



**The Faculty of Medicine of Harvard University
Format for the Curriculum Vitae**

Date Prepared: January 22, 2019
Name: Bo Li
Office Address: Broad Institute of MIT and Harvard
415 Main Street
Cambridge, MA 02142
Home Address: 5 Avalon Drive
Quincy, MA 02169
Work Phone: 608/609-3735
Work Email: libo@broadinstitute.org
Place of Birth: Shijiazhuang, Hebei Province, China

Education:

9/2004- 7/2008	BE	Computer Science and Engineering	Shanghai Jiao Tong University
9/2008- 5/2010	MS	Computer Sciences	University of Wisconsin- Madison
9/2008- 7/2013	PhD	Computer Sciences (Colin Dewey)	University of Wisconsin- Madison

Postdoctoral Training:

8/2013- 6/2017	Postdoctoral Researcher	Computational Biology (Lior Pachter)	University of California, Berkeley
7/2017-	Postdoctoral Associate	Computational Biology (Aviv Regev)	Broad Institute of MIT and Harvard

Appointments at Hospitals/Affiliated Institutions:

5/2019-	Research Staff (Assistant Investigator)	Center for Immunology and Inflammatory Diseases, Division of Rheumatology, Allergy and Immunology	Massachusetts General Hospital
---------	---	--	-----------------------------------

Other Professional Positions:

Major Administrative Leadership Positions:

Local

Regional

National

International

Committee Service:

Local

Regional

National

International

2016	Program Committee Member, RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-SEQ)	International Conference on Research in Computational Molecular Biology (RECOMB)
2016	Program Committee Member	International Joint Conferences on Artificial Intelligence

Professional Societies:

2012 & 2014	International Society for Computational Biology
----------------	---

Grant Review Activities:

Editorial Activities:**Ad hoc Reviewer**

BMC Bioinformatics

Bioinformatics

Nucleic Acids Research

Other Editorial Roles**Honors and Prizes:**

2004	First Place, Association for Computing Machinery - International Collegiate Programming (ACM-ICPC) Asia Programming Contest, Beijing Site	ACM-ICPC Committee	Programming Contest
2005 & 2006	OOCL Scholarship	Orient Overseas Container Line Corporation, China	
2005	Fifth Place, ACM-ICPC Asia Programming Contest, Taipei Site	ACM-ICPC Committee	Programming Contest
2005	Second Place, ACM-ICPC Asia Programming Contest, Beijing Site	ACM-ICPC Committee	Programming Contest
2006	Thirteenth Place, Google Code Jam, China	Google, China	Programming Contest
2007	State Scholarship	Ministry of Education, China	
2008	Alumni Scholarship	Computer Sciences Department, University of Wisconsin-Madison	
2010-2012	Morgridge Institute for Research Support for Senior Graduate Students in Computation and	Morgridge Institute for Research	Research

Informatics in Biology
and Medicine

2013	Finalist, Lane Fellows in Computational Biology	Carnegie Mellon University	Postdoctoral Fellowship
------	---	----------------------------	-------------------------

Report of Funded and Unfunded Projects

Funding Information:

Past

Current

Projects Submitted for Funding

Training Grants and Mentored Trainee Grants

Unfunded Current Projects

Report of Local Teaching and Training

Teaching of Students in Courses:

2006	Programming Practice of Computer Algorithms 2 nd year computer science students	Shanghai Jiao Tong University 1-hour sessions per week for 26 weeks
2007	Project Workshop of Compiler 2 nd year computer science students	Shanghai Jiao Tong University 1-hour sessions per month for 4 months
2008	Introduction to Bioinformatics	University of Wisconsin-Madison Teaching assistant, 1-hour office hour sessions per week for 14 weeks

senior undergraduates and 1st graduate students in computer science or biostatistics & medical informatics

Formal Teaching of Residents, Clinical Fellows and Research Fellows (post-docs):

Laboratory and Other Research Supervisory and Training Responsibilities:

Formally Mentored Harvard Medical, Dental and Graduate Students:

Other Mentored Trainees and Faculty:

12/2018- Yanay Rosen / Undergraduate student, University of California, Berkeley
1/2019
Career stage: research intern. *Mentoring role:* research intern mentor.
Accomplishments: Benchmarked run-time speeds of two single-cell RNA-Seq analysis tools, SCANPY and Seurat, on 1.3 million mouse brain single-cell data using Google Cloud.

