**Research Strategy:**

## 1. Background and Significance

**1.1. Heterogeneity is a common feature of cancer. A better understanding of this heterogeneity may present therapeutic opportunities:** Intratumor heterogeneity is a common feature across diverse cancer types[1,2,3]. Dynamic changes can be observed among intratumoral subclonal populations over time and following therapy, presenting challenges to current standards of cancer treatment[6,7,8]. Characterization of subclonal populations in cancer may enable precision medicine and the initiation of synergistic treatment combinations to target subclonal drivers and eliminate aggressive subpopulations to improve clinical outcome. Identification of subclonal driver mutations may also present new treatment options, particularly if these mutations fall within targetable pathways. Our proposal will yield innovative, novel statistical methods to enable the identification and characterization of subclonal populations in cancer using single cell RNA-seq data and yield open-source software that can be tailored and applied to diverse cancer types.

**1.2. Heterogeneity in CLL plays a role in clonal evolution to shape therapeutic resistance:** CLL is a slow-growing B cell malignancy that exhibits diverse combinations of clonal and subclonal somatic mutations along with a highly variable disease course among patients that remains poorly understood[3,4]. Our collaborators in the Wu group have recently established that the presence of particular subclonal mutations in CLL can be linked with adverse clinical outcomes using bulk samples and measurements[5]. Furthermore, these subclonal mutations change over time in response to therapy, suggesting an active evolutionary process, eventually leading to therapeutic resistance and relapse in many cases. While insights have been previously gained from bulk samples and measurements, further characterization on the single cell level is needed to more accurately dissect the pathway and regulatory features associated with subclonal mutations. Our proposal to analyze the transcriptomes of single CLL B cells derived from 3 CLL patients at various time points pre- and post-treatment and 4 additional CLL patients exhibiting different patterns of clonal and subclonal mutations will provide insights to the molecular mechanisms of relapse and progression in CLL.

**1.3. Statistical methods are needed to identify and connect genetic and transcriptional heterogeneity in single cells:** Transcriptional heterogeneity can be observed in normal cell types such as neural progenitor cells[23], and T cells[24], as well as aberrant cell types such as cancer[1,2]. Differential properties such as genetic differences among cells may be responsible for this heterogeneity but how it is regulated, along with its direct consequences on cellular behavior, remains unclear. Applying traditional bulk protein analysis methods on single cells has met with varied degrees of success due to the high levels of technical as well as biological stochasticity and noise inherent in single-cell measurements. Therefore, novel statistical methods are needed to identify and connect genetic and transcriptional heterogeneity in single cells as well as identify putative subpopulations. Our previous work demonstrates that integration of cell specific error models and probabilistic weighting of observations improves the ability to separate cell types within a mixed single cell sample when clustering cells based on gene expression[17,18]. Our proposal will apply these statistical approach as well as develop new approaches to improve characterization of genetic and transcriptional single cell heterogeneity and subsequently enhance our understanding of cellular variability and its connection to genetic differences as well as biological consequences.

## 2. Approach

The hallmarks of CLL make this cancer a particularly compelling model upon which to develop statistical methods for connecting genetic and transcriptional heterogeneity at the single cell level. Through my collaboration with the Wu lab, I have access to single-cell RNA-seq data for 7 CLL patient samples (CW14, CW106, CW84, CW236, MDA1, MDA2, MDA3) with known clonal and subclonal somatic mutations previously identified by bulk WES. Additional single-cell RNA-seq will also be generated as a part of separate research efforts. Here, I propose a series of single-cell studies to identify and connect patterns of genetic to transcriptional heterogeneity and associate clinical outcomes. First, I will develop a hierarchical Bayesian framework for to make probabilistic inferences on presence or absence of CNVs and SNVs inferred from single-cell RNAseq data. Second, I will reconstruct subclonal architectures, impute the order of genetic alterations incurred, and identify genetic subclones based on somatic mutations inferred from Aim 1 within CLL cases. Third, I will identify differentially expressed genes and pathways, with particular emphasis on pathways involved in RNA splicing, apoptosis, cell proliferation, cellular senescence, DNA damage repair, inflammation, Wnt and Notch signaling, to characterize these subpopulations. I will integrate treatment time course data for 3 patients (MDA1; 5 time points, MDA2; 3 time points, MDA3; 3 time points) to directly associate transcriptional features with treatment response and relapse.

