August 8, 2015

Dear CSR Division of Receipt and Referral,

Thank you for receiving my F31 application. The following information describes my submission:

1. Title: Computational Analysis of Subclonal Evolution in Chronic Lymphocytic Leukemia

2. FOA: PA-14-147

3. The following awarding component would be a good reviewer for this application:
   • National Cancer Institute – NCI

4. There are no individuals who should be excluded from reviewing this application.

5. N/A

6. N/A

7. N/A

8. Required agency approval documentation: N/A

9. List of referees:
   Catherine J. Wu, M.D, Dana Farber Cancer Institute
   Shamil Sunyaev, Ph.D, Division of Genetics, Harvard Medical School
   Soumya Raychaudhuri, M.D, Ph.D, Divisions of Genetics & Rheumatology, Harvard Medical School
   Chirag Patel, Ph.D, Department of Biomedical Informatics, Harvard Medical School
   Nils Gehlenborg, Ph.D, Department of Biomedical Informatics, Harvard Medical School

Thank you for your time and consideration.

Yours sincerely,

Jean Fan
PhD Candidate
Kharchenko Lab
Harvard Medical School
Intratumor genetic and transcriptional heterogeneity is a common feature across diverse cancer types, including CLL. CLL is a particular cancer that exhibits genetic and transcriptional heterogeneity along with a highly variable disease course among patients that remains poorly understood. Previous research has established that the presence of particular subclonal mutations in CLL can be linked with adverse clinical outcomes and that these subclonal mutations change over time in response to therapy. Therefore, genetic and transcriptional characterization of these subclonal populations will be paramount to enabling precision medicine and synergistic treatment combinations that target subclonal drivers and eliminate aggressive subpopulations thereby improving clinical outcome.

While bulk measurements and analysis has provided key insights into cancer biology, etiology, and prognosis in the past, this approach does not provide the resolution that is critical for understanding the interactions between different genetic events within the same environmental and genetic backgrounds to drive metastatic disease, drug resistance and disease progression. Single cell measurements are uniquely able to definitively unravel and connect these relationships. However, simultaneous extraction of DNA and RNA from the same single cells is currently not reliable. Therefore, new statistical methods and computational approaches are needed to identify and resolve genetic subpopulations using single cell transcriptional data alone.

In this proposed research, I will develop statistical methods and computational software to analyze single cell RNA-seq data derived from CLL patient samples. Specifically, I will develop methods to identify aspects of genetic heterogeneity, such as the presence of small single nucleotide mutations and regions of copy number variation, in single cells. I will then reconstruct the genetic subclonal architecture and characterize the gene expression profiles of identified subclonal populations. The proposed work will yield innovative statistical methods to enable the identification and characterization of subclonal populations in cancer and yield opensource software that can be tailored and applied to diverse cancer types. Ultimately, application of these developed methods to CLL will provide a better understanding of CLL development and progression.
August 8, 2015

Dear CSR Division of Receipt and Referral,

Thank you for receiving my F31 application. The following information describes my submission:

1. Title: Computational Analysis of Subclonal Evolution in Chronic Lymphocytic Leukemia

2. FOA: PA-14-147

3. The following awarding component would be a good reviewer for this application:
   • National Cancer Institute – NCI

4. There are no individuals who should be excluded from reviewing this application.

5. N/A

6. N/A

7. N/A

8. Required agency approval documentation: N/A

9. List of referees:
   Catherine J. Wu, M.D, Dana Farber Cancer Institute
   Shamil Sunyaev, Ph.D, Division of Genetics, Harvard Medical School
   Soumya Raychaudhuri, M.D, Ph.D, Divisions of Genetics & Rheumatology, Harvard Medical School
   Chirag Patel, Ph.D, Department of Biomedical Informatics, Harvard Medical School
   Nils Gehlenborg, Ph.D, Department of Biomedical Informatics, Harvard Medical School

Thank you for your time and consideration.

Yours sincerely,

Jean Fan
PhD Candidate
Kharchenko Lab
Harvard Medical School
Specific Aims:

Intratumor heterogeneity is a common feature across diverse cancer types\textsuperscript{1,2,3}. Chronic lymphocytic leukemia (CLL) exhibits particularly diverse combinations of clonal and subclonal somatic mutations and copy number variations (CNVs) along with a highly variable disease course among patients that remains poorly understood\textsuperscript{3,4}. Dynamic changes can be observed among intra-tumoral subclonal populations over time and following therapy, suggesting an active evolutionary process\textsuperscript{5}. Such tumoral evolution can lead to therapeutic resistance and relapse thereby presenting challenges to current standards of cancer treatment\textsuperscript{6,7,8}.

Transcriptional characterization of these subclonal populations is integral to understanding this evolutionary process. We hypothesize that different combinations of subclonal mutations in CLL will present different transcriptional changes that affect key pathways involved in RNA splicing, apoptosis, cell proliferation, cellular senescence, DNA damage repair, inflammation, Wnt and Notch signaling to ultimately provide particular subclones with enhanced tumorigenic efficiency. While bulk measurements and analysis has provided key insights into cancer biology, etiology, and prognosis in the past, this approach does not provide the resolution that is critical for understanding the interactions between different genetic events within the same environmental and genetic backgrounds to drive metastatic disease, drug resistance and disease progression. Single cell measurements are uniquely able to definitively unravel and connect these relationships. However, simultaneous extraction of DNA and RNA from the same single cells is currently not reliable. Therefore, new statistical methods and computational approaches are needed to identify and resolve genetic subpopulations using single cell transcriptional data alone.

I will take advantage of single-cell RNA-seq datasets generated by the Wu lab at the Dana-Farber Cancer Institute as a part of separate research efforts (1R01HL11645201, 1R01CA15501002) to study subclonal evolution in CLL. I will develop innovative statistical methods to resolve genetic subclonal populations and link transcriptional profiles at the single cell level.

Aim 1: Inferring somatic mutations from single-cell RNA-seq data. Somatic copy number variants (CNVs) and single nucleotide variants (SNVs) play an important role in cancer pathogenesis and progression\textsuperscript{9,10,11}. Inferring CNVs from transcriptomic data remains difficult due to uneven coverage across deletion and amplification sites\textsuperscript{12}. Similarly, inferring somatic mutations from transcriptomic data is liable to false negatives due to high rates of mono-allelic gene expression\textsuperscript{13,14}. Aim 1 will develop a Bayesian hierarchical approach to leverage information from across multiple genetic loci to make probabilistic inferences on presence or absence of CNVs and SNVs. My approach will take into consideration biases introduced by mono-allelic expression and other technical artifacts such as sequencing errors.

Aim 2: Reconstructing subclonal architecture and dissecting subclonal evolution on the single cell level. Intratumoral heterogeneity is a key factor in determining cancer progression, drug resistance, and clinical outcomes\textsuperscript{1,2,3,4,5}. Traditional bulk measurements are unable to resolve whether mutations are mutually exclusive or co-occurring\textsuperscript{15} and are subject to averaging artifacts\textsuperscript{16}. Aim 2 will develop statistical methods to reconstruct subclonal architectures, impute the order of genetic alterations incurred, and identify genetic subclones based on somatic mutations inferred from Aim 1 from within a probabilistic framework. These methods will be applied to primary, metastatic, pre- and post-treatment CLL samples to assess proportion, frequency, and evolution of subclonal populations and their impact on clinical outcome.

Aim 3: Transcriptomic characterization of genetic subclonal populations. While traditional bulk characterization of genetic clonal populations have revealed immense transcriptional differences putatively linked to particular somatic mutations, single-cell analysis of subclonal populations will allow us to characterize genetically distinct subpopulations within the same environment and genetic background to identify and tease out potentially more subtle or environment-dependent effects. Building on my lab’s previous work in characterizing transcriptional heterogeneity in single cells\textsuperscript{17,18}, Aim 3 will analyze the transcriptional state(s) of distinguishable genetic subclones to identify features associated with clonal growth rate, metastatic transition, and drug resistance. I will work closely with collaborators in the Wu lab to validate findings using in-vitro and in-vivo techniques.

Successful completion of this proposal will yield new insight into subclonal evolution in CLL and provide powerful new open-source computational software for identifying and characterizing subclonal populations that can be tailored and applied to diverse cancer types.
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. Dynamic changes can be observed among intratumoral subclonal populations over time and following therapy, presenting challenges to current standards of cancer treatment. 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. 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. 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, and T cells, as well as aberrant cell types such as cancer. 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. 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.

