 cord blood cells.

1.3 Single cell method validation

1.3.1 Assay validation

The correlation coefficient of each DNA copy number assay according to the manufacturer is at least 0.98. To assess the efficiency of each assay in our single cell system we used the standard curve method and data from single, two and three collectively sorted cells; 18 data points per group. As expected the correlation coefficient was not always above 0.98 in our system. However, only assays that presented curves with correlation coefficients of 0.90 and higher were used for DNA copy number analysis.

A control experiment was completed in a panel of 48 diploid cord blood cells to determine the frequency of false positive calls for SNVs and DNA copy number assays. The number of wells that presented a fusion gene, SNV or CNA indicated

the reliability of each assay and this was used as a threshold to estimate the error rate in sub-clonal population frequencies and define a cut-off for accepted populations (Supplementary Tables 2 and 3).

1.3.2 Detection of duplicate cells in a single well

Whilst it is not possible to identify the number of wells with two or more cells visually or by copy number analysis it is possible to estimate the amount of DNA in each well for each reference gene. With this information it is possible to identify those wells with above average amounts DNA indicating the presence of two or more cells and remove the data from further analysis. Briefly, one, two, three or four cells were sorted into consecutive wells of a 96 well plate for Q-PCR analysis and DNA quantification; eighteen for each group. We used the control gene assay *B2M*-1 to quantify the amount of DNA for each well and then considered the range for each cell number (Supplemental Figure 3). The results indicate a significant difference in average C_T between the eighteen wells containing one and two cells and two and three cells; the assay reaches saturation with DNA from four cells. Using this data cells sorted with the aim of obtaining a single cell but with a raw *B2M*-1 C_T value lower than the upper quartile obtained in this control experiment when two cells are sorted were removed from the final analysis as potentially two cells have been sorted into the same well. This ensures that only data from single target cells is used for phylogenetic analysis. The rate was approximately one well per 96 well plate but given our single cell visual assessment prior to sorting it was felt that this was an over estimation; however an acceptable loss to ensure robust data from single cells.

1.3.3 Data exclusion from Q-PCR and phylogenetic analysis

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 removed from further phylogenetic analysis. On average 75% of interrogated single cells generated complete comprehensive results. Supplemental Table 4 provides details and explanations of cell removed from each case experiment.

1.4 Phylogenetic analysis and clonal evolution

1.4.1 Phylogenetic characters and motifs

In order to analyse the clonal expansion and to infer the evolution of ALL we have associated the experimental observed data with a signature or a motif to label each cell. Each examined cell shows, for each unit assayed (i.e. gene), a determined genotype or a copy number alteration which can be concatenated in a linear array of *n* characters in length where *n* is the number of units assayed; each character (c) represents the observed unit state.

This array of *n* characters represents the genomic state of each cell, encloses signatures of the evolutionary history and thus, we can consider it as a phylogenetic motif:

$$C = \{c_1, \dots, c_n\}$$

where *C* is the observed finite motif of each cell and c_i represents each observation.

All cells with the same motif are then grouped together to form a clone or a taxon. Each character state of the motif may be phylogenetically informative and can be assumed as a phylogenetic character state; a set of mutually exclusive states with a fixed order of evolution. Each state directly evolves from another and a set of the observed character states can evolve from one ancestor character state known as the nearest common ancestor [4]. According to parsimonious analysis each character state, either observed or inferred, is assigned to a node or tip of an evolutionary tree [5].

1.4.2 Character state graph

Phylogenetic analysis is governed by assumptions given to each character state that determine its evolution[5]. For discrete characters with a limited number of possible states we can describe these assumptions using a visual representation adopted from the graph theory [6].

Consider a graph:

$$G = (V, E, \phi)$$

where:

(1) V is a n empty finite set called *vertices* of G (singular *vertex*);

(2) E is a consecutive set of e-elements of G called *edges* of G where the marginal e-elements are subsets of V;

(3) $V \cap E \neq \emptyset$;

(4) ϕ is a function with domain E and codomain $P(V)$ such that:

$$\phi\ (e) = \{V_i, V_j\}\ \forall\ e \in E,\ \forall\ V_i, V_j \in V.$$

Given a graph *G*, we denote it pictorially by drawing a dot (•) for each vertex in *V (G)*, and lines or arcs for each $e \in E\ (G)$ connecting the dots that represent vertices in ϕ (e).

We can, therefore assign an order to each vertex and a graphical direction to each edge to model a directional graph where each edge is a path carrying a direction from one vertex to the other; both directions can be allowed. To illustrate this principle consider the following graph:

$$G = (V, E, \phi)$$

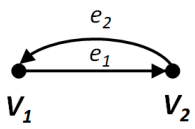
where:

$$V = \{V_1, V_2\};$$

$$E = \{e_1, e_2\};$$

$$\phi = \begin{pmatrix} e_1 & e_2 \\ \{V_1,V_2\} & \{V_2,V_1\} \end{pmatrix}.$$

This is a graph formally represented by a set of two vertices (V_1, V_2) connected by two edges (e_1, e_2). Each has a direction from V_1 to V_2 for the first and from V_2 to V_1 for the second:



This directional graph can be expanded to a set of *n* vertices where the main vertices represent the source or the origin of the graph while all others constitute the vertices of each edge. Assuming the vertex V_1 is the source, two different examples are given below:

1) Ordered and bidirectional graph:

$$G = (V, E, \phi)$$

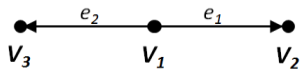
where:

$$V = \{V_3, V_1, V_2\};$$

$$E = \{e_1, e_2\};$$

$$\phi = \begin{pmatrix} e_1 & e_2 \\ \{V_1,V_2\} & \{V_1,V_3\} \end{pmatrix}.$$

where the source vertex V_1 is ordered in the central position and is connected to the remaining vertices V_2 and V_3 by two edges (e_1, and e_2,). The only two paths are from V_1 towards V_2 or V_3.



