SF 424 (R&R) APPLICATION FOR FEDERAL ASSISTANCE **Page 2**

14. PROJECT DIRECTOR/PRINCIPAL INVESTIGATOR CONTACT INFORMATION

Prefix: First Name*: Jean Middle Name: Last Name*: Fan Suffix:

Position/Title:	Graduate Student
Organization Name*:	Presidents and Fellows of Harvard College
Department:	Biomedical Informatics
Division:	
Street1*:	10 Shattuck Street #3
Street2:	Countway 336B
City*:	Boston
County:	
State*:	MA: Massachusetts
Province:	
Country*:	USA: UNITED STATES
ZIP / Postal Code*:	02115-0000

Phone Number*: 2403807477 Fax Number: Email*: jeanfan@fas.harvard.edu

15. ESTIMATED PROJECT FUNDING

a. Total Federal Funds Requested*	$380,044.00
b. Total Non-Federal Funds*	$0.00
c. Total Federal & Non-Federal Funds*	$380,044.00
d. Estimated Program Income*	$0.00

16.IS APPLICATION SUBJECT TO REVIEW BY STATE EXECUTIVE ORDER 12372 PROCESS?*

a. YES ○ THIS PREAPPLICATION/APPLICATION WAS MADE AVAILABLE TO THE STATE EXECUTIVE ORDER 12372 PROCESS FOR REVIEW ON:

DATE:

b. NO ● PROGRAM IS NOT COVERED BY E.O. 12372; OR

○ PROGRAM HAS NOT BEEN SELECTED BY STATE FOR REVIEW

17. By signing this application, I certify (1) to the statements contained in the list of certifications* and (2) that the statements herein are true, complete and accurate to the best of my knowledge. I also provide the required assurances * and agree to comply with any resulting terms if I accept an award. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties. (U.S. Code, Title 18, Section 1001)

● I agree*

** The list of certifications and assurances, or an Internet site where you may obtain this list, is contained in the announcement or agency specific instructions.*

18. SFLLL or OTHER EXPLANATORY DOCUMENTATION File Name:

19. AUTHORIZED REPRESENTATIVE

Prefix: First Name*: Jaclyn Middle Name: Last Name*: Lucas Suffix:

Position/Title*:	Sponsored Programs Officer
Organization Name*:	President and Fellows of Harvard College
Department:	Office for Sponsored Programs
Division:	
Street1*:	1033 Massachusetts Avenue, 5th Floor
Street2:	
City*:	Cambridge
County:	
State*:	MA: Massachusetts
Province:	
Country*:	USA: UNITED STATES
ZIP / Postal Code*:	021385369

Phone Number*: 617-495-1664 Fax Number: 617-496-2524 Email*: jaclyn_lucas@harvard.edu

Signature of Authorized Representative*	**Date Signed***
Completed on submission to Grants.gov	01/12/2017

20. PRE-APPLICATION File Name:

21. COVER LETTER ATTACHMENT File Name:Fan_Cover_Letter.pdf

424 R&R and PHS-398 Specific
Table Of Contents

OMB Number: 4040-0010
Expiration Date: 10/31/2019

Project/Performance Site Location(s)

Project/Performance Site Primary Location

○ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:	President and Fellows of Harvard College
Duns Number:	0823596910000
Street1*:	HARVARD UNIVERSITY
Street2:	Office for Sponsored Programs, 1033 Mass. Ave.
City*:	CAMBRIDGE
County:	
State*:	MA: Massachusetts
Province:	
Country*:	USA: UNITED STATES
Zip / Postal Code*:	021385369
Project/Performance Site Congressional District*:	MA-005

Project/Performance Site Location 1

○ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:	President and Fellows of Harvard College
DUNS Number:	0470063790000
Street1*:	Department of Biomedical Informatics
Street2:	10 Shattuck Street #3, Countway 336B
City*:	Boston
County:	
State*:	MA: Massachusetts
Province:	
Country*:	USA: UNITED STATES
Zip / Postal Code*:	021150000
Project/Performance Site Congressional District*:	MA-007

Additional Location(s) File Name:

OMB Number: 4040-0001
Expiration Date: 10/31/2019

RESEARCH & RELATED Other Project Information

1. Are Human Subjects Involved?* ● **Yes** ○ **No**

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? ○ Yes ● No

If YES, check appropriate exemption number: __ 1 __ 2 __ 3 __ 4 __ 5 __ 6

If NO, is the IRB review Pending? ● **Yes** ○ **No**

IRB Approval Date:

Human Subject Assurance Number 00004837

2. Are Vertebrate Animals Used?* ○ **Yes** ● **No**

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? ○ Yes ○ No

IACUC Approval Date:

Animal Welfare Assurance Number

3. Is proprietary/privileged information included in the application?* ○ **Yes** ● **No**

4.a. Does this project have an actual or potential impact - positive or negative - on the environment?* ○ **Yes** ● **No**

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? ○ Yes ○ No

4.d. If yes, please explain:

5. Is the research performance site designated, or eligible to be designated, as a historic place?* ○ **Yes** ● **No**

5.a. If yes, please explain:

6. Does this project involve activities outside the United States or partnership with international collaborators?* ○ **Yes** ● **No**

6.a. If yes, identify countries:

6.b. Optional Explanation:

	Filename
7. Project Summary/Abstract*	Fan_Project_Summary.pdf
8. Project Narrative*	Fan_Project_Narrative.pdf
9. Bibliography & References Cited	Fan_Bibliography_Cited.pdf
10.Facilities & Other Resources	Fan_Facilities_and_Resources.pdf
11.Equipment	Fan_Equipment.pdf
12. Other Attachments	Fan_Nomination.pdf

Project Summary

Chronic lymphocytic leukemia (CLL) is a cancer that exhibits genetic and transcriptional heterogeneity along with a highly variable disease course among patients that remains poorly understood. Previous research has highlighted vast inter- and intra-patient genetic heterogeneity, with subclonal evolution commonly occurring in treatment settings leading to therapeutic resistance and relapse in many cases. In addition, our understanding of the role of co-existing non-cancer cells in the tumor-microenvironment remains limited. Therefore, characterization of these subclonal populations and their corresponding microenvironment will be paramount to enabling precision medicine and synergistic treatment combinations that target subclonal drivers and eliminate aggressive subpopulations thereby improving clinical outcome. In order to accurately dissect the genetic landscape and reconstruct the underlying subclonal architecture in CLL, measurements must be made on the single cell level.

In the F99-phase of this proposed research, Jean Fan will continue developing statistical methods and computational software to analyze single cell RNA-seq data derived from CLL patient samples. Specifically, Jean 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. Jean will then reconstruct the genetic subclonal architecture and characterize the gene expression profiles of identified subclonal populations.

In the K00-phase of this proposed research, Jean will characterize heterogeneity in the tumor-microenvironment and develop methods to assess potential reciprocal interactions between subclones and their microenvironment over time in response to therapy.

The proposed work will yield innovative statistical methods to enable the identification and characterization of subpopulations in cancer and yield open-source 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.

Project Narrative

Intratumor genetic and transcriptional heterogeneity is a common feature across diverse cancer types, including chronic lymphocytic leukemia (CLL). Understanding how heterogeneity both within CLL cells and in relation to the composition and state of co-existing non-CLL cells in the micro-environment impacts clinical outcome and shapes therapeutic resistance is paramount to improving treatment strategies and enabling more personalized cancer treatments. This research 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.

Bibliography Cited

1. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. New England Journal of Medicine. 2012 [accessed 2017 Jan 25];366(10):883–892. http://www.nejm.org/doi/abs/10.1056/NEJMoa1113205
2. Wu CJ. CLL clonal heterogeneity: an ecology of competing subpopulations. Blood. 2012;120(20).
3. Mroz EA, Tward AM, Hammon RJ, Ren Y, Rocco JW. Intra-tumor Genetic Heterogeneity and Mortality in Head and Neck Cancer: Analysis of Data from The Cancer Genome Atlas. PLoS medicine. 2015 [accessed 2015 Feb 11];12(2):e1001786. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001786
4. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481(7382):506–510.
5. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012 [accessed 2017 Jan 25];486(7403):395. http://www.nature.com/doifinder/10.1038/nature10933
6. Brennecke P, Anders S, Kim JK, Ko\lodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, et al. Accounting for technical noise in single-cell {RNA-seq} experiments.
7. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nature methods. 2014 [accessed 2014 Apr 28];(October 2013). http://www.ncbi.nlm.nih.gov/pubmed/24747814
8. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nature medicine. 2013 [accessed 2016 Aug 17];19(11):1423–37. http://www.ncbi.nlm.nih.gov/pubmed/24202395
9. Swann JB, Smyth MJ. Immune surveillance of tumors. The Journal of clinical investigation. 2007 [accessed 2017 Jan 31];117(5):1137–46. http://www.ncbi.nlm.nih.gov/pubmed/17476343
10. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 [accessed 2016 Aug 17];152(4):714–26. http://www.ncbi.nlm.nih.gov/pubmed/23415222
11. Schuh A, Becq J, Humphray S, Alexa A, Burns A, Clifford R, Feller SM, Grocock R, Henderson S, Khrebtukova I, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. Blood. 2012;120(20).
12. Burger JA, Landau DA, Taylor-Weiner A, Bozic I, Zhang H, Sarosiek K, Wang L, Stewart C, Fan J, Hoellenriegel J, et al. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. Nature communications. 2016 [accessed 2016 Aug 17];7:11589. http://www.ncbi.nlm.nih.gov/pubmed/27199251
13. Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, Stevenson K, Sougnez C, Wang L, Li S, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer cell. 2014 [accessed 2017 Jan 25];26(6):813–25. http://www.ncbi.nlm.nih.gov/pubmed/25490447
14. Ramsay AG, Johnson AJ, Lee AM, Gorgün G, Le Dieu R, Blum W, Byrd JC, Gribben JG. Chronic lymphocytic leukemia T cells show impaired immunological synapse formation that can be reversed with an immunomodulating drug. Journal of Clinical Investigation. 2008 [accessed 2017 Jan 31];118(7):2427–37. http://www.ncbi.nlm.nih.gov/pubmed/18551193
15. Forconi F, Moss P. Perturbation of the normal immune system in patients with CLL. Blood. 2015;126(5).
16. Kawaguchi A, Ikawa T, Kasukawa T, Ueda HR, Kurimoto K, Saitou M, Matsuzaki F. Single-cell gene profiling defines differential progenitor subclasses in mammalian neurogenesis. Development (Cambridge, England). 2008 [accessed 2014 May 21];135(18):3113–24. http://www.ncbi.nlm.nih.gov/pubmed/18725516
17. Ma C, Fan R, Ahmad H, Shi Q, Comin-Anduix B, Chodon T, Koya RC, Liu C-C, Kwong GA, Radu CG, et al. A clinical microchip for evaluation of single immune cells reveals high functional heterogeneity in phenotypically similar T cells. Nature Medicine. 2011 [accessed 2017 Jan 25];17(6):738–743. http://www.ncbi.nlm.nih.gov/pubmed/21602800

18. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan J-B, Zhang K, Chun J, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nature methods. 2016 [accessed 2016 Aug 17];13(3):241–4. http://www.ncbi.nlm.nih.gov/pubmed/26780092
19. Kharchenko P V, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature methods. 2014;11(7):740–2.
20. Zhang X, Chen MH, Wu X, Kodani A, Fan J, Doan R, Ozawa M, Ma J, Yoshida N, Reiter JF, et al. Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell. 2016 [accessed 2017 Jan 25];166(5):1147–1162.e15. http://linkinghub.elsevier.com/retrieve/pii/S0092867416309321
21. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science (New York, N.Y.). 2014 [accessed 2016 Sep 7];343(6167):193–6. http://www.ncbi.nlm.nih.gov/pubmed/24408435
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010 [accessed 2016 Sep 7];20(9):1297–303. http://www.ncbi.nlm.nih.gov/pubmed/20644199
23. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013 [accessed 2017 Jan 25];31(3):213–219. http://www.nature.com/doifinder/10.1038/nbt.2514
24. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics (Oxford, England). 2012 [accessed 2016 Aug 17];28(3):423–5. http://www.ncbi.nlm.nih.gov/pubmed/22155870
25. Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics (Oxford, England). 2011 [accessed 2016 Aug 18];27(2):268–9. http://www.ncbi.nlm.nih.gov/pubmed/21081509
26. Exome Aggregation Consortium.
27. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature Biotechnology. 2012 [accessed 2017 Jan 25];30(5):413–421. http://www.nature.com/doifinder/10.1038/nbt.2203
28. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed B V, Curry WT, Martuza RL, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (New York, N.Y.). 2014 [accessed 2014 Jul 9];344(6190):1396–401. http://www.ncbi.nlm.nih.gov/pubmed/24925914
29. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016 [accessed 2017 Jan 25];539(7628):309–313. http://www.nature.com/doifinder/10.1038/nature20123
30. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005 [accessed 2016 Aug 17];102(43):15545–50. http://www.ncbi.nlm.nih.gov/pubmed/16199517
31. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. The New England journal of medicine. 2011;365(26):2497–506.
32. Irish JM, Myklebust JH, Alizadeh AA, Houot R, Sharman JP, Czerwinski DK, Nolan GP, Levy R. B-cell signaling networks reveal a negative prognostic human lymphoma cell subset that emerges during tumor progression. Proceedings of the National Academy of Sciences. 2010 [accessed 2017 Jan 25];107(29):12747–12754. http://www.ncbi.nlm.nih.gov/pubmed/20543139
33. Wang L, Shalek AK, Lawrence M, Ding R, Gaublomme JT, Pochet N, Stojanov P, Sougnez C, Shukla SA, Stevenson KE, et al. Somatic mutation as a mechanism of Wnt/β-catenin pathway activation in

{CLL}. Blood. 2014;124(7):1089–1098.

34. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015;161(5):1187–1201.
35. Kharchenko P V, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature methods. 2014;11(7):740–2.

Facilities and Other Resources

Department of Biomedical Informatics at Harvard Medical School

Located on the third and fourth floors of the Francis A. Countway Library of Medicine, one of the largest medical and health libraries in the United States, the Department of Biomedical Informatics (DBMI) promotes and facilitates collaborative activities in biomedical informatics among researchers at Harvard Medical School (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. DBMI 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). As such, DBMI provides an excellent environment to pursue computational biology investigations and collaborations with the HMS-affiliated institutions.

Office

The Department of Biomedical Informatics is located within the Countway Library at 10 Shattuck Street on the Harvard Medical School campus in Boston, Massachusetts. The Countway is housed in a building of 139,782 square feet, of which 114,446 square feet are occupied by the library. The Department of Biomedical Informatics takes up 25,336 square feet of this total area, including a large, newly-renovated section comprising five shared offices, four private offices, more than 40 cubicles, and several study areas. The building is situated prominently on the campus of the Harvard Medical Area, where the Harvard Medical, Dental, and Public Health Schools are located. This location also places the DBMI in very close proximity to several of the affiliated teaching hospitals, including Brigham and Women's, Children's, and Beth Israel Deaconess, as well as the adjacent School of Public Health. Photocopiers, scanners, computers, printers (color and B&W), telephones, and fax machines are available for staff use, in addition to the other standard office equipment. The library building houses a studio with production equipment (a NewTek TriCaster, Digital Rapids encoders, Yamaha 01V96i mixer, and Cisco TelePresence SX-80) and lighting designed for video and audio recording, and large and small meeting rooms equipped with state-of-the-art technology for web-conferencing, presentations, and collaborative work.

In particular, two newly-renovated fourth floor conference rooms offer a Cisco Video Conferencing system, including wireless PC connection, SX80 Codec, Cisco Touch 10 control panel, and two Cisco SpeakerTrack 60 HD cameras featuring voice recognition and facial recognition. In addition, both rooms offer 80-90" LED flat panel monitors, a SMART PODIUM allowing for annotation and other interactive capability, and ceiling mounted microphones for video conferencing and audio capture.

Computational Infrastructure

Fully loaded computers, including PC and Macintosh, are available to the PIs and their groups from the DBMI. PI computers typically have i7 processors, 16 GB RAM, and 512 GB or 1 TB Solid State Drives. Laptops for administrators and other staff are typically similar to Dell Latitude E7240 or Apple MacBook Air/Pro, with Windows 7/Mac OSX, i5 processors, 8 GB RAM, and 256 GB Solid State Drives. Desktops are similar to Dell Optiplex 9020 or Apple iMac 21.5", with Windows 7/Mac OSX, i5 processors, 8 GB RAM, and 500GB or 1 TB Hard Drives. All have high-speed access to the Internet and e-mail. The entire building also provides wireless access.

All personal computers are networked and well-supported by the Harvard Medical School information technology division.

DBMI itself also has a dedicated technology with expertise in cloud infrastructure, such as various Software-as-a-service (SaaS) and the dominant Infrastructure-As-A-Service (IaaS) -- Amazon Web Services (AWS).

Elastic Compute Resources/AWS "Cluster"

DBMI and Harvard University have business associate agreements in place with Amazon, Inc., for HIPAA-compliant and on-demand compute resources.

At this time DBMI mostly uses AWS as its on-demand infrastructure, including an instance (Undiagnosed Diseases Network) developed for a project requiring FISMA moderate certification. We use a thorough and complex DevOps workflow to spin up infrastructure and secure it on demand. Because we do everything in code, we are able to ensure consistency and testability in our environment. AWS provides us the ability to scale down to tiny amounts of infrastructure and up to huge amounts of infrastructure.

We have pre-existing workflows in our DevOps for:

- Large scale web-services (servicing thousands of simultaneous users)
- Large scale genomics (GATK Best-practices to thousands of simultaneous machines)
- Serving and uploading very large files (up to 1 TB in size) securely and efficiently
- Starting and maintaining i2b2/TransMart stacks and their wrapper services

Any of the above workflows can be spun up (and kept updated) in minutes of work.

All of these services log to environments like Splunk and CloudCheckr to maintain compliance and have an offsite audit trail (and alerting).

We are able to maintain thousands of simultaneous machines using this technology with a single DevOps employee. By using code to maintain everything, scaling is as simple as code, not labor.

DBMI 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.

Research and Security Compliance

The HMS DBMI compliance function consists of the combined compliance capacity of Harvard Medical School and Harvard University while drawing upon the compliance resources of the Harvard affiliated teaching hospitals. Scott Edmiston, Director of Research Privacy and Security Compliance, oversees compliance activities for DBMI in collaboration with internal DBMI staff trained in implementation of leading privacy and security frameworks. He is a licensed attorney and holds CIPP/US and CITI human subjects protections certifications. Harvard Medical School and Harvard University maintain comprehensive in-house counsel departments, IT departments, technology transfer specialists, and dedicated offices of research compliance totaling more than 200 FTEs that are routinely used to fulfill regulatory, ethical and legal obligations associated with large multi-site biomedical informatics research projects. HMS also operates a coordinating center that oversees the authentication and authorization of users in complex data sharing networks through Amazon cloud services and other third party computing tools.

DBMI's David Bernick, Director of Technology, holds a CISSP certification for security.

Administrative Support

DBMI has full time expert, dedicated grants administration, project management, and operations staffs that support our many and varied research projects and report to the Executive Director of the Department, who brings over 30 years experience managing and supporting large, complex research initiatives.

Kharchenko Lab

In addition to the computational resources provided by the HMS and DBMI, Kharchenko lab has three computational servers, with 24-36 processing cores each, 500-1000GB of RAM per server, and total of 250TB of replicated storage space. These servers are used for algorithm development and data analysis. The lab also has access to a full stall of wetlab space and all equipment necessary for carrying out droplet microfluidics experiments, including indrop measurements.

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.

Harvard University
The GRADUATE SCHOOL *of* ARTS AND SCIENCES
OFFICE OF THE DEAN

February 8, 2017

To Whom It May Concern:

I write in my capacity as Dean of the Harvard Graduate School of Arts and Sciences (GSAS) to enthusiastically nominate Ms. Jean Fan for the National Cancer Institute F99/K00 award under the primary sponsorship of Prof. Peter Kharchenko and confirm her eligibility for this award. Ms. Fan is currently a student in GSAS enrolled in the Bioinformatics and Genomics PhD program in the Division of Medical Sciences based at Harvard Medical School. She is a U.S. citizen and is in good standing.

While the specific training requirements are determined by each program, GSAS is committed to ensuring that our PhD students have the resources needed to successfully complete their doctoral work and gain the critical thinking, oral and written communication skills, and other professional development that is important for their success during their PhD training and beyond. To this end all GSAS students have access to a number of services focused on their academic and professional development including the Center for Writing and the Communication of Ideas, the Derek C. Bok Center for Teaching and Learning, and the Office of Career Services. Ms. Fan has full access to all of these resources, in addition to those provided by the Division of Medical Sciences and thus is well poised to achieve her career goals.

With regards to our selection process, a request for nominations was sent out to departments and programs across GSAS whose student's are doing research broadly aligned with NCI's mission. Nominations were reviewed by a faculty committee with expertise in the area of cancer biology. Ms. Fan was unanimously ranked as the top candidate across Harvard to nominate for this competition. She was selected based on her already outstanding publication record, her commitment to research, the breadth of her research experience in both computational and experimental approaches, and her strong interest in pursuing a career in cancer research. In addition, her student leadership and STEM outreach also impressed the committee. Obtaining this F99/K00 award will help support this outstanding young scientist as she completes her PhD work and in the transition towards independence as she goes on to pursue her post-doctoral training.