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 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).

**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 mon-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\textsuperscript{27}, identify putative somatic variants and heterozygous germline SNPs using MuTect\textsuperscript{28}. We will also call for somatic SNVs from single-cell RNA-seq using GATK\textsuperscript{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.

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\textsuperscript{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.
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\textsuperscript{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 post-treatment 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 subclonal 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. We propose modifying one such approach, BEAST, 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.

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 or ABSOLUTE. 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:

2.3.1a. Statistical model for single-cell RNA-seq data and Bayesian test identifies robustly differentially expressed genes. 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
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\textsuperscript{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)\textsuperscript{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\textsuperscript{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.

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\textsuperscript{17} to identify differentially upregulated and downregulated genes associated with each subclone. We will use gene set enrichment analysis\textsuperscript{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\textsuperscript{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\textsuperscript{36}, subclonal activation of Wnt signaling\textsuperscript{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\textsuperscript{34} or ABSOLUTE\textsuperscript{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.
Resource Sharing Plan:

Software

The proposed project will develop statistical and computational approaches to be made available for broader use as open source R software packages. These packages will be made freely downloadable through publically accessible online repositories such as GitHub and BioConductor thereby enabling other groups to apply these methods to other single-cell studies. Mailing lists, issue reporting, wiki documentation, and other infrastructure will be set up to connect and assist researchers applying these methods in their investigations.

My advisor, Dr. Peter Kharchenko and I, have excellent track records in delivering computational tools to the scientific community, including software for inferring spatial localization of gene sets (Fan et al, manuscript in review), pathway and gene set overdispersion analysis (Fan et al, manuscript in review), single-cell differential expression analysis (Kharchenko et al, Nature Methods 2014), analysis of ChIP-seq data (Kharchenko et al, Nature Biotechnol. 2008), analysis of repetitive elements (Day et al, Genome Biol. 2010), and identification of transposable element insertions (Lee E et al, Science 2012).

Publications

All results generated in this project will be published in peer-reviewed journals and will be made available to the scientific community. In such publications, as well as in related presentations or press releases, we will fully acknowledge the support provided by NIH in conducting this work.
References:

28. Cibulskis K, Lawrence MS,
A. Personal Statement

My longterm research interests involve the development of a comprehensive understanding of key genetic, epigenetic, and other regulatory mechanisms driving cellular identity and heterogeneity within cellular groups, tissues, and organs. I am particularly interesting in heterogeneity in the context of cancer and how this heterogeneity shapes tumor progression, therapeutic resistance, and ultimately clinical impact.

My extensive scientific research experience has exposed me to both the wet lab, empirical approaches as well as the dry lab, computational approaches in tackling different sides of the same biological questions. As a high school student, I conducted wet lab research at the National Cancer Institute to identify putative oncogenes within the 8p11-12 amplicon driving breast cancer pathogenesis. As an undergraduate, I developed computational algorithms in Rachel Karchin’s lab to predict the deleterious impact of mutations based on sequence conservation. I also assessed genetic variation and population structures on an organismal level in Shamil Sunyaev’s lab. For my doctoral training, I will focus on developing statistical and computational methods for analyzing genomics, not on an organismal level, but on a single cell level. Since the start of my doctoral training, under the guidance and mentorship of Peter Kharchenko as well as my collaborators, I have developed software for inferring spatial localization of gene sets (Fan et al, manuscript in review) and pathway and gene set overdispersion analysis (Fan et al, manuscript in review) as well as applied my computational skills to analyze locally disordered methylation (Landau et al, Cell 2014) and clonal evolution in developing drug resistance (Burger et al, Cancer Discovery 2015) in chronic lymphocytic leukemia.

I hope to continue developing powerful statistical methods with user-friendly computation software in close collaboration with wet lab researchers and oncologists to enable more personalized cancer therapies in this era of precision medicine.
B. Positions and Honors

<table>
<thead>
<tr>
<th>ACTIVITY/OCCUPATION</th>
<th>START DATE (mm/yy)</th>
<th>ENDING DATE (mm/yy)</th>
<th>FIELD</th>
<th>INSTITUTION/COMPANY</th>
<th>SUPERVISOR/EMPLOYER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer research intern</td>
<td>06/2008</td>
<td>09/2008</td>
<td>Cancer Biology</td>
<td>National Cancer Institute, National Institutes of Health</td>
<td>Paul Meltzer and Liang Ciao</td>
</tr>
<tr>
<td>Undergraduate research scientist</td>
<td>08/2009</td>
<td>05/2013</td>
<td>Bioinformatics, Evolutionary biology, Genomics</td>
<td>Institute for Computational Medicine, Johns Hopkins University</td>
<td>Rachel Karchin</td>
</tr>
<tr>
<td>Summer research intern</td>
<td>06/2012</td>
<td>08/2012</td>
<td>Population genetics</td>
<td>Harvard Medical School</td>
<td>Shamil Sunyaev</td>
</tr>
<tr>
<td>Teaching assistant</td>
<td>08/2012</td>
<td>12/2012</td>
<td>Mathematics</td>
<td>Johns Hopkins University</td>
<td>Donniell Fishkind</td>
</tr>
<tr>
<td>Teaching fellow</td>
<td>10/2014</td>
<td>10/2014</td>
<td>Genomics, Transcriptomics</td>
<td>Harvard Stem Cell Institute</td>
<td>Peter Kharchenko</td>
</tr>
</tbody>
</table>

Awards and Honors

- Siemens Competition in Math, Science and Technology SemiFinalist (2008)
- Intel Science Talent Search SemiFinalist (2009)
- DC-AMS Scholarship Winner (2009)
- Posse Scholarship Semi-Finalist (2009)
- Johns Hopkins University Dean’s list (2009-2013)
- Provost’s Undergraduate Research Award Winner (2012)
- National Science Foundation Graduate Research Fellowship Program awardee (2013)

Memberships

- American Society of Human Genetics
- International Society for Computational Biology
- American Association for the Advancement of Science
C. Contributions to Science (for predoctoral students and more advanced candidates only; high school students, undergraduates, and postbaccalaureates should skip this section)

A more complete understanding of chronic lymphocytic leukemia

Advancements in high-throughput sequencing technologies have uncovered tremendous genetic, epigenetic, and transcriptional heterogeneity in chronic lymphocytic leukemia (CLL) but its impact on clinical course is not well understood. I have established a close collaboration with the Wu lab at the Dana-Farber Cancer Institute, where I have focused on developing and applying bioinformatics methods for (1) assessing variability of single cell gene expression, (2) calling mutations from single cell qt-qPCR data, (3) differential expression and gene set enrichment tests for both bulk and single cell, RNA-sequencing and targeted qt-qPCR data. Our collaboration has led to many scientific findings that contribute to a more complete understanding of CLL:

- Pleiotropic effects of splice variants generated by SF3B1 mutations in chronic lymphocytic leukemia (manuscript in preparation)

Statistical methods and software for analyses of single cell data

While heterogeneity within cellular systems has long been widely recognized, only recently have technological advances enabled measurements to be made on a single cell level. Applying traditional bulk 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 characterize heterogeneity in single cells. In the Kharchenko lab, I have focused on developing methods for analyzing single cell data, including differential expression analysis methods that takes into account sources of technical noise inherent to single cell RNA-seq data, clustering methods to identify pathways and gene sets that exhibit coordinated variability, and methods for spatial placement of cell subpopulations based on expression signatures. This work has led to the development of various statistical methods available as software for the scientific community:


STEM Outreach

Women are underrepresented in science, technology, engineering, and math (STEM) fields. Improving the representation of women in STEM is pertinent to workplace diversity, gender equality, and American innovation. To help address this issue, I have been involved in a number of outreach efforts:

- I was the lead software engineer for the BioHazardz 3D video games (http://bioinfor.me/), which teach students the fundamentals of protein evolution through an intuitive and attractive gaming environment.
- I founded the 501(c)3 non-profit cuSTEMized (http://www.custemized.org/) and developed the website, software, graphics, and content to enable parents to generate and download free personalized motivational ebooks to help girls envision themselves in science, technology, engineering, and math.
- I am the co-chair of the Harvard Graduate Women in Science and Engineering student group. I manage and lead organization of networking, professional development, and mentoring events for women in natural sciences, social sciences, and engineering at Harvard University.
### D. Scholastic Performance