2) Ordered and unidirectional graph:

$$G = (V, E, \phi)$$

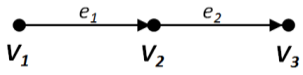
where:

$$V = \{V_1, V_2, V_3\};$$

$$E = \{e_1, e_2\};$$

$$\phi = \begin{pmatrix} e_1 & e_2 \\ \{V_1,V_2\} & \{V_2,V_3\} \end{pmatrix}.$$

where the source vertex V_1 is ordered at the left of the linear graph. In this case only one direction is allowed starting from V_1 and ending with V_3 but via V_2.



We then applied the graph theory to each unit assayed (i.e. gene). Each edge represents a character state transition or a step (e.g. DNA copy number loss) and

each vertex corresponds to a character state (e.g. one, two or three DNA copies). Therefore each graph can model the assumptions made for each gene. The source vertex (V_1) represents the ancestral state of 2 copies, while the other vertices represent copy number alterations. For example, suppose the experimental data from a gene indicates that single cell DNA copy number alterations range from 0 to 4, if we assume that reverse alterations are not possible, the directional graph can be drawn as:

$$G_{gene1} = (V, E, \phi)$$

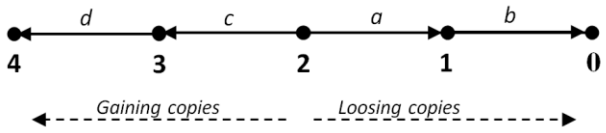
where:

$$V = \{4, 3, 2, 1, 0\};$$

$$E = \{a, b, c, d\};$$

$$\phi = \begin{pmatrix} a & b & c & d \\ \{2,1\} & \{1,0\} & \{2,3\} & \{3,4\} \end{pmatrix}.$$

The graph is linear and bidirectional where the ancestral state occupies the centre. The direction allowed is that which leads to DNA copy number loss (edges *a* and *b*) or DNA copy number gain (edges *c* and *d*); the reverse direction is not allowed. The graph is as follows:



In the case of SNVs, there are only three possible character states: ancestral (*anc*), one mutation (heterozygous) and two mutations (homozygous). Although the environment in which the clonal expansion arises is under selection[7] we cannot exclude back mutations and thus, we need a graph that considers multidirectional character states:

$$G_{gene2} = (V, E, \phi)$$

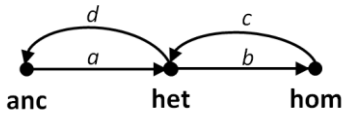
where:

$$V = \{anc, het, hom\};$$

$$E=\{a, b, c, d,\};$$

$$\phi = \begin{pmatrix} a & b & c & d \\ \{anc,het\} & \{het,hom\} & \{hom,het\} & \{het,anc\} \end{pmatrix}.$$

The graph is linear and multidirectional where the ancestral state occupies the left of the graph. All paths are allowed except the from the ancestor to homozygous and back, which are assumed to be impossible:



1.4.3 Step Matrix

Once a character state graph has been defined which imposes an order and direction upon each phylogenetic assumption, we can use a matrix to represent the evolutionary cost for each transition or step from one character state to another. Consider a set of numbers arranged in a rectangular array containing *n* rows and *m* columns. If *n*, *m* ≥ 1, a matrix of size *n* × *m* is a map of {1, . . . , *n* } × {1, . . . , *m* } values. Each entry or component of the matrix is designed as m_{ij} indicating the position of the element at the intersection of the i^{th} row and the j^{th} column. Therefore, if **M** is the matrix of order *n* by *m* the $\mathbf{M}_{nxm}$ matrix is denoted as follow:

$$\boldsymbol{M} = \begin{bmatrix} m_{11} & \cdots & m_{1m} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nm} \end{bmatrix}$$

When *m* = *n* the resulting matrix is a *matrix* of order *n* ($\mathbf{M}_n$), and is called n-*square matrix*[8].

We can use an n-*square matrix* to describe the evolutionary distances between each character state. Distances within this matrix represent the "cost" of each genetic alteration; the higher the cost the less likely an alteration will occur. We can then assign a cost to each edge. In order to optimize this step-cost approach for each criteria, we expressed the cost using an algebraic equation that describes each cost as a function of the step. As the character state graph is linear the cost of each step is given by the linear equation:

$$y_i = mx_i$$

where x_i represents the step of the *i* edge and y_i represents the cost for that step; *m* is a constant (the slope). We assigned the natural number two to *m* as it allows the smallest total cost; therefore the above equation becomes:

$$y_i = 2x_i$$

We did not assign the natural number of zero to *m* as zero this would suggest no step. For the first evolutionary step (initiating leukaemia specific fusion *ETV6-RUNX1*) we assumed *m*=1 paired with *x*=1 (one step). Therefore, the following smallest natural number to be assigned to *m* is two and consequent steps are described by the equation $y_i=2x_i$. This criterion programs the matrix with an order and the first triggering evolutionary event.

The character graph approach for modelling the step and cost is extremely powerful. If higher cost was attributed to a step and the number of steps increased an exponential relationship may also be employed instead of a linear one. This is expressed by the following equation:

$$y_i = 2^{x_i}$$

where x_i represents the step of the *i* edge and y_i represents the cost for that step; 2 is the constant for the point slope. The first two steps for both the linear and exponential equations yield the same cost. Once a cost is assigned to each step of the character graph, a step matrix is built.

