Supplementary Information Appendix

Supplementary Materials and Methods

Patient Samples

Six diagnostic bone marrow and saliva samples collected during remission from children with B-ALL diagnosed at Lucile Packard Children's Hospital were collected on protocol 11062 approved by the Stanford University Institutional Review Board. Informed consent was obtained prior to specimen collection and samples were deidentified before use in our study, in accordance with the Declaration of Helsinki. Mononuclear cells were isolated using Ficoll-Paque (GE Life Sciences) according to manufacturer's instructions, followed by cryopreservation in 90% fetal calf serum and 10% DMSO, as previously described (1).

Identification of Putative Variants in Exome Sequencing Data

DNA was extracted from tumor and saliva samples using the Qiamp DNA mini kit with RNAse A treatment according to the manufacturer's instructions (Qiagen). Exome-enriched sequencing libraries were prepared using the Nextera Expanded Exome Kit (Illumina) and were sequencing on a HiSeq 2000 or 2500 with 2X100 paired-end reads (Illumina). Adapter sequences and poor quality bases were trimmed using Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), and aligned to human reference hg19 using BWA (http://bio-bwa.sourceforge.net). Duplicates were marked using Picard (http://picard.sourceforge.net), followed by local realignment and base recalibration with GATK (http://www.broadinstitute.org/gatk/). Putative SNVs were called by comparing germline and leukemia samples using MuTect (http://www.broadinstitute.org/cancer/cga/mutect), and annotated using Annovar (http://www.openbioinformatics.org/annovar/). Putative somatic Indels and locations with loss of heterozygosity were called with VarScan2 (http://varscan.sourceforge.net). On target coverage was calculated using Picard CalculateHsMetrics. Mutation motifs were determined by downloading the sequencing using the UCSC table browser (http://genome.ucsc.edu/), followed by sequence stacking using Weblogo (http://weblogo.berkeley.edu).

Single-Cell Capture and WGA

Cells from the samples that underwent bulk sequencing were thawed quickly in a 37 degree C water bath followed by dilution in RPMI supplemented with 10% FBS. The cells were then washed 5 times with C1 DNA-seq wash buffer (Fluidigm). Cells were counted and loaded in small C1 DNA-seq chips according to the manufacturer's instructions using an on chip LIVE/DEAD viability stain (Invitrogen). Each capture site was imaged using a Leica microscope where phase contrast, as well as fluorescent images with GFP and Y3 filters were acquired to determine the number of cells captured, as well as the viability of each of the captured cells, as previously described (2). The cells then underwent lysis, neutralization, and MDA WGA according to the manufacturer's instructions (Fluidigm) using the GenomePhiv2 MDA kit (GE Life Sciences). Three C1 chips were run per patient.

Targeted Resquencing to Confirm SNVs in Bulk Samples

All putative coding SNVs, as well as all other SNVs with greater than 5 supporting reads underwent validation using microfluidic PCR-based targeted resequencing of bulk DNA with the Access Array System. In addition, the same methods were used to identify confirmed SNVs in the single cells. Target-specific assays were designed by Fluidigm (Patients 1,2,6), as well as using primer3plus (http://probes.pw.usda.gov/batchprimer3/) (Patients 3,4,5), followed by oligo purchase from IDT and multiplexing according to guidelines in the Access Array manual (Fluidigm). All samples were loaded with the Access Array loader and underwent PCR cycling in FC1 system, followed by sample-specific barcoding using standard PCR, all according to the manufacturer's instructions (Fluidigm). Amplicons were run on the MiSeq using 2X150bp paired-end reads (Illumina) using custom sequencing primers according the Access Array manual (Fluidigm). All data underwent quality trimming, as well as adapter removal using Trimmomatic. Reads were then aligned to hg19 using BWA, followed by sorting, compressing, and indexing using Picard. An mpileup file was created using samtools (http://samtools.sourceforge.net/), followed by putative SNV, InDel, and LOH calls using VarScan2 with options (--min-var-freq 0.005 --min-coverage-normal 5 --min-coverage 5 --p-value 0.1 --min-avg-qual 35) to maximize loci capture before applying more stringent filtering criteria at later steps. The mutation calls were then annotated using Annovar. Those putative calls were compared to the exome sequencing sample using custom Bash scripts that required concordance of the location and base change between the exome and confirmation data, as well as a minimum of 3 reads comprising more than 1% of all reads at that position to support the variant call.

Estimating ADO

Allele-discriminating taqman assays were used to call polymorphisms and determine ADO rate at 46 commonly heterozygous loci, as previously described (3). In addition, Access Array resequencing assays for 96 loci were performed as described above to estimate the ADO rate. A custom R script was used to call heterozygous sites in the bulk sample if they had two alleles that each had at least 3 reads that comprised at least 10% of all reads at that location in at least 80% of replicates. The sites that were found to be heterozygous for each patient were then assayed in each of the single cells where we required the same base change be detected in at least 2 reads, which must also comprise at least 1% of the reads. These thresholds were determined by evaluating the upper limit of read count and percent of reads that were found in locations known to be absent to maximize the sensitivity by allowing for some allelic imbalance while minimizing false positive variant calls. The ADO rate for each cell was then calculated using 1-(number of alleles detected/(number of heterozygous sites identified in the bulk sample*2)). Cells with an estimated ADO rate less than 30% were retained for further analyses (Fig. S1).

Relating mixture of multivariate Bernoulli Distributions to Clonal Structures

Probabilistic modeling can be used to estimate an unknown probability distribution based on a finite set of data. The estimated probability distribution gives insights into the process that generated the data. The advantage of a mixture model for clonal analysis in single cells is that its components can represent different clones that makes up the true distribution, which would be impossible to estimate by a single parametric distribution.

We concentrated on mixtures of multivariate Bernoulli distributions, because our single cell mutational profile was binary in nature.

Single cell mutational profiles can be presented as binary vectors x ∈ {0, 1} , in which 1 denotes the presence of a mutation and 0 a normal base. The probabilities of the outcomes of a single cell observation $x = (x_1 , . . . , x_d)$ are modeled as $\theta_i = P(x_i = 1)$, $i = 1, . . . , d$. Since θ represents the success rate of observing the mutation, 1- θ gives an estimation of the allele dropout rate.

The probability of the observed single cell mutational profile x is estimated using multivariate Bernoulli distribution:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i} \tag{1}$$

The finite mixture of multivariate Bernoulli distributions representing different clones is defined as:

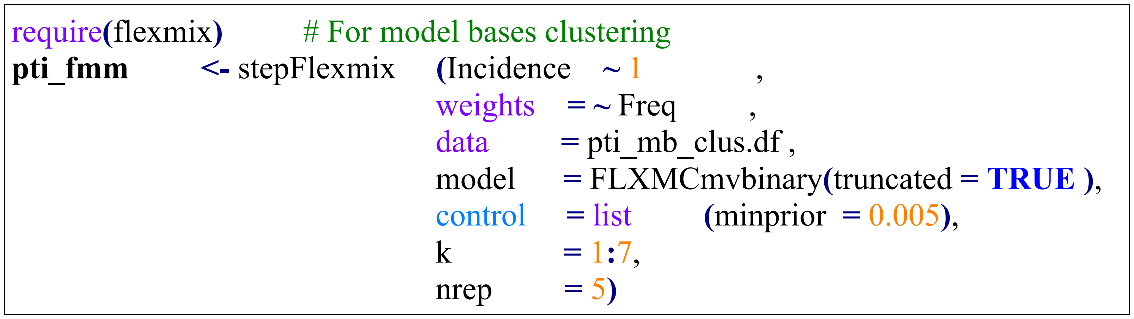
$$p(\boldsymbol{x}|\Theta) = \sum_{j=1}^{J} \pi_j \, p(\boldsymbol{x}|\boldsymbol{\theta}_j) = \sum_{j=1}^{J} \theta_i^{x_i}(1-\theta_i)^{1-x_i} \tag{2}$$

where π_j are the proportions of clones such that $\pi_j \geq 0$ and $\sum_{j=1}^{J} \pi_j = 1$ .When the number of different clones J is fixed and we have mutational profiles from N cells. The log-likelihood of the parameters $\{\pi_j, \boldsymbol{\theta}_j\}_{j=1}^{J}$ can then be written as:

$$l = \sum_{n=1}^{N} \log[\sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_{ni}} (1-\theta_{ji})^{1-x_{ni}}] \tag{3}$$

Finite Mixture Model based clustering of binary data

To infer the parameters of the mathematical model, optimization was carried out using EM algorithm. EM based clustering of single cells were implemented in R using *flexmix* package.



```
require(flexmix)        # For model bases clustering
pti_fmm       <- stepFlexmix    (Incidence   ~ 1              ,
                                 weights     = ~ Freq         ,
                                 data        = pti_mb_clus.df ,
                                 model       = FLXMCmvbinary(truncated = TRUE ),
                                 control     = list       (minprior  = 0.005),
                                 k           = 1:7,
                                 nrep        = 5)
```

The *stepFlexmix* function was used to perform the model-based clustering in steps. Each step the number of clones is reflected in the k parameter which was specified to range from 1 to 7 clones. The probability of the data fitting the model for each of the k parameter is stored and will be used for inferring the number of clones.