<table>
<thead>
<tr>
<th>YEAR</th>
<th>SCIENCE COURSE TITLE</th>
<th>GRADE</th>
<th>YEAR</th>
<th>OTHER COURSE TITLE</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>General Physics I</td>
<td>A</td>
<td>2009</td>
<td>Honors Multivariable Calculus</td>
<td>A</td>
</tr>
<tr>
<td>2009</td>
<td>General Physics Lab</td>
<td>A+</td>
<td>2009</td>
<td>BME Modeling and Design</td>
<td>A</td>
</tr>
<tr>
<td>2010</td>
<td>General Physics II</td>
<td>A-</td>
<td>2009</td>
<td>State and Family: Revised Classical Perspective</td>
<td>A+</td>
</tr>
<tr>
<td>2010</td>
<td>General Physics Lab II</td>
<td>A</td>
<td>2009</td>
<td>Philosophic Classics</td>
<td>A</td>
</tr>
<tr>
<td>2010</td>
<td>Organic Chemistry I</td>
<td>A-</td>
<td>2010</td>
<td>Linear Algebra</td>
<td>A-</td>
</tr>
<tr>
<td>2010</td>
<td>Molecules and Cells</td>
<td>A-</td>
<td>2010</td>
<td>Discrete Mathematics</td>
<td>B</td>
</tr>
<tr>
<td>2010</td>
<td>Independent Research</td>
<td>A</td>
<td>2010</td>
<td>BME Design Group</td>
<td>B+</td>
</tr>
<tr>
<td>2011</td>
<td>Statistical Mechanics/Thermo</td>
<td>A-</td>
<td>2010</td>
<td>Differential Equations/Applications</td>
<td>A</td>
</tr>
<tr>
<td>2011</td>
<td>Independent Research</td>
<td>A</td>
<td>2010</td>
<td>Data Structures</td>
<td>B+</td>
</tr>
<tr>
<td>2012</td>
<td>Independent Study - BME</td>
<td>A</td>
<td>2010</td>
<td>Responsible Conduct of Research</td>
<td>Sat</td>
</tr>
<tr>
<td>2013</td>
<td>Senior Design Project I</td>
<td>A</td>
<td>2011</td>
<td>Intro to Probability</td>
<td>A-</td>
</tr>
<tr>
<td>2013</td>
<td>Senior Design Project II</td>
<td>A-</td>
<td>2011</td>
<td>Systems and Controls</td>
<td>B+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2011</td>
<td>Models and Simulations</td>
<td>B+</td>
</tr>
<tr>
<td>2013</td>
<td>Principles of Genetics</td>
<td>B+</td>
<td>2011</td>
<td>System Bioengineering I</td>
<td>B+</td>
</tr>
<tr>
<td>2013</td>
<td>Analysis of the Biological Literature</td>
<td>B+</td>
<td>2011</td>
<td>System Bioengineering Lab I</td>
<td>A</td>
</tr>
<tr>
<td>2013</td>
<td>Quantitative Genomics</td>
<td>B+</td>
<td>2011</td>
<td>Women, Family, &amp; the Body Politic</td>
<td>A-</td>
</tr>
<tr>
<td>2014</td>
<td>Computational and Functional Genomics</td>
<td>A-</td>
<td>2012</td>
<td>Intro to Stochastic Processes</td>
<td>A</td>
</tr>
<tr>
<td>2014</td>
<td>Biological Macromolecules: Structure, Function and Pathways</td>
<td>A-</td>
<td>2012</td>
<td>Intro to Statistics</td>
<td>A</td>
</tr>
<tr>
<td>2014</td>
<td>Introduction to Biomedical Informatics I</td>
<td>A</td>
<td>2012</td>
<td>Systems Bioengineering II</td>
<td>A-</td>
</tr>
<tr>
<td>2014</td>
<td>Conduct of Science</td>
<td>Sat</td>
<td>2012</td>
<td>Systems Bioengineering Lab II</td>
<td>A</td>
</tr>
<tr>
<td>2015</td>
<td>Introduction to Biomedical Informatics II</td>
<td>A</td>
<td>2012</td>
<td>Business Law I</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2012</td>
<td>Computational Molecular Medicine</td>
<td>A-</td>
</tr>
<tr>
<td>YEAR</td>
<td>SCIENCE COURSE TITLE</td>
<td>GRADE</td>
<td>YEAR</td>
<td>OTHER COURSE TITLE</td>
<td>GRADE</td>
</tr>
<tr>
<td>------</td>
<td>------------------------</td>
<td>-------</td>
<td>------</td>
<td>-------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2012</td>
<td>Systems Bioengineering III</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>GRE Quant</td>
<td>163/170</td>
<td>2012</td>
<td>Business Law II</td>
<td>B+</td>
</tr>
<tr>
<td></td>
<td>GRE Verbal</td>
<td>163/170</td>
<td>2013</td>
<td>History of Modern Medicine</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>GRE Writing</td>
<td>5.5/6</td>
<td>2013</td>
<td>Justice and Health</td>
<td>B+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2013</td>
<td>Medical Informatics</td>
<td>A-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2013</td>
<td>Machine Learning in Complex Domains</td>
<td>B-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2013</td>
<td>Law and the Internet</td>
<td>B+</td>
</tr>
<tr>
<td></td>
<td>Harvard University</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistical Inference</td>
<td></td>
<td>2013</td>
<td></td>
<td>B-</td>
</tr>
<tr>
<td></td>
<td>Selected Topics in High Dimensional Analysis</td>
<td>2014</td>
<td>Sat</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayesian Methodology in Biostatistics</td>
<td>2014</td>
<td>B+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Doctoral Dissertation and Other Research Experience:

Summer High School Research Intern  
June 2008 - Oct 2008  
Mentored by Paul Meltzer and Liang Ciao at the National Cancer Institute, National Institutes of Health  
Project: Delineating the Role of BRF2 in Breast Cancer Pathogenesis  
The 8p\textsubscript{11-12} amplicon is a region of genetic amplification present in 10-15\% breast cancers and is associated with poor prognosis. While putative driving oncogenes have been proposed, genes within this amplicon had yet to be definitively implicated in cancer growth, survival, or pathogenesis. I evaluated the role of BRF2, one gene located within this amplicon, in breast cancer growth and development. I performed immunoblot analysis and used previous CGH and expression array studies to establish a strong correlation between BRF2 gene amplification and BRF2 protein overexpression, a trait consistent with an oncogenic role. I employed lentiviral-mediated gene transfer to deliver BRF2 shRNA into breast cancer cells with 8p\textsubscript{11-12} amplification and subsequently established long term stable cell lines with shRNA constructs targeting BRF2, leading to marked reduction of BRF2. Using clonogenic assays, I determined that BRF2 inhibition resulted in impeded growth and proliferation rates as well as increased cell death. My findings suggest that BRF2 is a relevant oncogene in the 8p\textsubscript{11-12} amplicon and may play a role in breast cancer growth and pathogenesis.  

Undergraduate Research Scientist  
Aug 2009 - May 2013  
Mentored by Rachel Karchin at the Institute for Computational Medicine, Johns Hopkins University  
Project: Computational Assessment of the Utility of Limiting Orthologous Sequence Depth in Mutation Impact Prediction Performance  
To predict for the function impact of mutations, current computational models often use sets of orthologous sequences, which are presumed to originate from a common ancestor such that their differences can be attributed to mutation and selective pressures. However, the extent to which these orthologous sequences have been subjected to the same selective pressures and subsequently the validity of using overly distant orthologous sequences remains unknown. I devised a SVM classifier approach as well as implemented published approaches such as SIFT and PolyPhen2 to assess the utility of limiting orthologous sequence depth in mutation impact prediction performance in 33 Mendelian disease-related genes. I developed the feature scores used by the SVM classifier to capture information concerning the physiochemical differences between reference and variant amino acid residues as well as the evolutionary conservation of amino acid residues up to a certain phylogenetic distance depth limit. I measured the overall performance of predictions using standard protocols for statistical learning including calculation of ROC and AUC. My results suggested an orthologous sequence depth limit at the divergence point between vertebrates and invertebrates that may improve mutation impact prediction performance.  
• Fan J, Karchin R. Computational Assessment of the Utility of Limiting Orthologous Sequence Depth in Mutation Impact Prediction Performance. International Congress of Human Genetics/American Society of Human Genetics Conference, Montreal, 2011 (Poster), the BME Undergraduate Research Day, Johns Hopkins University, 2012 (Poster), and Provost’s Undergraduate Research Poster Session, Johns Hopkins University, 2012 (Poster)