### 2.1. Aim 1: Inferring somatic mutations from single-cell RNA-seq data.

### 2.1.1. Preliminary data:

2.1.1.a. Intratumoral genetic heterogeneity can be observed in CLL and is linked with adverse clinical outcome. My collaborators in the Wu group have previously revealed that the presence of subclonal mutations in CLL can be linked with adverse clinical outcomes[5]. The Wu group and other investigators have identified several novel putative CLL drivers, including the splice factor SF3B1, LCP1, and WNK1[26]. The mechanisms by which these mutations confer impacts CLL biology is unknown.

2.1.1.b. SNVs called from single-cell RNA-seq can be used to distinguish cell lines. Despite being limited to variants within the expressed exons, SNVs derived from RNA-seq can still be used to separate genetically distinct single cells. Previously, using single-cell RNA-seq data and a benchmark variant sets identified from WES for GM12878 and K562 cell lines, we evaluated the sensitivity and precision of such RNA-based SNV calls, comparing various combinations of aligners and variant callers. We found that sufficiently high performance can be achieved for SNVs within highly expressed genes (Fig. 1a). Using simulated mixtures of GM1282 and K562 single cells, we are able to separate these genetically distinct cell types based on a small fraction of SNVs called from single-cell RNA-seq data (Fig. 1b) . Single cells from the same CLL patient sample will come from the same genetic background and harbor less distinctive subclonal SNVs, thus creating a more challenging problem in need of additional statistical methods and alternative data integration such as CNVs.

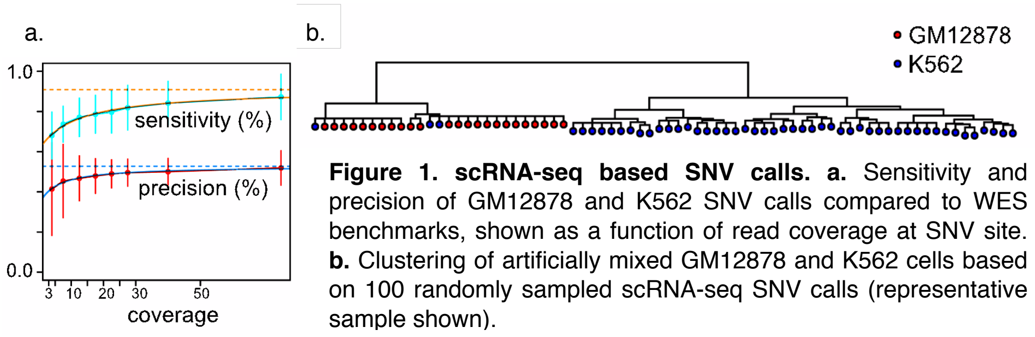


**Figure 1. scRNA-seq based SNV calls. a.** Sensitivity and precision of GM12878 and K562 SNV calls compared to WES benchmarks, shown as a function of read coverage at SNV site. **b.** Clustering of artificially mixed GM12878 and K562 cells based on 100 randomly sampled scRNA-seq SNV calls (representative sample shown).

2.1.1.c. Biased allele expression can be observed within CNV regions for single-cell RNA-seq data. Our previous analysis of clonal deletion regions in multiple myeloma revealed distinct patterns in the detection of known heterozygous germline single nucleotide polymorphisms (SNPs) identified by WES within regions affected by CNV in single cells. For each heterozygous germline SNP within a candidate CNV region, we infer which allele is affected by the CNV based on deviations away from the expected 1:1 allele ratios for heterozygous variants observed in bulk. As expected, for deletion regions, only non-deleted allele variants are observed within the deleted region (Fig. 2). Most of SNPs within the CNV neutral regions also exhibit highly biased allele ratios, but the direction of the bias varies between cells. This suggests that despite prevalent mono-allelic and biased expression, because the direction of bias is random within CNV neutral regions, we should be able to detect CNVs based on observations of persistent directional bias of expression. However, additional statistical methods are needed to quantify the probability of such observations, taking into consideration potential sequencing errors or RNA-processing.

