Li Song

https://mourisl.github.io/

Li.Song@dartmouth.edu

EDUCATION

2012-2018: Ph.D., Computer Science, Johns Hopkins University (GPA: 4.0/4.0. Advisor: Liliana Florea.
Thesis committee: Liliana Florea, Ben Langmead, Sarven Sabunciyan)

2016-2017: M.Se., Applied Mathematics and Statistics, Johns Hopkins University (GPA: 4.0/4.0)

2009-2011: M.Sc, Computer Science, Michigan Technological University (GPA: 3.95/4.0)

2005-2009: B.E., Computer Science and Technology, Tongji University, China (Excellent Graduate)

PROFESSIONAL EXPERIENCE

2022-Present: Assistant Professor
Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College

2018-2022: Research Fellow
X. Shirley Liu Lab, Department of Data Science, Dana-Farber Cancer Institute

2021-2022: Research Fellow
Heng Li Lab, Department of Data Science, Dana-Farber Cancer Institute

RESEARCH INTERESTS

Development of algorithms and software tools for analyzing next generation sequencing data; immune receptor analysis; design and analysis of algorithms; parallel computing.

SOFTWARE

TRUST4: TCR/BCR assembler for RNA-seq data. TRUST4 can be applied on either bulk or single-cell RNA-seq data. In addition to report CDR3s, TRUST4 also assembles full-length TCRs/BCRs. [Github]

T1K: Genotyper for highly polymorphic genes including KIR and HLA. T1K is versatile and works with RNA-seq, WGS and WES data. T1K also identifies novel SNPs and is compatible with single-cell RNA-seq data. [Github]

CLASS/CLASS2: Efficient and accurate transcript assemblers for RNA-seq data that detect more fine-grained alternative splice variants. The programs combine linear programming algorithms to detect exons from read coverage levels, with splice graph representations of genes and their splice variants, and memory efficient optimization algorithms for transcript selection. [SourceForge]

PsiCLASS: Simultaneous multi-sample transcript assembler for RNA-seq data. It builds a global data structure representing the structure of the transcripts, from which each sample generates its expressed transcripts. The global information allows accurate sample-wise assemblies and final meta-assembly. [Github]

Lighter: Fast and memory-efficient k-mer-based software to correct the sequencing errors from whole genome sequencing data without counting. It samples the k-mers in the data set and uses two memory-efficient Bloom filters to obtain solid k-mers. [Github]

Rcorrector: Efficient and accurate k-mer-based error correction software for Illumina RNA-seq reads. It can also be applied to data sets where the read coverage is non-uniform, such as single-cell sequencing. [Github]

Rascaf: Scaffolding with RNA-seq read alignment. It uses information from paired-end and split reads to improve the completeness and contiguity of a draft genome assembly, particularly in the gene regions.

[Github]

Centrifuge: Fast and memory-efficient classifier for metagenomics sequences using an FM-index. It requires only 4.2 Gb memory for a database containing ~4300 prokaryotic genomes. [Github]

Chromap: Ultrafast alignment and preprocessing for chromatin profiling sequencing data, including ChIP-seq, ATAC-seq and Hi-C. It supports both bulk and single-cell platforms, and is more than 10 times faster than traditional workflows without sacrificing alignment accuracy. [Github]

JOURNAL PUBLICATIONS

Song L, Bai G, Liu XS, Li B and Li H, *Efficient and accurate KIR and HLA genotyping with massively parallel sequencing data*, Genome Res. 2023 May 11;gr.277585.122. [PubMed]

Yang L, Wang J, Altreuter J, Jhaveri A, …, **Song L**, …, Liu Y, Liu XS, *Tutorial: integrative computational analysis of bulk RNA-sequencing data to characterize tumor immunity using RIMA*. Nat Protoc. 2023 Jun 30. [PubMed]

Song L, Ouyang Z, Cohen D, Yang C, …, Liu XS, *Comprehensive characterizations of immune receptor repertoire in tumors and cancer immunotherapy studies*, Cancer Immunol Res. 2022 Jul 1;10(7):788-799. [PubMed] [Spotlight of the journal]

Song L, Cohen D, Ouyang Z, Cao Y, Hu X, and Liu XS, *TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data*. Nat Methods. 2021 Jun;18(6):627-630. [PubMed]