Additional projects: Estimating the Phylogenetic Distance Between Target Organisms, Missense Mutation Trends in PIK3CA, Investigation of Pseudogenes as Potential Confounders of Mutation Function Prediction, Critical Assessment of Genome Interpretation - The Personal Genomes Challenge

Summer Undergraduate Research Intern  
June 2012 – August 2012  
Mentored by Shamil Sunyaev for the Harvard-MIT HST and i2b2 BIG Program  
Project: Detecting Synergistic Epistasis in Humans  
The prevalence of sexual reproduction, despite its inherent two-fold cost disadvantage, suggests that sexual reproduction must confer some compensatory evolutionary advantage. The deterministic mutation hypothesis for the evolution of sex posits that such an evolutionary advantage may be achieved contingent on synergistic epistasis, whereby accumulations of deleterious mutations lead to larger decreases in relative fitness. We devised a theoretical test using variance-mean ratios of mutations accumulated since the out-of-Africa migration to detect synergistic epistasis in humans. I applied this test to Genome of the Netherlands (GoNL) data and compared various functional classes of mutations, hypothesizing that variance will be depleted for deleterious mutations but not for benign or neutral mutations. I devised and conducted statistical tests including nonparametric bootstrap, ANOVA, and principal component analysis to assess the significance of results and performed quality control tests to assess for potential batch and flow-cell effects. While detection of synergistic epistasis in humans remains inconclusive, my results did suggest segregation in variance-mean ratios between benign and damaging mutations.
Rotation Student  
**June 2013 - Sept 2013**

**Mentored by Peter Kharchenko at the Center for Biomedical Informatics, Harvard University**

**Project: Transcriptional heterogeneity in mouse neural progenitor cells**

Recent technological advances have revealed tremendous transcriptional heterogeneity among single cells. But how this transcriptional heterogeneity plays a role in cell behavior, fate, and function is still not well understood. We developed PAGODA to resolve multiple, potentially overlapping aspects of transcriptional heterogeneity by identifying known pathways or novel gene sets that show significant excess of coordinated variability among the measured cells. We demonstrate that PAGODA effectively recovers the subpopulations and their corresponding functional characteristics in a variety of single-cell samples, and use it to characterize transcriptional diversity of neuronal progenitors in the developing mouse cortex. Specifically, I contributed to the development of various clustering approaches to identify de novo gene sets that exhibit coordinated variability across cells and ultimately cluster cells into putative subpopulations. Integrating data from the Allen Brain atlas, I also developed an R package to spatially location cells based on their gene expression signatures. Our work resulted in the development of software that can be readily applied to diverse single cell RNA-seq datasets to assess transcriptional heterogeneity.


**Rotation Student  
**Sept 2013 – Nov 2013**

**Mentored by Nir Hacohen and Catherine Wu at the Broad Institute**

**Project: Locally disordered methylation in chronic lymphocytic leukemia**

Intratumoral heterogeneity plays a critical role in tumor evolution. How DNA methylation contributes to this heterogeneity is not well understood. We performed genome-scale bisulfite sequencing of 104 primary chronic lymphocytic leukemias (CLLs) in bulk. We found that, compared with normal B cell samples, CLLs consistently displayed higher intrasample variability of DNA methylation patterns across the genome. I helped perform transcriptome analysis of single CLL cells revealed that methylation disorder was linked to low-level expression.


**Additional projects:** Machine learning algorithm for inferring VDJ-recombination from SNP6-array data

**Graduate Student  
**December 2013 – Present**

**Mentored by Peter Kharchenko at the Center for Biomedical Informatics, Harvard University**

**Project: Precise dissection of genetic and transcriptional heterogeneity in chronic lymphocytic leukemia by single cell analysis**

Intratumoral genetic heterogeneity is the basis of tumor cell plasticity. To more accurately dissect this heterogeneity, detect subclones, define phylogenetic relationships, and to directly uncover genotype-phenotype relationships, we developed a versatile approach based on qPCR for simultaneous targeted mutation and gene expression detection from single cells. I developed the computational methods for mutation calling from raw florescence readouts. We have applied and will continue to apply this method to study chronic lymphocytic leukemia, revealing distinct genetic subclones at different stages of CLL progression.


- Pleiotropic effects of splice variants generated by SF3B1 mutations in chronic lymphocytic leukemia (manuscript in preparation)

**Additional projects:** Transcriptional heterogeneity and similarity in ferret and human neural progenitor cells, Genetic and transcriptional heterogeneity in multiple myeloma, Progenitor origins in chronic lymphocytic leukemia, Mouse model for SF3B1 mutation
Goals for Fellowship Training and Career

My career goal is to be a leading scientist in either an academic research laboratory or comparable industry-based or non-profit research institution in the field of cancer biology. My long-term research interests involve the development of a comprehensive understanding of key genetic, epigenetic, and other regulatory mechanisms driving cellular identity and heterogeneity within cellular groups, tissues, and organs, particularly in the context of cancer and how this heterogeneity shapes tumor progression, therapeutic resistance, and ultimately clinical outcome. I believe that in order to understand this heterogeneity, novel statistical methods with user-friendly computational software must be developed to enable researchers to harness the power of big high-throughput sequencing datasets currently being generated. My goal for the F31 is to gain the skills and knowledge necessary for a lifetime career in the development of statistical methods and computational software for the analysis of high-throughput sequencing cancer data to uncover the biological mechanisms driving cancer development and prognosis. My advisor, Dr. Peter Kharchenko, and I have developed a training plan to work toward the following goals:

Build a strong quantitative and computational skill set: I have taken undergraduate and graduate courses that provide me with the necessary technical background for developing statistical methods and computational software including applied mathematics courses such as stochastic processes and Bayesian inference along with computational courses in data structures and machine learning. Throughout this fellowship, I will continue to learn statistical methods and computational approaches through hands-on research experience, attending conferences and meetings, and reading scientific literature. Specifically through this proposed research, I will learn to formulate problems analytically into a Bayesian framework, build Bayesian models, optimize model fitting, and assess model performance.

Strengthen background in the life sciences: In order to better interpret the biological significance of my computational analyses as well as improve communication with wet lab collaborators, I strive to become more familiar with topics in life sciences disciplines, with particular emphasis on immunology and cancer biology. I have taken undergraduate and graduate courses that have provided me with a strong biological background in molecular biology and genetics. In addition, I will attend the Immunology Seminar Series at Harvard Medical School and continue collaborating closely with Dr. Lili Wang of the Wu lab to strengthen my background in immunology and cancer biology. Dr. Wang has extensive experience as a trained immunologist working with chronic lymphocytic leukemia and her mentorship will provide me with a better understanding of the biological background and interpretation of my computational analyses.

Apply computational skills to gain biological insight from high-throughput sequencing data: I will apply the skills and knowledge acquired from the previous two goals to this proposed research project. This proposed research will also expose me to various sources of technical errors and considerations common to sequencing data that I might encounter again in the future in order to design and test new analysis pipelines as technological advances continue to develop.

Learn to independently lead and manage a scientific project: While my sponsors, collaborators, and fellow lab members will provide advice and suggestions as I complete the aims outlined in this project proposal, I will be in charge of developing and completing the proposed research. The experience I gain from managing this long-term project will be very valuable for my development as an independent researcher.

Manage and mentor students: Throughout my training, I will have the opportunity to mentor summer interns and more junior graduate students. I hope to gain experience managing and mentoring students by providing them with small projects related to the research outlined in this proposal such as applying the developed methods to novel datasets or creating simulations. Through mentoring students, I will develop skills needed to help train the next generation of scientists.

Improve communication and presentation skills: In order to effectively communicate my research with the broader scientific community, I strive to improve my scientific writing and oral presentation skills. In terms of scientific writing, I will continue gaining experience in preparing and submitting manuscripts to peer-reviewed journals. I will be practicing my oral presentation skills by presenting my research in regular lab meetings, program and institutional retreats, and local and national conferences. Furthermore, I will seek the mentorship of Dr. Nils Gehlenborg on proper visualization of my data. Dr. Gehlenborg has published extensively on data visualization and his mentorship will help me improve my graphical data representation and presentation skills.
Selection of Sponsor and Institution:

Selection of Sponsor:

I rotated in Peter Kharchenko’s lab the summer prior to the start of my graduate program. I worked closely with Dr. Kharchenko to build upon his previous method for single cell differential expression, published in Nature Methods, in order to identify cell subpopulations based on gene expression. This was my first introduction to single cell technologies and I was quickly enthralled with the statistical and computational methods behind the single cell analysis and enticed by the opportunities presented in this emerging field. As his first, and currently only graduate student, I receive a lot of personalized attention and hands-on mentorship. For example, Dr. Kharchenko takes time to sit down and debug with me. I have already improved immensely as a programmer simply by observing Dr. Kharchenko and programing with him. Dr. Kharchenko and I have also attended conferences, written grants, and reviewed manuscripts together. I am confident that Dr. Kharchenko will be able to provide the mentorship and guidance I need to accomplish the goals in this proposal.

