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: Mark Chaisson

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

POSITION TITLE: Assistant Professor

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
University of California, San Diego, La Jolla, CA	B.S.	5/2000	Computer Science
University of California, San Diego, La Jolla, CA	Ph.D.	10/2008	Bioinformatics
University of Washington, Seattle, WA	PostDoc	8/2017	Genetics

A. Personal Statement

My research focuses on method development and genetic analysis using novel sequencing types, including developing methods for *de novo* assembly, and detecting structural variation (SV). This research has been enabled by my training in Computer Science and Bioinformatics with Dr. Pavel Pevzner at the University of California, San Diego, and in Genetics with Dr. Evan Eichler at the University of Washington. As a Ph.D. student, I developed the first de Bruijn assembler for short read data, which became the dominant paradigm for short-read assembly. Next, as an employee of Pacific Biosciences, I developed and published the first method to rapidly align the long, noisy single-molecule sequencing (SMS) reads. Finally, as a postdoctoral scholar I studied human SV that may be detected using SMS, demonstrating that an order of magnitude more variation may be detected in this approach than using typical short-read based studies. This work has been followed by the Human Genome Structural Variation Consortium (HGSVC), where as a postdoc I have been leading research cataloging SV in three trios of diverse descent using a set of six diverse sequencing platforms. The HGSVC comprises of over 80 researchers worldwide, and my role in organizing such an effort makes me uniquely suited to continue this research while starting my independent research program at the University of Southern California.

B. Positions and Honors

1996-2001 B.S. University of California, San Diego, La Jolla, CA
2001-2002 Employee, Q.E.D. Labs, San Jose, CA
2002-2008 Ph.D. University of California, San Diego, La Jolla, CA
2008-2009 Postdoctoral scholar, University of California, San Diego, La Jolla, CA
2009-2012 Sr. Algorithms Engineer, Pacific Biosciences, Menlo Park, CA
2012-2017 Postdoctoral Scholar, University of Washington, Seattle, WA
2017- present Assistant Professor, University of Southern California, Los Angeles, CA

C. Contributions to Science

1. My Ph.D. research focused on developing methods for *de novo* assembly of early short-read sequencing data. The then new types of sequences produced by Illumina and 454 Life Sciences were too short (22 nt) to be assembled using the overlap-layout-consensus assembly techniques developed for Sanger reads, and too error prone to use existing de Bruijn based methods for assembly. To enable de Bruijn based assembly, I developed novel error correction routines for preprocessing reads and improving assembly quality and enabled assembly of genes from bacterial and viral genomes. Importantly, nearly all other algorithmic developments by other researchers for short read assembly followed this algorithmic framework.

    a. Medvedev, P.; Pham, S.; Chaisson, M.; Tesler, G.; Pevzner, P.; Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers, *Journal of Computational Biology* 18,11, 1625-1634, 2011.
    b. Chaisson, M.J.; Brinza, D.; Pevzner, P.A.; De novo fragment assembly with short mate-paired reads: Does the read length matter *Genome Research* 19, 2, 336-346, 2009.
    c. Chaisson, M.J.; Pevzner, P..A.; Short read fragment assembly of bacterial genomes, *Genome Research* 18,2,324-330,2008.
    d. Chaisson, M.; Pevzner, P.; Tang, H.; Fragment assembly with short reads, *Bioinformatics* 20, 13, 2067-2074, 2004.

2. As an employee of Pacific Biosciences, I developed a method for mapping long and noisy single-molecule sequencing reads to genomes. Similar to the dilemma facing assembly of short-read sequencing data, existing methods for mapping these reads were not computationally efficient enough to scale to large data sizes, or were designed for short and accurate Illumina reads, and had low mapping sensitivity. When publishing this method, I also included a study of the empirical and theoretical distribution of exact matching sequences in these long and noisy reads. The mapping method, BLASR, has been in widespread use by investigators, and the study on the distribution of exact matches has been referenced in the development of overlap algorithms for two SMS *de novo* assembly methods: MHAP (Berlin et al. 2015) and daligner (Myers, 2014).

    a. Chaisson, M.J.; Tesler, G.P.; Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory, *BMC Bioinformatics*, 13, 1, 238, 2012.

3. During my postdoctoral studies, I developed methods to detect SV using SMS and applied these methods to detect variation in haploid human genomes, between humans and gorillas. To detect SV in SMS sequences are assembled either after aligning to a reference, for detecting SV in humans, or *de novo*, for detecting SV between humans and gorillas. SV is detected by mapping the assembled sequences back to the human reference. In humans, this approach yields an order of magnitude more variants in a typical SMS genome than a short-read sequenced genome; 25-30 thousand versus 4-6 thousand, with most of the gain as small-scale variation between 50 and 2000 basepairs. Because the variation is detected using assembled sequences, the breakpoints are accurate to the limit of the consensus sequence of the assembly, typically > 99.9% accurate. This enables the short-read alignment based genotyping of SMS-detected variants in large groups at a two-fold increase in sensitivity for deletion, and completely enabling insertion genotyping relative to previous methods.

    I have also developed a novel approach to *de novo* assembly of SMS sequences for resolving long and highly identical sequences known as segmental duplications. These duplications are longer than most SMS reads making the overlaps between reads from paralogous copies of a duplication indistinct. I have shown that while duplications are highly identical, they are not exact, and the duplication content

may be disentangled using polyploid phasing.

a. Chaisson, M.J.P. ; Huddleston, J.; Dennis, M.Y.; Sudmant, P.H.; Malig, M., *et al.*; Resolving the complexity of the human genome using single-molecule sequencing, *Nature* 517, 7536, 608-611, 2015
b. Chaisson, M.J., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, *et al*., 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communication*s, 10.
c. Vollger, M.R., Dishuck, P.C., Sorensen, M., Welch, A.M. Dang, V., *et al*., 2019. Long-read sequence and assembly of segmental duplications. *Nature methods*, 16(1), p.88.
d. Chaisson, M.J., Mukherjee, S., Kannan, S., Eichler, E.E.; Resolving multicopy duplications de novo using polyploid phasing." *International Conference on Research in Computational Molecular Biology*. Springer, Cham, 2017.

A complete list of publications is available at http://chaissonlab.usc.edu/ .

D. Additional Information: Research Support and/or Scholastic Performance