Sincerely,

Xiao-Li Meng, Ph.D.
Dean, Graduate School of Arts and Sciences
Whipple V.N. Jones Professor of Statistics
Harvard University

Office for Sponsored Programs
Harvard University

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

PROFILE - Project Director/Principal Investigator

Prefix: First Name*: Jean Middle Name Last Name*: Fan Suffix:

Position/Title*: Graduate Student
Organization Name*: Presidents and Fellows of Harvard College
Department: Biomedical Informatics
Division:
Street1*: 10 Shattuck Street #3
Street2: Countway 336B
City*: Boston
County:
State*: MA: Massachusetts
Province:
Country*: USA: UNITED STATES
Zip / Postal Code*: 02115-0000

Phone Number*: 2403807477 Fax Number:

E-Mail*: jeanfan@fas.harvard.edu

Credential, e.g., agency login: JEAN_FAN

Project Role*: PD/PI Other Project Role Category:

Degree Type: BS Degree Year: 2013

Attach Biographical Sketch*: File Name: Fan_Biosketch.pdf

Attach Current & Pending Support: File Name:

PROFILE - Senior/Key Person

Prefix: Prof. First Name*: PETER Middle Name Vasili Last Name*: KHARCHENKO Suffix: Ph.D

Position/Title*: Assistant Professor
Organization Name*: President and Fellows of Harvard College
Department: Biomedical Informatics
Division:
Street1*: Harvard Medical School
Street2: Countway Medical Library
City*: Boston
County:
State*: MA: Massachusetts
Province:
Country*: USA: UNITED STATES
Zip / Postal Code*: 021150000

Phone Number*: 6174327377 Fax Number:

E-Mail*: peter.kharchenko@post.harvard.edu

Credential, e.g., agency login: kharchenko

Project Role*: Other (Specify) Other Project Role Category: Sponsor

Degree Type: PHD Degree Year: 2005

Attach Biographical Sketch*: File Name: Kharchenko_Bio.pdf

Attach Current & Pending Support: File Name:

PROFILE - Senior/Key Person

Prefix: First Name*: CATHERINE Middle Name Ju-Ying Last Name*: WU Suffix:

Position/Title*: Associate Professor
Organization Name*: DANA FARBER CANCER INSTITUTE
Department:
Division:
Street1*: 450 Brookline Avenue, Dana 540B
Street2:
City*: BOSTON
County:
State*: MA: Massachusetts
Province:
Country*: USA: UNITED STATES
Zip / Postal Code*: 022150000

Phone Number*: (617) 632-5943 Fax Number: (617) 632-3351

E-Mail*: cwu@partners.org

Credential, e.g., agency login: CWU001

Project Role*: Other (Specify) Other Project Role Category: Co-Sponsor

Degree Type: MD Degree Year: 1994

Attach Biographical Sketch*: File Name: Wu_Biosketch.pdf

Attach Current & Pending Support: File Name:

APPLICANT BIOGRAPHICAL SKETCH

Use only for individual predoctoral and postdoctoral fellowships, dissertation research grants (R36),and Research Supplements to Promote Diversity in Health-Related Research (Admin Suppl). DO NOT EXCEED FIVE PAGES.

NAME OF APPLICANT: Jean Fan

eRA COMMONS USER NAME (credential, e.g., agency login): JEAN_FAN

POSITION TITLE: Graduate Student

EDUCATION/TRAINING *(Most applicants will begin with baccalaureate or other initial professional education, such as nursing. Include postdoctoral training and residency training if applicable. High school students should list their current institution and associated information. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	START DATE MM/YYYY	END DATE *(or expected end date)* MM/YYYY	FIELD OF STUDY
Johns Hopkins University, Baltimore, MD	BS	08/2009	05/2013	Biomedical Engineering, Applied Mathematics and Statistics
Harvard University, Boston, MA	PhD	06/2013	present	Bioinformatics and Integrative Genomics

A. Personal Statement

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. I am particularly interesting in heterogeneity in the context of cancer and how this heterogeneity shapes tumor progression, therapeutic resistance, and ultimately clinical impact. Support through The NCI Predoctoral to Postdoctoral Fellow Transition Award (F99/K00) will provide me with the financial independent to pursue my scientific objects and prepare me in establishing as an independent investigator and cancer researcher.

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 in cancer research. 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 have focused on developing statistical and computational methods for analyzing sequencing data, 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 single cells and pathway and gene set overdispersion analysis to identify and characterize transcriptional subpopulations (Fan et al, Nature Methods 2016) as well as applied my computational skills to analyze locally disordered methylation (Landau et al, Cell 2014), clonal evolution in developing drug resistance (Burger et al, Cancer Discovery 2015), and impact of *SF3B1* mutation (Wang et al, Cancer Cell 2016) in chronic lymphocytic leukemia. Currently I am working on computational approaches to link transcriptional and genetic heterogeneity at the single cell level using a Bayesian hierarchical models for inferring copy number alteration from single cell RNA-seq data with applications to multiple myeloma (Fan and Lee et al. manuscript in progress).

In additional to achieving excellence in research, I strive to effectively train the next generation of scientific professionals who will carry on and effectuate the same vision and mission. In my mentoring, I seek to empower students with the capacity for self-generation of ideas, self-direction, and self-monitoring through challenging and purposeful project-driven learning. I have already mentored one undergraduate student to develop a pure client-side implementation of various bioinformatics analyses to enable scalable, barrier-free bioinformatics analysis. We are currently preparing a manuscript (Fan and Fan et al, manuscript in progress). I look forward to mentoring more students during my post-doctoral training and future scientific career.

With support from The NCI Predoctoral to Postdoctoral Fellow Transition Award (F99/K00), 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

ACTIVITY/ OCCUPATION	START DATE (mm/yy)	ENDING DATE (mm/yy)	FIELD	INSTITUTION/ COMPANY	SUPERVISOR/ EMPLOYER
Summer research intern	06/2008	09/2008	Cancer Biology	National Cancer Institute, National Institutes of Health	Paul Meltzer and Liang Ciao
Undergraduate research scientist	08/2009	05/2013	Bioinformatics, Evolutionary biology, Genomics	Institute for Computational Medicine, Johns Hopkins University	Rachel Karchin
Summer research intern	06/2012	08/2012	Population genetics	Harvard Medical School	Shamil Sunyaev
Teaching assistant	08/2012	12/2012	Mathematics	Johns Hopkins University	Donniell Fishkind
Teaching fellow	10/2014 10/2015 10/2016	10/2014 10/2015 10/2016	Genomics, Transcriptomics	Harvard Stem Cell Institute	Peter Kharchenko
Teaching fellow	06/2016	08/2016	Bioinformatics and Integrative Genomics	Harvard Medical School	Susanne Churchill

Awards and Honors

- Siemens Competition in Math, Science and Technology Semi-Finalist (2008)
- Intel Science Talent Search Semi-Finalist (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)
- Department of Defense American Society for Engineering Education National Defense Science and Engineering Graduate Fellowship awardee (2013)
- ASH Abstract Achievement Award Recipient (2015)
- NSF MRS Outreach Award Recipient (2016)
- Ruth L. Kirschstein Predoctoral Individual National Research Service Award F31 Grant recipient, CA206236-01 (2016)

Memberships

- American Society of Hematology
- American Society of Human Genetics
- International Society for Computational Biology
- American Association for the Advancement of Science

C. Contributions to Science -- Graduate Research.

A more complete understanding of chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) is a cancer of the blood and bone marrow with a highly variable clinical course. Advancements in high-throughput sequencing technologies have uncovered tremendous genetic, epigenetic, and transcriptional heterogeneity in CLL. To study the impact of this heterogeneity on clinical course, I have established a close collaboration with the group of Catherine Wu (DFCI). The Wu group pioneered investigations of subclonal mutations in CLL. Their access to patient samples and other wet lab resources has enabled me to focus on bioinformatics analyses of their data. Our collaboration has led to many scientific findings that contribute to a more complete understanding of CLL:

- Landau DA, Clement K, Ziller MJ, Boyle P, **Fan J**, Gu H, Stevenson K, Sougnez C, Wang L, Li S, Kotliar D, Zhang W, Ghandi M, Garraway L, Fernandes SM, Livak KJ, Gabriel S, Gnirke A, Lander ES, Brown JR, Neuberg D, Kharchenko PV, Hacohen N, Getz G, Meissner A, and Wu CJ. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell 2014, Dec 8; 26(6):813-25
- Burger JA*, Landau DA*, Taylor-Weiner A*, Bozic I*, Zhang H*, Sarosiek K, Wang L, Stewart C, **Fan J**, Hoellenriegel J, Sivina M, Dubuc AM, Fraser C, Han Y, Li S, Livak KJ, Zou L, Wan Y, Konoplev S, Sougnez C, Brown JR, Abruzzo LV, Carter SL, Keating MJ, Davids MS, Wierda WG, Cibulskis K, Zenz T, Werner L, Dal Cin P, Kharchencko P, Neuberg D, Kantarjian H, Lander E, Gabriel S, O'Brien S, Letai A, Weitz DA, Nowak MA, Getz G, Wu CJ. Clonal evolution in patients with chronic lymphocytic leukemia developing resistance to BTK inhibition. Nature Communications 2016, May 20. doi: 10.1038/ncomms11589.
- Wang L*, Brooks AN*, **Fan J*,** Wan Y*, Gambe R, Li S, Hergert S, Yin S, et al. Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. Cancer Cell 2016, Nov 3. doi: 10.1016/j.ccell.2016.10.005.
- Wang L*, **Fan J*,** Zhang CZ, Francis JM, Georghiou G, Hergert S, Shuqiang Li, Gambe R, Zhou CW, Yang C, Xiao S, Cin PD, Bowden M, Kotliar D, Shukla SA, Brown JR, Neuberg D, Alessi DR, Khachenko PV, Livak KJ, Wu CJ. Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. Genome Research (in revision).

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 analysis of single cell RNA-seq data. This work has led to the development of various statistical methods available as software for the scientific community:

- **Fan J,** Salathia N, Liu R, Kaeser G, Yung Y, Herman J, Kaper F, Fan JB, Zhang K, Chun J, and Kharchenko PV. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nature Methods 2016 Mar;13(3):241-4. doi: 10.1038/nmeth.3734.
- **Fan J**, Fan D, Slowikowski K, Gehlenborg N, Kharchenko PV. UBiT2: a client-side web-application for gene expression data analysis. (manuscript in preparation).
- **Fan J*,** Lee H*, Lee S, Ryu D, Lee S, Kim SJ, Kim K, Park P, Park WY, Kharchenko PV. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq. (manuscript in preparation).

Improving the representation of Women in STEM through mentoring and 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 and enthusiasts the fundamentals of protein evolution. To make the material accessible and exciting, lessons are conveyed through an intuitive and attractive gaming environment and framed into the context of common human health concerns and topics.
- I was the co-chair for the Harvard Graduate Women in Science and Engineering student group for 2 years. I oversaw operations of our mentoring program, held meetings with deans of the university, led board

meetings, led events including coffee hours with female faculty and leadership workshops, and organized our 10th year anniversary WISE Beyond Your Years conference.

- I founded the non-profit (501c3) CuSTEMized (http://www.custemized.org/) which engages, encourages, and empowers young girls in Science, Technology, Engineering, and Math (STEM) by providing them with tangible products and educational experiences that foster a positive scientific identity from a young age. I developed the website, software, graphics, and content to enable generation of free personalized motivational books and manage a team of approximately 10 high-schoolers, post-bacs, and social activists to accomplish our outreach programs. I ran a successful Kickstart campaign which raised over $6,000 (nearly 400% over our asking amount). Since our launch, over 10,000 personalized ebooks have been created using our platform with over 250 free printed books provided to girls and classrooms from underserved communities. I also led the organization of 3 STEM-enrichment events around Boston in collaboration with local libraries and after-school programs which have hosted 9 female scientific speakers to engage over 150 young girls in hands-on learning activities.