Selection of Co-Sponsor:

Dr. Kharchenko completed his graduate research in George Church’s lab. I have taken an ethics course with Dr. Church and find him to be an inspiring visionary with ethical character. I believe Dr. Church will provide guidance to both myself and Dr. Kharchenko in the execution of this proposal.

Selection of Institution:

The Bioinformatics and Integrative Genomics (BIG) program at Harvard University is one of the premier places to train as a bioinformatics graduate student. Based at the Center for Biomedical Informatics (CBMI) at Harvard Medical School (HMS), BIG provides interdisciplinary training in biological as well as quantitative methods. HMS offers additional opportunities for training outside the classroom, ranging from seminars to career panels and workshops. The Harvard medical area fosters a collaborative atmosphere between doctors and researchers and provides ample access to high quality datasets.

In addition, HMS provides access to significant computing resources in the shared research cluster (“Orchestra”), which was recently expanded with a three million dollar ARRA grant. The cluster comprises more than 400 Linux compute nodes and 4500 cores, provides access to 10TB of disk space on a high performance Isilon storage cluster for ongoing computational tasks, and allows for multiple job submissions via Platform Computing’s LSF resource management system, thereby allowing for powerful computations to be completed quickly. Multiple core facilities in the Harvard Medical Area including the Broad Institute provide access to next-generation sequencing platforms including the Fluidigm C1 machine for singlecell RNA sequencing.
Equipment:

No experimental equipment is required for the proposed project. All computation and software development will be done on an iMac Desktop with an 3.2 GHz Intel Core i5 Processor, 24 GB 1600 MHz DDR3 memory, and connection to shared computing facilities at HMS.
Facilities and Other Resources:

**Laboratory:** Trainees have access to laboratories located in the Longwood Medical Area (which is comprised of Harvard Medical School (HMS), Harvard School of Public Health, the Dana-Farber Cancer Institute, Boston Children’s Hospital, Beth Israel Deaconess Medical Center, Brigham and Women’s Hospital, CBR Institute for Biomedical Research, Joslin Diabetes Center), the Harvard University Cambridge campus, McLean Hospital, New England Primate Research Center, Harvard Institute of Proteomics, Institute of Chemistry and Cell Biology, Harvard Center of Neurological Diseases, and laboratories at the Massachusetts General Hospital, including the Massachusetts Eye and Ear Infirmary. Major new research facilities include Harvard Medical School’s 430,000 square foot New Research Building (NRB), the 295,000 square foot Karp Research Laboratories at Boston Children’s Hospital, and the 705,000 square foot Center for Life Sciences building. The medical area laboratories are all within easy walking distance of each other. The lab of my advisor Dr. Peter Kharchenko is located within the Center for Biomedical Informatics at HMS. The Center for Biomedical Informatics (CBMI) promotes and facilitates collaborative activities in biomedical informatics among researchers at HMS and its affiliated institutions. Its core faculty members conduct research at the intersection of biomedicine and information sciences, including bioinformatics, functional genomics, translational medicine, and clinical knowledge management. The center also hosts the Bioinformatics & Integrative Genomics program (a graduate training program, sponsored by NHGRI) and the Biomedical Informatics Research Training program (a consortium of informatics laboratories at Harvard and MIT, sponsored by NLM). As such, CBMI provides an excellent environment to pursue computational biology investigations and collaborations with the HMS-affiliated institutions.

**Office:** CBMI has ample office space, including the desk space for all trainees and faculty. The PI has been given a spacious cubicle. The cubicle is fully networked with high-speed ethernet connected to the Harvard Medical School’s backbone network. In addition, two conference rooms are adjacent to the lab, for small group project meetings, full weekly lab meetings, and weekly journal club meetings. Shared departmental administrative space is staffed by departmental support (Sponsored Research, Finance, Academic Appointments, etc.) and outfitted with computers, printers, fax, and photocopier.

**Computer:** WiFi service is broadly available on campus. Students have access to both Macs and PCs that are part of a school-wide network and provide online access to various biomedical databases, MEDLINE databases, and the Internet. The Research Computing Center (RCC) at HMS offers comprehensive bioinformatics training for Harvard students, faculty and staff. In addition, students have computer access to library holdings via such in-house programs as HOLLIS, Countway Plus, etc. HMS provides access to significant computing resources in the shared research cluster ("Orchestra"), which was recently expanded with a three million dollar ARRA grant. The cluster comprises more than 400 Linux compute nodes and 4500 cores. The cluster also provides access to 10TB of disk space on a high-performance Isilon storage cluster for ongoing computational tasks. Jobs are submitted to the shared cluster via Platform Computing’s LSF resource management system. CBMI and HMS IT have a great track record of deploying research web applications and maintaining them over extended periods of time, far exceeding the funding periods.

**Other:** The Division of Medical Sciences offices are housed in the Medical Education Center to provide students with logistical and other support. In addition, each graduate program hosts staff dedicated to student advising. Students also receive supplementary travel funds from DMS to ensure their ability to travel to at least one scientific conference relevant to their area of research interest. The Division makes available staff and financial resources to support this training program. Through DMS, the Longwood Medical Area hosts a number of graduate programs. Hence, a vibrant student community is part of Harvard Medical School’s campus. A wide variety of weekly journal clubs and seminar series provide opportunities for students to interact with peers, post-docs and faculty across disciplines. In addition, students have the opportunity to interact with the general public as teachers and mentors through outreach programs offered by the Office for Diversity Inclusion and Community Partnership.
Activities Planned Under This Award:

Percentage of time to be spent under each category:

<table>
<thead>
<tr>
<th></th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research</strong></td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

The majority of my time will be dedicated to the proposed research project during the funding period of this fellowship. However, I will also engage in the following activities to supplement my training:

All course work and teaching requirements of my PhD program have been fulfilled during my first two years of graduate study. However, I will continue to enroll in the responsible conduct of research training curriculum at Harvard Medical School (HMS).

To stay informed about scientific progress in bioinformatics and biomedicine and to develop my communication and presentation skills, I will be attending a variety of seminars, workshops, and retreats hosted by HMS and other affiliated institutions. My PhD program organizes a work-in-process seminar series in which students from different laboratories can meet and share their research on a regular basis. Regular journal clubs and faculty talks are also available at HMS. I will travel to local and national conferences related to my research, including the ASHG Annual Meeting and the Single Cell Analysis Investigators Meeting.

Timeline of research:

<table>
<thead>
<tr>
<th>Year 1:</th>
<th>Year 2:</th>
<th>Year 3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Gather and process datasets</td>
<td>- Apply hierarchical Bayesian models for CNV and SNV inference to CLL datasets (Aim 2)</td>
<td>- Functional validation of differential gene expression using targeted rt-qPCR and other methods at the discretion of collaborators (Aim 3)</td>
</tr>
<tr>
<td>- Code hierarchical Bayesian models for CNV and SNV inference (Aim 1)</td>
<td>- Benchmark model sensitivity and precision using simulations and artificial mixtures of K562 and GM12878 cells (Aim 1)</td>
<td>- Finish analysis</td>
</tr>
<tr>
<td>- Benchmark model sensitivity and precision using simulations and artificial mixtures of K562 and GM12878 cells (Aim 1)</td>
<td>- Code Bayesian model phylogenetic reconstruction model</td>
<td>- Package up code</td>
</tr>
<tr>
<td>- Code Bayesian model phylogenetic reconstruction model</td>
<td>- Benchmark model against regular hierarchical clustering using simulations and artificial mixtures of K562 and GM12878 cells (Aim 2)</td>
<td>- Create website and supporting features</td>
</tr>
<tr>
<td>- Benchmark model against regular hierarchical clustering using simulations and artificial mixtures of K562 and GM12878 cells (Aim 2)</td>
<td>- Apply hierarchical Bayesian models for CNV and SNV inference to CLL datasets (Aim 2)</td>
<td>- Prepare manuscript</td>
</tr>
<tr>
<td></td>
<td>- Benchmark model against regular hierarchical clustering using simulations and artificial mixtures of K562 and GM12878 cells (Aim 2)</td>
<td>- Submit manuscript</td>
</tr>
<tr>
<td></td>
<td>- Differential expression and gene set enrichment analysis on putative CLL subpopulations (Aim 3)</td>
<td>- Submit software to BioConductor and other code repositories</td>
</tr>
<tr>
<td></td>
<td>- Compare genes and gene sets across patient and across time points (Aim 3)</td>
<td>- Perform additional analysis, revise manuscript, and resubmit as needed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Responsible Conduct of Research:

PLAN FOR INSTRUCTION IN RESPONSIBLE CONDUCT OF RESEARCH (RCR):

Overview: In accordance with NIH guidelines (NOT-OD-10-019), the Division of Medical Sciences (DMS) has produced two courses on RCR that must be taken by all students. The first course occurs their second year, and the second in their fifth/sixth year. In 2011, the RCR curriculum was expanded and revised in concert with Harvard’s Vice Provost for Research. The course’s leader is a member of Harvard’s RCR Working Group.