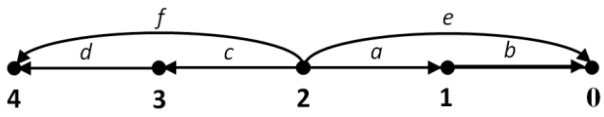
$$G = (V, E, \phi)$$

where:

$$V = \{4, 3, 2, 1, 0\};$$

$$E = \{a, b, c, d, e, f\};$$

$$\phi = \begin{pmatrix} a & b & c & d & e & f \\ \{2,1\} & \{1,0\} & \{2,3\} & \{3,4\} & \{2,0\} & \{2,4\} \end{pmatrix}.$$



where each *vertex* represents the observed copy number and each e*dge* the step or character change from one condition state to the other. The character graph is ordinate and bidirectional.

The 5-*square matrix* of *5* rows and *5* columns (**M**) is follows:

$$M = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} . & i & i & i & i \\ c_b & . & i & i & i \\ c_e & c_a & . & c_c & c_f \\ i & i & i & . & c_d \\ i & i & i & i & . \end{bmatrix} \end{array}$$

The cost for changing from one state to another is denoted by c_j where *j* represents the edge; each cost will result from the equation $y_i=2x_i$. The character *i* in the matrix indicates infinity, representing an infinite cost and an impossible evolutionary change or disallowed transformation. The dot character represents no evolutionary changes.

Within this matrix we cannot ignore reverse mutations, which restore the previous ancestral conditions, especially when deleterious mutations occur. Reverse mutations for an SNV occur at a very low rate but for some types of DNA alterations, such as large deletions, reversal is nearly impossible [9]. In the case of SNV we could not ignore the potential for reverse mutations and a 1:5 (forward:reverse) cost was used indicating that each reverse mutation costs five times more than a forward mutation.

Estimating branch length s under parsimony criterion: Sankoff algorithm

In order to infer the phylogeny of the clonal expansion, given the step matrix, we needed an algorithm that considered the number of character changes required by any given tree. The Sankoff's algorithm[10, 11], is an algorithm that counts the number of evolutionary changes for a specific site in a phylogenetic tree and evaluates all possible character reconstructions. Given a cost matrix C = [c_{jz}], in which the cost for changing from state *j* to *z* can be read, the Sankoff algorithm computes the total cost of the combinations of η events for each character. For each node (ϑ) of the tree, a character state *j* is assigned and the cost vector $S_\vartheta\ (j)$ is computed. This reflects the minimum cost of events (state changes) from ϑ to the root of the tree. Sankoff algorithm calculates this at each node starting from the tips of the tree moving towards to the root. Initially, the $S_\vartheta\ (j)$ at the inner nodes are unknown while those at

the tips are computed assigning the cost 0 to the observed state *j* and infinity (outlined as *i*) to the rest (Supplemental Figure 4). If a copy number of 1 is an observed character state for a specific cell or cub-clone, the cost would be $S_\vartheta(0) = i$, $S_\vartheta(1) = 0$, $S_\vartheta(2) = i$, assuming three character states (0, 1, and 2) are observed within the sample. Then for the node *α* representing the immediate common ancestor S_α[4] is calculated according to the following equation:

$$S_\alpha(j) = \min_z \left[c_{jz} + S_l(z)\right] + \min_k \left[c_{zk} + S_r(k)\right]$$

where S_α is the actual node in state *j*, $S_l(z)$ is the left descendant in state *z* and $S_r(k)$ is the right descendant in state *k*. This means that the cost of the character state *j* for the node *α* is the cost c_{jz} of changing from character state *j* to *z* in the left descendant lineage plus the cost $S_l(z)$ of having reached state *z* at the node *l*. Character *z* is selected to minimize this sum. The same procedure is then applied to character *k* in the right descendant lineage. The sum of the minimum *z* and minimum *k* is the smallest possible cost for the sub-tree above node *α*, given the node *α* in state *j*. The equation is applied to all nodes of the tree, up to the root (node *0*):

$$S = \min_j S_0(j)$$

where *S* represents the minimum number of evolutionary changes for a given tree with character state *j* at the root.

Searching for optimal trees: heuristic search

Felsestein (1993) suggests applying heuristic approaches, when trees are constructed on the basis of ten or more sequences[12, 13]. Alternatively, when the data set is smaller, an exhaustive search can be applied. In this study we analysed two datasets, a smaller set of four taxa (Case A) and a larger one of seven taxa (Case

B). In order to infer the most parsimonious tree, we employed a heuristic search to find optimal trees, using branch swapping of trees constructed by stepwise addition of taxa. This searching algorithm is capable of generating all possible tree topologies within an efficient computation time and negates the computational hardness of exhaustive searches [14]. However, as our data sets were small we also employed the branch and bound algorithm used in exhaustive searches [15] to phylogenetically analyse our data. The results generated were the same as those achieved using the heuristic search.

The step wise addition algorithm begins joining three taxa in an initial tree of three branches representing the taxa and one internal node representing the common ancestor [16]. Each remaining unplaced taxon is then added to the tree one at a time. The algorithm stops when all taxa have been joined to the tree. The algorithm needs to be instructed as to how to determine which three taxa will be initially joined and which one of the unplaced taxa will be connected to the tree during each step. In order to search for the largest number of possible trees, the best approach is to allow the algorithm to test as many tree topologies as possible using random addition of taxa. This approach may not be very effective in terms of stepwise addition but is extremely appropriate in obtaining different starting points for branch swapping. All heuristic algorithms are susceptible to the problem of entrapment in local optima but a tree topology search from a variety of starting points increases the probability of escaping from the local optimum. In particular, stepwise addition, is an extremely greedy algorithm and is highly susceptible to local-optima problems; by initiating branch swapping repeatedly from different starting trees we increase the probability of the heuristic search finding the optimum tree [5, 17].

Branch swapping [16] is a tree perturbation method that involves cutting off one or more pieces of a tree (sub-trees) and reassembling them in a way that is locally different from the original tree. This increases the effectiveness of searching the global optimum. Different studies have shown that a non-branch swapping approach yields significantly lower support estimates than analyses using some kind of branch swapping approach [18-20]. The phylogenetic package, PAUP*, implements the branch swapping method using different algorithms. We employed the tree bisection and reconnection (TBR) algorithm which is the most extensive rearrangement strategy available in PAUP*.