D. Scholastic Performance

YEAR	MATH/SCIENCE COURSE TITLE	GRADE	YEAR	OTHER COURSE TITLE	GRADE
	Johns Hopkins University			**Johns Hopkins University**	
2009	General Physics I	A	2009	Honors Multivariable Calculus	A
2009	General Physics Lab	A+	2009	BME Modeling and Design	A
2010	General Physics II	A-	2010	Linear Algebra	A-
2010	General Physics Lab II	A	2010	Discrete Mathematics	B
2010	Organic Chemistry I	A-	2010	BME Design Group	B+
2010	Molecules and Cells	A-	2010	Differential Equations/Applications	A
2011	Statistical Mechanics/Thermo	A-	2010	Data Structures	B+
			2010	Responsible Conduct of Research	Sat
			2011	Intro to Probability	A-
			2011	Systems and Controls	B+
			2011	Models and Simulations	B+
			2011	Intro to Optimization	A
			2011	System Bioengineering I	B+
			2011	System Bioengineering Lab I	A
			2012	Intro to Stochastic Processes	A
			2012	Intro to Statistics	A
			2012	Systems Bioengineering II	A-
			2012	Systems Bioengineering Lab II	A
			2012	Computational Molecular Medicine	A-
			2012	Systems Bioengineering III	A
			2013	History of Modern Medicine	A
	GRE Quant	**163/170**	2013	Justice and Health	B+
	GRE Verbal	**163/170**	2013	Medical Informatics	A-
	GRE Writing	**5.5/6**	2013	Machine Learning in Complex Domains	B-
	Harvard University				
2013	Principles of Genetics	B+	2016	Social Entrepreneurship	A-

YEAR	MATH/SCIENCE COURSE TITLE	GRADE	YEAR	OTHER COURSE TITLE	GRADE
2013	Analysis of the Biological Literature	B+			
2013	Quantitative Genomics	B+			
2013	Statistical Inference	B-			
2014	Computational and Functional Genomics	A-			
2014	Biological Macromolecules: Structure, Function and Pathways	A-			
2014	Selected Topics in High Dimensional Analysis	Sat			
2014	Bayesian Methodology in Biostatistics	B+			
2014	Introduction to Biomedical Informatics I	A			
2014	Conduct of Science	Sat			
2015	Introduction to Biomedical Informatics II	A			

SAT/UNS The grade of Satisfactory includes letter grades from A to C–; the grade of Unsatisfactory represents work below C– and is considered a failing grade. No students enrolled in courses graded SAT/UNS may receive letter grades in those courses.

PHS Fellowship Supplemental Form

OMB Number: 0925-0001
Expiration Date: 10/31/2018

Introduction

1. Introduction
(RESUBMISSION)

Fellowship Applicant Section

2. Applicant's Background and Goals for Fellowship Training*	Fan_Applicant_Background_and_Goals_for_Fellowship_Training.pdf

Research Training Plan Section

3. Specific Aims*	Fan_Specific_Aims.pdf
4. Research Strategy*	Fan_Research_Strategy.pdf
5. Respective Contributions*	Fan_Respective_Contributions.pdf
6. Selection of Sponsor and Institution*	Fan_Selection_of_Sponsor_and_Institution.pdf
7. Progress Report Publication List (RENEWAL)	
8. Training in the Responsible Conduct of Research*	Fan_Responsible_Conduct_of_Research.pdf

Sponsor(s), Collaborator(s) and Consultant(s) Section

9. Sponsor and Co-Sponsor Statements	Fan_Sponsor_Cosponsor_Statements.pdf
10. Letters of Support from Collaborators, Contributors and Consultants	Fan_Letters_of_support_from_collaborators.pdf

Institutional Environment and Commitment to Training Section

11. Description of Institutional Environment and Commitment to Training	Fan_Institutional_Environment.pdf

Other Research Training Plan Section

Human Subjects

Please note. The following item is taken from the Research & Related Other Project Information form. The response provided on that page, regarding the involvement of human subjects, is repeated here for your reference as you provide related responses for this Fellowship application. If you wish to change the answer to the item shown below, please do so on the Research & Related Other Project Information form; you will not be able to edit the response here.

Are Human Subjects Involved? ☑ Yes ☐ No

12. Human Subjects Involvement Indefinite?	☐ Yes ☑ No
13. Clinical Trial?	☐ Yes ☑ No
14. Agency-Defined Phase III Clinical Trial?	
15. Protection of Human Subjects	Fan_Protection_of_Human_Subjects.pdf
16. Data Safety Monitoring Plan	
17. Inclusion of Women and Minorities	Fan_Inclusion_Women_Minorities.pdf
18. Inclusion of Children	Fan_Inclusion_Children.pdf

Vertebrate Animals

The following item is taken from the Research & Related Other Project Information form and repeated here for your reference. Any change to this item must be made on the Research & Related Other Project Information form.

Are Vertebrate Animals Used? ☐ Yes ☑ No

19. Vertebrate Animals Use Indefinite?

PHS Fellowship Supplemental Form

20. Are vertebrate animals euthanized?	
If "Yes" to euthanasia Is method consistent with American Veterinary Medical Association (AVMA) guidelines?	
If "No" to AVMA guidelines, describe method and provide scientific justification	
21. Vertebrate Animals	
Other Research Training Plan Information	
22. Select Agent Research	
23. Resource Sharing Plan	Fan_Resource_Sharing_Plan.pdf
24. Authentication of Key Biological and/or Chemical Resources	

PHS Fellowship Supplemental Form

Additional Information Section

25. Human Embryonic Stem Cells

Does the proposed project involve human embryonic stem cells?* □ Yes ☑ No

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s), using the registry information provided within the agency instructions. Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Specific stem cell line cannot be referenced at this time. One from the registry will be used.

Cell Line(s):

26. Alternate Phone Number:

27. Degree Sought During Proposed Award:

Degree: If "other", please indicate degree type: Expected Completion Date (month/year):

28. Field of Training for Current Proposal*: 102 Bioinformatics

29. Current Or Prior Kirschstein-NRSA Support?* ☑ Yes □ No

If yes, please identify current and prior Kirschstein-NRSA support below:

Level*	Type*	Start Date (if known)	End Date (if known)	Grant Number (if known)
Predoctoral	Individual			

30. Applications for Concurrent Support?* □ Yes ☑ No

If yes, please describe in an attached file:

31. Citizenship*

U.S. Citizen U.S. Citizen or Non-Citizen National? ☑ Yes □ No

Non-U.S. Citizen □ With a Permanent U.S. Resident Visa

□ With a Temporary U.S. Visa

If you are a non-U.S. citizen with a temporary visa who has applied for permanent resident status and expect to hold a permanent resident visa by the earliest possible start date of the award, please also check here. □

32. Change of Sponsoring Institution Name of Former Institution:*

PHS Fellowship Supplemental Form

Budget Section

All Fellowship Applicants:

1. Tuition and Fees*:

☐ None Requested ☑ Funds Requested

Year 1	$6,872.00
Year 2	$6,872.00
Year 3	
Year 4	
Year 5	
Year 6 (when applicable)	
Total Funds Requested:	$13,744.00

Senior Fellowship Applicants Only:

2. Present Institutional Base Salary:	Amount	Academic Period	Number of Months

3. Stipends/Salary During First Year of Proposed Fellowship:

a. Federal Stipend Requested:	Amount	Number of Months
b. Supplementation from other sources:	Amount	Number of Months
	Type (sabbatical leave, salary, etc.)	
	Source	

Appendix

Applicant’s Background and Goals for Fellowship Training

A. Doctoral Dissertation and Research Experience

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 in understanding cancer pathogenesis and progression. These experiences, summarized below, along with experiences to be gained during the completion of the F99/K00 plans will prepare me for a long-term career as an independent cancer investigator.

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$_{11--12}$ 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$_{11--12}$ 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$_{11--12}$ amplicon and may play a role in breast cancer growth and pathogenesis.

- Fan J, Yu Y, Meltzer PS, Cao L. Delineating the Role of BRF2 in Breast Cancer Pathogenesis. HURJ 2011. 14, 53-55

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 non-parametric 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 Pathway And Geneset OverDispersion Analysis (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.

- **Fan J,** Salathia N, Liu R, Kaeser G, Yung Y, Herman J, Kaper F, Fan JB, Zhang K, Chun J, and Kharchenko PV. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nature Methods 2016 Mar;13(3):241-4. doi: 10.1038/nmeth.3734. □

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.

- Landau DA, Clement K, Ziller MJ, Boyle P, **Fan J**, Gu H, Stevenson K, Sougnez C, Wang L, Li S, Kotliar D, Zhang W, Ghandi M, Garraway L, Fernandes SM, Livak KJ, Gabriel S, Gnirke A, Lander ES, Brown JR, Neuberg D, Kharchenko PV, Hacohen N, Getz G, Meissner A and Wu CJ. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell 2014, Dec 8; 26(6):813-25

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

Graduate Student **December 2013 – Present**
Mentored by Peter Kharchenko at the Department for Biomedical Informatics, Harvard University
In collaboration with Catherine Wu at the Dana-Farber Cancer Institute
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

fluorescence 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.

- Burger JA*, Landau DA*, Taylor-Weiner A*, Bozic I*, Zhang H*, Sarosiek K, Wang L, Stewart C, **Fan J**, Hoellenriegel J, Sivina M, Dubuc AM, Fraser C, Han Y, Li S, Livak KJ, Zou L, Wan Y, Konoplev S, Sougnez C, Brown JR, Abruzzo LV, Carter SL, Keating MJ, Davids MS, Wierda WG, Cibulskis K, Zenz T, Werner L, Dal Cin P, Kharchencko P, Neuberg D, Kantarjian H, Lander E, Gabriel S, O'Brien S, Letai A, Weitz DA, Nowak MA, Getz G, Wu CJ. Clonal evolution in patients with chronic lymphocytic leukemia developing resistance to BTK inhibition. Nature Communications 2016, May 20. doi: 10.1038/ncomms11589. □
- Wang L*, Brooks AN*, **Fan J*,** Wan Y*, Gambe R, Li S, Hergert S, Yin S, et al. Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. Cancer Cell 2016, Nov 3. doi: 10.1016/j.ccell.2016.10.005. □
- Wang L*, **Fan J*,** Zhang CZ, Francis JM, Georghiou G, Hergert S, Shuqiang Li, Gambe R, Zhou CW, Yang C, Xiao S, Cin PD, Bowden M, Kotliar D, Shukla SA, Brown JR, Neuberg D, Alessi DR, Khachenko PV, Livak KJ, Wu CJ. Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. Genome Research (in revision). □
- **Fan J**, Fan D, Slowikowski K, Gehlenborg N, Kharchenko PV. UBiT2: a client-side web-application for gene expression data analysis. (manuscript in preparation). □
- **Fan J*,** Lee H*, Lee S, Ryu D, Lee S, Kim SJ, Kim K, Park P, Park WY, Kharchenko PV. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq. (manuscript in preparation). □

Additional projects: Progenitor cell origins in chronic lymphocytic leukemia, analysis of SF3B1 mutation in a novel mouse model of CLL

B. Training Goals and Objectives

My career goal is to be a leading scientist in either an academic research laboratory or comparable 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 impact. I believe that in order to understand this heterogeneity, novel statistical methods and user-friendly computational software must be developed to enable researchers to harness the power of big –omics datasets currently being generated. My goal for the F99/K00 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 cancer –omics data to uncover the biological mechanisms driving clinical prognosis in cancer progression. My sponsors, Dr. Peter Kharchenko and Dr. Catherine Wu, and I have developed the following set of training goals and objectives:

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. I took 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 about various statistical methods and computational approaches through hands-on research experience, attending conferences and meetings, and reading scientific literature. Most importantly, I will develop my ability to think quantitatively and critically as new research questions emerge. The F99-phase will focus more on building a strong quantitative and computational skill set under the guidance and mentorship of my sponsors while the K00-phase will focus more on refining these skill sets independently.

