BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Li C. Xia, Ph.D

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

POSITION TITLE: Instructor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Fudan University (Shanghai, China)	BS	06/2003	Electronics Engineering
Fudan University (Shanghai, China)	MS	06/2006	Physics
University of Southern California (Los Angeles, CA)	MS	12/2008	Computer Science
University of Southern California (Los Angeles, CA)	MS	12/2012	Statistics
University of Southern California (Los Angeles, CA)	PhD	05/2013	Bioinformatics Computational Biology
Stanford University (Stanford, CA)	Postdoctoral Researcher	04/2018	Genomics and Precision Medicine

A. Personal Statement

I am an *Instructor of Medicine* at Stanford University. My research focus is on developing and applying statistical and algorithmic methods to model, analyze and make inferences from large-scale genomics data. My current research involves analyzing somatic structural aberrations, human microbiota and immune signatures to understand their roles in cancer. I have published many open source bioinformatics tools, *e.g. SWAN* **[a]**, *ZoomX* **[b]**, *SVEngine* **[c]** and *CoreProbe* **[d]**, for characterizing structural aberrations and microbiota in low frequency cancer genome mixtures. Prior to 2011, I wrote under the name Li Xia. Since then, I have been using the middle name 'Charlie' as abbreviated in **Li C. Xia** to reduce name collisions. In the citations below, my peer-reviewed publications were sorted in the order that they were cited. A ***** next to my name denotes my role as the corresponding author for the publication.

a) **Li C. Xia**, S Sakshuwong, ES Hopmans, JM Bell, SM Grimes, DO Siegmund, HP Ji, NR Zhang. A genome-wide approach for detecting novel insertion-deletion variants of mid-range size. *Nucleic Acids Research* (**2016**) 44(15): e126 [PMCID: PMC5009736]
b) **Li C. Xia**, J Bell, C Wood- Bouwens, J Chen, NR Zhang, HP Ji. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Research* (**2018**) 46(4): e19 [PMCID: PMC5829571]
c) **Li C. Xia**, D Ai, H Lee, N Andor, C Li, NR Zhang, HP Ji. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *GigaScience*, 7, giy081, (**2018**) [PMCID: PMC6057526]
d) D Ai*, H Pan, R Huang, **Li C. Xia***. *CoreProbe*: A novel algorithm for estimating relative abundance based on metagenomic reads. *Genes* (**2018**) 9 (7), 314 (***corresponding**)

B. Positions and Honors

Positions and Employment

2018- Instructor, Division of Oncology, Department of Medicine, Stanford University
2018 Research Scientist, Division of Oncology, Department of Medicine, Stanford University
2013-2018 Postdoc Scholar, Division of Oncology, Department of Medicine, Stanford University
2013- Visiting Scholar, Department of Statistics, University of Pennsylvania
2012 Statistical Consultant, School of Social Work, University of Southern California
2005 Research Internship, IBM

Other Experience and Professional Memberships

2015 Co-chair, COMMANMD Workshop at the IEEE's Bioinformatics and Biomedicine Conf. (BIBM)
2009- Member, International Society for Computational Biology
2012- Member, American Statistical Association
2015- Associate Member, American Association for Cancer Research
2015- Member, American Society of Human Genetics
2010- Reviewer for journals: *Plos Medicine, Bioinformatics, Briefings in Bioinformatics, Database, Plos One, BMC Bioinformatics, BMC Res Notes, Frontiers in Microbiology, Stat Methods Med Res, Peerj, Stat Appl Genet Mol Biol, Comp Biol in Medicine, Evol Bioinform, Clinical Epidemiology*
2009- Reviewer for meetings: *IEEE Int'l Conference on Bioinformatics and Biomedicine (BIBM), Asian Pacific Bioinformatics Conference (APBC), Conference on Research in Computational Molecular Biology (RECOMB), Intelligent Systems for Molecular Biology (ISMB)*

Honors

2018 American Cancer Society (ACS) Postdoc Fellow
2018 'Get Your Rear in Gear Philadelphia' Scholar-in-Training Awards (AACR)
2017 Finalist Award for Pacific Biosciences 'Open Your Eyes to Isoform Diversity' Challenge
2016 Travel Fellowship for Alzheimer's Association International Conference (AAIC)
2015 Reviewer's Choice Best Abstract of American Society of Human Genetics Meeting (ASHG)
2014 Travel Award for Bayer International Computational Biology Workshop
2012 University of Southern California Dissertation Year Fellowship
2006-07 University of Southern California Merit Fellowship