Estimation of Number of Clones

One main challenge in analyzing these single cell data is the determination of the 'correct' number of clones. With the underlying probability model driving the clustering process, the challenge of determining the number of clones are reduce to model selection problems in the probability framework. Essentially, using the code above, we are using the variable k to change the number of clones in the statistical model. We are interested in selecting the statistical model with the associated number of clones k that the data most likely originate from. The EM-algorithm implemented allows us to obtain the likelihood of data given the model. However, the choice of number of clones cannot be entirely assessed in this manner. By increasing the number of clones, we can always make the model fit better. The Bayesian information criterion (BIC) penalizes models with larger numbers of free parameters, and in this case the number of clones. Consequently, the model with the lowest BIC was chosen as the model with the minimum number of clones required to explain the data. With the selected model, we will be able to assign each cell to their respective clonal population.

```
# Perform the information criterion for clonal number analysis----
my_ic_fun      <- function(pt_fmm){
  bic.df <- data.frame(nos_clus=seq(1:length(BIC(pt_fmm))),IC=BIC(pt_fmm))
  aic.df <- data.frame(nos_clus=seq(1:length(AIC(pt_fmm))),IC=AIC(pt_fmm))
  bic.df$measure <- rep("bic",dim(bic.df)[1])
  aic.df$measure <- rep("aic",dim(aic.df)[1])
  ic.df  <- rbind(bic.df,aic.df)
  g <- ggplot(ic.df,aes(x=nos_clus,y=IC,group=measure,color=measure))+
       geom_point(data=subset(ic.df,IC%in%c(min(bic.df$IC),min(aic.df$IC))),
                        color="black",size=7.5,alpha=0.2)+
       geom_point(size=3)+
       geom_line(alpha=0.7)+
       scale_colour_discrete(name  ="Measure",
                         breaks=c("aic", "bic"),
                         labels=c("Akaike", "Bayesian"))+
       ggtitle("Number of Clones Selection Using Bayesian/Akaike Information
Criterion")+
    xlab  ("Number of clusters")  +
    ylab  ("Information Criterion")+
    theme_bw()
 g
 return(g)
}
```

Hierarchical Clustering of single cell mutational profile

Cluster analysis methods try to quantify the similarity between two single cell mutational profile and then try to group the cells so as to maximize within class similarity. The challenge is in finding an appropriate measure of similarity. In our case of binary data, we borrowed the jaccard distance used extensively in ecological studies. Hierarchical clustering with jaccard distance is implemented in R using the *hclust* and *vegan* package:

```
# Generate the heatmaps----
# Returns: Heatmap[[1]] -> EM clustering heatmap
# Returns: Heatmap[[2]] -> EM contrast with hclust
# Returns: vector       -> hclust results
# ---------------------
my_heatmap     <- function(pt_cluster , pt_fmm_best , nos_clust,nos_mut){
 bin.mat        <-  as.matrix(pt_cluster[,-which(colnames(pt_cluster)=="clusters")])
 EM_cluster     <- factor       (flexmix::clusters (pt_fmm_best))
 nos_cluster    <- max          (as.numeric(flexmix::clusters(pt_fmm_best)))
 row_order      <- order        (as.numeric(flexmix::clusters(pt_fmm_best)))
 data.dist      <- vegdist      (bin.mat     , method = "jaccard")
 data.dist.g    <- vegdist      (t(bin.mat)  , method = "jaccard")
 row.clus        <- hclust       (data.dist   , "ward.D2")
 col.clus        <- hclust       (data.dist.g, "ward.D2")
 hclust_ass     <- cutree       (row.clus    , nos_clust)
 color_scheme <- colorRampPalette(c("white", "#660000"), space = "rgb")(2)
 p           <- annHeatmap2(bin.mat[row_order,],
                scale  = "none", col    = color_scheme, breaks = 2,
                legend = 3,
                dendrogram = list(Col   = list(dendro = as.dendrogram(col.clus)),
                                  Row = list(status = "no") ),
                cluster    = list(Col = list(cuth    = col.clus$height[length(col.clus$height)-
nos_mut+1]),Row = list(grp     = EM_cluster[row_order],col     =
brewer.pal(nos_cluster,"Set2")[seq(1,nos_cluster,by=1)]) ),
                ann        = list(Row =
list(data=data.frame(EM_cluster=EM_cluster[row_order]))))
```

```
 p_row       <- annHeatmap2(bin.mat,
                scale  = "none",
                col    = color_scheme, breaks = 2,
                legend = 3,
                dendrogram = list(Col = list(dendro = as.dendrogram(col.clus)),
                                  Row = list(dendro = as.dendrogram(row.clus)) ),
                cluster    = list(Col = list(cuth   = col.clus$height[length(col.clus$height)-
nos_mut+1]    ),
                                  Row = list(cuth   = row.clus$height[length(row.clus$height)-
nos_clust+1]) ),
                ann        = list(Row = list(data   = data.frame(EM_cluster=EM_cluster)))
                )
  return(list(p,p_row,hclust_ass))
}
```

Multiple Correspondence Analysis of Single cell mutational profiles

Multiple correspondence analysis (MCA) is a data analysis technique which can be considered as the counterpart of principal component analysis for categorical data such as the binary nature of our single cell profiles. It can be used to detect and represent underlying structures in our single cell profile by representing single cells as points in a 2-dimensional Euclidean space. MCA is implemented in R using the *MCA* function in *factoMineR* package:

```
# Perform MCA on the binary data----
res.mca     <- MCA(bin.mat, graph = TRUE)
```

Directed Minimum Spanning Tree between Clones

Based on the clonal profile generated by the EM based clustering method, mutational age of the individual clone can be ordered by quantifying the number of mutations detected in each clone. The directed minimum spanning tree between is then generated in R using the *seqtrack* function in the *adegenet* package.

```r
# Perform the tree generation algorithm
my_run_tree     <- function(clone_gen_dis,clone_time){
  clone_gen_dis <- as.matrix (vegdist(clone_gen_dis,method="jaccard"))
  sqtk.res.add  <- seqTrack  (clone_gen_dis,
                              x.names = row.names(clone_gen_dis),
                              x.dates = clone_time)
  g_pri         <- plot(sqtk.res.add,vertex.size=4)
  return(list(g_pri,sqtk.res.add))
}
```

Performance of EM based clustering Method on Simulated Data

At each iteration of simulated data, 1 to 5 different clonal profiles are generated. Single cell mutational profile are generated by drawing from the multivariate Bernoulli describe above with different allele dropout rates. EM based clustering methods as described above are then applied and the number of inferred clones at different parameters are compared with the ground truth.

Analysis of the sensitivity of the analysis methodology

To deduce the sensitivity of this analysis methodology, we set the allele dropout rate to 0.2 and simulated single cell mutational profile originating from two clones: One being a dominant clone at a higher proportion compared to the other. Using such setup, we varied the proportion of the smaller clone from 1% to 7% of the total number of cells. Detection of the smaller clone would then depend on the number of mutations assayed as well as the total number of cells. The simulation results plot reflects the parameter space that can affect sensitivity.

Calculation of P-value for Clonality between two clones

The null hypothesis is that when comparing between two clones, the mutations count drawn from this pair have the same proportion as the clonal size. The alternative is that this proportion is different in at least one of the mutations. This is analogous to applying the chi-square statistic on a multi-sample Bernoulli model. The test statistic is given by:

$$Test\ Statistic = \sum_{i=1}^{m} \sum_{j=0}^{1} \frac{(O_{i,j} - e_{i,j})^2}{e_{i,j}} \qquad (4)$$

Where $j \in \{0,1\}$ with 0 representing one clone and 1 the other. m is the number of mutations. $O_{i,j}$ is the observed counts of the mutation i from the clone j. $e_{i,j}$ is the expected counts of the mutation i from the clone j. The expected counts can be calculated with the success probability being the ratio of clone sizes being compared. Implementation of the test was done using the *prop.test* function in R.

Identification and Confirmation of Deletions

To identify putative deletions in our exome sequencing data, we identified locations where 5 contiguous LOH calls had been made in VarScan with a Fischer's Exact Test <0.01. Depending on the number of heterozygous locations within the putative deletions, up to five heterozygous sites within those regions were used to confirm the deletions in the bulk and germline samples using the targeted resequencing approach outlined above where we required at least 30 reads in each of the two alleles, and a decrease of at least 10 reads that resulted in a minimum 20% decrease in read percent in the putative allele with LOH. The assays that were confirmed in the bulk sample were then used to call deletions in the single cells. This was accomplished by first phasing each of the alleles based on their LOH for each deletion in the bulk sample using a custom R function. We then created a binary matrix where the putative LOH allele was considered absent if there were less than 10 reads that made up less than 5% of the total reads at each of those sites.

To detect deletions in each clone, we required at least half of the LOH assays in that clone support each deletion.