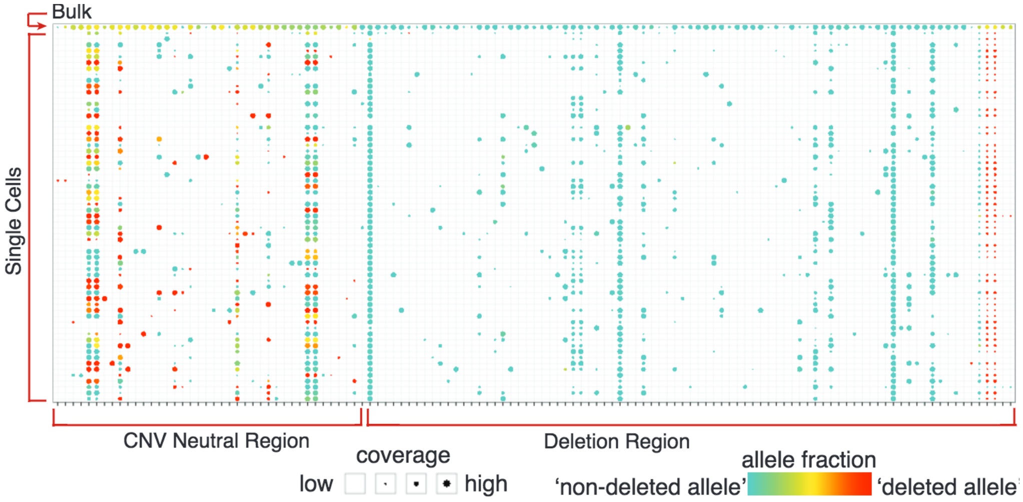
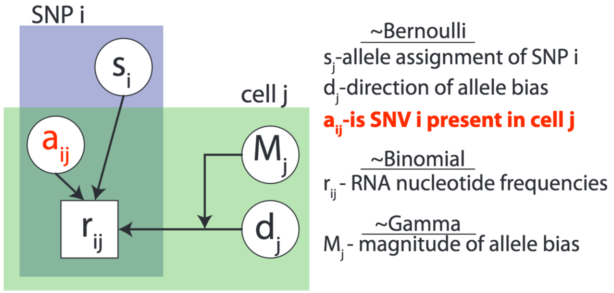


**Figure 2. Biased allele expression within and outside of CNV regions.** Heterozygous germline SNPs (columns) for single cells (rows) inferred from single-cell RNA-seq is biased away from the expected allele fraction of 0.5 for heterozygous variants due to mono-allelic expression within CNV neutral regions and due to clonal deletion status within deletion regions. In this example, all single cells exhibit a deletion in the known deletion region based on clonal deletion status inferred from bulk WES.

**2.1.2. Research design:** Here, we propose integrating prior knowledge acquired from bulk WES along with single-cell RNA-seq to infer the presence of somatic mutations on a single-cell level. Specifically, from bulk WES, we will identify candidate regions of CNV using Control-FREEC[27], identify putative somatic variants and heterozygous germline SNPs using MuTect[28]. We will also call for somatic SNVs from single-cell RNA-seq using GATK[29] to identify additional rarer somatic SNVs that may be not be present at sufficient frequencies to be detectable in bulk. We will then use the following hierarchical Bayesian models to assess the posterior probability of the presence of candidate SNVs and CNVs in single cells.

2.1.2.a. Bayesian approach to SNV inference. Inference of subclonal architecture relies on detection of subclonal variants such as SNVs. However, mono-allelic expression poses a major challenge to SNV detection, since a SNV may not be observed in the sequenced reads but can actually be present and simply not detected due to mono-allelic expression of the non-SNV carrying allele, thus hindering further analysis.