C. Contribution to Science

i. ***Novel statistical algorithms for analyzing structural aberration and genetic heterogeneity.***
It has become increasingly known that structural variation and genetic heterogeneity contribute significantly to the susceptibility to and the development of many diseases. The lack of precise tools to detect and characterize structural variants has dragged our understanding of them behind. Addressing those deficiencies, my work at Stanford had resulted in several new bioinformatics tools for analyzing structural variations **[1]**, somatic rearrangements **[2]** and tumor heterogeneity **[3,4]**. They are applicable to other diseases, such as the Alzheimer's. For instance, I developed the Statistical Structural Variants Analysis for NGS (***SWAN***), the first structural variant detection tool to statistically integrates read-depth, read-pair and split-read signals. ***SWAN*** significantly improved the accuracy of detecting small- to mid-scale (<10Kb) structural variants in low frequency genomic mixtures **[1]**. ***SWAN*** won the Reviewer's Choice Best Abstract for the ASHG meeting. It was adopted by large-scale disease sequencing projects, e.g., the Alzheimer's Disease Sequencing Project (ADSP). I was invited to present the ***SWAN*** analysis of the ADSP data in a panel talk at the AAIC meeting. I then developed a novel structural variant caller named ***ZoomX*** for linked-read sequencing data. Utilizing the unique high weight molecule identifier within the data, ***ZoomX*** achieved significantly better performance in detecting, genotyping and phasing complex rearrangements **[2]**. I was invited then gave a oral presentation of ***ZoomX*** at the ASHG meeting.

[1] **Li C. Xia**, S Sakshuwong, ES Hopmans, JM Bell, SM Grimes, DO Siegmund, HP Ji, NR Zhang. A genome-wide approach for detecting novel insertion-deletion variants of mid-range size. *Nucleic Acids Research* (**2016**) 44(15): e126 [PMCID: PMC5009736]

[2] **Li C. Xia**, J Bell, C Wood- Bouwens, J Chen, NR Zhang, HP Ji. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Research* (**2018**) 46(4): e19 [PMCID: PMC5829571]

[3] N Andor, T Graham, M Jansen, **Li C. Xia**, C Aktipis, C Petritsch, H Ji, C Maley. Pan-cancer analysis of the extent and consequences of intra-tumor heterogeneity. *Nature Medicine* (**2016**) 22:105-113 [PMCID: PMC4830693]
[4] JM Bell, BT Lau, SU Greer, C Wood, **Li C. Xia**, ID Connolly, MH Gephart, HP Ji. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Research* (**2017**) 45(19): e162 [PMCID: PMC5737808]

ii. ***Novel computational and statistical methods for modeling large scale multiomics data.*** Genomics data generated from various high-throughput technologies are typically high dimensional. High dimensional data trouble statistical analyses with high false positive rate, slow computation and other problems. I have been working to address those difficulties since my graduate school. The work has resulted in several useful tools. I developed an efficient and accurate asymptotic p-value method based on random walk theories that replaced slow permutation procedure and enabled local similarity network analysis for tens of thousands of genes **[5]**. I developed an efficient asymptotic p-value method based on Markov chain theories that enabled such analysis for large-scale discretized data **[6]**. With Drs Siegmund, Yakir and Zhang, I developed a probabilistic framework to control for false discovery rate (q-value) in *Poisson* random field scans **[7]**, which sped up a wide range of bioinformatics tools used for microarray, RNA-seq, whole genome and exome sequencing data analyses. I also collaboratively developed an efficient data mining tools for pattern mining in proteomics networks **[8]**.

[5] **Li C. Xia**, D Ai, J Cram, J Fuhrman, F Sun. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* (**2013**) 29(2): 230-237 [PMCID: PMC4990825]
[6] **Li C. Xia**, D Ai, J Cram, J Fuhrman, F Sun. Statistical significance approximation in local trend analysis of high-throughput time-series data using the theory of Markov chains. *BMC Bioinformatics* (**2015**) 16:301 [PMCID: PMC4578688]
[7] NR Zhang, B Yakir, **Li C. Xia,** D Siegmund. Scan statistics on Poisson random fields with applications in genomics. *Annals of Applied Statistics* (**2016**) 10 (2) 726-755 [doi: 10.1214/15-AOAS892]
[8] S Zhang, YJ Li, **Li C. Xia**, Q Pan. PPLook: an automated data-mining tool for protein-protein interaction. *BMC Bioinformatics* (**2010**) 11:326 [PMCID: PMC2906489]