Exhaustive explanations of the above algorithms are described in many phylogenetics treatises and will not be explained here.

Evaluating the reliability of inferred trees

Phylogenetic reconstruction can determine the evolutionary history of taxa based on characters states. However, different ways of assembling data such as taxon sampling, alignment and data concatenation may bias the phylogenetic reconstruction. A reliable phylogenetic tree is a tree where a small modification in the data should not drastically change the phylogeny inferred or at least if it does, it should only do so with a small probability. An inferred phylogeny without this property is weak and inefficient. How can we estimate the reliability of the trees we have inferred?

In the context of parsimony analyses two basic types of re-sampling methods are used to assess the reliability of the inferred phylogenetic tree: bootstrap [21-23] and jackknife [24, 25] which are both re-sampling statistical methods for error estimation [26]. Re-sampling procedures are considered to be an essential component of rigorous

parsimonious phylogenetic analysis, offering support to any branch node of the tree. Both bootstrap and jackknife can be used to quantifying any tree branch reliability.

The bootstrap approach is one of the most popular re-sampling methods to place confidence in phylogenies but recently Simmons and Freudenstein (2011) indicate that jackknife re-sampling should be used rather than bootstrap re-sampling [27]. Because of the step matrix method employed where for each unit (or gene) the proper matrix is applied, a bootstrap analysis cannot be computed for our data set. However, we did not want to ignore this approach and wrote an R in house script that mimics the bootstrap re-sampling. Due to the short length of the data (only eight characters) and aiming to keep the informative, complete data set, the script samples without replacement, generating eight character replicas of the observed clones but with units in another order. This approach, therefore has allowed us to test if data concatenation could mislead the inferred phylogeny. Each replica obtained using this script, has then been used in PAUP* to infer the replica phylogenetic tree. We conducted this approach for each sample and the results were the same for all replicas; a single tree found for case A and two identical trees found for case B. These results support the evidence that the order of the concatenated character does not bias our approach.

In order to evaluate the stability of the inferred phylogenetic trees and to support each node, we applied a jackknife re-sampling approach. Jackknifing repeatedly calculates the statistics of interest, missing out one or more characters in turn and preserving their orders in the original data. This procedure does not conflict with the step-matrix as bootstrap does. However, this approach has been criticised as single character deletions from large dataset would produce very similar trees from the respective replicates and would not provide any effective measure of

support. Deleting a larger proportion of the characters has consequently been adopted to increase the performance of the jackknife algorithm to that of the bootstrap [28]. However, Farris et al., (1996) investigated jackknifing further and concluded that deleting 50% of the characters was too severe.

Both Cases investigated consist of eight characters generating four sub-clones for Case A and seven sub-clones for Case B; both are small datasets. Considering the small size of our datasets and to keep as much data as possible we chose to delete 12.5% of characters at random within each iteration. We then jackknifed both datasets which resulted in a jackknife 50% majority rule consensus tree [29]. We also tested our dataset with jackknifing at 25% and 50%. Considering the size of our dataset we agree with Farris et al., (1996) that 50% jackknifing may be too severe but a 25% jackknifing is still robust. Results are shown in Supplementary Table 7.

*PAUP**

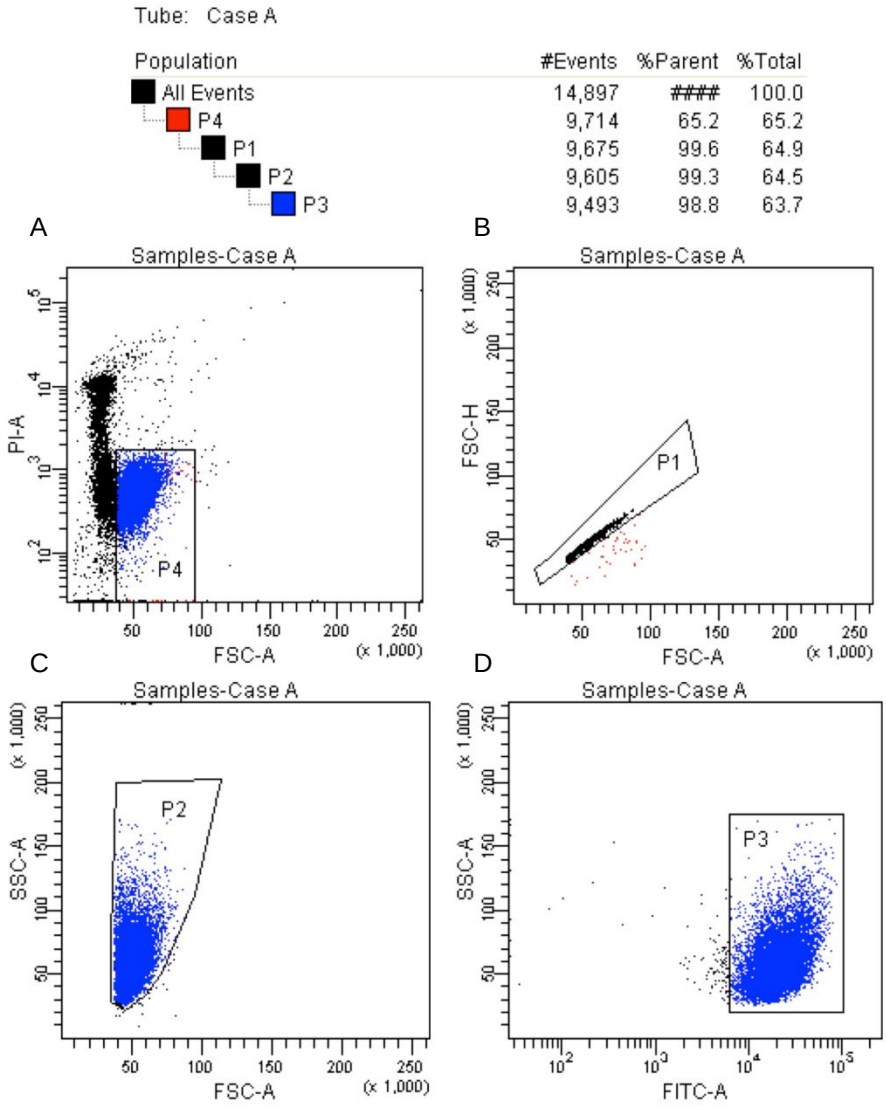
The program PAUP* [17] is one of the most widely used software packages for inferring evolutionary trees. The program is particularly proficient in inferring phylogenies using parsimony. PAUP* also implements jackknife to support tree nodes.