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. For the F99-phase, I will continue collaborating closely with Dr. Lili Wang of the Wu Lab. Dr. Wang has extensive experience as a trained immunologist and her mentorship will provide me with a better understanding of the biological background and interpretation of my computational analyses. For the K00-phase, I will identify a lab environment that has both a wet and dry lab component to enable me to more closely interact with bench biologists on a regular basis.

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. For the F99-phase, I will still be depending on datasets acquired as part of separate research efforts. For the K00-phase, I intend to pursue my own independent dataset design and sequencing efforts. Such an experience will ensure I am able to establish collaborations and guide collaborators in the initial batch design, a necessary skill for an independent bioinformatics investigator.

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. Specifically, in the F99-phase, I plan on mentoring an undergraduate student as part of the BIG HST Summer Institute in Biomedical Informatics on a summer project to be written up and presented at the BIG summer undergraduate student conference and further submitted to an undergraduate research journal if applicable. Through mentoring students, I will develop skills needed to effectively train the next generation of scientists.

Improve communication and presentation skills: Two key components of this goal are scientific writing and oral presentation. In terms of scientific writing, I hope to continue gaining experience in preparing and submitting manuscripts to peer-reviewed journals. I have also previously written project proposals for courses and for grants such as the NSF GRFP and F31. 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.

C. Activities Planned Under This Award

Percentage of time to be spent under each category:

	F99 – Year 1	F99 – Year 2	K00 – Year 1	K00 – Year 2	K00 – Year 3	K00 – Year 4
Research	90%	85%	85%	80%	75%	70%
Mentoring	5%	5%	10%	10%	15%	15%
Conferences	5%	10%	5%	10%	10%	15%

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 in preparation for an independent career:

To gain experience in mentoring students, I will be directly mentoring and supervising at least 1 undergraduate student as part of the DBMI Summer Program for Biomedical Informatics. I will also be serving as a teaching assistant for this summer program, participating in weekly lectures and encouraging students to ask questions.

To stay informed about scientific progress in bioinformatics and cancer research and to develop my communication and presentation skills, I will be attending a variety of seminars, workshops, and retreats hosted by HMS and other organizations. 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 speaking opportunities are also available at HMS. I will travel to local and national conferences to network and present my research with my sponsor and co-sponsor, including the ASH, AACR, and the Single Cell Analysis Investigators Meeting.

As I progress in my academic training, I expect to decrease the amount of time directly spent on research and shift more towards mentoring and supervising students to carry out specific elements of the proposed research project. I will also increase the amount of time dedicating to networking and presenting at conferences.

Timeline of research:

F99 – Year 1	F99 – Year 2	K00 – Year 1	K00 – Year 2	K00 – Year 3	K00 – Year 4
• Continue doctoral dissertation work (Aim 2) • Prepare and submit works on all current projects	• Begin informational interviews to identify post-doctoral lab • Revise and resubmit works on current project as needed	• Gather and process single cell epigenetic datasets (Aim 3) • Develop methods for epigenetic subpopulations discovery • Benchmark using artificial mixtures and simulations	• Apply methods to real datasets • Develop methods for connecting epigenetic subpopulations to transcriptional subpopulations • Mentor students on transcriptomic portion of the project	• Finish analysis • Prepare and submit manuscript • Mentor students • Apply for K99 • Network and present at conferences	• Perform additional analysis, revise manuscript, and resubmit as needed • Network and present at conferences for leads on faculty positions

Specific Aims

Intra-tumor heterogeneity is a common feature across diverse cancer types. Dynamic changes can be observed among intra-tumoral subclonal populations over time and following therapy, presenting challenges to current standards of cancer treatment[1–5]. Our current understanding of how distinct subclones may compete or collaborate with each other in response to microenvironmental or developmental changes is limited and warrants further investigation. The emergence of single-cell technologies for analyzing cancer cells has highlighted the potential to discover novel cellular subpopulations and states within heterogeneous cell populations. However, such single-cell measurements are variable and sparse, owing to both the stochasticity of biological processes and inconsequential technical noise[6,7]. Therefore, new statistical methods and computational tools are needed to analyze and model the inherently statistical nature of these single-cell measurements and dissect the subclonal architecture from within a probabilistic framework.

In this proposed research, I will develop statistical methods and computational software to characterize single-cell heterogeneity in cancer. For the F99 phase, I will continue focusing on characterizing intra-tumoral heterogeneity. Specifically, I will develop methods to reconstruct the phylogenetic relationship among individual cancer cells based on genetic information inferred from single-cell RNA-seq data. I will build on my previous work on transcriptional heterogeneity to characterize the gene expression profiles of identified subclonal populations and assess correspondence to identified transcriptional subpopulations. For the K00 phase, I will shift focus to characterizing heterogeneity in the tumor microenvironment and develop approaches to ultimately assess the correspondence between cancer cell subpopulations in relation to the composition and state of co-existing non-cancer cells in the microenvironment over time in response to therapy. Although the proposed methods will be applicable across diverse cancer types, I will focus on demonstrating utility in chronic lymphocytic leukemia (CLL). I will take advantage of single-cell datasets generated by the Wu lab at the Dana-Farber Cancer Institute as a part of separate research efforts (1R01HL11645201, 1R01CA15501002), as well as public datasets.

Aim 1: The Dissertation Research Project: progress thus far – identification and characterization of transcriptional subpopulations. The ability to separate a heterogeneous cell sample into functionally relevant subpopulations is the key advantage of a single-cell approach. My previous work has shown that given high levels of variability inherent to the single-cell measurements, probabilistic modeling of the underlying processes and integration of prior knowledge through pathway annotations provide marked improvements in the ability to discern and interpret cell subpopulations. This work has led to open-source software available to the scientific community[19] and has been applied to identify and characterize subpopulations in the developing mouse brain, as well as in cancers such as CLL.

Aim 2: The Dissertation Research Project: work to be completed - reconstructing subclonal architecture and dissecting subclonal evolution. Because traditional bulk measurements are unable to resolve whether mutations are mutually exclusive or co-occurring and are subject to averaging artifacts, Aim 2 will develop statistical approaches to reconstruct subclonal architectures, impute the order of genetic alterations incurred, and identify genetic subclones based on somatic CNVs and mutations inferred at the single-cell level from single-cell RNA-seq data. These methods will be applied to primary, metastatic, pre and post-treatment CLL samples. I will then analyze the transcriptional state(s) of distinguishable genetic subclones to identify features associated with clonal growth rate, metastatic transition, and drug resistance as well as assess the correspondence of genetic heterogeneity with transcriptional heterogeneity.

Aim 3: The Postdoctoral Research Direction – characterizing the composition and state of the tumor microenvironment. Reciprocal interactions between the tumor cell and its microenvironment can influence cancer progression and treatment response[8,9]. Thus, for the K00-phase, I will shift focus to investigate the role of the tumor microenvironment in cancer progression and relapse. To identify dynamic changes in subpopulations that may shed mechanistic insight to clinical responses, I will examine single-cell RNA-seq data for cancer and non-cancer cells in multiple patients over multiple time points. Aim 3 will take skills and insights gained from the F99-phase to separate cancer from non-cancer cells based on inferred somatic alterations and transcriptional profiles, then assess changes in cell population proportions over time, and develop new methods for identifying potential changes in particular T or myeloid cell regulatory networks and pathways over time.

Successful completion of this proposal will yield new insight into subclonal evolution in CLL and provide new open-source computational software for identifying and characterizing aspects of tumor heterogeneity 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: Intra-tumor heterogeneity is commonly observed across diverse cancer types and presents challenges to current standards of cancer treatment[1–5]. Characterization of subclonal populations in cancer may enable precision medicine and the initiation of synergistic treatment combinations to target and eliminate aggressive subpopulations to improve clinical outcome. Beyond intra-tumoral heterogeneity, differences in composition and state of the tumor microenvironment may also play a role in cancer progression and treatment response[8,9]. Our proposal will yield novel statistical methods to identify and characterize genetic and transcriptional intra-tumoral heterogeneity for the F99-phase and assess correspondence to changes in microenvironment in the K00-phase using single-cell data, yielding 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 genetic and epigenetic heterogeneity along with a highly variable disease course among patients that remains poorly understood[2,10–13]. Our collaborators in the Wu group have recently established that the presence of particular subclonal mutations[10] and disordered methylation[13] in CLL can be linked with adverse clinical outcomes using bulk 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[12]. While insights have been previously gained from bulk measurements, further characterization on the single-cell level is needed to more accurately dissect the pathway and regulatory features associated with subclonal mutations. In the F99-phase, I will 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 to provide insights to the molecular mechanisms of relapse and progression in CLL. CLL also provides an exemplary model system of a microenvironment-dependent tumor in which neoplastic cells co-evolve together with host immune cells within blood, marrow or lymph nodes[14,15]. Thus, in the K00-phase, I will examine CLL and non-CLL cells in multiple patients over multiple time points to characterize subclonal heterogeneity in CLL cells in relation to the composition and state of co-existing non-CLL cells in the microenvironment over time in response to therapy.

1.3. Statistical methods are needed to identify and connect aspects of heterogeneity: Transcriptional heterogeneity can be observed in normal cell types such as neural progenitor cells[16], and T cells[17], as well as aberrant cell types such as cancer[1–5]. Differential properties such as genetic and epigenetic differences among cells may be responsible for this heterogeneity but how it is regulated, along with its direct consequences on cellular behavior, in particular in relation to the tumor microenvironment, remains unclear. 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[6,7]. Therefore, novel statistical methods are needed to identify and connect genetic, transcriptional, and microenvironmental differences in single cells as well as identify putative subpopulations. Our previous work demonstrates that integration of cell-specific error models, probabilistic weighting of observations, and prior knowledge such as annotated pathways improves the ability to separate cell types within a mixed single-cell sample when clustering cells based on gene expression[18,19]. Our proposal will apply these statistical approaches as well as develop new approaches to improve characterization of genetic and transcriptional single-cell heterogeneity and subsequently enhance our understanding of transcriptional variability and its connection to genetic and microenvironmental differences.

1.4. CLL cells impact and are impacted by their microenvironment: Reciprocal interactions between the tumor cell and its microenvironment may influence cancer progression and treatment response. Indeed, immune cellular elements can eliminate cancer through immunosurveillance mechanisms[9]. In turn, tumor cells can subvert physiological immune responses to through the generation of an immunosuppressive milieu to escape from immune recognition[15]. CLL provides an exemplary model system of a microenvironment-dependent tumor in which neoplastic cells co-evolve together with host immune cells within blood, marrow or lymph nodes. However, how distinct CLL subclones impact the composition and interaction with host immune cells and vice versa is unknown.

2. Approach:

Although the proposed methods to be developed will be applicable to diverse cancers, our approach focuses on CLL. Previous studies on bulk CLL samples have demonstrated the association of genetic heterogeneity[10] and composition of tumor microenvironment[14,15] in clinical outcome, thus making CLL 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.