iii. ***Statistical modeling of next generation sequence data***.
Statistical modeling of next generation sequencing (NGS) data is at the heart of many genome technologies. I have contributed several tools to this field, addressing the fundamental probabilistic nature of NGS short read, and real-world DNA and protein sequences. I developed *SVEngine* a sequence data simulator, which has locus-specific allelic-fraction control mimicking cancer clonal evolution **[9]**. I developed a *Dirichlet* model and *Gibbs* sampling-based Markov Chain Monte Carlo (MCMC) method ***CoreProbe***, which adjusts for NGS mapping ambiguity to reach efficient and accurate microbiome abundance estimation **[10]**. In earlier years, I have studied sequence unique reconstruction - a math question abstracted from the sequencing-by-hybridization technology. It asks if one will be able to recover the original biological sequence by piecing together the k-mer words derived from consecutive sliding windows. I developed a program that finds the critical value k-mer size, above which the sequences become mostly reconstrutable. The result was described in my first first-author publication **[11]**. I also studied taxon-specific sequences across microbial genomes, which became a useful way to distinguish specific strains from peer species in a community **[12]**.

[9] **Li C. Xia**, D Ai, H Lee, N Andor, C Li, NR Zhang, HP Ji. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *GigaScience*, 7, giy081, (**2018**)
[10] D Ai*, H Pan, R Huang, **Li C. Xia***. *CoreProbe*: A novel algorithm for estimating relative abundance based on metagenomic reads. *Genes* (**2018**) 9 (7), 314 (***corresponding**)
[11] **Li C. Xia**, C Zhou. Phase transition in sequence unique reconstruction. *Journal of Systems Science and Complexity* (**2007**) 20 (1), 18-29
[12] PA He, **L C. Xia.** Oligonucleotide profiling for discriminating bacteria in bacterial communities. *Combinatorial Chemistry & High Throughput Screening* (**2007**) 10(4), 247-255

iv. ***Novel computational and statistical methods for analyzing microbiome data.*** I also worked to address the emergent need for accurate and efficient methods for analyzing microbiome data. Human microbiota

represents a diverse mixture of symbiotic species and many are obedient to lab culture. Metagenomics is a revolutionary approach to capture the entire diversity using shotgun sequencing. However, faithfully recover species composition from ambiguous short read alignments presented a critical roadblock for metagenomics. In response to that, I developed ***GRAMMy***, which was the first tool to introduce statistical mixture modeling to metagenomics data analysis **[13]**. Using an *Expectation-Maximization* based algorithm, ***GRAMMy*** accurately estimates species relative abundance. I also dealt with the challenge to identify species-species interactions using the time series data of metagenomic samples. I developed the Extended Local Similarity Analysis (***ELSA***) tool **[14]** – a novel method for co-occurrence or -expression network analysis. ***ELSA*** is exceptional in capturing time-dependent (such as delayed or interval restricted) associations which elude all other methods. Both ***GRAMMy*** and ***ELSA*** were provided as open source software and received great acceptance. I recently integrated triplet liquid association analysis into the ***ELSA*** tool that enables the discovery of association mediators in microbial community **[15]**. Because of my contributions, I was invited to co-author two chapters in *Springer*'s *Encyclopedia of Metagenomics*, a staple reference book for metagenomics research. Collaboratively, I also made several other contributions such as in quantitatively delineating microsatellite instability **[16]**.

[13] **Li C. Xia**, JA Cram, T Chen, JA Fuhrman, F Sun. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One* (**2011**) 6(12): e27992 [PMCID: PMC3232206]

[14] **Li C. Xia**, JA Steele, JA Cram, ZG Cardon, SL Simmons, JJ Vallino, JA Fuhrman, F Sun. Extended local similarity analysis of microbial community and other time series data with replicates. *BMC Systems Biology* (**2011**) 5: S15 [PMCID: PMC3287481]

[15] D Ai, X Li, H Pan, **Li C. Xia***. Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. *BMC Genomics* (**2018**) [Accepted] (***corresponding**)