Formal Teaching of Peers (e.g., CME and other continuing education courses):

No presentations below were sponsored by outside entities

Those presentations below sponsored by outside entities are so noted and the sponsor(s) is (are) identified.

Local Invited Presentations:

No presentations below were sponsored by outside entities

2015	RNA-Seq Transcript Quantification with RSEM: A Detailed Tutorial with Common Use Cases / Invited Speaker Core Skills in Computational Biology, Center for Computational Biology, University of California, Berkeley, Berkeley, CA
2016	PROBer: A General Toolkit for Analyzing Sequencing-based 'Toeprinting' Assays / Invited Speaker Center for RNA Systems Biology, University of California, Berkeley, Berkeley, CA

2018 How to build a successful human cell atlas? Lessons learned from 1.7 million single immune cells / Invited Speaker
Immunology Seminar Series, Massachusetts General Hospital, Boston, MA

Report of Regional, National and International Invited Teaching and Presentations

No presentations below were sponsored by outside entities

Regional

2016 Quantifying RNA Information from Transcriptome-wide Chemical Probing Experiments / Invited Speaker
Next Generation Sequencing (NGS) Data Analysis and Informatics Conference, San Diego, CA

National

2017 Taming Big Sequencing Data for RNA Biology: From Transcript Abundance Estimation to 'Epitranscriptomic' Mark Detection / Invited Speaker
Toyota Technological Institute at Chicago, Chicago, IL

2017 Taming Big Sequencing Data for RNA Biology: From Transcript Abundance Estimation to 'Epitranscriptomic' Mark Detection / Invited Speaker
Department of Computer Science, University of Illinois at Urbana-Champaign, IL

2017 Taming Big Sequencing Data for RNA Biology: From Transcript Abundance Estimation to 'Epitranscriptomic' Mark Detection / Invited Speaker
Department of Computer Science, Duke University, Durham, NC

International

2016 PROBer: A General Toolkit for Analyzing Sequencing-based 'Toeprinting' Assays (selected oral abstract)
21st Annual Meeting of the RNA Society, Kyoto, Japan

Report of Teaching and Education Innovations

Report of Technological and Other Scientific Innovations

Methods and compositions for multiplexing single cell and single nuclei sequencing	<p>US Patent Application, 62/770,580, filed December 7, 2018, pending</p> <p>As a member of the Regev lab at the Broad Institute of MIT and Harvard, my colleagues and I developed methods to multiplexing single cells and single nuclei from different samples experimentally and demultiplex sequencing data to sample-specific cells and nuclei computationally. These methods has</p>
--	--

the potential of significantly reducing the experimental cost of single cell and single nucleus RNA-Seq experiments internationally. My contribution is the development of computational methods for demultiplexing the pooled sequencing data.

Report of Education of Patients and Service to the Community

No presentations below were sponsored by outside entities

Those presentations below sponsored by outside entities are so noted and the sponsor(s) is (are) identified.

Activities

Educational Material for Patients and the Lay Community:

Books, monographs, articles and presentations in other media (e.g., video and websites, movies, television and radio) that educate the public about medicine, health or biomedical sciences

Educational material or curricula developed for non-professional students

Patient educational material

Recognition:

Report of Scholarship

Peer-Reviewed Scholarship in print or other media:

Research Investigations

1. **Li B**, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 2010; 26(4):493-500.
2. Chung D, Kuan PF, **Li B**, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keleş S. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput Biol.* 2011; 7(7):e1002111.
 - In Brief. *Nature Reviews Genetics.* 2011; 12:588.

3. **Li B**, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011; 12:323.
4. Haas BJ*, Papanicolaou A*, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, **Li B**, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013; 8(8):1494-512.
5. **Li B***, Fillmore N*, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN. Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. *Genome Biol*. 2014; 15(12):553.
6. Zeng X, **Li B**, Welch R, Rojo C, Zheng Y, Dewey CN, Keleş S. Perm-seq: Mapping Protein-DNA Interactions in Segmental Duplication and Highly Repetitive Regions of Genomes with Prior-Enhanced Read Mapping. *PLoS Comput Biol*. 2015; 11(10):e1004491.
7. Choudhary K, Shih NP, Deng F, Ledda M, **Li B**, Aviran S. Metrics for rapid quality control in RNA structure probing experiments. *Bioinformatics* 2016; 32(23):3575-3583.
8. **Li B**, Tambe A, Aviran S, Pachter L. PROBER Provides a General Toolkit for Analyzing Sequencing-Based Toeprinting Assays. *Cell Syst*. 2017; 4(5):568-574.