The problem of multiple Equally Parsimonious Reconstructions

In many situations, competing equally parsimonious reconstructions can result from the inferred phylogeny. These trees may have different implications for the evolutionary hypothesis under investigation and discharging one or more alternative equally parsimonious reconstructions can strongly affect the conclusions.

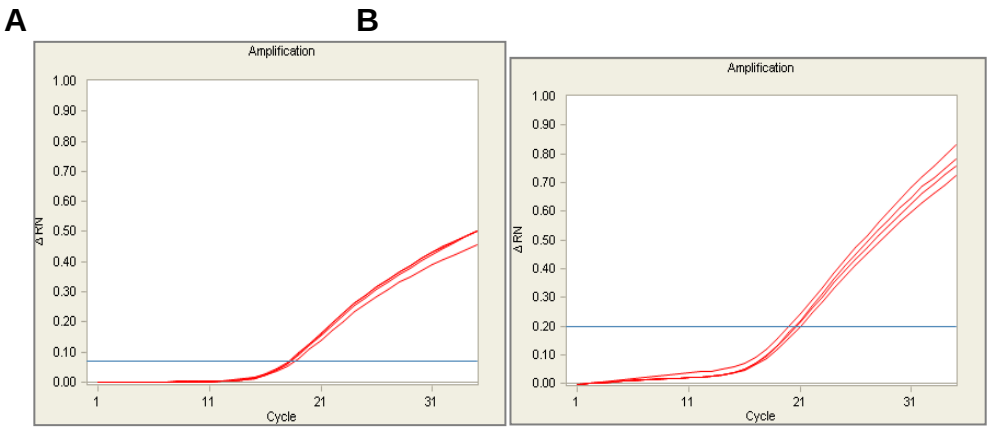
Consequently in this manuscript we have kept all parsimonious trees; one most parsimonious tree for the case A and two equally parsimonious trees for the case B.

SUPPLEMENTAL FIGURES

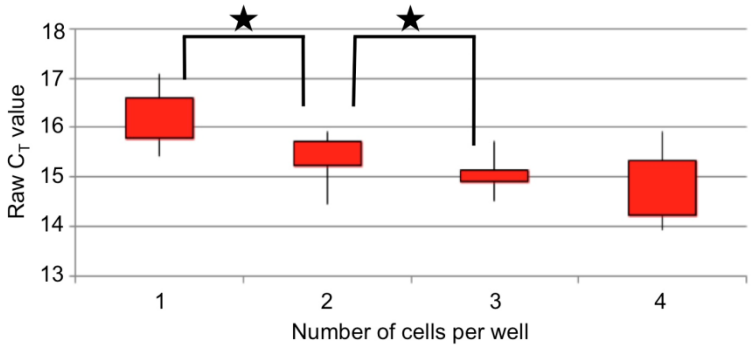


Supplemental Figure 1. Fluorescence-activated cell sorting plots illustrating gating approaches to efficiently isolate single cell. (A) This figure shows all events collected by the BDFACSAria1. Propidium iodide staining distinguishes dead and live cells; the P4 gate encompasses live cells that lack staining. (B) This figure displays gating (P1) for single cells only avoiding clumps identified by lower forward-scattered light height (FSC-H) and broad forward-scattered light area (FSC-A); these events were gated from the P4 population. (C) This figure displays only events gated by P1 but

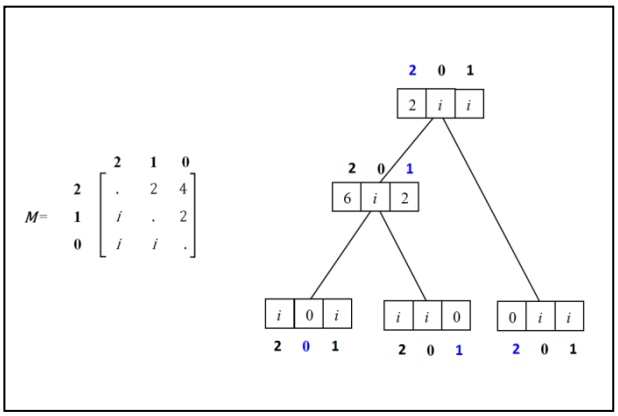
confirms cell isolation by size and internal complexity (FSC-A and side-scattered light (SSC)); P2 gate. (D) Fluorescein isothiocyanate (FITC) marks those cells that successfully take-up the viability marker carboxyfluorescin diacetate, succinimidyl ester (CFSE). Viable single cells were sorted according to the P3 gate ensuring random sampling of the leukaemia; P3 was gated from P2.



Supplemental Figure 2. *EPOR* SNP assay amplification curves generate by Q-PCR for a single REH cell**.** (A) Amplification curve generated from the probe labelled with VIC and complementary to the SNP sequence. (B) Amplification curve generated from the probe labelled with FAM and complementary to the wild-type sequence.