Single-Cell IgH Sequencing and Alignment

Single–cell IgH sequencing was performed on each of the bulk and single cell samples using the BIOMED-2 primers in the Access Array System (4). We added a common sequence to the primers for a subsequent barcoding PCR, according the manufacturer's instructions (Fluidigm). The samples were then sequenced on a MiSeq with 2X150 paried-end reads. Custom Bash scripts were used to collect the IgH reads and mark the sample of origin of each read. Sequences were again trimmed using Trimmomatic, and files were converted to fasta format using prinseq (http://prinseq.sourceforge.net/). Reads were then aligned to the IgH locus using IgBlast (ftp://ftp.ncbi.nih.gov/blast/executables/igblast/release/) with default options. Custom Bash scripts were again used to parse the output file and call VH segments from the 5' reads, as well as D and J segments from the 3' read. The reads were required to have 130 bases aligned to the V segment, 5 to the D segment, and 10 to the J segment. The best alignment was retained for each read, and only *01 alleles for the V segments were retained to prevent ambiguous calls. The VDJ calls were then joined for each paired read. The sequence diversity was examined in the bulk samples. In addition, a consensus VDJ call, as well as the minimum number of VH segment mutations were determined for each cell. Finally, to determine if there were multiple high frequency VH segments used in each sample, we classified up to two V segments as predominant sequences if the second VH segment was present in 75% of the number of cells of the most abundant VH segment. All less abundant VH segments were classified as VH-replacement products.

Supplementary References

1. Gawad C*, et al.* (2012) Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood* 120(22):4407-4417.
2. Treutlein B*, et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*.
3. Fan HC, Wang J, Potanina A, & Quake SR (2011) Whole-genome molecular haplotyping of single cells. *Nature biotechnology* 29(1):51-57.
4. van Dongen JJ*, et al.* (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* 17(12):2257-2317.

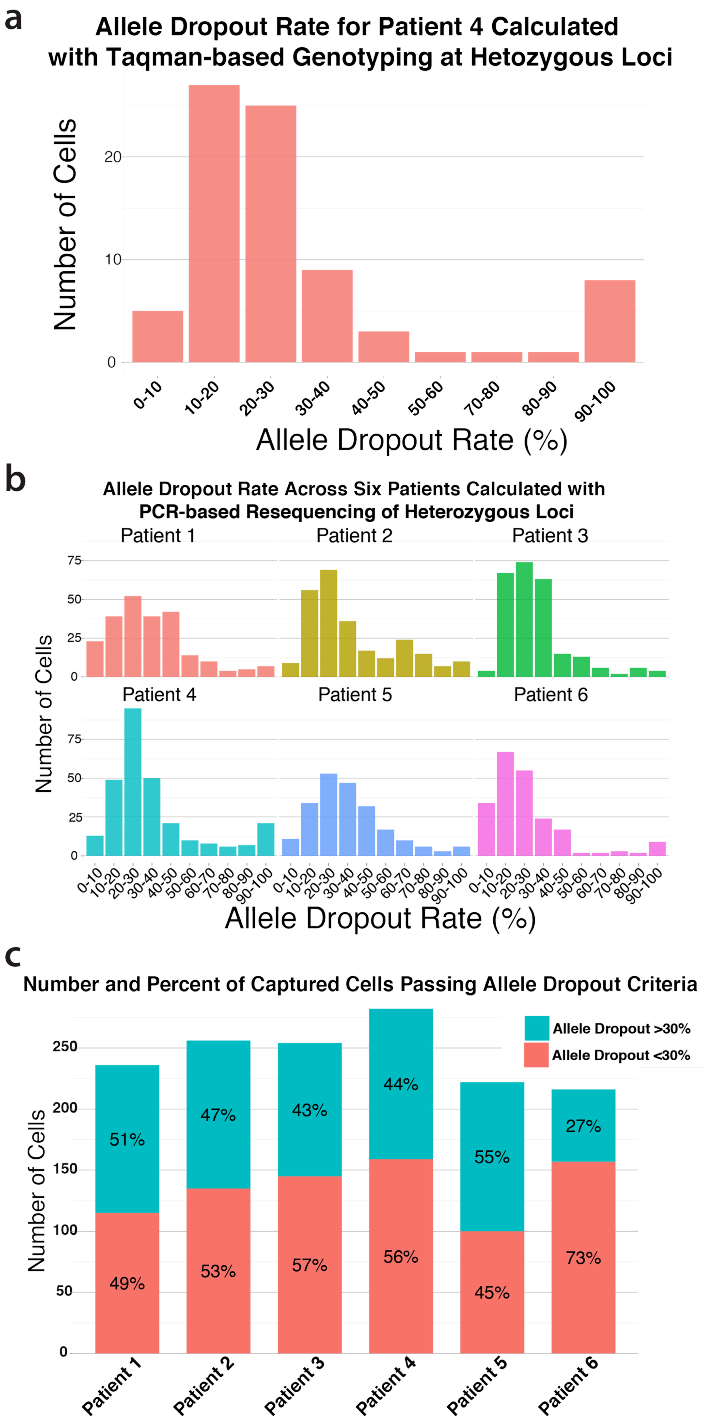
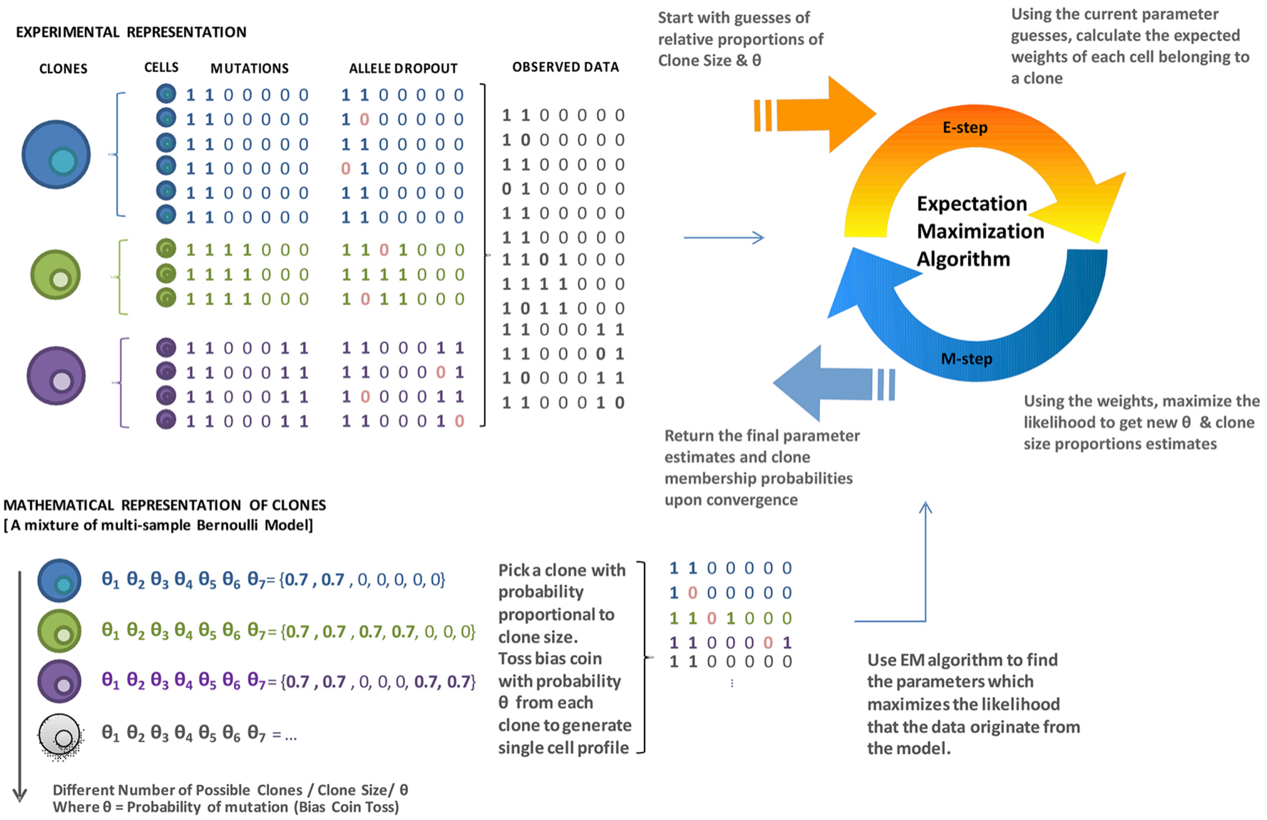


Figure S1: Determination of ADO Rate and Identification of High Quality Cells in Single-Cell Sequencing Data

a) ADO rate measured in patient 4 using qPCR for 46 commonly heterozygous loci b) ADO rate for all 6 patients using targeted resequencing of 96 commonly heterozygous loci c) Number and percent of cells above and below 30% ADO for each patient based on targeted resequencing



Start with guesses of relative proportions of Clone Size & θ

Using the current parameter guesses, calculate the expected weights of each cell belonging to a clone

Using the weights, maximize the likelihood to get new θ & clone size proportions estimates

Return the final parameter estimates and clone membership probabilities upon convergence

Pick a clone with probability proportional to clone size. Toss bias coin with probability θ from each clone to generate single cell profile

Use EM algorithm to find the parameters which maximizes the likelihood that the data originate from the model.