* contributed equally

Other peer-reviewed scholarship

Scholarship without named authorship

Non-peer reviewed scholarship in print or other media:

Proceedings of meetings or other non-peer reviewed scholarship

Reviews, chapters, monographs and editorials

Books/textbooks for the medical or scientific community

Case reports

Letters to the Editor

Professional educational materials or reports, in print or other media:

Thesis:

Li B. Computational analysis of RNA-Seq data in the absence of a known genome [dissertation].
Madison, Wisconsin: University of Wisconsin-Madison; July 2013.

Abstracts, Poster Presentations and Exhibits Presented at Professional Meetings:

Narrative Report

My area of Excellence is Investigation. My research focuses on developing reliable computational tools to empower novel analyses of various Next-Generation Sequencing (NGS) data, and in particular, single-cell genomics data.

My most significant accomplishment so far is the development of the world-renowned transcript abundance estimation software, RSEM. RSEM is the gold standard for quantifying gene expression levels from RNA sequencing (RNA-Seq) data, and RSEM papers have over 5,100 citations since 2010. RSEM is selected as the most preferable RNA-Seq expression quantification tool by ENCODE (ENCyclopedia Of DNA Elements) and TCGA (The Cancer Genome Atlas), which are nationally funded consortia aiming at creating standardized procedures for biology research. RSEM is also recommended by HCA (Human Cell Atlas) as a five-star single-cell RNA-Seq (scRNA-Seq) expression estimation tool to process plate-based SMART-Seq2 data.

I also developed the first reference-free de novo transcriptome assembly evaluator, RSEM-EVAL, that can select the best assembly without knowing the ground truth transcript sequences. RSEM-EVAL is used by Trinity, the most popular de novo transcriptome assembler, to evaluate its assembling results. I am experienced in analyzing a variety of NGS data besides RNA-Seq. I developed a general-purpose tool PROBER to analyze RNA structure, RNA modification, and RNA-protein interaction data. I also contributed to developing tools that utilize ChIP-Seq (protein-DNA interaction) and DNase-Seq (chromatin accessibility) data, allowing the detection of transcription factor binding sites.

With respect to single-cell genomics, I have developed scCloud, a cloud-based single-cell data analysis pipeline to scale up single-cell data analysis to millions of cells. I am also the computational lead of the human immune cell atlas project, in which we have profiled 1.7 million single cells from a variety of human tissue types, such as bone marrow, cord blood and peripheral blood. Together with my colleague, Jellert Gaublomme, we developed nucleus-hashing, a novel single-nuclei multiplexing protocol and demuxEM, a novel demultiplexing algorithm.

As a new faculty member of The Center for Immunology and Inflammatory Diseases (CIID), my research will focus on developing standard operating procedures (SOPs) for single-cell genomics data analysis, building on state-of-the-art big data, machine learning and graphics processing unit (GPU) computing technologies. For example, I will work on re-engineering my previously developed RSEM software to handle multi-mapping reads for droplet-based scRNA-Seq data. Upon finish, my new RSEM will enable the single-cell genomics community to better utilize their single-cell sequencing data. I will also work on developing novel GPU-based read sequence aligners that are substantially faster than existing CPU-based aligners, and training novel deep learning architectures to identify immune cell types automatically based on gene expression profiles for every single cell. Besides developing novel tools, I will also help CIID to set up the computing infrastructure to fulfill CIID faculty's needs on large scale genomics data analyses. Collaborating closely with immunologists, I and other faculty at CIID will

work together towards answering key immunology questions by leveraging a wide-range of computational tools my lab has developed.

As an assistant professor of medicine at Harvard Medical School, I have strong interest in teaching courses related to computational biology / bioinformatics at either upper division undergraduate or graduate level. As an instructor for bioinformatics courses, I will supplement the classic materials such as sequence alignment and phylogeny with my research expertise, e.g., I will introduce up-to-date algorithms for analyzing next generation sequencing data (ChIP-Seq, RNA-Seq) and single-cell genomics data (scRNA-Seq, CITE-Seq). My educational goal is to help students 1) understand that research on biology has stepped into the “big data” era; 2) learn essential computer science and statistics skills for biological data analysis; 3) acquire abilities solving practical bioinformatics problems in their own research. To achieve my educational goal, I will 1) tailor course contents according to students’ backgrounds; 2) emphasize on course project and undergraduate research; 3) listen to students and adapt teaching strategies to the needs of students accordingly.