Supplemental Figure 3. Box and whisker plots to illustrate the range of raw C_T values generated by Q-PCR from wells that contained either one, two, three or four cells with a normal diploid karyotype. The upper and lower quartiles and minimum and maximum values are displayed. The median values for the groups with either one and two or two and three cells were significantly different ($p \leq 0.05$); three and four were not indicating assay saturation.



Supplemental Figure 4. Example of Sankoff's algorithm. Possible character states are bold, those infered or observed are blue. Each rectangle represents a node or a tip and the digit within represents the cost of being in state 2, 1 or 0. The step matrix is shown on the left. Each character state transition cost is indicated by a number; *i* stands for infinite cost.

SUPPLEMENTAL TABLES

Supplemental Table 1A. QC metrics for Affymetrix® Cytogenetics Whole Genomic 2.7M arrays.

	QC	SNPQC	MAPD	Antigenomic Ratio	Waviness seg count
CaseAdiagnosis.cychp	true	2.376	0.17	0.25	52
CaseAremission.cychp	true	2.316	0.18	0.23	3
CaseBdiagnosis.cychp	true	2.658	0.16	0.23	567
CaseBremission.cychp	true	2.502	0.17	0.21	4

Supplemental Table 1B. QC metrics for Affymetrix® Genome-Wide Human SNP Array 6.0.

File	Computed Gender	Call Rate	Contrast QC
DS-ALL.birdseed-v2.chp	female	98.98	2.27
REH.birdseed-v2.chp	female	99.45	2.83

Supplemental Table 2. Custom and designed assays for each sample specific mutation used in this study

	Mutation Custom Assays	Error Rate in diploid cells
IL7R	Forward Primer TGCATGGCTACTGAATGCTC Reverse Primer CCCACACAATCACCCTCTTT Probe1 ATGGATGGCTGTCTGGTCAT Probe2 CTGATGGTTAGTAAGATAGGATCCATC	0%
EPOR SNP assay	rs318720 (Life Technologies)	0%
PIK3R1	Forward Primer AGAACAGTGCCAGACCCAAGAG Reverse Primer CAAGGGAAACACCAACCTTTGT Probe1 AGGCAATCAGAAAGA Probe2 AGGCAATGAGAAAGA	0%
DAXX	Forward Primer GGAAATGTCCGTCTCCACAGA Reverse Primer CTGACCCTGGAGGAAGAAAGC Probe1 AAGAGCTAAGACACAG Probe2 TCAAAGAGCTGAGACAC	0%
EZH2	Forward Primer AGCGGCTCCACAAGTAAGACA Reverse Primer TGCAAAGCACAGTGCAACAC Probe1 TAGCACAGGCACTG Probe2 TAGCACGGGCACTGC	2%
BCHE	Forward Primer GCCAGAAACTTGCCATCATAAAC Reverse Primer ACCTAAACCAAAAAATGCCACTGT Probe1 ACCACCATAAGTCC Probe2 CCACCATAAATCCA	0%
BAZ2A	Forward Primer AAGGAAGTCCCCAAGGTGAAA Reverse Primer GTCTTGTTCAATAGCTCAGTGATTTTG Probe1 TCGAGGTCGGCTAC Probe2 TCGAGGTCGGCCAC	2%
RB1	Forward Primer CCTAGTTCACCCTTACGGATTCC Reverse Primer TGTTGGCAGACCTTCTGAAATTT Probe1 CCCTGAACAGTCCAT Probe2 CCCTGAAGAGTCCAT	0%
KRAS	Forward Primer TGGTCCTGCACCAGTAATATGC Reverse Primer AAGGCCTGCTGAAAATGACTGA Probe1 CTACGCCACCAGCT Probe2 TACGCCACAAGCT	0%
SFRS11	Forward Primer AATCAAGCTTTGTATTTTAGCGAACA Reverse Primer ACACACAAAAACACCCAGAAAATG Probe1 TACTTTCAACAAACTGAG Probe2 TTCAACAAACTCAGGTGG	0%
P2RY8-CRLF2	Forward Primer CTCTGAGCTCCATGGTTCGT Reverse Primer CAAGCCACCCTTCCTTTAAT Probe1 TCTCGAACTCCTGACCTCGT	0%
ETV6-RUNX1 REH	Forward Primer TGGAGGTTGTAATGAGCCAAGA Reverse Primer CCCACCCCGACATAAATTCA Probe1 AGCCTGGGCAACAGAGTGATACTCTCCC	REH: 0%
ETV6-RUNX1 Case A	Forward Primer GTGTATACACATATAGTGTATGTGCGTGTAC Reverse Primer CCTCCTGCCATTGCTTTCTC Probe1 AAGAGTCTGGAGGCATA	Case A: 0%
ETV6-RUNX1 Case B	Forward Primer TCACTCCCAACCTCTAGAACA Reverse Primer TGTGTGCATGTGTGTAAGATGGA Probe1 ACACTAATTTTCTCAGGTTGC	Case B: 0%

Supplemental Table 3. Copy number assays for each sample specific alteration in this study

	Copy Number Assays	Error Rate (% loss/gain in diploid cells)
B2M -1	Hs 00128408	-
B2M -2	Hs 00112422	
B2M -3	Hs 03896400	
CCNC -1	Hs 02941667	+/- 6.6%
CCNC -2	Hs 02942602	
CCNC -3	Hs 06148409	
CDKN2A -1	Hs 03724208	+/- 4.5%
CDKN2A -2	Hs 03700684	
CDKN2A -3	Hs 03704181	
DPF3 -1	Hs 07066526	+/- 7.1%
DPF3 -2	Hs 07074482	
DPF3 -3	Hs 0795788	
MX1 -1	Hs 05557497	+/- 4.5%
MX1 -2	Hs 02954936	
MX1 -3	Hs 05528846	
PAX5 -1	Hs 01885952	+/- 4.3%
PAX5 -2	Hs 02165423	
PAX5 -3	Hs 06837891	
TBL1X -1	Hs 05633780	+/- 4.7%
TBL1X -2	Hs 05614341	
TBL1X -3	Hs 02806412	
VPREB1[§] -1	Hs 06690444	+/- 6.3%
VPREB1 -2	Hs 02879734	

[§]Only two *VPREB1* assays were used as only a small number of commercial assays were available in this region.