Figure S2 Overview of Expectation Maximization Algorithm on a Multivariate Bernoulli Model

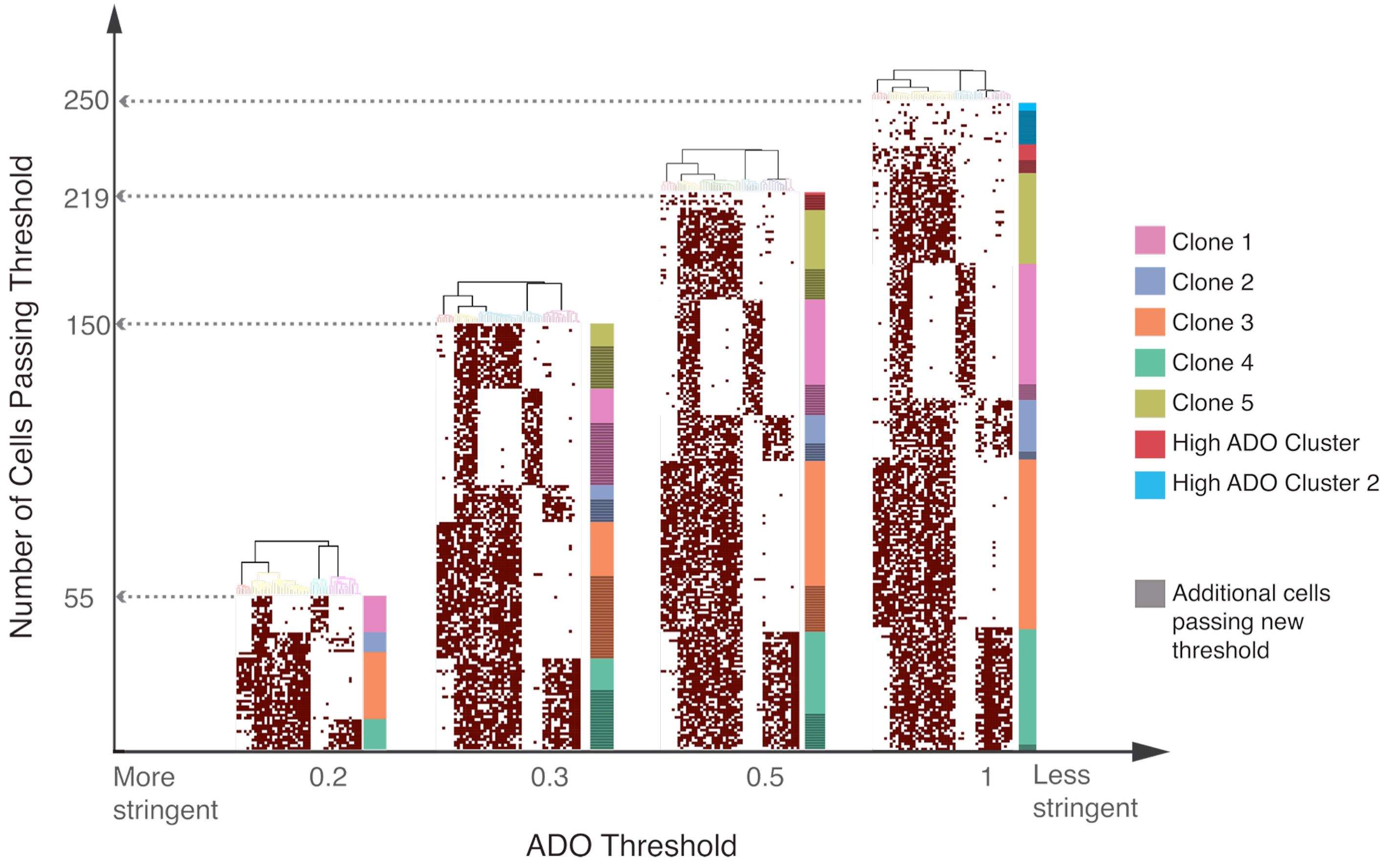


Figure S3: Determination of the Stability of Clonal Structures of Patient 3 After Increasing the Measured ADO Threshold

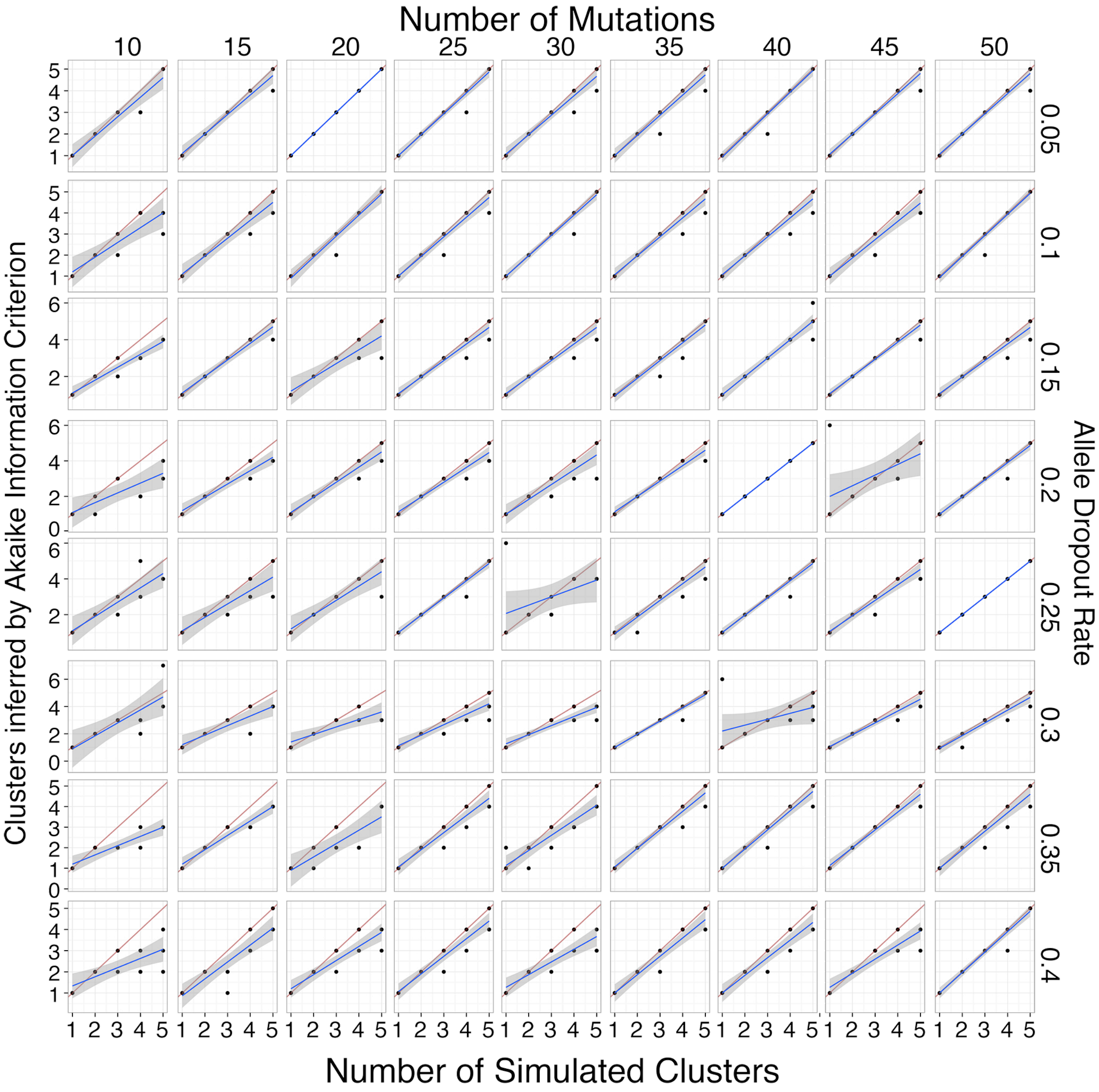


Figure S4: Simulation of Clone Number Estimates at Increasing ADO Rates and Mutations Numbers

Number of simulated clusters are compared to estimated number of clusters using randomly generated data. X=Y line (red) is included to show with the values would reside if there was complete agreement between the two cluster measurements, and standard error is represented in grey. As the number of mutations decrease or ADO rate increases, the Akaike Infromation Criterion underestimate the true number of clones. Our data have an ADO rate of 0.2 and use between 10 and 105 mutations (median 46).