[16] GW Shin, SM Grimes, HJ Lee, BT Lau, **Li C. Xia**, HP Ji. CRISPR–Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nature Communications* (**2017**) 8:14291 [PMCID: PMC5309709]

v. ***New scientific insights gained from integrative analysis of genomics big data.*** My work included aggregating and analyzing large-scale genomics data to derive new scientific knowledge. For instance, pathologists have been debating multiple etiology models of periodontitis and there were oral cavity metagenomics data supporting of very different models. To our better understanding, I integrated many datasets and perform a thorough bioinformatics and statistical analysis using ***GRAMMy*** and logistic regression models. I was able to first describe an association between periodontitis and oral microbiome loss of diversity **[17]**. I also identified the pathogen pivoting model as the most probable one from competing models. My collaborative works have resulted in many other scientific discoveries. Working with Dr. Jed Fuhrman's marine lab, we used ***ELSA*** to perform a cross-depth analysis of marine community and found diverse partner-switching mechanisms of microbes in response to nutrients, temperature and light **[18]**. We used ***ELSA*** to integrative analyze bacteria, eukaryotes, archae species data and was the first describe cross-kingdom taxonomical and functional associations in the sea **[19]**. Those works received significant attention from the International Society of Microbial Ecology and its official journal *The ISME Journal*. A recent collaborative benchmark study of network analysis tools by a leading microbiome researcher, Dr. Rob Knight, ranked ELSA the best method in almost all analysis scenarios **[20]**.

[17] D Ai, R Huang, J Wen, C Li, J Zhu, **Li C. Xia***. Integrated metagenomic data analysis demonstrates that a loss of diversity in oral microbiota is associated with periodontitis. *BMC Genomics* (**2017**) 18:1041 (***corresponding**) [PMCID: PMC5310281]

[18] JA Cram, **Li C. Xia**, DM Needham, R Sachdeva, F Sun, JA Fuhrman. Cross-depth analysis of marine bacterial networks suggests downward propagation of temporal changes *The ISME Journal* (**2015**) 9(12): 2573-86 [PMCID: PMC4817623]

[19] JA Steele, PD Countway, **Li C. Xia**, … (11 authors), F Sun, DA Caron, JA Fuhrman. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal* (**2011**) 5:1414-1425 [PMCID: PMC3160682]

[20] S Weiss, VW Treuren, C Lozupone, K Faust, J Friedman, Y Deng, **Li C. Xia**, Z Xu, L Ursell, E Alm, A Birmingham, J Cram, J Fuhrman, J Raes, F Sun, J Zhou, R Knight. Correlation detection strategies in microbial datasets vary widely in sensitivity and precision. *The ISME Journal* (**2016**) 10(7): 1669-1681 [PMCID: PMC4918442]

The link below has the full list of my 25 peer-reviewed journal papers and books chapters:
http://www.ncbi.nlm.nih.gov/sites/myncbi/1JkzpAIoc50QD/bibliography/44867884/public

D. Research Support

ACS 132922-PF-18-184-01-TBG *"Postdoc Fellowship"* ***Role: PI*** *current*
Dr. Xia is the PI of this grant. Dr. Xia will develop novel statistical and algorithmic methods to characterize the genomic determinants of immune-escaping tumors. Dr. Xia expect to deliver open source tools and database that facilitate single-molecule and single-cell based analysis of tumor and microenvironmental cellular mixtures.

NIH R01 HG006137-07 *"Genomic and Cellular Variation from Single Molecules to Single Cells" (Co-PI: Nancy Zhang at U Penn and Hanlee Ji at Stanford)* ***Role: Key Person*** *current*
Dr. Xia is a key person of this grant. Dr. Xia developed *ZoomX* and *SVEngine*, new statistical methods that enable accurate characterization of cellular mixtures exhibiting both DNA and RNA variations. Dr. Xia expect to develop new bioinformatics tools for single cell genomics and they will be released as open source software.

NIH R01 HG006137-04 *"Statistical Models and Analysis of Complex Genomic Variation in Clonal Mixtures" (Co-PI: Nancy Zhang at U Penn and Hanlee Ji at Stanford)* ***Role: Key Person*** *2014/7/1 - 2017/6/30*
Dr. Xia was a key person of this grant. Dr. Xia worked closely with Drs. Zhang and Ji and developed SWAN a robust, sensitive statistical procedures to delineate complex variations such as genomic rearrangements and other structural variations in cancer clonal populations.