For the F99 phase, I have already developed methods to analyze transcriptional heterogeneity and have applied these methods to CW14, CW106, CW84, CW236. I propose an additional set 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 to make probabilistic inferences on presence or absence of CNVs and SNVs inferred from single-cell RNA-seq data. Second, I will reconstruct subclonal architectures, impute the order of genetic alterations incurred, and identify genetic subclones based on somatic mutations 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 also 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. I will also assess the correspondence of previously identified transcriptional subpopulations with genetic subclones.

For the K00 phase, I will characterize heterogeneity in CLL cells in relation to the composition and state of co-existing non-CLL cells in the tumor microenvironment to gain insights into how distinct subclones may co-evolve together with host immune cells within blood, marrow or lymph nodes over time in response to therapy.

2.1. Aim 1: Identification and characterization of transcriptional subpopulations:

2.1.1. Results:

2.1.1a. Statistical error modeling and pathway-based approach identifies and characterizes transcriptional subpopulations in single-cell RNA-seq data. Single-cell transcriptomic measurements via single-cell RNA-seq is complicated by high levels of technical and biological noise[6,7,19]. To accommodate these different sources of noise, we have built on our previous work modeling the measurement of each cell as a mixture of two probabilistic processes[19]. Because detection of individual genes may be complicated by drop-out and stochastic bursting, we sought to improve power by pooling information across genes within annotated pathways. We developed pathway and geneset overdispersion analysis (PAGODA)[18] to build on error models and identify significant aspects of coordinated variability within annotated pathways and well as *de novo* gene sets (Fig. 1). We have shown that such an approach robustly identifies multiple potentially over-lapping aspects of transcriptional heterogeneity to characterize biologically relevant transcriptional subpopulations[18,20].

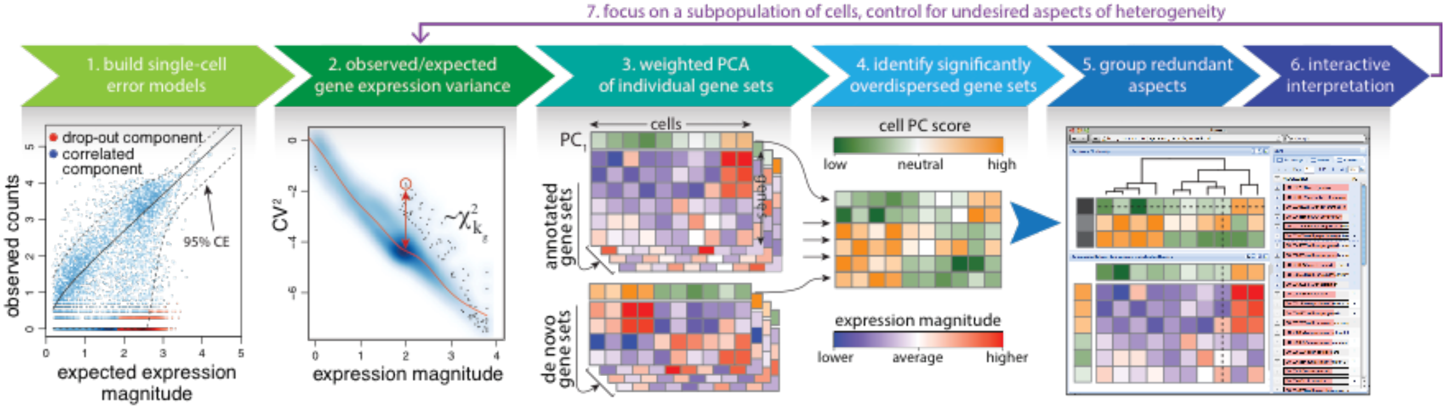

7. focus on a subpopulation of cells, control for undesired aspects of heterogeneity
1. build single-cell error models
2. observed/expected gene expression variance
3. weighted PCA of individual gene sets
4. identify significantly overdispersed gene sets
5. group redundant aspects
6. interactive interpretation
drop-out component
correlated component
95% CE
observed counts
expected expression magnitude
0
1
2
3
4
5
CV^2
$\sim\chi^2_{k_g}$
expression magnitude
cells
PC_1
genes
annotated gene sets
de novo gene sets
cell PC score
low
neutral
high
expression magnitude
lower
average
higher

Figure 1. Overview of pathway and geneset overdispersion analysis (PAGODA). Briefly, 1. cell error model are fit to quantify the dependency of amplification noise and drop-out probabilities on expression magnitude; 2. Variance normalization takes into account the uncertainty in variance estimates of each gene by determining effective degrees of freedom; 3. Weighted PCA is performed on each functionally-annotated gene sets, as well as *de novo* gene sets determined based on correlated expression; 4. If the amount of variance explained by a principal component of a given gene set is significantly higher than expected, the gene set is called overdispersed, and the cell scores defined by that principal component are included as one of the significant aspects of heterogeneity in the dataset; 5. Redundant aspects are grouped; 6. Users explore results in an interactive application.

2.1.1b. Previous unbiased transcriptional characterization of CLL reveals transcriptional heterogeneity. Applying our developed PAGODA approach, preliminary analysis of 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. 2). However, how these transcriptionally distinct subpopulations relate to genetically distinct subclones is still not well understood.

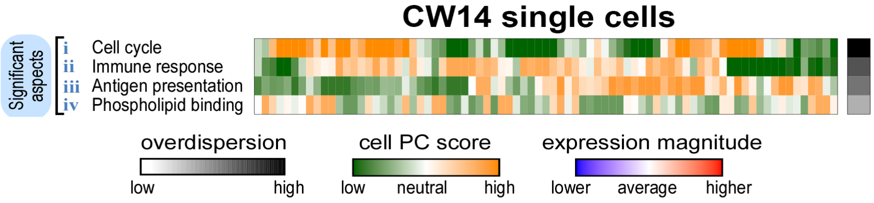


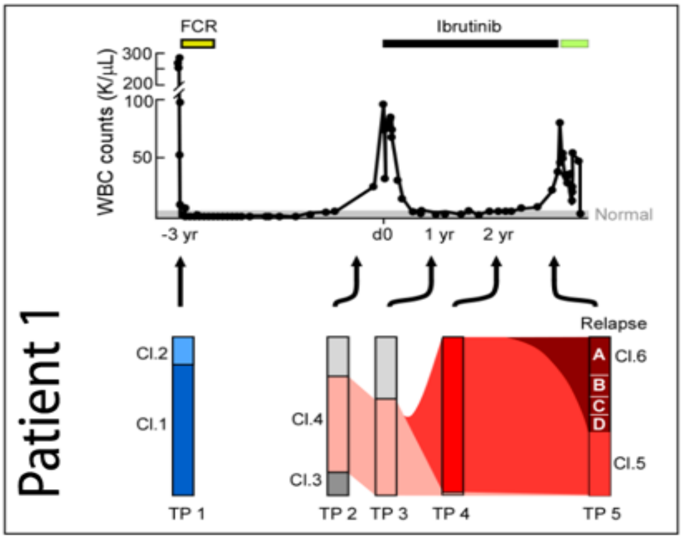
Figure 2. Transcriptional heterogeneity in a CLL sample. The heat map illustrates single cells as columns and consensus gene expression within pathway clusters as rows. Analysis of a scRNA-seq data from a CLL tumor illustrates presence of transcriptionally distinct sub-sets, separating along functionally relevant criteria such as immune response pathways.

2.1.2. Skills and techniques acquired: Working closely under the guidance and mentorship of Dr. Peter Kharchenko and Dr. Catherine Wu, completion of Aim 1 has introduced me to the fundamentals of Bayesian model design and execution, concepts in benchmarking and using simulations, and provided me with hands-on experience in methods development and manuscript writing. I have also been able to present this work as a poster at the Single Cell Investigators meeting at the NIH and at the American Society of Hematology meeting, as a talk at the Cold Spring Harbor single cell workshop and the Festival of Genomics conference, and at other local engagements to practice and refine my presentation and communication skills.

2.2. Aim 2: Reconstructing subclonal architecture and dissecting subclonal evolution:

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[12], suggesting an active evolutionary process (Fig. 3). 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.



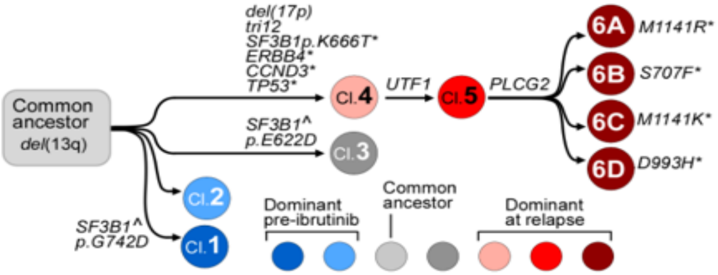


Figure 3. Genetic evolution in CLL in response to ibrutinib treatment. Bulk samples were collected and sequenced for each patient at various time points pre- and post-treatment as indicated by the black arrows. Analysis of cancer cell fractions by ABSOLUTE reveals different subclonal populations at each time point (TP1 to TP5). In particular, dominant subclonal populations in relapsing CLL cases can be observed as minor subclasses pre-treatment (e.g. cl.5), suggesting an active, branched evolutionary process in CLL clonal expansion.

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. We found that sufficiently high performance can be achieved for SNVs within highly expressed genes

(Fig. 4A). 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. 4B). Cells from the same CLL patient sample will come from the same genetic background and thus harbor less distinctive subclonal SNVs, creating a more challenging problem in need of alternative data integration such as CNVs.

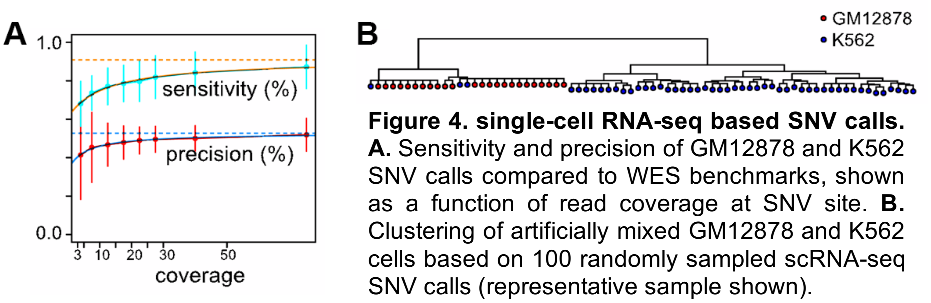


Figure 4. single-cell RNA-seq based SNV calls. A. Sensitivity and precision of GM12878 and K562 SNV calls compared to WES benchmarks, shown as a function of read coverage at SNV site. **B.** Clustering of artificially mixed GM12878 and K562 cells based on 100 randomly sampled scRNA-seq SNV calls (representative sample shown).

2.1.1.c. Biased allele expression can be observed within CNV regions for single-cell RNA-seq data. Our previous analysis of clonal deletion regions in multiple myeloma has revealed distinct detection patterns for known germline heterozygous single nucleotide polymorphisms (SNPs) identified from WES within regions affected by CNV in single cells. As expected, for deletion regions, all detected alleles originate from the same non-deleted allele (Fig. 5). However, even SNPs within CNV neutral regions exhibit highly biased allele ratios due to prevelant mono-allelic detection[21]. However, the direction of the bias varies between cells, suggesting that despite prevalent mono-allelic and biased expression, the random direction of bias within CNV neutral regions may enable us to detect CNVs based on the observation 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.