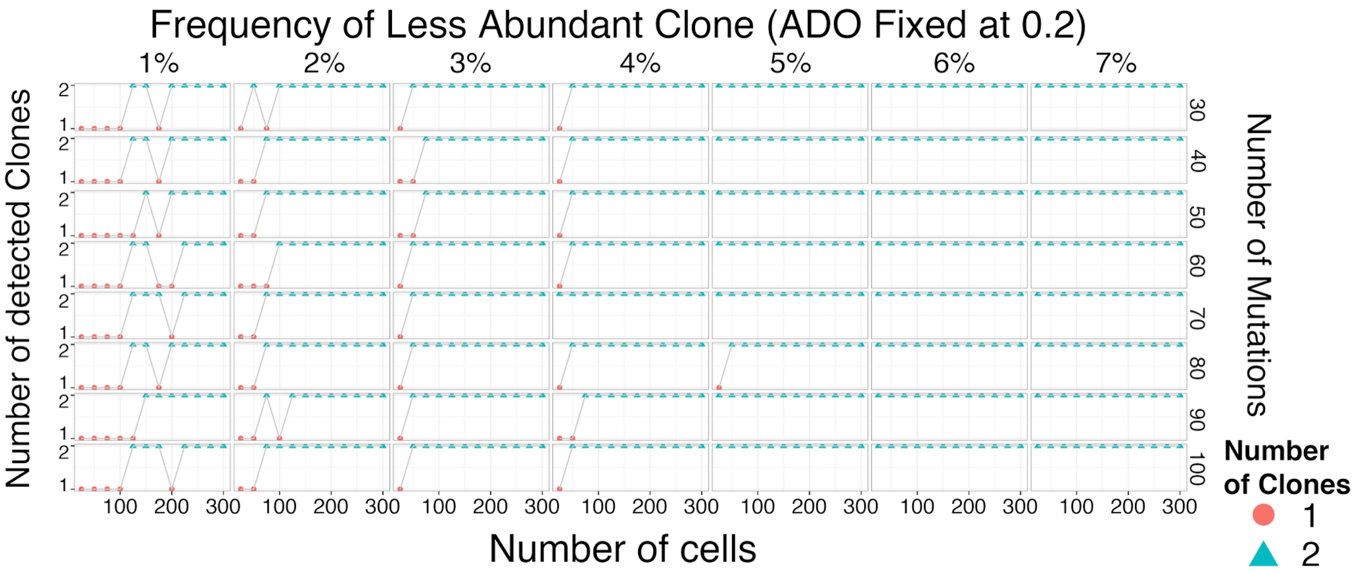


Figure S5: Simulation of Number of Cells Needed to Evaluate to Identify Lower Frequency Clone with Varied Number of Input Mutations

Fixing our data at our estimated ADO rate of 0.2, we varied the number of mutations and cells evaluated to determine when we could reliably detect a minor clonal population. At the median number of mutations we evaluated (40), we could detect a 1% clone with 200 cells, 2% with 75 cells, and 4% with 50 cells. Thus, on average, we would need to identify at least two to three different cells from the same clone to accurately detect that population,

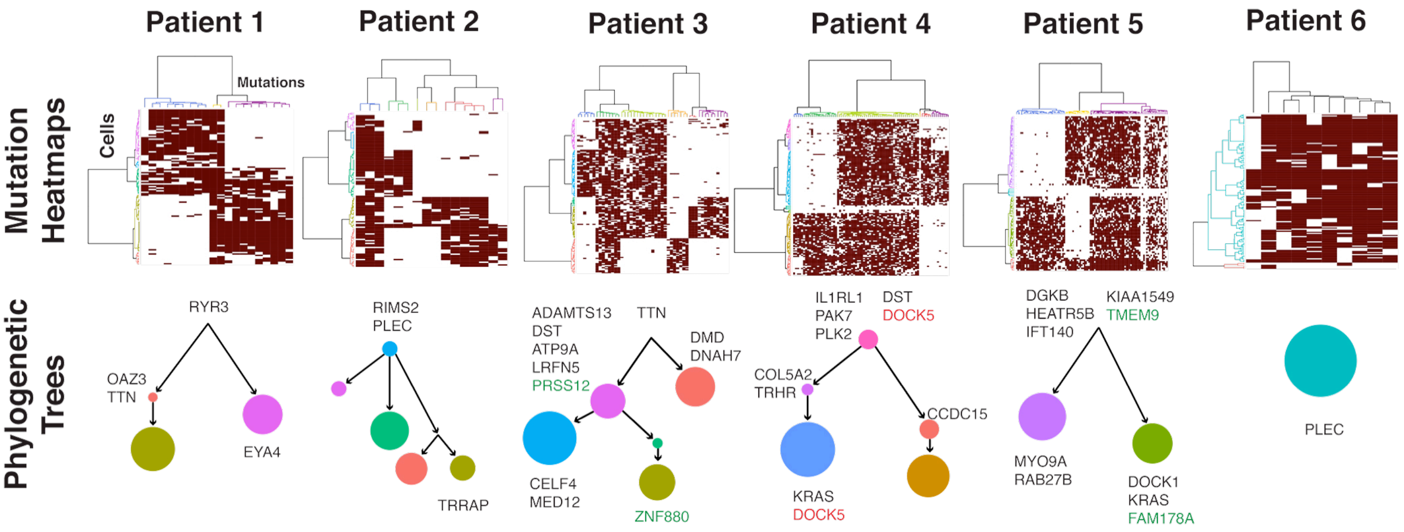


Figure S6: Unsupervised Clustering of Single-Cell Mutation Calls to Identify Clonal Populations and Reconstruct Tumor Phylogenies in ALL Samples
Cells were clustered on the y-axes and mutations on the x-axes, both by Jaccard distance. Mutation calls are represented by maroon boxes, and mutation or cell clusters are represented by different colors. The identification of distinct clusters of cells and mutations enabled the resolving of distinct clones, as well as the inference of inter-clonal relationships and undetectable ancestors as measured in the minimum spanning trees. The size of each clone is proportional to its relative abundance, and the length of edges are proportional to the Jaccard distance between clones. Recurrently mutated genes in *ETV-RUNX1* leukemias are shown in the clones where they were acquired, and green genes are mutated more than once in the same clone while red genes are mutated more than once in the same patient but in different clones.

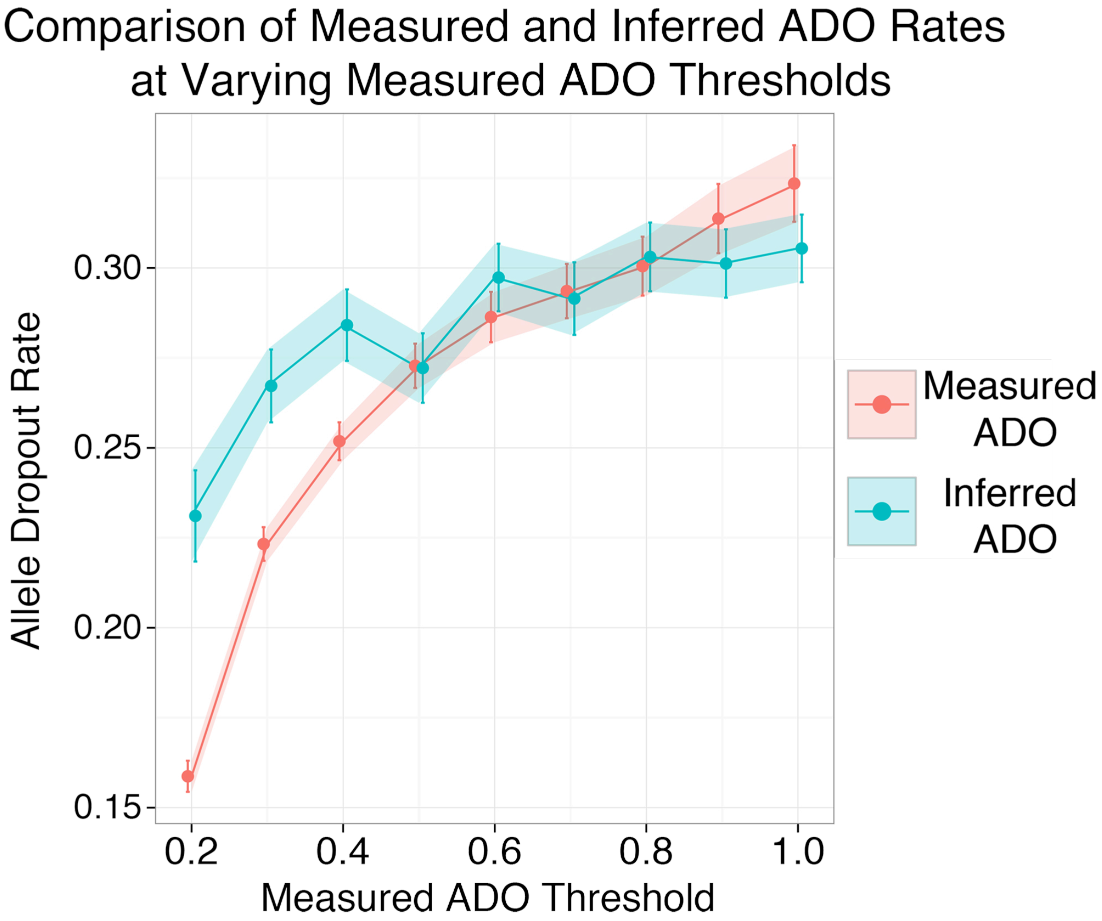


Figure S7: Comparing the Median Measured and Inferred ADO Rates for Patient 3 While Increasing the Measured ADO Threshold