Here, we propose an alternative approach to overcome this uncertainty by first establishing that both alleles are indeed expressed in given cell, or, even more specifically, that the allele carrying the somatic variant is expressed, by looking at neighboring heterozygous SNPs. Specifically, we will take advantage of prevalent mono-allelic expression to derive probabilistic models of SNP (germline) and SNV (somatic) phasing, using them to increase certainty in the SNV presence/absence calls (Fig. 3). For example, if both alleles are observed for neighboring heterozygous SNPs, we will have greater certainty that mono-allelic expression is not a factors and that the SNV absence is a true negative. Our hierarchical Bayesian approach allows us to derive posterior probabilities on the presence of SNVs to quantify the uncertainty in our calls. Initial testing indicates that such approach is very effective at recovering phasing of SNVs with germline variants, allowing us to confidently infer SNV absence in approximately half of the ambiguous cases.



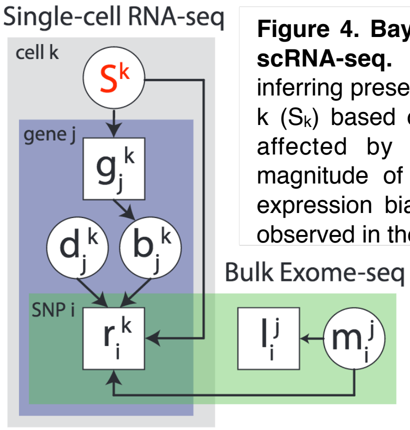
SNP i
$s_i$
cell j
$a_{ij}$
$M_j$
$r_{ij}$
$d_j$
~Bernoulli
$s_j$-allele assignment of SNP i
$d_j$-direction of allele bias
$a_{ij}$-is SNV i present in cell j
~Binomial
$r_{ij}$ - RNA nucleotide frequencies
~Gamma
$M_j$ - magnitude of allele bias

**Figure 3. Bayesian model for inferring SNV presence from single-cell RNA-seq.** A hierarchical Bayesian model is shown for inferring presence or absence of a candidate SNV i in a given cell j ($a_{ij}$), based on the inferred phasing with germline variants ($s_i$), and inferred allele expression bias ($M_i$ and $d_i$).

To model the rate of allele bias magnitude as a parameter in our model, we will look at heterozygous SNVs in known CNV neutral regions based on bulk WES. We expect to be able to observe both alleles are equal proportions unless there is mono-allelic expression or allele bias. We can then assess for the probability or rate of mono-allelic expression and allele bias as a function of gene length, gene expression, or other factors. Likewise, to determine the effective error rate due to reverse transcription, amplification, and sequencing, we will look at homozygous SNVs in known CNV neutral regions based on bulk WES. SNVs observed that are not of the expected allele can be attributed to error. We can then assess error as a function of coverage and other factors.

We will benchmark our method by calling SNVs in clonal mutant and normal samples. In this manner, samples with clonal mutations are used as true positive benchmarks while normal samples, which should harbor no mutations, are used as true negative benchmarks.

2.1.2.b. Bayesian inference of CNV absence/presence from single-cell RNA-seq data. Detection of CNVs provide larger somatic changes that can be used for more robust inference of subclonal architecture. Previous efforts to infer CNVs on a single cell level from transcriptomic data have been limited to whole chromosome and chromosome arm level changes[12]. Here, we propose an alternative approach to enable detection of smaller CNVs, taking advantage of heterozygous SNPs within CNV regions. Intuitively, if a cell has the deletion, then we expect there to be only expression from the non-deleted allele. Which allele is the deleted allele can be inferred probabilistically using bulk WES



Single-cell RNA-seq
cell k
$S^k$
gene j
$g_j^k$
$d_j^k$
$b_j^k$
SNP i
$r_i^k$
Bulk Exome-seq
$l_i^j$
$m_i^j$
~Bernoulli
$S^k$-presence of deletion in cell *k*
$d_j^k$-bias-affected allele
$m_i^j$-CNV-affected allele
~Binomial
$r_i^k$,$l_i^j$- RNA/exome allele frequencies
~Normal
$g_j^k$- expression level of gene *j* in cell *k*