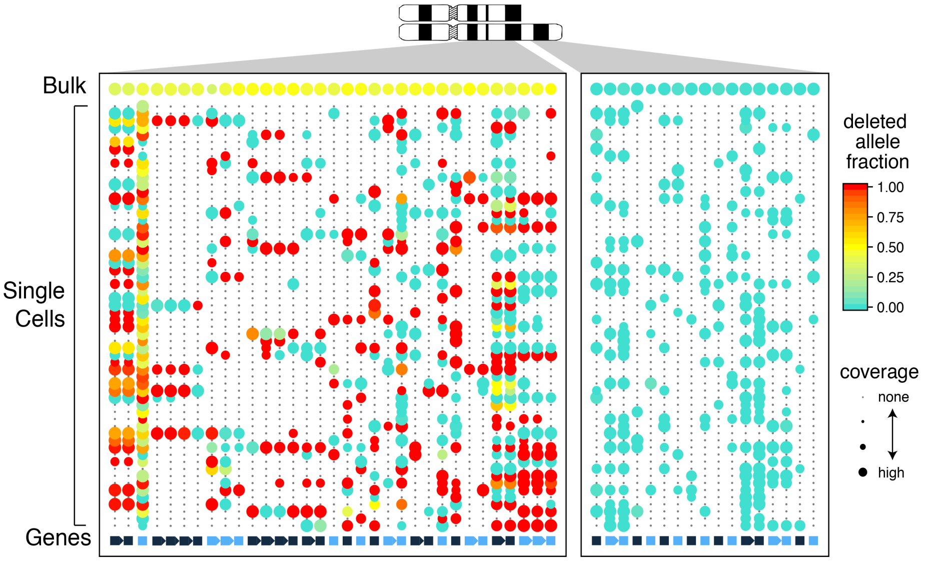


Figure 5. Patterns of allele expression outside and within CNV regions. Heterozygous germline SNPs (columns) for single cells (rows) inferred from single-cell RNA-seq is biased away from the expected 1:1 allele ratios for heterozygous variants due to mono-allelic expression within CNV neutral regions and due to clonal deletion status within deletion regions. In this example, all single cells exhibit a clonal deletion. Coordinated mono-allelic expression can be observed within genes in the CNV neutral region.

2.2.2. Skills and techniques to be acquired: Execution of this Aim will provide more independent practice and application of skills obtained from Aim 1, namely methods development, benchmarking, software development, working with collaborators, manuscript writing, manuscript submission, revision, software and data release, and presentation. Specifically, for presentations, I will travel with Dr. Wu to the American Society of Hematology conference to present the biological findings from this Aim, and to the Keystone Symposia Single Cell Omics Conference or Single Cell Genomics conference to present the computational methods. I will also mentor an undergraduate summer research student in packaging and testing the software developed in this Aim to ensure ease of use for the scientific community and gain experience in managing students.

2.2.3. Research design: Here, we propose integrating RNA-level and DNA-level information at the single-cell level by directly inferring the presence of somatic alterations from single-cell RNA-seq data. We will take advantage of prior knowledge acquired from bulk WES where possible. Specifically, from bulk WES, we will identify putative heterozygous germline SNPs using GATK[22] and MuTect[23]. We will also identify candidate regions of CNV using Control-FREEC[24,25]. Where bulk WES is not available, we will take advantage of public databases of common heterozygous SNPs[26] to identify putative patient-specific heterozygous sites. We will also further apply CNV on pooled single cells to identify candidate regions of CNV. We will then use the following hierarchical Bayesian model to assess the posterior probability of the presence of candidate CNVs in single cells. We will then reconstruct the subclonal architecture of single cells within each sample and further infer the temporal ordering of SNVs and CNVs following maximum parsimony assumptions. We will apply this method to reconstruct the subclonal architectures for 2 multiple myeloma patients (in collaboration with Dr. Peter Park, see letter of support) and 3 CLL patients at multiple time points, pre- and post-chemo and ibrutinib treatment (Fig. 3). We anticipate that our single-cell CNV detection method will recapitulate cancer cell fraction estimates from bulk WES by ABSOLUTE[27]. Having established the subclonal architecture, we will perform further transcriptomic characterization as described below.

2.2.3.a. 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[28,29] and highly dependent on the normalization reference employed. Here, we propose an alternative approach to enable detection of smaller CNVs, taking advantage of heterozygous SNPs within CNV regions. We previous observed that complete allelic imbalance at germline heterozygous SNPs suggests presence of a deletion or LOH. However, such patterns of allelic imbalance 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. 6). 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.

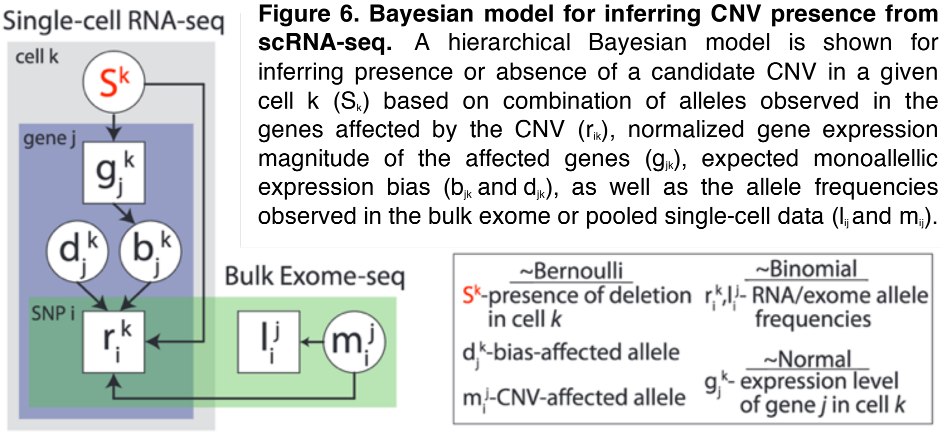

Single-cell RNA-seq
cell k
S^k
gene j
g_j^k
d_j^k
b_j^k
SNP i
r_i^k
Bulk Exome-seq
l_i^j
m_i^j
~Bernoulli
S^k-presence of deletion in cell k
d_j^k-bias-affected allele
m_i^j-CNV-affected allele
~Binomial
r_i^k,l_i^j- RNA/exome allele frequencies
~Normal
g_j^k- expression level of gene j in cell k

Figure 6. Bayesian model for inferring CNV presence from scRNA-seq. A hierarchical Bayesian model is shown for inferring presence or absence of a candidate CNV in a given cell k (S_k) based on combination of alleles observed in the genes affected by the CNV (r_{ik}), normalized gene expression magnitude of the affected genes (g_{jk}), expected monoallelic expression bias (b_{jk} and d_{jk}), as well as the allele frequencies observed in the bulk exome or pooled single-cell data (l_{ij} and m_{ij}).

2.2.3.b. Transcriptomic characterization of genetic subclonal populations. Having identified subclonal populations using somatic mutations, we will assess the transcriptional profiles of each subclonal populations. For each intra-patient subpopulation, we will apply single-cell differential expression analysis[19] to identify differentially upregulated and downregulated genes associated with each subclone. We will use gene set enrichment analysis[30] 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. 3) 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[31]. 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[32], subclonal activation of Wnt signaling[33] and other pathways implicated in CLL-B-cell expansion of potential relevance to CLL progression. We will also apply an unbiased approach to assessing transcriptional heterogeneity using results from Aim 1. We will then assess whether the observed patterns of transcriptional heterogeneity are associated with particular somatic mutations or alternative modes of regulation.

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, particularly when CNVs are present. 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 to avoid potential false signals.

In addition, 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 ABSOLUTE[27] to establish correspondence.

2.3. Aim 3: characterizing the composition and state of the tumor microenvironment:

2.3.1. Preliminary data:

2.3.1.a. CNV and transcriptomic profiling separates cancer from non-cancer cells. Advancements in single-cell techniques such as with the InDrop method[34] now enables simultaneous sequencing of thousands of single cells to examine the heterogeneity in cellular composition and state of entire samples. Such samples thus contain both cancer and non-cancer cells. Preliminary results (Fig. 7) show that methods such as PAGODA[18] can separate cancer and non-cancer cells based on transcriptional signatures while identified subpopulations can be then identified as either cancer or non-cancer based on inferred CNVs.

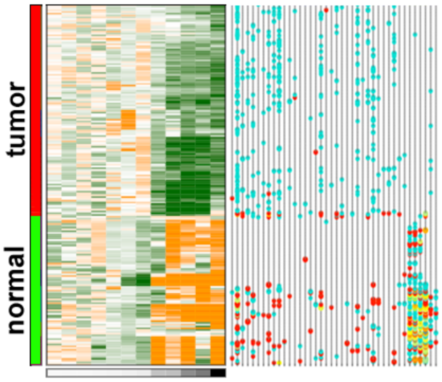


Figure 7. Separating tumor from normal cells. PAGODA analysis of 396 cells from a breast cancer sample (left) identifies two major transcriptional subpopulations. CNV profiling identifies multiple clonal deletion (representative region shown) affecting one subpopulation, which is then inferred to be tumor, while the other normal.

2.3.2. Research design: To identify dynamic changes in subpopulations that may shed mechanistic insight to clinical responses in CLL, I will examine single-cell RNA-seq data for CLL and non-CLL cells in multiple patients over multiple time points collected as part of separate research efforts. I will apply methods I've previously developed from Aims 1 and 2 to identify putative genetic subclones, and transcriptional subpopulations, cell types, and cell states. I will then apply SCDE[35] to identify significantly differentially expressed genes marking each subpopulation. To assess changes in frequency and proportion of subpopulations over time, I will train a support vector machine or other supervised machine learning algorithms on identified marker genes for each subpopulation identified at each time point to then connect the same putative subpopulation at the next time point. Visual inspection of marker genes will also be used to establish correspondence of subpopulations at different time points. Subpopulations may shift over time over time, particularly following therapy. For example, Ibrutinib, with its immunostimulatory effects, may further alter the cellular state of T cells to have less dampened function or in a manner different from chlorambucil exposure. Significantly differentially expressed genes for corresponding subpopulations at different time points will also be identified and validated by applying custom-designed quantitative PCR assays to cDNA derived from the originating samples. To identify potential changes in particular T or myeloid cell regulatory networks and pathways over time, I will apply a Bayesian network approach to derive causal influences among genes. Integration of prior knowledge through analysis of bulk and public data to establish prior expectations on the expected result of treatments on gene expression must be integrated for robustness of network reconstruction.

2.3.3. Potential problems and alternative solutions: Gene expression changes as a result of biologically functional subpopulation shifts must be teased apart from inherently confounding batch effects. Similarly, while novel subpopulations may also arise and expand over time, potential batch and quality differences may confound discovery. Therefore, I will analyze all datasets jointly with batch correction as well as independently in addition to validating findings using orthogonal approaches such as quantitative PCR to ensure that findings are not attributable to batch effects.