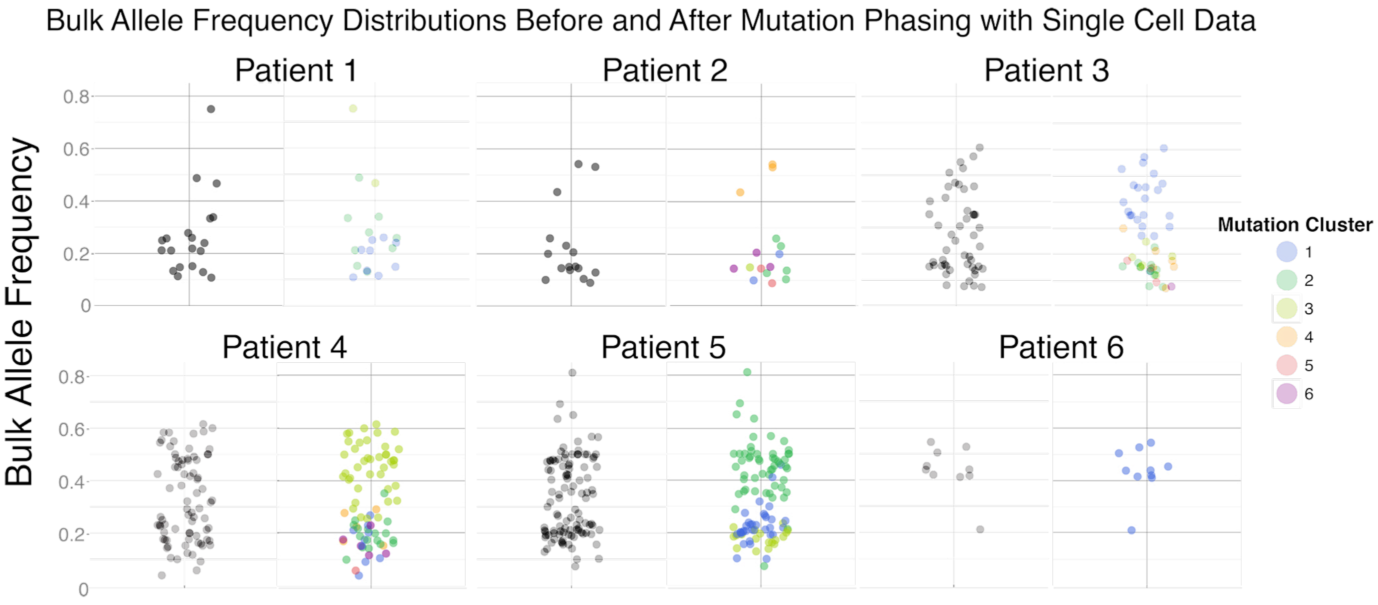


Figure S8: Phasing of Mutations into Cells Based on Single-Cell Cooccurrence Measurements

Bulk allele frequency distributions are depicted before (in black) and after (in color) phasing of mutations by clustering single-cell mutation profiles. Most of the clusters of lower frequency mutations have overlapping allele frequency distributions, which precludes resolving them into distinct clones based on the bulk allele frequency data alone.

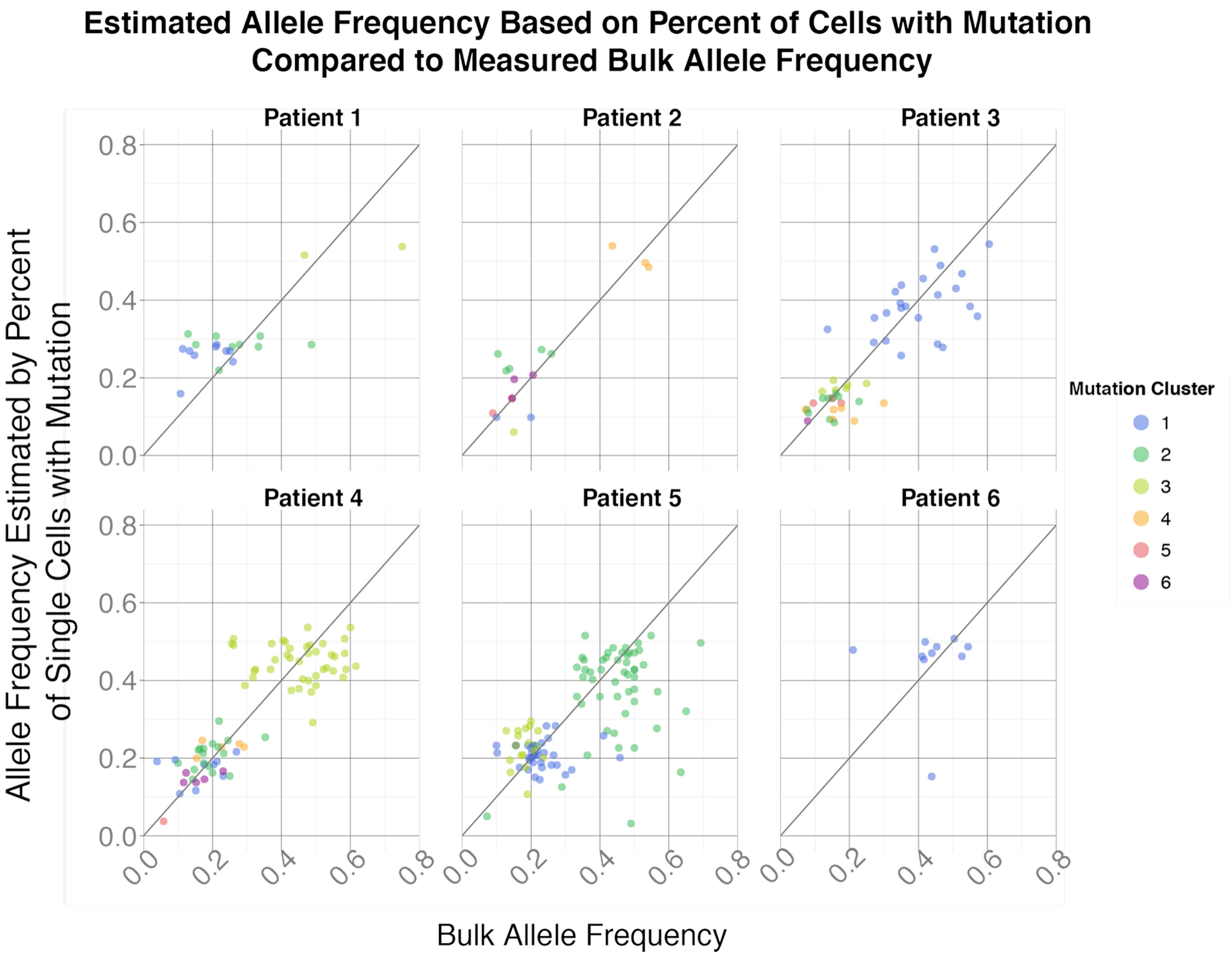


Figure S9: Comparing Allele Frequency Measured in Bulk Sample to Percent of Cells with each Mutation

There is a strong correlation between the bulk allele frequency and percent of cells with a detected mutation. In addition, mutation clusters identified in figure S6 group at the same allele frequencies measured by bulk and single cell approaches. Single cell allele frequencies are corrected for ADO by dividing by (1-ADO rate measured for each patient).