**Figure 4. Bayesian model for inferring CNV presence from scRNA-seq.** A hierarchical Bayesian model is shown for inferring presence or absence of a candidate CNV in a given cell k ($S_k$) based on combination of alleles observed in the genes affected by the CNV ($r_{ik}$), normalized gene expression magnitude of the affected genes ($g_{jk}$), expected monoallelic expression bias ($b_{jk}$ and $d_{jk}$), as well as the allele frequencies observed in the bulk exome data ($l_{ij}$ and $m_{ij}$).

data. If a cell does not have the deletion, a number of scenarios may occur. If there is no mono-allelic expression, then we expect to be able to observe both SNPs with approximately equal probability with some deviation expected due to biased allele expression. If we consistently observe only expression from the non--deleted allele across SNPs within multiple genes, then the cell most likely has a deletion. However, if we observe only expression from the non-deleted allele in one gene and there is a high probability of mono-allelic expression, such patterns may also be explained by mono-allelic expression, increasing uncertainty in our deletion status inference. Similarly for amplifications, we would rely on allelic imbalance and higher expression from the amplified allele in comparable ratios across heterozygous SNVs within the amplification region.

Our hierarchical Bayesian approach allows us to incorporate the uncertainty in each detected allele in the single-cell RNA-seq data, in the bulk WES data, gene expression magnitude, mono-allelic expression, and sequencing error to assess the joint likelihood that the CNV is present in a given cell (Fig. 4). The proposed model thus infers the posterior probability on the presence/absence of a single CNV in a given cell. Again, we will integrate mono-allelic expression and effective error as done in the SNV model and benchmark our method by calling CNVs in clonal deletion and normal samples.

**2.1.3. Potential problems and alternative solutions.** While the preliminary testing of the proposed approaches demonstrates their performance on well-defined cases such as cell lines and clonal samples, additional development will be necessary to accommodate more common experimental designs and improve overall performance. Specifically, the proposed design relies on the availability of the WES data, which is used to infer candidate CNVs, candidate SNVs, and heterozygous germline SNPs. We find that the exact boundaries of the CNVs can differ from those detected by the CNV prediction algorithms. Furthermore, some CNV boundaries vary among the clones. A boundary refinement step can be used to correct for such cases by focusing on the smallest shared region or trimming edges. An alternative HMM-like application of the current model will be evaluated in order to detect subclonal CNVs, but evaluating joint “emission” probability of both single-cell RNA-seq and WES data. Similarly, SNV analysis proposed currently avoids all SNVs that fall within CNVs detected in bulk. While this restriction can in principle be relaxed for the SNVs within amplified regions, additional provisions will have to be made to exclude SNVs found within common subclonal CNVs.

## 2.2. Aim 2: Reconstructing subclonal architecture and dissecting subclonal evolution on the single-cell level.

### 2.2.1. Preliminary data

2.2.1.a. Active genetic evolutionary process is observed in CLL in response to treatment. Recent advancements in the understanding of the role of B cell receptor signaling in CLL pathogenesis have led to the development Ibrutinib, an irreversible inhibitor of Bruton's tyrosine kinase, that has demonstrated prolonged responses in heavily pretreated and refractory patients. In a detailed study of 3 cases treated with ibrutinib at multiple time points both pre- and post-treatment, using bulk WES, my collaborators in the Wu group and I have identified distinct subclonal populations marked by mutually exclusive somatic mutations that change in population frequency and proportion at each time point[22], suggesting an active evolutionary process (Fig. 5). An in-depth characterization of such samples, such as that offered by single-cell RNA-seq, will provide definitive information on the mechanisms underlying clonal dynamics of CLL and their relation to therapeutic resistance.