2.3.4. Skills and techniques to be acquired: Execution of this Aim will provide me with the experiences needed to start my career as an independent investigator, namely being able to obtain data, interact with a core facility for sequencing, work with collaborators, oversee finances, and mentor students, in addition to refining and practicing skills developed in Aims 1 and 2.

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++. Dr. Wu has also been extremely supportive and generous with her lab resources, providing me with the means to experimentally validate findings and obtain additional samples or sequencing as needed.

I completed background research and preliminary analyses with input from Dr. Kharchenko and Dr. Wu. I designed the structure and aims of this proposal with their feedback. I also wrote this research proposal and designed all figures.

I am submitting this research proposal with both Dr. Kharchenko 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.

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 written manuscripts together. I am confident that Dr. Kharchenko will be able to provide the mentorship and guidance I need to accomplish the methodological and analytical goals in this proposal.

Selection of Co-Sponsor

I collaborate extensively with Catherine Wu's lab to apply our developed methods to understanding chronic lymphocytic leukemia. Our collaboration has already led to many important biological findings and joint publications. Through my collaboration with Dr. Wu and others in her lab, I have gained exposure to bench techniques available for downstream validation as well as gained insights into the potential causes of batch effects at the experimental step. Dr. Wu has encouraged me to present at conferences, apply for travel awards, and has always been willing to introduce me to her colleagues and assist in my networking. Dr. Wu and I have also attended conferences and written manuscripts together. I am confident that Dr. Wu will be able to provide the mentorship and guidance I need to accomplish the biologically-driven goals in this proposal and ensure that I develop the skills necessary to eventually become an independent cancer investigator like herself.

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 Department of Biomedical Informatics (DBMI) 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 single-cell RNA sequencing.

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: 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. I have already taken Med Sci 300qc and will be taking Med Sci 302qc during the term of this fellowship. **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 this Conduct of Science course in the fall of 2013.

Personal: In addition, in my weekly meetings with my sponsor Dr. Peter Kharchenko, we discuss responsible conduct of research including the topics of authorship, ethics, confidentiality, and more. In addition, Dr. Wu and I discuss responsible conduct of research with respect to human subjects, vertebrate animals, patient privacy, and ethics. 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.

Institutional Environment and Commitment to Training

The F99-phase of the proposed research will be conducted at the Department of Biomedical Informatics (DBMI), Harvard Medical School (HMS) under the mentorship of Dr. Kharchenko as part of the Bioinformatics and Integrative Genomics (BIG) PhD program.

DBMI

DBMI is located in the Longwood Medical Area in Boston and is housed in the Francis A. Countway Medical Library of Medicine of HMS, which provides immediate access to more than 3,000 current journal titles and over 600,000 volumes in the biomedical field. The eight core faculty members of DBMI conduct research at the intersection of biomedicine and information sciences, including bioinformatics (e.g. Dr. Kohane, Dr. Park, Dr. Tonellato, Dr. Kharchenko, Dr. Wall), functional genomics (e.g. Dr. Park, Dr. Kharchenko, Dr. Alterovitz, Dr. Wall), translational medicine (Dr. Kohane, Dr. Tonellato), and clinical knowledge management (Dr. Kohane, Dr. McCray). The center also hosts the Bioinformatics & Integrative Genomics program (a graduate training program, sponsored by NHGRI, headed by co-directors Dr. Isaac Kohane and Dr. Peter Park) and the Biomedical Informatics Research Training program (a consortium of informatics laboratories at Harvard and MIT, sponsored by NLM).

BIG PhD Program

The interdepartmental Bioinformatics and Integrative (BIG) PhD program trains future leaders in the field of bioinformatics and genomics. Our mission is to provide BIG graduate students with the tools to conduct original research and the ability to develop novel approaches and new technologies to address fundamental biological questions.

The BIG program is administered by the Department of Biomedical Informatics at Harvard Medical School, located in Boston, Massachusetts. The students have a substantial flexibility in choosing their advisors--they can choose practically any faculty members who have a primary appointment at Harvard, in addition to the core faculty members listed on this site. Our core faculty is unparalleled in its record high-quality research in many areas of genomics and bioinformatics, both in basic science and translational research.

We believe that the students need to be experts not only in computational analysis but also have broad background in the biological sciences and expertise in an area of application, whether that is a basic biological phenomenon or a disease area. The students therefore take courses in genetics and molecular biology as well as a biological literature reading course, with graduate students in other programs at the medical school. For those interested in advanced quantitative courses, there are numerous courses at nearby institutions, e.g., Bayesian statistics at Harvard School of Public Health, machine learning at the Harvard main campus, or signal processing at MIT (10 minutes to MIT and 20 minutes to the main campus by a shuttle bus).

The Longwood Medical Area

HMS, its affiliated hospitals (Brigham and Women’s Hospital, Beth Israel Deaconess Hospital, Children’s Hospital), research institutes (Dana-Farber Cancer Institute, Broad Institute) and the Harvard School of Public Health adjacent to DBMI, have over 8000 full-time faculty members, covering most current areas of biomedical research. This provides unparalleled opportunities for collaboration on experimental and quantitative investigations of biological problems. In addition, a free shuttle bus links the area with MIT and Harvard Square in Cambridge.

All of the aforementioned institutions provide a stimulating scientific environment by offering an extensive program of courses and seminars that are open to members of the HMS community. These include Brigham and Women’s Translational Genetics Program Seminar, the Dana-Farber Center for Cancer Computational Biology Seminar, the Dana-Farber Biostatistics and Computational Biology Seminar Series and the Harvard School of Public Health Program in Quantitative Genomics Seminar. These seminars enjoy an excellent list of national and international speakers and are well attended by researchers from the Longwood Medical Area. These events offer excellent opportunities for gaining a deeper understanding of the relevant research problems and establishing further collaborations.

Protection of Human Subjects

1. **Risk to Human Subjects**
   a. Human Subjects Involvement and Characteristics: We propose to generate single cell transcriptome data from leukemia cells (either marrow or peripheral blood derived) of adult CLL patients, enrolled on clinical trials at DFCI or other collaborating centers (including MD Anderson Cancer Center, and centers associated with the CLL Research Consortium). IRB-approval for research performed on these samples is covered on DFHCC 2332-14 (PI Wu; Clonal heterogeneity and Evolution in CLL; approval pending). Consented individuals on these protocols provide extra samples of peripheral blood or marrow for research when they are undergoing collection at clinically indicated times for research studies including genomic analyses. No extra risks to subjects will be posed by the studies since they are performed on cryopreserved samples that were procured at clinically indicated times.
   b. Sources of Materials: Mononuclear cells from peripheral blood or marrow from patients with CLL will be used for the proposed studies. Since we will try to link clinical status (i.e. disease stage, history of prior therapy) with the results of analyses used in the proposed aims, clinical information will be linked with laboratory data. Tumor samples will therefore be de-identified for the purposes of laboratory studies, and identifying characteristics of patients will be known only to the principal and collaborating investigators, who will keep this information in a password protected database available only to key study staff, not including the principal investigator Jean Fan.
   c. Potential Risks: The only physical risks to subjects include the minimal risks associated with blood draw and marrow biopsy/aspirate (pain, bleeding, bruising, infection, and rarely, nerve damage).
2. **Adequacy of Protection Against Risks.**
   a. Recruitment and Informed consent. All subjects included under DFHCC protocol 2332-14 have CLL and have undergone informed consent during which they have agreed to allow surplus or unused banked material is available for research studies.
   b. Protection against risk. Because the proposed studies use banked material or extra material obtained at the time of clinically indicated sampling there is minimal risk to the human subject. Laboratory studies are undertaken using de-identified samples, and linked clinical information is known to only limited study staff, not including the principal investigator Jean Fan, and is recorded in a password-protected database.
3. **Potential Benefits of Proposed Research to Human Subjects and Others.** CLL is incurable by conventional chemotherapy. In particular, patients who have chemo-refractory disease, or who have received more than 3 prior regimens have poor prognosis. It has been estimated that patients who have failed to respond to fludarabine or other purine analogs have a median survival of 8 months, and a 1 year survival of 40% (Keating et al., 2002). The proposed studies to understand the impact of the subclonal structure of CLL on clinical outcome are anticipated to reveal insights into the basic biology of this disease and the mechanisms by which standard chemoimmunotherapy or targeted therapy directed against B cell receptor signaling affects disease evolution and the development of treatment relapse and resistance. Understanding the basis by which treatment relapse and resistance develops (for example, whether or not clonal dynamics and evolutionary selective pressures play a role) has great impact on the development of alternative potentially more efficacious therapies. This would include developing rational combinations that target those subclonal populations that are ‘destined’ to generate resistance, or possibly targeting subpopulations to diminish the impact of aggressive subclones while maintaining the ‘trunk.’ A further potential benefit of the proposed studies is the development of novel prognostic markers that can predict the composition and rapidity of relapse.
4. **Importance of the Knowledge to be Gained.** This study will contribute to greater understanding of the biology of CLL. As we are using cryopreserved samples or samples procured in real time at clinically indicated sampling times, sample collection poses only minimal risks to human subjects. As stated in the previous section, we anticipate that our findings will improve prognosis and development of treatment for patients with CLL.

Inclusion of Women and Minorities

This study does not exclude any individuals on the basis of gender. We will consider available leukemia samples from patients with CLL that are enrolled on clinical studies at DFCI or collaborating center. These studies also do not exclude any individuals on the basis of race or ethnicity.

Since CLL occurs with a slight male preponderance (according to the estimates of the American Cancer Society, approximately 8,100 new cases of CLL were diagnosed in 2000, 4,600 in men and 3,500 in women), it is anticipated a somewhat greater male enrollment compared to female enrollment. Within the United States, CLL affects African-Americans as frequently as it does Caucasians.

PHS Inclusion Enrollment Report

This report format should NOT be used for collecting data from study participants.

OMB Number:0925-0001 and 0925-0002
Expiration Date: 10/31/2018

***Study Title:** Statistical Methods for Characterizing Tumor Heterogeneity at the Single Cell Level

***Delayed Onset Study?** ☐ Yes ☑ No

If study is not delayed onset, the following selections are required:

Enrollment Type ☑ Planned ☐ Cumulative (Actual)

Using an Existing Dataset or Resource ☑ Yes ☐ No

Enrollment Location ☑ Domestic ☐ Foreign

Clinical Trial ☐ Yes ☑ No

NIH-Defined Phase III Clinical Trial ☐ Yes ☑ No

Comments:

Racial Categories	Ethnic Categories									Total
	Not Hispanic or Latino			Hispanic or Latino			Unknown/Not Reported Ethnicity			
	Female	Male	Unknown/ Not Reported	Female	Male	Unknown/ Not Reported	Female	Male	Unknown/ Not Reported	
American Indian/Alaska Native	0	0		0	0					0
Asian	0	0		0	0					0
Native Hawaiian or Other Pacific Islander	0	0		0	0					0
Black or African American	0	0		1	2					3
White	10	12		0	0					22
More than One Race	0	0		0	0					0
Unknown or Not Reported										
Total	10	12		1	2					25

Report 1 of 1

Inclusion of Children

CLL occurs only very rarely in children. For this reason, we do not anticipate the analysis of any samples collected from children.

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.

Myy 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*, Nature Methods 2016), pathway and gene set overdispersion analysis (Fan *et al*, Nature Methods 2016), 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.