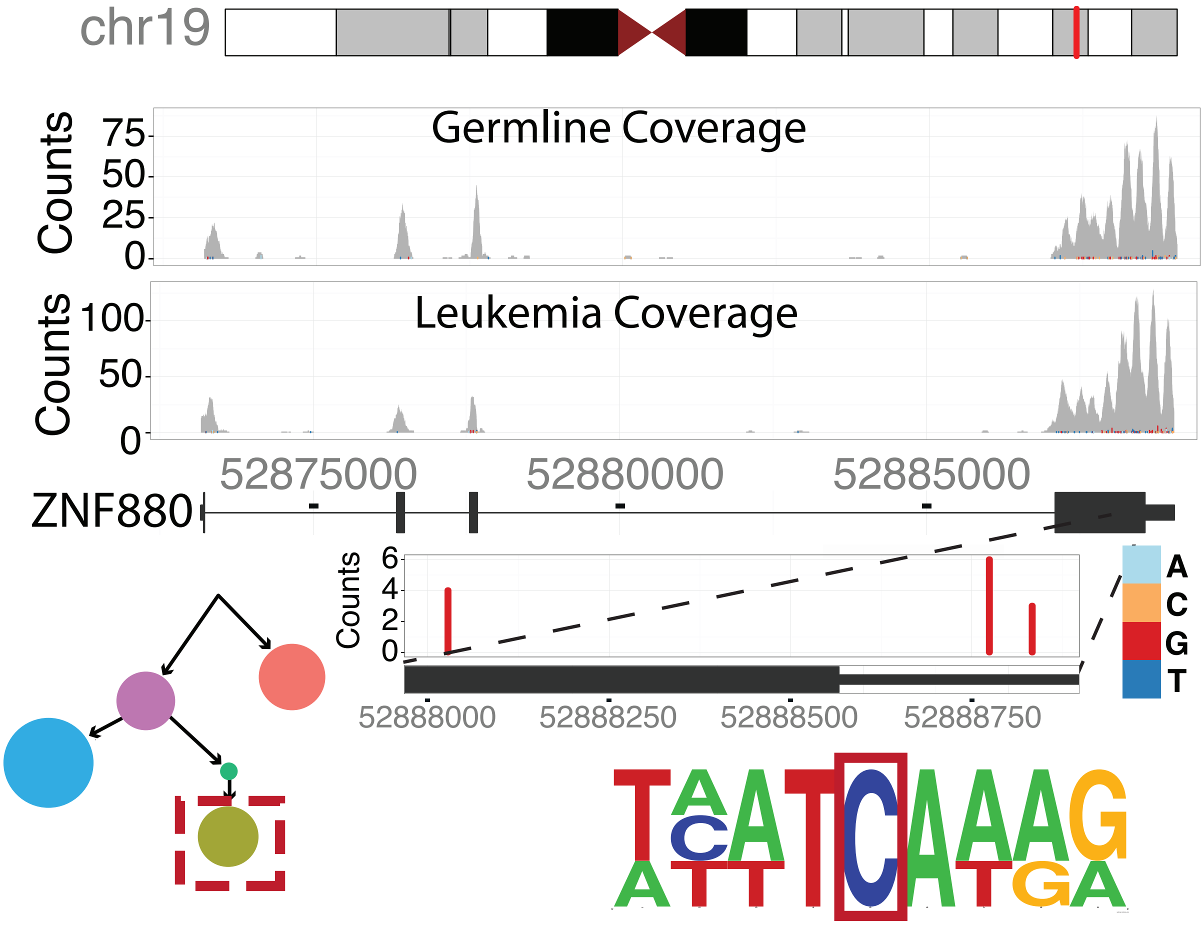


Figure S10: Identification of Clone-Specific Punctuated Cytosine Mutagenesis Three mutations in a single clone of patient 4 (dashed red box) are localized to a 750bp stretch of a single exon and were acquired in the same clone. All three are C->G mutations, and all three have a TC̲A motif.

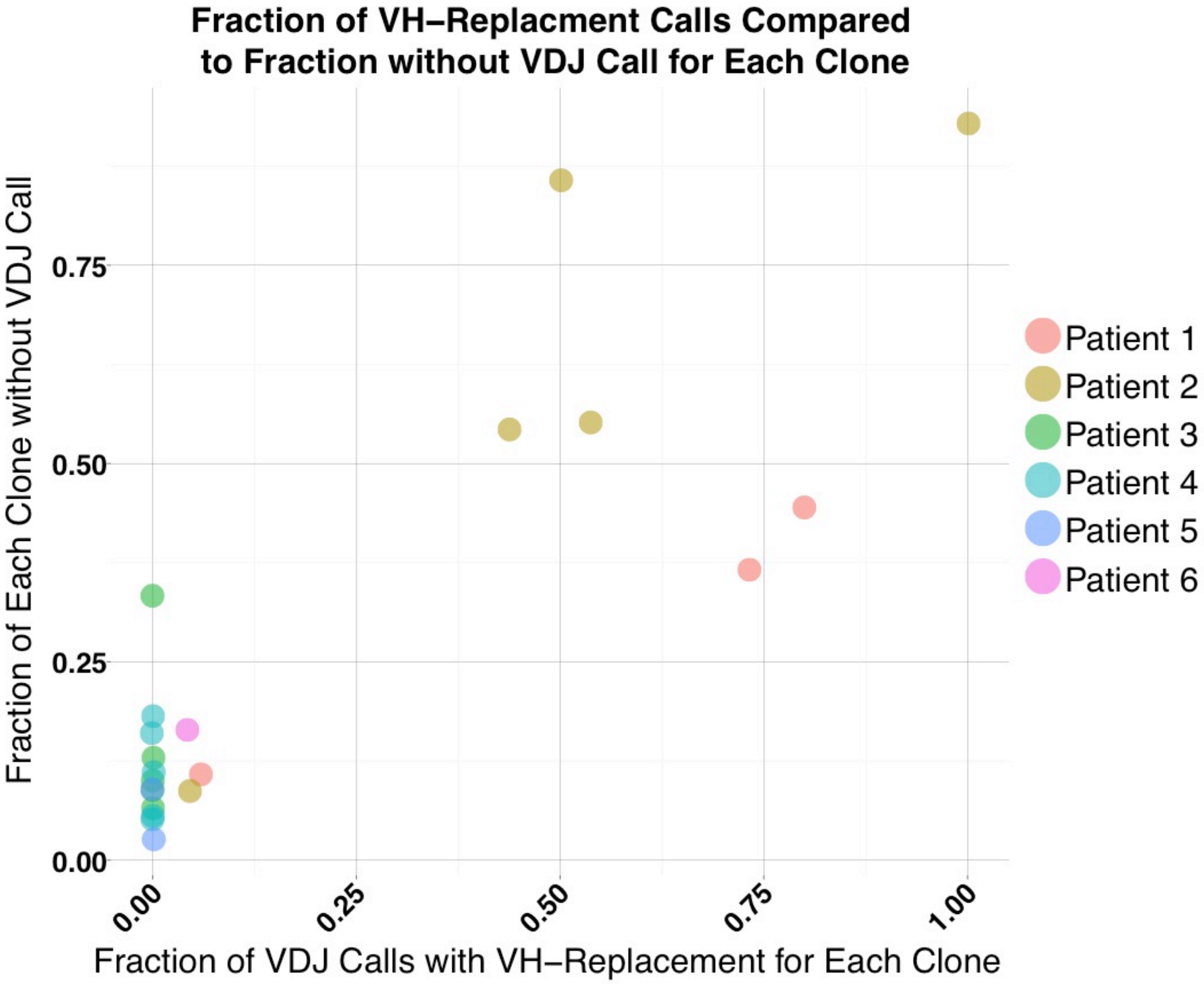


Figure S11: Fraction of cells Without a Called VDJ Sequence for each Clone Compared to the Fraction of Cells with VH Replacement

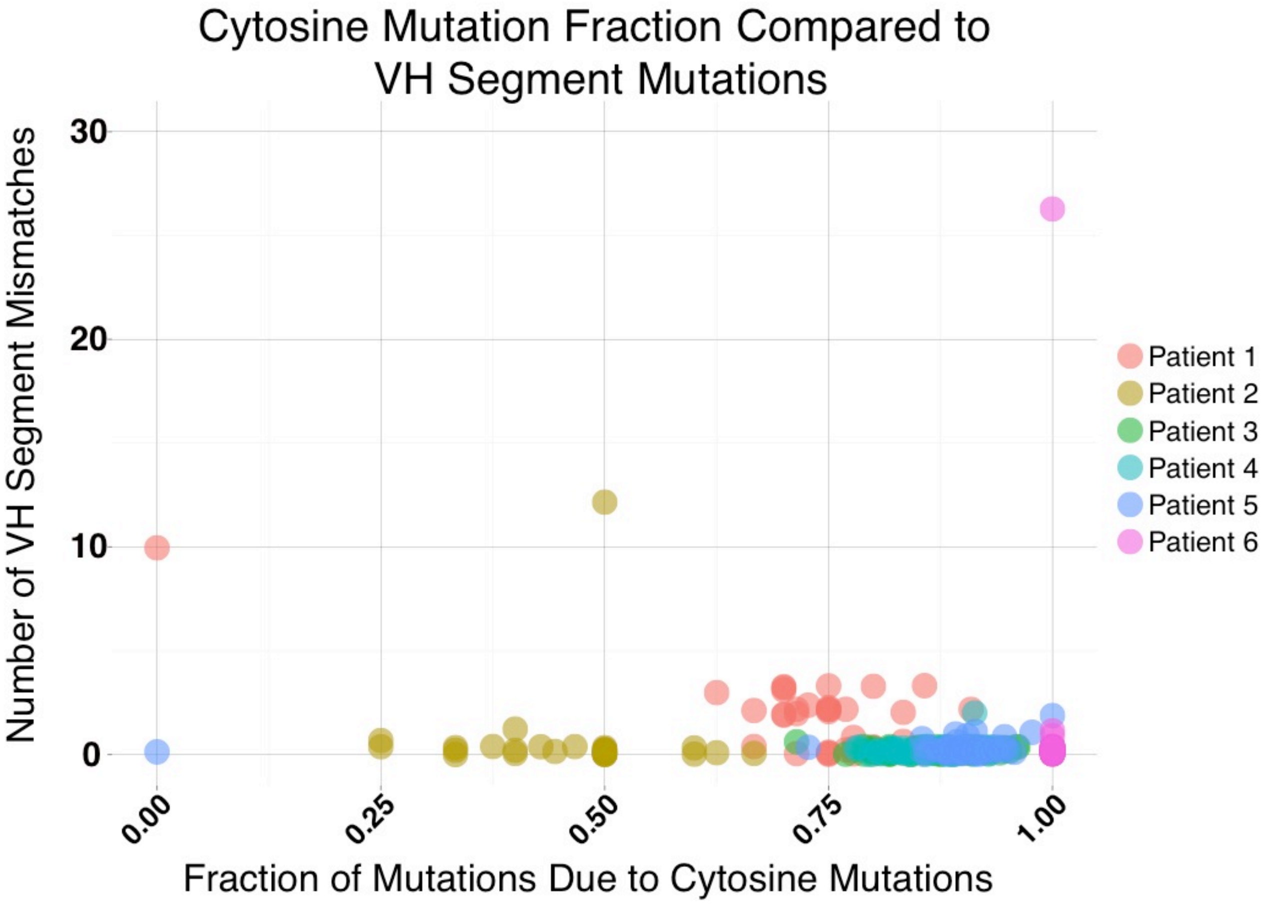


Figure S12: Number of VH-Segment Mutations Detected and Correlation Between VH-Segment Mutations and Percent Cytosine Mutations for each Cell