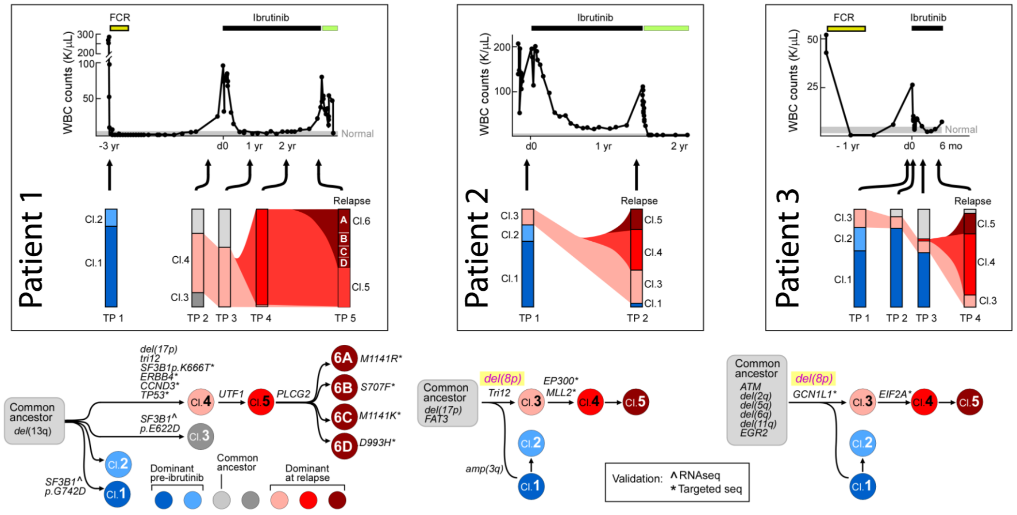


**Figure 5. Genetic evolution in CLL in response to ibrutinib treatment.** Bulk samples were collected and sequenced for each patient at various time points pre and posttreatment as indicated by the black arrows. Analysis of cancer cell fractions by ABSOLUTE reveals different subclonal populations at each time point (TP1-TP5). In particular, dominant subclonal populations in relapsing CLL cases can be observed as minor subclasses pretreatment (e.g. cl. 5 in Patient 1), suggesting an active, branched evolutionary process in CLL clonal expansion.

**2.2.2. Research design:** We will build upon SNVs and CNVs identified in Aim 1 to reconstruct the sub clonal architecture of single cells within each sample and further infer the temporal ordering of somatic mutations. Intuitively, if cells within a tumor carry several shared somatic mutations, then they must be derived from the same single ancestral cell that also harbored these mutations. The probability that cells acquired the same mutations independently is unlikely. We can thus use these somatic mutations such as SNVs and CNVs to reconstruct the underlying subclonal architecture and identify subpopulations. However, such reconstruction must be done from within a probabilistic framework due to the inherent uncertainty associated with detection of each SNV and CNV.

Phylogenetic tree reconstruction from sequence data is a well-studied problem. A number of statistical likelihood-based and fully Bayesian approaches for phylogenetic tree reconstruction are already available[30,31,32]. We propose modifying one such approach, BEAST[32], to incorporate uncertainty in the observed genotypes. We will benchmark these approaches as we have done previously in 2.1.1.c. We will apply this method to reconstruct the subclonal architectures for 3 CLL patients at multiple time points, pre and post chemo and ibrutinib treatment (Fig. 5). We anticipate that our single cell CNV detection method will recapitulate cancer cell fraction estimates and proportions previously estimated from bulk WES by ABSOLUTE[15].

The reconstructed phylogenetic tree will also give us information on the order in which somatic mutations were acquired. To benchmark the accuracy of our inferred temporal orderings, we will compare our the ordering with the dynamics of subclonal architecture architecture reconstructed from bulk WES using methods such as PhyloWGS[34] or ABSOLUTE[15]. Furthermore, for the 3 multi-time-point CLL samples, we will compare the inferred temporal ordering at each time point.