1. Format: The two required courses: Medical Sciences 300qc and Med Sci 302qc each have two components: didactic; and small group interaction with case studies. Med Sci 300qc is the introductory course taken by 2nd year students and Med Sci 302 qc is the advance course taken by upper level students. The courses are case-based; designed to maximize interaction among students and faculty on matters of responsible scientific practice and ethics. Students prepare case materials and readings in advance of each session and then meet to present and discuss these readings. They utilize extensive resource materials, including articles, essays, prepared example cases and mini-cases. Each discussion group is led by a member of the DMS faculty and is composed of 4–10 students purposely mixed from among the Division’s programs thereby allowing the students to experience a wide range of views. The Med Sci 302 qc course in addition to lectures and small group activities, permits for the advanced students to share their RCR reflections with the 2nd year students, based on their graduate school experience.

2. Subject Matter: The subject matter for the RCR curriculum is consistent with the guidelines in NOT-OD-10-019. Each student is provided with a 90 page Course Guide containing cases and readings organized thematically around the main topics of the RCR curriculum. Additionally, students are required to read the National Academy of Sciences publication On Being a Scientist (3rd Edition) which also covers this curriculum. At the end of the semester, each student will be required to write a case of their own that will be evaluated by the faculty member. During Med Sci 300qc and Med Sci 302qc, students explore a wide range of topics including rules of the scientific method and practice, the use of animals, human trials, writing and publication issues, relationships with colleagues and mentors, fraud and misconduct, and philosophy of science; issues of science in society, including genetic screening, environmental, political, social, and news media issues; and the interface between the scientific community and society, including patents, conflict of interest, animal rights, whistle-blowing, and regulation of research. Med Sci 302qc is particularly geared to address the specific issues that students have confronted during their training. Each year, the RCR faculty will review the curriculum to ensure that all topics have been represented adequately.

3. Faculty Participation: a. 23 professors taught Med Sci last year. Each faculty member facilitates participation of every student in these sessions. Participation in all sessions is mandatory and students who miss a session are required to make up that session by attending another small group session lead by a different faculty member.

4. Duration of Instruction: G2s take Med Sci 300qc, which is comprised of three 1.25hr lectures and six 1.5hr small group meetings over a period of one semester. G5/6s take Med Sci 302qc, which is comprised of three 1.25hr lectures and three 1.5hr small group meetings over a period of one semester. Total: 21hrs.

5. Frequency of Instruction: The two RCR courses encompass the approximately 6 years of graduate training. Since much of RCR deals with aspects relating to research in the laboratory, we feel that the students gain greater benefit from taking the course in their second year. The second RCR course is taken in their fifth/sixth year. It is anticipated that the design of these courses will emphasize the importance placed upon the RCR curriculum, will allow it develop as the students mature scientifically, and will keep it fresh in the minds of the students. I took Med Sci 300qc in the fall of 2014 and I will take Med Sci 302qc in the fall of 2017.

Personal: In addition, in my weekly meetings with my advisor Dr. Peter Kharchenko, we discuss responsible conduct of research including the topics of authorship, ethics, confidentiality, and many more. In addition to documenting all analyses in a digital lab notebook available to the whole lab, I also strive to publish all code to online repositories associated with figures in our published papers.
Respective Contributions:

Project proposal:

I entered the lab of Dr. Peter Kharchenko with a great interest on applying computational approaches to analyze single cell data. Dr. Kharchenko, along with my collaborators in Dr. Catherine Wu’s lab, have proposed several different research directions for me that were based on training potential, interest, importance to the field, and feasibility for completion during my time in graduate school. After careful reading and assessment of available datasets and resources, I chose to design and pursue a research plan to develop statistical methods and computational tools to study subclonal evolution in chronic lymphocytic leukemia, which is the topic of this research proposal.

Dr. Kharchenko has been a very supportive and attentive mentor who has provided me with extensive training and expertise in statistical analysis, method benchmarking, method validation, software development, software documentation, computational parallelization, and various programming languages such as R and C++. He encouraged me to write this proposal. I completed background research and preliminary analyses with input from Dr. Kharchenko and Dr. Wu. I designed the structure and aims of this proposal. I also wrote this research proposal and designed all figures.

Additionally, Dr. George Church, Dr. Lili Wang, and Dr. Nils Gehlenborg provided comments and feedback on this proposal.

I am submitting this research proposal with Dr. Kharchenko, Dr. Church, and Dr. Wu’s review and approval.

Data Sources:

Through collaboration with Dr. Catherine Wu’s group at the Dana-Farber Cancer Research Institute, we were able to obtain the single cell RNA-seq and previous bulk whole exome-seq datasets that will be used in this project. I have met with Dr. Wu to discuss the datasets and will be able to contact her in the future if specific questions about these data arise while carrying out the proposed project.
SUMMARY STATEMENT
(Privileged Communication)

Application Number: 1 F31 CA206236-01

Fan, Jean
President and Fellows of Harvard College
260 Longwood Ave
TMEC 432
Boston, MA 02115-5701

Review Group: ZRG1 F09A-D (20)
Center for Scientific Review Special Emphasis Panel
Fellowships: Oncology

Meeting Date: 11/05/2015
Council: JAN 2016
Requested Start: JAN 2016
PCC: O6TR

Project Title: Computational Analysis of Subclonal Evolution in Chronic Lymphocytic Leukemia

Requested: 3 years

Sponsor: Kharchenko, Peter V
Department: Division of Medical Sciences
Organization: HARVARD MEDICAL SCHOOL
City, State: BOSTON MASSACHUSETTS

SRG Action: Impact Score: 26  Percentile: 19
Human Subjects: 10-No human subjects involved
Animal Subjects: 10-No live vertebrate animals involved for competing appl.
RESUME AND SUMMARY OF DISCUSSION: This F31 fellowship proposal is submitted by a very promising student Jean Fan, who proposes to develop statistical methods and computational software to analyze single cell RNA-seq data derived from CLL patient samples. The exceptional and highly motivated applicant is recipient of many awards, has stellar academic and excellent publication records. The reference letters are extremely laudable, emphasizing high level of her creativity, research commitment and strong leadership drive. Strong computational training environment and resources at Harvard were the additional score-driving strengths. The sponsor is an expert in computational analysis of genomic and epigenetic data sets and has solid funding. Very limited sponsor’s training record is mitigated by the addition of a highly accomplished co-sponsor on the mentoring team. The proposed research plan leverages sponsors’ and collaborators’ funded work and extends it to an important new area. One of the assigned reviewers raised concerns that the proposed aims are entirely phenomenological and do not put the specific hypothesis to any experimental validation test, and that there is only incremental benefit from training perspective for the applicant who already possesses very strong computational skills. During the thorough discussion the panel remained divided and enthusiasm for the proposal ranged from very good to exceptional. The committee voted and concluded that a fellowship award will likely have a high impact on the future scientific career of Ms. Fan as an independent investigator.

DESCRIPTION (provided by applicant): Intratumor genetic and transcriptional heterogeneity is a common feature across diverse cancer types, including, chronic lymphocytic leukemia (CLL). Understanding how this heterogeneity impacts clinical outcome and shapes therapeutic resistance is paramount to improving treatment strategies and enabling more personalized cancer treatments. This research
The proposal will develop statistical methods and computational software to analyze and connect these different aspects of heterogeneity to provide a better understanding of cancer development and progression, using CLL as a primary focus.