Supplemental Table 4. Details and explanations of cell data removed from each individual case single cell experiment

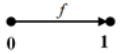
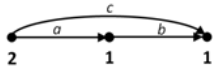
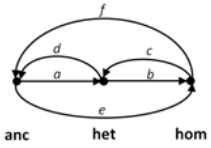
	REH - 2x96 plate	DS-ALL - 2x96 plate	ALL Case A - 4x96	All Case B - 4x96
Total number of single cells sorted	190	190	380	380
Total number of control cells sorted (11 per plate)	22	22	44	44
Total number of target cells sorted	168	168	336	336
Number of wells that were blank - no cell	6	9	10	6
Number of wells with data that suggested the cell was damaged	25	24	31	28
Number of wells with high DNA levels suggesting two cells	2	1	5	2
Number of cells constituting minor sub-clones below error rates	12	25	35	53
Successful data collected from target cell	126	115	261	254
Successful data collected from experiment	145	133	299	291
Percent of data removed because of failure	17.37	17.89	12.11	9.47
Percent of data removed as part of sub-clonal populations below error rates	7.14	14.88	10.42	15.77
Percent of successful data collected form experiment	76.32	70.00	78.68	76.58

Supplemental Table 5. FISH probes and scoring frequencies for each sample specific alteration in this study

Gene/Chromosome region	BAC clone	Nuclei with normal signal pattern (%) (expected signals)	Number of gene copies in each sample tested
MX1	WI2-2208G5	100 (2C)	REH: 4C-5%, 3C-39%, 2C-49%, 1C-7%
CDKN2A	W12-1034K10	99 (2C)	REH: 1C-2%, 0C-98% DS-ALL: 2C-8%, 1C-78%, 0C-14% Case B: 2C-21%, 1C-79%
P2RY8-CRLF2	RP13-309C18, RP11-309M23	98 (0F)	DS-ALL: 1F-96%
ETV6-RUNX1	Vysis LSI ETV6(TEL)/RUNX1(AML1) ES Dual Colour Translocation Probe Set	98 (0F)	REH: 1F-100% Case A: 1F-96% Case B: 1F-91%
TBL1X	RP11930D21	98 (2C)	Case B: >2C-3%, 2C-7%, 1C-90%
CCNC	RP11-484P14	97 (2C)	Case A: 2C-10%, 1C-90%
PAX5	WI2-500F8	99 (2C)	Case B: 2C-19%, 1C-81%
DPF3	RP11-73K12	97 (2C)	Case B: 2C-27%, 1C-73%
VPREB1	RP11-24N11	100 (2C)	Case B: 2C-8%, 1C(minor signal)-8%, 0C-84%

[§]C=copy number
F= gene fusion

Supplemental Table 6. Character states graphs and corresponding matrices employed in this manuscript

Gene	Genomic alteration	Graph	Graphic visualization	Matrix
Case A and B: *ETV6-RUNX1*	Fusion	$G_{fusion} = (V, E, \varphi)$ where: V = {0, 1,}; E={f}; $\phi = \begin{pmatrix} f \\ \{0,1\} \end{pmatrix}$		$y=x$ $M_{fusion} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} . & 1 \\ i & . \end{bmatrix} \end{array}$
Case A: *CCNC* *TBL1X* Case B: *CDKN2A* *DPF3* *PAX5* *VPREB1*	CNA	$G_{CNA} = (V, E, \varphi)$ where: V = {2, 1, 0}; E={a, b, c}; $\phi = \begin{pmatrix} a & b & c \\ \{2,1\} & \{1,0\} & \{2,0\} \end{pmatrix}$.		*Forward:* $y_i=2x_i$ *Back: not allowed* (i) x_a= x_b= 1 step x_c= 2 steps $M_{CNA} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} . & i & i \\ 2 & . & i \\ 4 & 2 & . \end{bmatrix} \end{array}$
Case A: *BCHE* *BAZ2A* *DAXX* *EZH2* *PI3KR1* Case B: *KRAS* *RB1* *SFRS11*	SNPs	$G_{SNPs} = (V, E, \varphi)$ where: V = {anc, het, hom}; E={a, b, c, d, e, f}; $\phi = \begin{pmatrix} a & b & c & d & e & f \\ \{anc,het\} & \{het,hom\} & \{hom,het\} & \{het,anc\} & \{anc,het\} & \{hom,anc\} \end{pmatrix}$.		*Forward:* $y_i=2x_i$ *Back:* $y_i=(2x_i)*5$ x_a= x_b= x_c= x_d= 1 step x_e= x_f= 2 steps $M_{SNPs} = \begin{array}{c} \\ \text{anc} \\ \text{het} \\ \text{hom} \end{array} \begin{array}{c} \begin{array}{ccc} \text{anc} & \text{het} & \text{hom} \end{array} \\ \begin{bmatrix} . & 2 & 4 \\ 10 & . & 2 \\ 20 & 10 & . \end{bmatrix} \end{array}$

Supplemental Table 7. Jackknife analyses within the parsimonious trees of patient A and B employing three different percentages of character deletion

% Characters Deleted	12.5	25	50
Case A			
1 node	100	100	92
2 node	85	74	49
Case B			
Tree B1: 1 node	100	96	81
Tree B1: 2 node	87	77	50
Tree B1: 3 node	88	75	48
Tree B1: 4 node	51	49	40
Tree B1: 5 node	75	54	21
Tree B2: 1 node	100	96	79
Tree B2: 2 node	87	74	48
Tree B2: 3 node	86	74	49
Tree B2: 4 node	51	48	41
Tree B2: 5 node	73	54	24

SUPPLEMENTAL APPENDIX

Basic terms and concepts in phylogenetics applied to cancer evolution

Clade—a monophyletic group, which includes all the descendants of an ancestor.