Zhang H, **Song L***, …, Liu XS, Li H, Fast alignment and preprocessing of chromatin profiles with Chromap. Nat Commun. 2021 Nov 12;12(1):6566. [PubMed] (*co-first author)

Song L, Sabunciyan S, Yang G and Florea L, *A multi-sample approach increases the accuracy of transcript assembly*. Nat Commun. 2019 Nov 1;10(1):5000. [PubMed]

Zhang J, Hu X, …, **Song L,** …, Liu XS, *Immune receptor repertoires in pediatric and adult acute myeloid leukemia.* Genome Med. 2019 Nov 26;11(1):73. [PubMed]

You Y, **Song L**, Nonyane BAS, Price LB and Silbergeld EK, *Genomic differences between nasal Staphylococcus aureus from hog slaughterhouse workers and their communities*, PLoS One. 2018 Mar 6;13(3):e0193820. [PubMed]

Miller JR, Zhou P, Mudge J, …, **Song L**, ..., Silverstein KAT, *Hybrid assembly with long and short reads improves discovery of gene family expansions*, BMC Genomics. 2017;18(1):541. [PubMed]

Song L, Sabunciyan S and Florea L, *CLASS2: accurate and efficient splice variant annotation from RNA-seq reads,* Nucleic Acids Res. 2016;44(10):e98. [PubMed]

Song L, Shankar D and Florea L, *Rascaf: improving genome assembly with RNA-seq data,* Plant Genome. 2016;9(3). [PubMed]

Kim D, **Song L***, Breitweiser FP and Salzberg SL, *Centrifuge: rapid and sensitive classification of metagenomic sequences,* Genome Res. 2016; 26(12): 1721–1729. [PubMed] (*co-first author)

Song L and Florea L, *Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads*, Gigascience. 2015;4:48. [PubMed]

Song L, Florea L and Langmead B, *Lighter: fast and memory-efficient sequencing error correction without Counting*, Genome Biol. 2014;15(11):509. [PubMed]

Song L and Florea L, *CLASS: constrained transcript assembly of RNA-seq reads*, Third Annual RECOMB Satellite Workshop on Massively Parallel Sequencing - RECOMB-SEQ 2013, BMC Bioinformatics 14(Suppl 5):S14. [PubMed]

Florea L, **Song L** and Salzberg SL, *Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues*, F1000 Research 2013, 2:188. [Full text]

CONFERENCE PARTICIPATION

Song L, Bai G, Liu XS, Li B and Li H, *T1K: efficient and accurate KIR and HLA genotyping with next-generation sequencing data*, Cold Spring Harbor Laboratory Meeting - Biology of Genomes 2023, Cold Spring Harbor, NY. (Poster)

Song L, Hu X, Zhang J, …, Liu J, Li B, Liu XS, *Characterize Tumor Infiltrating Immune Repositories with TRUST*, ITCR, 2019, Park City, UT. (talk, poster)

Song L and Florea L, *ClassX—scalable simultaneous transcript assembly of multiple RNA-seq data sets*, Cold Spring Harbor Laboratory Meeting - Genome Informatics 2017, Cold Spring Harbor, NY. (poster)

Song L and Florea L, *Rascaf: improving genome assembly with RNA-seq data*, Plant and Animal Genome Conference - PAG XXIV, 2016, San Diego, CA. (poster)

Song L, Langmead B and Florea L, *Lighter and Rcorrector: tools for next generation sequencing error correction*, Plant and Animal Genome Conference - PAG XXIV, 2016, San Diego, CA. (poster)

Song L, Langmead B and Florea L, *Lighter and Rcorrector: a suite for next generation sequencing error correction*, Cold Spring Harbor Laboratory Meeting - Biology of Genomes 2015, Cold Spring Harbor, NY. (poster)

Kim D, **Song L**, Breitweiser FP and Salzberg SL, *Centrifuge: rapid and accurate classification of metagenomic sequence*, Cold Spring Harbor Laboratory Meeting - Biology of Genomes 2014, Cold Spring Harbor, NY. (poster)

Song L, and Florea L, *Rcorrector: error correction for Illumina RNA-seq reads*, Cold Spring Harbor Laboratory Meeting - Biological Data Science 2014, Cold Spring Harbor, NY. (poster)