**CRITIQUE 1:**

- **Fellowship Applicant:** 1
- **Sponsors, Collaborators, and Consultants:** 1
- **Research Training Plan:** 2
- **Training Potential:** 2
- **Institutional Environment & Commitment to Training:** 1

**Overall Impact/Merit:** This is an exciting proposal to develop new statistical and computational approaches to analyze single cell RNA-seq data. The methods developed by the applicant will be used to determine the presence of SNVs and CNAs, and with those data to reconstruct subclonal architectures for CLL. The proposed research, the applicant's previous training and prior research experience, and the training environment that are afforded by the host institution, sponsors and collaborators will position Ms. Fan for a career as an independent investigator. The applicant's commitment to research and the collaborations already established by the applicant and her sponsors afford a superb training opportunity. The overall impact is high.

1. **Fellowship Applicant**

   **Strengths**
   - Strong background in mathematics and remarkable productivity in ~ 2 years of PhD studies.
   - Impressive GRE scores.
   - Excellent undergraduate research experience.
   - Involvement in STEM and specifically leadership for Harvard Graduate Women in Science and Engineering group.
   - Highly laudatory letters that point to the applicant's potential for an independent research career.

   **Weaknesses**
   - No letter of support from undergraduate laboratory (Rachel Karchin) or other Johns Hopkins University faculty.

2. **Sponsors, Collaborators, and Consultants**

   **Strengths**
   - Outstanding co-sponsors/collaborator team.
   - Dr. Church has mentored graduate students and post-docs who have risen to positions of prominence in the genetics/genomics/bioinformatics community.
   - Dr. Kharchenko, although junior, has a track record in training.
   - Collaboration with Wu laboratory is firmly established.

   **Weaknesses**
Uncertainty as to the role doctoral students play in the Kharchenko laboratory.

3. Research Training Plan:

Strengths

- CLL experimental system well-suited to proposed work to develop novel statistical and bioinformatics tools to better understand heterogeneity and clonal evolution based on single cell expression data.
- Data sets already in existence or will soon be available for methods development and hypothesis testing work proposed (Wu single cell RNA-seq data sets) and as such the proposed research will almost certainly be completed during the remaining doctoral training period.
- The proposed research leverages sponsors' and collaborators' funded work and extends it to an important new area.

Weaknesses

- At times the written proposal has the voice of a PI (mentor), and the applicant's role within the larger research group is at times difficult to discern.

4. Training Potential:

Strengths

- Goals for graduate training and beyond spelled out clearly.
- Attention to strengthening background in life sciences is recognized by applicant as a priority.
- The intention to interact collaboratively with Wu Lab (cancer data sets being investigate) and with the Wang collaboration will ensure 'life sciences' context.

Weaknesses

- It is unclear which seminar series are currently attended and what areas/topics are priority for the applicant as she attends seminars and/or joint lab meetings.

5. Institutional Environment & Commitment to Training:

Strengths

- Outstanding environment.

Weaknesses

- None noted.

Protection for Human Subjects:
Not Applicable (No Human Subjects)

Vertebrate Animals:
Not Applicable (No Vertebrate Animals)

Biohazards:
Not Applicable (No Biohazards)

**Training in the Responsible Conduct of Research:**
Acceptable
Comments on Format (Required):
- Acceptable
Comments on Subject Matter (Required):
- Acceptable
Comments on Faculty Participation (Required):
- Acceptable
Comments on Duration (Required):
- Acceptable
Comments on Frequency (Required):
- Acceptable

**Select Agents:**
Not Applicable (No Select Agents)

**Resource Sharing Plans:**
Acceptable

**Budget and Period of Support:**
Recommend as Requested

**CRITIQUE 2:**

Fellowship Applicant: 1
Sponsors, Collaborators, and Consultants: 3
Research Training Plan: 2
Training Potential: 1
Institutional Environment & Commitment to Training: 1

**Overall Impact/Merit:** The candidate earns the “1” reserved for truly exceptional candidates. The sponsoring team is judged to be outstanding, with proper measures taken to account for the early career stage of the primary sponsor. However, the co-sponsor’s letter and offered level of input came across as weak and distant. The project is exciting and has realistic potential to reveal important insights in CLL. More broadly, the methods established upon completion of the proposed work, when shared freely, have potential to impact our understanding of the clonal evolution of CLL and other malignancies. The training plan will expand Jean’s already impressive and respected repertoire of
statistical methods applied to understanding clonal architecture. The institutional environment and commitment to training are judged as outstanding.

1. Fellowship Applicant

Strengths

- Excellent undergraduate academic record and GRE scores.
- Contributing author on recent, high impact studies. (*Cancer Cell* and *Cancer Discovery*)
- Active in outreach for women scientists. (established a non-profit, cuSTEMized, to encourage girls to envision themselves in science, technology, engineering, and math)

Weaknesses

- None noted.

2. Sponsors, Collaborators, and Consultants

Strengths

- Sponsor and applicant have an established rapport and work style that appears to be highly mutually beneficial.
- Sponsor has NSF and NIH funding through 2020.
- Sponsor is very straight about “I’m new, with limited experience with training – this is my first grad student”.

Weaknesses

- Jean is sponsor’s first graduate student; co-sponsor selection mitigates this already minor concern; lab personnel includes several postdocs.
- Co-sponsor’s letter and offered level of input came across as weak and distant; seemed to be there to prevent criticism of a young sponsor. This is a proposal that could benefit from a co-sponsor bringing some other expertise and with a way to push this young woman’s training. Dr. Church’s credentials are spectacular; no argument there. I raise this issue as respectfully as possible and hope that I worded it non-offensively.

3. Research Training Plan

Strengths

- Cutting edge question at the interface of CLL biology and computational modeling.
- One of the clearest and most professional applications I’ve read in quite some time (clerical error in first sentence of Abstract had me worried, but it was one of the only blips: “…, including.”) Also, please see comment below in this section’s “Weaknesses”.
- Release of findings/methods as open source is a valuable deliverable.

Weaknesses

- Please re-visit your use of the word “metastatic” when referring to leukemia; I presume you mean “relapsed” or “advanced”. This set off a bit of a firestorm that had to be extinguished.

4. Training Potential
Strengths

- The candidate is already publishing her work in Nature Methods and other very high-tier journals; the best move seems to be to let her and her advisor decide on what they consider the best further training, and this project suits the bill.

- Opportunity for applicant to train and involve junior colleagues and summer students is in line with her passion and advocacy for women in science.

Weaknesses

- I found myself wanting to suggest linking the computational work to a greater investment by Jean in “wet lab” based study. That may be asking too much, but if the applicant wants to run her own lab and be as broadly prepared as possible (and can make the time), I do think this would be worth her time. Of course, the counter argument is “Why don’t all of the experimentalists learn to do computational analysis?” I think they should, at least to a level that helps drive project and career.

- Relatedly, I worry a little that Jean could be viewed as being trained toward almost a super data analyst – great team member - rather than a principal investigator. Her “plans” are a bit vague. I suggest that she and sponsor keep an open dialogue on this issue. The analogy that comes to mind is x-ray crystallographers – some do great work but wait for projects to come, while the very best (Eric Gouaux, for example) drive their fields and even pioneer new fields. Maybe you have that kind of talent…

5. Institutional Environment & Commitment to Training

Strengths

- Appropriate, supportive institutional environment and demonstrated commitment to training.

- Clear plan for Jean’s training within the program elaborated in application.

- Detailed (highly positive) comments regarding Jean’s preliminary qualifying exam demonstrate a deep commitment by the faculty.

Weaknesses

- None noted.

Protections for Human Subjects:
Not Applicable (No Human Subjects)

Vertebrate Animals:
Not Applicable (No Vertebrate Animals)

Biohazards:
Not Applicable (No Biohazards)

Training in the Responsible Conduct of Research:
Acceptable

Comments on Format (Required):
adequate

Comments on Subject Matter (Required):
- adequate

Comments on Faculty Participation (Required):
- adequate

Comments on Duration (Required):
- adequate

Comments on Frequency (Required):
- adequate

Select Agents:
Not Applicable (No Select Agents)

Resource Sharing Plans:
Acceptable

Budget and Period of Support:
Recommend as Requested

Additional Comments to Applicant (Optional):
- You are doing really impressive things at an early career stage.