Character—an attribute of a cell (or group of cells); it may be a morphological attribute, a nucleotide position, a chromosomal insertion/deletion etc.

Phylogeny—evolutionary history of a group of cells or clones.

Phylogenetic tree—a mathematical structure depicting the evolutionary history of a group of cells.

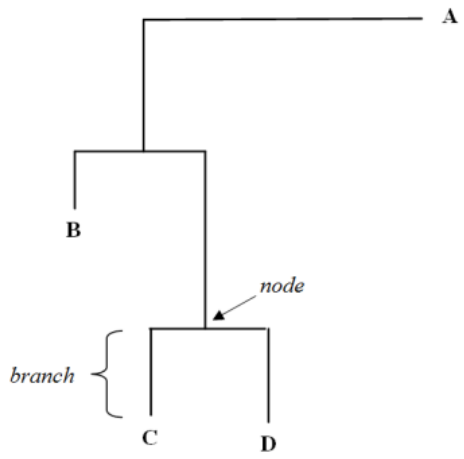
Node— a branching point on a phylogenetic tree. *Internal nodes* represent inferred or observed ancestors.

Branch—connections between nodes representing a process of evolution in which mutations inherited by descendants accumulate.

Root—the ancestor of all taxa include in the tree.

Taxon—a group of related cells or clones.

In the figure below, *A*–*D* represent related taxa as indicated by the branches. *A* represents also the root. *C* and *D* are sister taxa and are thus a clade; *B*, *C* and *D* form a larger clade Each node, or internal branching point, represents an ancestor of the clade that lies “above” it. The node indicated by the arrow is the common ancestor of taxa *C* and *D*.


A
B
node
branch
C
D

SUPPLEMENTAL REFERENCES

1. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
2. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915 (2010).
3. Weaver, S. et al. Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods* **50**, 271-276 (2010).
4. Farris, J.S. Methods for computing Wanger Trees, Vol. 19. (Systematic Zoology, 1970).
5. Page, R.M.D. & Holmes, E.C. Molecular evolution: a phylogenetic approach (Wiley-Blackwell 1998).
6. Bender, E.A. & Williamson, G.S. Mathematics for Algorithm and Systems. (Dover Publications 2009).
7. Greaves, M. & Maley, C.C. Clonal evolution in cancer. *Nature* **481**, 306-313 (2012).
8. Serre, D. Matrices: Theory and applications; 2nd edition (Springer, 2010).
9. Lande, R. Risk of population extinction from fixation of deleterious and reverse mutations. *Genetica* **102-103**, 21-27 (1998).
10. Sankoff, D. & Rousseau, P. Locating the vertices of a steiner tree in an arbitrary metric space. *Math Program* **9**, 240-256 (1975).
11. Sankoff, D. Minimal mutation trees of sequences. *SIAM J ApplMath* **28**, 35-42 (1975).
12. Felsenstein, J. PHYLIP: Phylogeny inference package. (University of Washington, Seattle, 1993).
13. Swofford, D. & Sullivan, J. in The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing; 2nd Edition. (ed. P.S. Lemey, M. Vandamme, A.) 267-312 (Cambridge Universit Press, 2009).
14. Giribet, G. Efficient tree searches with available algorithms. *Evol Bioinform Online* **3**, 341-356 (2007).
15. MD, H. & D, P. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* **59**, 277-290 (1982).
16. Swofford, D.L., Olsen, G.J., Waddell, P.J. & Hillis, D.M. in Molecular Systematics, 2nd Edition. (eds. M.H. DM, M. C & M. BK) (Sinauer Associates, 1996).
17. Swofford, D. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) Version 4.0 Beta 10. (Sinauer Associates, Inc. Sunderland, 2005).
18. Mort, M.E., Soltis, P.S., Soltis, D.E. & Mabry, M.L. Comparison of three methods for estimating internal support on phylogenetic trees. *Syst Biol* **49**, 160-171 (2000).
19. Sanderson, M.J. Objections to Bootstrapping Phylogenies: A Critique. *Syst Biol* **44**, 299-320 (1995).
20. Salamin, N., Chase, M.W., Hodkinson, T.R. & Savolainen, V. Assessing internal support with large phylogenetic DNA matrices. *Mol Phylogenet Evol* **27**, 528-539 (2003).

21. Wiemels, J.L. & Greaves, M. Structure and possible mechanisms of TEL-AML1 gene fusions in childhood acute lymphoblastic leukemia. *Cancer Res* **59**, 4075-4082 (1999).
22. Efron, B., Halloran, E. & Holmes, S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A* **93**, 13429-13434 (1996).
23. Holmes, S. Bootstrapping Phylogenetic Trees: Theory and Methods. *Statistical sciene* **18**, 241-255 (2003).
24. Mueller, L.D. & Ayala, F.J. Estimation and interpretation of genetic distance in empirical studies. *Genetical Research* **40**, 127-137 (1982).
25. Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, M. & Kluge, A.G. Parsimony jackknifing outperforms neigbor-joining *Cladistics* **12**, 99-124 (1996).
26. Efron, B. & G, G. A leisurely look at the Bootstrap, the Jackknife, and cross-validation. *The American Statistician* **37**, 36-48 (1983).
27. Simmons, M.P. & Freudenstein, J.V. Spurious 99% bootstrap and jackknife support for unsupported clades. *Mol Phylogenet Evol* **61**, 177-191 (2011).
28. Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **39**, 783-791 (1985).
29. Margush, T. & McMorris, F.R. Consensus n-trees. *Bull Math Biol.* **43**, 239-244 (1981).