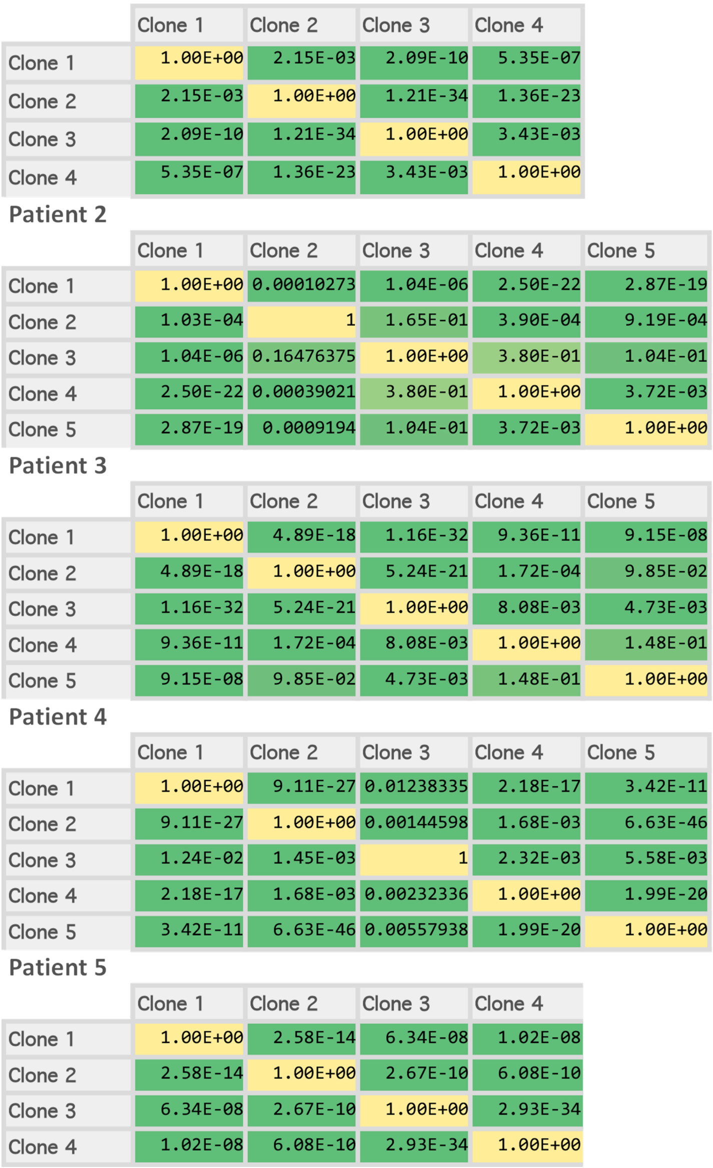
	Sex	Age Dx	Initial WBC	ETV6-RUNX1	% Cell FISH+	Karyotype	Remission	Day 29 MRD	Normal Coverage	Tumor Coverage
Patient 1	M	6	44	N	N/A	46 XY	Y	Negative	19.3	42.6
Patient 2	M	3	0.6	Y	82	cryptic t(12:21), trisomy 21	Y	unk	24.3	47.5
Patient 3	F	4	163	Y	87	del7p13-15(8/20), -X 2/20?	Y	Negative	36.7	49.4
Patient 4	F	6	16	Y	90	46 XX	Y	Negative	37.5	55.7
Patient 5	F	5	11	Y	88	46XX, del 5q13, add 12p11, del 14q24	Y	Positive	34.9	55.7
Patient 6	F	3	140	Y	73	46 XX	Y	Negative	31.6	56

Table S1: Summary of Patient Characteristics

ADO Estimate Method	Sample	ADO Rate	ADO Rate After Removing Low Quality Cells
Taqman-based genotyping of 46 loci	Patient 4	22.70%	N/A
Taqman-based genotyping of 46 loci	Lymphoblastoid Cell Line	15.60%	N/A
Targeted resequencing of 96 loci	Patients 1-6	33.30%	18.90%
Dropout of wildtype allele with mutation call	Patients 1-6	24.40%	19.70%

Table S2: Summary of ADO Dropout Data Based on 3 Methods

Patient 1



	Clone 1	Clone 2	Clone 3	Clone 4
Clone 1	1.00E+00	2.15E-03	2.09E-10	5.35E-07
Clone 2	2.15E-03	1.00E+00	1.21E-34	1.36E-23
Clone 3	2.09E-10	1.21E-34	1.00E+00	3.43E-03
Clone 4	5.35E-07	1.36E-23	3.43E-03	1.00E+00

Patient 2

	Clone 1	Clone 2	Clone 3	Clone 4	Clone 5
Clone 1	1.00E+00	0.00010273	1.04E-06	2.50E-22	2.87E-19
Clone 2	1.03E-04	1	1.65E-01	3.90E-04	9.19E-04
Clone 3	1.04E-06	0.16476375	1.00E+00	3.80E-01	1.04E-01
Clone 4	2.50E-22	0.00039021	3.80E-01	1.00E+00	3.72E-03
Clone 5	2.87E-19	0.0009194	1.04E-01	3.72E-03	1.00E+00

Patient 3

	Clone 1	Clone 2	Clone 3	Clone 4	Clone 5
Clone 1	1.00E+00	4.89E-18	1.16E-32	9.36E-11	9.15E-08
Clone 2	4.89E-18	1.00E+00	5.24E-21	1.72E-04	9.85E-02
Clone 3	1.16E-32	5.24E-21	1.00E+00	8.08E-03	4.73E-03
Clone 4	9.36E-11	1.72E-04	8.08E-03	1.00E+00	1.48E-01
Clone 5	9.15E-08	9.85E-02	4.73E-03	1.48E-01	1.00E+00

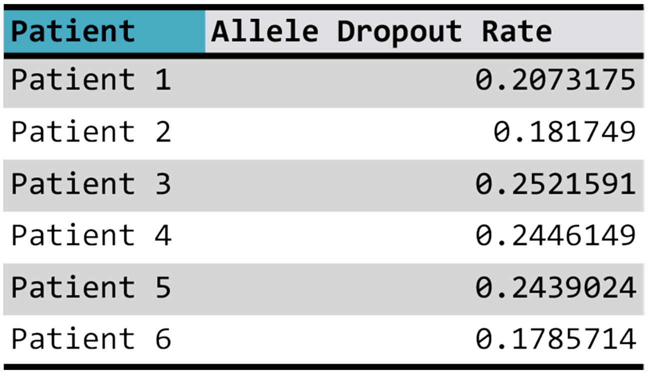
Patient 4

	Clone 1	Clone 2	Clone 3	Clone 4	Clone 5
Clone 1	1.00E+00	9.11E-27	0.01238335	2.18E-17	3.42E-11
Clone 2	9.11E-27	1.00E+00	0.00144598	1.68E-03	6.63E-46
Clone 3	1.24E-02	1.45E-03	1	2.32E-03	5.58E-03
Clone 4	2.18E-17	1.68E-03	0.00232336	1.00E+00	1.99E-20
Clone 5	3.42E-11	6.63E-46	0.00557938	1.99E-20	1.00E+00

Patient 5

	Clone 1	Clone 2	Clone 3	Clone 4
Clone 1	1.00E+00	2.58E-14	6.34E-08	1.02E-08
Clone 2	2.58E-14	1.00E+00	2.67E-10	6.08E-10
Clone 3	6.34E-08	2.67E-10	1.00E+00	2.93E-34
Clone 4	1.02E-08	6.08E-10	2.93E-34	1.00E+00

Table S3: Pairwise Inter-clonal P-Value Calculated by Comparing to the Null Hypothesis that the Clones are Identical



Patient	Allele Dropout Rate
Patient 1	0.2073175
Patient 2	0.181749
Patient 3	0.2521591
Patient 4	0.2446149
Patient 5	0.2439024
Patient 6	0.1785714

Table S4: ADO Estimated by Measuring the Intra-clonal Variant Call Loss Rate