CRITIQUE 3:

Fellowship Applicant: 2
Sponsors, Collaborators, and Consultants: 4
Research Training Plan: 4
Training Potential: 2
Institutional Environment & Commitment to Training: 1

Overall Impact/Merit: Strong applicant with an expert mentor. Lack of training history and questions about the role of co-mentors and independence of primary sponsor reduce enthusiasm. The proposal tackles significant technical data analysis problems, but the clinical importance and potential impact of the proposed studies is questioned.

1. Fellowship Applicant

Strengths
- Top notch student.
- Track record of publishing productivity.
STEM outreach and other activities show evidence of balance and potential leadership.

Weaknesses
- None noted.

2. Sponsors, Collaborators, and Consultants

Strengths
- Dr. Kharchenko is an expert in computational analysis of genomic, expression, and epigenetic data sets.
- The sponsor was trained in world leading laboratories.
- Assuming that adequate, dedicated time is spent with the applicant, the involvement of Dr. Church is a significant strength.
- Numerous opportunities for mentorship or training are noted, many of which are informal and attest to a collaborative environment.

Weaknesses
- Unclear as to the specific role for the co-mentor. One stated rationale is that the applicant finds Dr. Church “to be an inspiring visionary with ethical character”.
- Given that Dr. Kharchenko was trained in the Church lab, this raises some concern about independence of the primary sponsor.
- Primary sponsor has a very limited training record.

3. Research Training Plan

Strengths
- Hypothesis regarding roles of Wnt and Notch signaling is potentially important.
- Leverages existing data sets.
- Development of new computational methods could have a significant enabling effect on the research community.

Weaknesses
- The proposed aims are entirely phenomenological and do not put the specific hypothesis to any critical experimental test.
- Although intratumor heterogeneity is an interesting general question, it is not at all clear that intraclonal heterogeneity is a clinically significant feature in CLL, and thus the overall impact of this proposal is questionable.
- From a clinical perspective, it is not obvious that there is a pressing clinical need given the recent success of BTK inhibitors (e.g. ibrutinib) in treating CLL subsets.

4. Training Potential

Strengths
- Strong computational training environment and program.

Weaknesses
Experimental context and specific mentorship is lacking.

5. Institutional Environment & Commitment to Training

Strengths
- Excellent training programs, resources, and environment.

Weaknesses
- None noted.

Protections for Human Subjects:
Not Applicable (No Human Subjects)

Vertebrate Animals:
Not Applicable (No Vertebrate Animals)

Biohazards:
Not Applicable (No Biohazards)

Training in the Responsible Conduct of Research:
Acceptable
Comments on Format (Required):
- structured courses

Comments on Subject Matter (Required):
- broad, standard

Comments on Faculty Participation (Required):
- widespread (23 professors)

Comments on Duration (Required):
- multiple 1.25- 1.5 hour sessions in structured or small group format

Comments on Frequency (Required):
- throughout 6 yr program of study

Select Agents:
Not Applicable (No Select Agents)

Resource Sharing Plans:
Acceptable

Budget and Period of Support:
Recommend as Requested
THE FOLLOWING SECTIONS WERE PREPARED BY THE SCIENTIFIC REVIEW OFFICER TO SUMMARIZE THE OUTCOME OF DISCUSSIONS OF THE REVIEW COMMITTEE, OR REVIEWERS' WRITTEN CRITIQUES, ON THE FOLLOWING ISSUES:

COMMITTEE BUDGET RECOMMENDATIONS: The budget was recommended as requested.

NIH has modified its policy regarding the receipt of resubmissions (amended applications). See Guide Notice NOT-OD-14-074 at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-074.html. The impact/priority score is calculated after discussion of an application by averaging the overall scores (1-9) given by all voting reviewers on the committee and multiplying by 10. The criterion scores are submitted prior to the meeting by the individual reviewers assigned to an application, and are not discussed specifically at the review meeting or calculated into the overall impact score. Some applications also receive a percentile ranking. For details on the review process, see http://grants.nih.gov/grants/peer_review_process.htm#scoring.
MEETING ROSTER

Center for Scientific Review Special Emphasis Panel
CENTER FOR SCIENTIFIC REVIEW
Fellowships: Oncology
ZRG1 F09A-D (20) L
November 05, 2015 - November 06, 2015

CHAIRPERSON
RIEGEL, ANNA TATE, PHD
PROFESSOR
DEPARTMENTS OF ONCOLOGY AND PHARMACOLOGY
GEORGETOWN UNIVERSITY
WASHINGTON, DC 20007

MEMBERS
BLACK, JENNIFER D., PHD
PROFESSOR
DEPARTMENT OF GENETICS, CELL BIOLOGY, AND ANATOMY
EPPLEY INSTITUTE FOR RESEARCH IN CANCER AND ALLIED DISEASES
UNIVERSITY OF NEBRASKA MEDICAL CENTER
OMAHA, NE 68198

CARRIER, FRANCE , PHD
ASSOCIATE PROFESSOR
DEPARTMENT OF RADIATION ONCOLOGY
UNIVERSITY OF MARYLAND, BALTIMORE
BALTIMORE, MD 21201

DOEBELE, ROBERT C, MD, PHD
ASSOCIATE PROFESSOR
DIVISION OF MEDICAL ONCOLOGY
UNIVERSITY OF COLORADO
ANSCHUTZ MEDICAL CAMPUS
AURORA, CO 80045

EKLUND, ELIZABETH ANN, MD
PROFESSOR
DIVISION OF HEMATOLOGY AND ONCOLOGY
FEINBERG SCHOOL OF MEDICINE
NORTHWESTERN UNIVERSITY
CHICAGO, IL 60611

GERMAIN, DORIS A, PHD
ASSOCIATE PROFESSOR
DEPARTMENT OF MEDICINE
TISCH CANCER INSTITUTE
MOUNT SINAI SCHOOL OF MEDICINE
NEW YORK, NY 10029

GEWIRTZ, DAVID A., PHD
PROFESSOR
DEPARTMENT OF PHARMACOLOGY AND TOXICOLOGY
VIRGINIA COMMONWEALTH UNIVERSITY
RICHMOND, VA 23298

GMEINER, WILLIAM H., PHD
PROFESSOR
DEPARTMENT OF CANCER BIOLOGY
SCHOOL OF MEDICINE
WAKE FOREST UNIVERSITY
WINSTON-SALEM, NC 27157

GOODFELLOW, PAUL JOSEPH, PHD
PROFESSOR
COLLEGE OF MEDICINE
OHIO STATE UNIVERSITY COMPREHENSIVE CANCER CENTER
THE OHIO STATE UNIVERSITY MEDICAL CENTER
COLUMBUS, OH 43210

JANZ, SIEGFRIED , MD, DSC
PROFESSOR
DEPARTMENT OF PATHOLOGY
CARVER COLLEGE OF MEDICINE
UNIVERSITY OF IOWA
IOWA CITY, IA 52242

KLINGE, CAROLYN M., PHD
PROFESSOR
DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY
UNIVERSITY OF LOUISVILLE SCHOOL OF MEDICINE
LOUISVILLE, KY 40202

KOGAN, SCOTT C, MD
PROFESSOR
DEPARTMENT OF LABORATORY MEDICINE
HELEN DILLER FAMILY COMPREHENSIVE CANCER CENTER
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
SAN FRANCISCO, CA 94143

KROLEWSKI, JOHN J., MD, PHD
PROFESSOR
CENTER FOR PERSONALIZED MEDICINE
DEPARTMENT OF CANCER GENETICS
ROSWELL PARK CANCER INSTITUTE
BUFFALO, NY 14263

MATTINGLY, RAYMOND R, PHD
PROFESSOR
DEPARTMENTS OF PHARMACOLOGY AND ONCOLOGY
WAYNE STATE UNIVERSITY SCHOOL OF MEDICINE
DETROIT, MI 48201

MILLS, ALEA A, PHD
PROFESSOR
DEPARTMENT OF MOLECULAR AND CANCER GENETICS
COLD SPRING HARBOR LABORATORY
COLD SPRING HARBOR, NY 11724

MILLS, KEVIN D, PHD
ASSOCIATE PROFESSOR
THE JACKSON LABORATORY
BAR HARBOR, ME 04609

MULDER, KATHLEEN M, PHD
PROFESSOR
DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY
COLLEGE OF MEDICINE
PENN STATE UNIVERSITY
HERSHEY, PA 17033
Consultants are required to absent themselves from the room during the review of any application if their presence would constitute or appear to constitute a conflict of interest.