**2.2.4. Potential problems and alternative solutions.** Classification of subclonal structure of the single-cell samples is critical for the proposed analysis. Our preliminary results indicate that high coverage achieved for many genes in single-cell RNA-seq measurements provides sufficient information to examine subclonal structure. In samples where such analysis will be limited by noise/coverage, we will restrict subclonal architecture reconstruction to the somatic variants detected in the bulk WES data alone and remove putative gremlin variants based on prior knowledge from bulk WES data.

### 2.3. Transcriptomic characterization of genetic subclonal populations.

### 2.3.1. Preliminary data:

<u>2.3.1a. Statistical model for single-cell RNA-seq data and Bayesian test identifies robustly differentially expressed genes.</u> Single-cell transcriptomic measurements via single-cell RNA-seq is complicated by high levels of technical and biological noise. Losses during the reverse transcription step of library preparation along

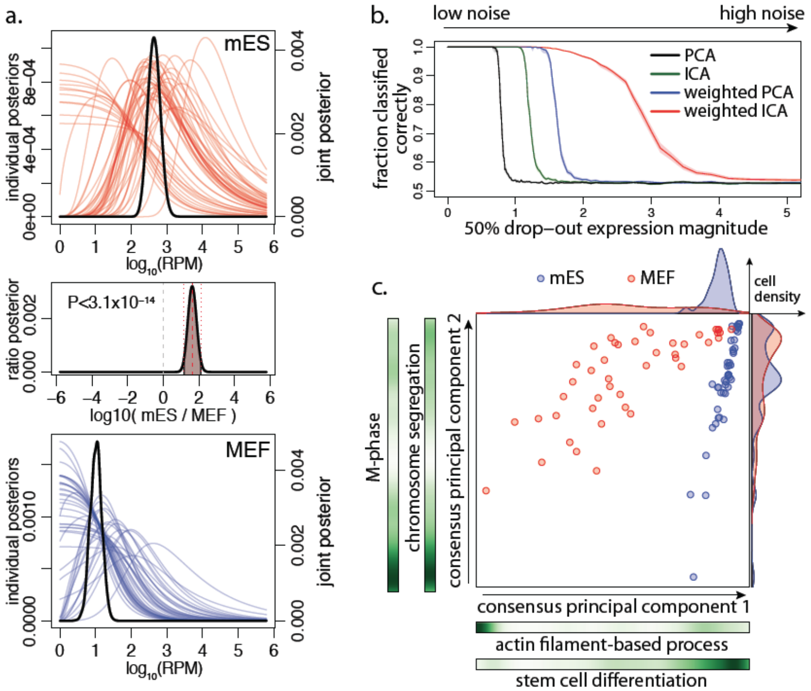


**Figure 6. Bayesian analysis of differential expression. a.** The error model of each single cell is used to estimate the expression magnitude posterior (red/blue curves) given the observed data. The approach estimates joint posterior distribution for the overall level with each cell type (black curves), and the expression fold difference between the cell types (middle plot). The example demonstrates expression differences of Sox2 between mES and MEF cells. **b.** Probabilistic weighting delays breakdown of common multivariate techniques. Ability of PCA and ICA to correctly separate two simulated cell types (3000 genes, 150 differentially expressed) drops sharply as the frequency of dropout events increases (x axis: log10 RPM at which 50% of genes fail to be detected). Probabilistic weighting of observations using dropout probabilities predicted by the error models allow PCA and ICA to distinguish subpopulations at much higher levels of noise. **c.** Principal component analysis separating mES and MEF single cells. The sidebars show (density) of genes from different GO categories based on their loading in the corresponding principal component.

with stochastic transcriptional bursting can lead to “drop-out” events, where a gene is observed at moderate or even high expression level in one cell but is not detected in another cell even though expression may be present but simply low[17,18]. To accommodate these abundant drop-out events along with the high variability of single-cell data, we model the measurement of each cell as a mixture of two probabilistic processes – one in which the transcript is amplified and detected at a level correlating with its abundance (modeled using a negative binomial distribution), and the other where a gene fails to amplify or is not detected for other reasons (modeled as a low-level Poisson background)[17,18]. We have further implemented a Bayesian method for such differential expression analysis that uses these error models to estimate the likelihood of a gene being expressed at any given average level in each of the single-cell subpopulations, as well as the likelihood of expression fold change between them (Fig. 6). We find that such an approach shows improved specificity/ sensitivity compared to other common RNA-seq analysis methods[17].