Song L, and Florea L, *CLASS—a program for accurate reconstruction of genes and alternative splicing variations from RNA-seq data*, Cold Spring Harbor Laboratory Meeting - Genome Informatics 2013, Cold Spring Harbor, NY. (poster)

Song L, and Florea L, *CLASS and ASprofile: resources for alternative splicing annotation from RNA-seq data*, International Plant and Animal Genomes Meeting XXI, 2013, San Diego, CA. (poster)

Song L, and Seidel S, *User defined data distributions in UPC*, PGAS 12: Proceedings of the Sixth Conference on Partitioned Global Address Space Programming Models, 2012, Santa Barbara, CA. (poster)

Song L, and Seidel S, *A fast longest common subsequence algorithm for finding similar sequences in a*

genome database, Great Lakes Bioinformatics Conference – GLBIO 2011, Athens, OH. (talk)

INVITED TALKS

Immune receptor repertoire analysis in tumor immunology. Biostatistics Seminar Series, Fred Hutchinson Cancer Research Center, 2022

Comprehensive characterizations of immune receptor repertoire in tumors and cancer immunotherapy studies. Bioinformatics Group Meeting, CIMACs-CIDC, 2022

Immune receptor repertoire analysis in tumor immunology. Immune-Mechanism and Recognition Working Group Meeting, NCI IOTN, 2021

TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Bioinformatics Group Meeting, CIMACs-CIDC, 2021

THESIS

Song L, *Improving Genome Annotation with RNA-Seq Data*, Ph.D. Thesis, 2018.
Song L, Liu G and Jiang C, *Incidence matrix based methods for computing repetitive vectors and siphons of Petri nets*, Bachelor Thesis, 2009. (Excellent Graduation Thesis)

REVIEW EXPERIENCE

Genome Medicine, GigaScience, Bioinformatics, BMC Genomics, BMC bioinformatics, Horticulture Research, PeerJ, Algorithms for Molecular Biology, IEEE/ACM Transactions on Computational Biology and Bioinformatics

MENTORSHIP

Miaomiao Fu, Master student, Dartmouth College (2023 summer)
Jennifer Liu, Master student, Dartmouth College (2023 summer)
McKenzy Wall, Undergrad, Harvard College (2020-2022)
Gali Bai, Bioinformatics Analyst, Dana-Farber Cancer Institute (2021-2022)
David Cohen, Bioinformatics Analyst, Dana-Farber Cancer Institute (2019-2021)
Haowen Zhang, Visiting Scholar, Georgia Institute of Technology (2020 Summer)
Yang Cao, Visiting Scholar, Sichuan University (2020)
Zhangyi Ouyang, Visiting Scholar, Beijing Institute of Radiation Medicine (2019-2020)

TEACHING EXPERIENCE

STAT115 Introduction to Computational Biology and Bioinformatics, Harvard (Sprint 2019, guest lecture, videos)
CS5321 Advanced Algorithms, MTU (Fall 2010, Fall 2011, TA)
CS4321 Introduction to Algorithms, MTU (Fall 2011, Spring 2012, TA)
CS3311 Formal Models of Computation, MTU (Fall 2010, Spring 2011, TA)
CS4121 Programming Language, MTU (Fall 2009, Spring 2010, TA)
CS4461 Computer Networks, MTU (Spring 2012, TA)
CS4611 Computer Graphics, MTU (Fall 2009, TA)

INTERNSHIP

2014 Summer Internship at J. Craig Venter Institute (Mentor: Jason Miller): Implemented a tool set to

validate the assembly method “Alpaca” developed at JCVI for several Medicago species genomes. The tool set adapts the longest common subsequence algorithm to find possible improvements over other assembly software, as well as misassemblies and genome rearrangements against other reference Medicago species genomes.

HONORS

1st place, 11th Northern Michigan University Programming Contest, 2010
Silver Award, ACM/ICPC Shanghai Invitational Contest, 2009
First Prize (Rank 2nd), Tongji University Programming Contest, 2009
First Prize (Rank 3rd), Tongji University Programming Contest, 2008
Second Prize, Tongji University Programming Contest, 2007

LANGUAGE AND TECHNOLOGIES

C/C++, Python, Perl, R, ActionScript, Shell script, UPC, Matlab, Lua, other.

MISCELLANEOUS

Solved more than 800 problems on Peking University Judge Online for ACM/ICPC (poj.org)

(July 15, 2023)