2.3.1b. Previous unbiased transcriptional characterization of CLL reveals transcriptional heterogeneity. Preliminary analysis of low-coverage single-cell RNA-seq data from 4 CLL tumor samples (CW14, CW106, CW84, CW236) illustrate the presence of intra-tumoral as well as inter-tumoral transcriptionally distinct sub-sets, separating along functionally relevant criteria such as immune response pathways (Fig. 7). However, how these transcriptionally distinct subpopulations relate to genetically distinct subclones is not known.

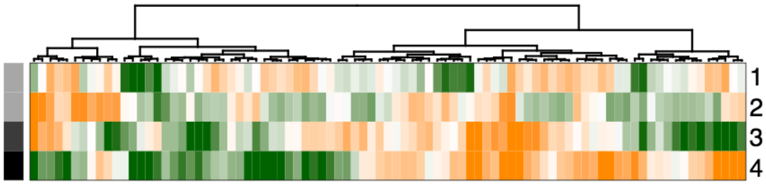


**Figure 7. Transcriptional heterogeneity in a CLL sample.** The heat map illustrates single cells as columns and consensus gene expression within pathway clusters as rows. Preliminary analysis of low-coverage scRNA-seq data from a CLL tumor illustrates presence of transcriptionally distinct sub-sets, separating along functionally relevant criteria such as immune response pathways.

**2.3.2. Research design:** Having identified subclonal populations using somatic mutations in Aim 1 and Aim 2, we will assess the transcriptional profiles of each subclonal populations.

For each intra-patient subpopulation, we will apply single-cell differential expression analysis[17] to identify differentially upregulated and downregulated genes associated with each subclone. We will use gene set enrichment analysis[35] to determine if differentially expressed genes genes are enriched for particular pathways or gene sets.

Additionally, the ability to assay multiple time points in CLL patients (Fig. 5) provides a rare opportunity to observe expansion, contraction, and evolution of tumor subpopulations following therapeutic interventions. By comparing single-cell RNA-seq data from different time points we will identify: 1) Transcriptional features such as unregulated and down regulated genes and gene sets accompanying subclonal expansions (in relapse and metastatic samples) following treatment, 2) transcriptional features predictive of subclonal dynamics (expansion or contraction), and 3) persistent aspects of transcriptional heterogeneity not tied to the underlying genetic shifts in the subclonal architecture.

We will focus on assessing transcriptional heterogeneity of key regulatory pathways and downstream targets of signaling pathways previously identified by our collaborators in the Wu lab to be associated with CLL development, therapeutic response, and remission including RNA splicing, apoptosis, cell proliferation, cellular senescence, DNA damage repair, inflammation, Wnt and Notch signaling[26]. In this manner, we will examine the potential impact of presence of genetic subpopulations in which these various pathways as well as the pathway directed targeted by the administered drug are inhibited on the subsequent disease progression. Similarly, we will test for potential association with different modes of B-cell receptor signaling[36], subclonal activation of Wnt signaling[37] and other pathways implicated in CLL-B-cell expansion of potential relevance to CLL progression.

**2.3.3. Potential problems and alternative solutions.** The comparison of subclonal populations will focus on the major (high posterior probability) splits in the phylogeny. However, in the cases when such subpopulations will not be obvious, or when the subclone correspondence cannot be established between serial samples from the same individual, we will refer back to the WES data, using predictions from methods such as PhyloWGS[34] or ABSOLUTE[15] to establish correspondence. We will also apply an unbiased approach to assessing transcriptional heterogeneity using the pathway and gene set over dispersion analysis method I previously developed. We will then assess whether particular somatic mutations are associated with the observed patterns of transcriptional heterogeneity.