Haoyu Cheng
haoyu.cheng@yale.edu

EDUCATION/TRAINING

Postdoctoral Fellow, Dana Farber Cancer Institute, Harvard Medical School, USA 2019-2024
Mentor: Heng Li

Ph.D., Computer Science, University of Science and Technology of China, China 2013-2019
Mentor: Yun Xu

B.E., Computer Science, Hebei University of Technology, China 2009-2013

POSITIONS AND EMPLOYMENT

Assistant Professor, Biomedical Informatics and Data Science, Yale University, USA 2024-present

Postdoctoral Fellow, Dana Farber Cancer Institute, Harvard Medical School, USA 2019-2024

PERSONAL STATEMENT

I am a tenure-track Assistant Professor at the Department of Biomedical Informatics and Data Science (BIDS) at Yale University, starting in July 2024. My research primarily focuses on developing computational methods for genomic data, with a particular emphasis on *de novo* genome assembly and its applications. I have developed a suite of algorithms for *de novo* genome assembly, including hifiasm, hifiasm (Hi-C) and hifiasm (UL). These algorithms have been widely used by numerous large-scale sequencing projects, such as the Human Pangenome Reference Consortium (HPRC), the Genome in a Bottle (GIAB), the Vertebrate Genomes Project (VGP), and the Darwin Tree of Life project (DToL). Since its initial publication in 2021, these algorithms have been cited thousands of times and have been funded by the K99/R00 Pathway award from the National Institutes of Health (NIH). In addition to my experience in developing computational methods, I worked closely with collaborators to explore the applications of genome assemblies. As an active member of the ongoing Telomere-to-Telomere (T2T) consortium and the Human Pangenome Reference Consortium (HPRC), I contributed to the creation of the first human pangenome reference, leveraging haplotype-resolved assemblies produced by hifiasm. Furthermore, I collaborated with the Vertebrate Genomes Project (VGP) and the Darwin Tree of Life project (DToL) teams in utilizing hifiasm to assemble challenging genomes.

1. **Cheng H**, Qu H, McKenzie S, Lawrence KR, Windsor R, Vella M, Park PJ, Li H. (2025) Efficient near telomere-to-telomere assembly of Nanopore Simplex reads. ***bioRxiv***, 2025.04.14.648685.
2. **Cheng H**, Asri M, Lucas J, Koren S, Li H. (2024) Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. ***Nat Methods***, 21(6):967-970. PMCID: PMC11214949.
3. **Cheng H**, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, Li H. (2022) Haplotype-resolved assembly of diploid genomes without parental data. ***Nat Biotechnol***, 40(9):1332-1335. PMCID: PMC9464699. (>400 citations)
4. **Cheng H**, Concepcion GT, Feng X, Zhang H, Li H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. ***Nat Methods***, 18(2):170-175. PMCID: PMC7961889. (>4,000 citations)

RESEARCH GRANTS

K99/R00 Pathway to Independence Award, National Institutes of Health (NIH) 2023–2028
Robust and cost-effective computational methods for haplotype-resolved genome assemblies
Role: PI; Project Number: K99HG012798

Essential Open Source Software for Science, The Chan Zuckerberg Initiative (CZI) 2021–2023
Improving computational methods for high-throughput sequence data analysis
Role: Key Personnel; Project Number: 2021-237653

OTHER EXPERIENCE

2025 Talk, London Calling 2025 (*Title: Efficient telomere-to-telomere genome assembly with nanopore reads using hifiasm*)
2025 Poster, Biology of Genomes 2025 (*Title: Efficient telomere-to-telomere assembly of ONT simplex reads using hifiasm (ONT)*)
2024 Talk, the Telomere-to-Telomere "Face-to-Face" (T2T-F2F) Meeting (*Title: Fast and cost-efficient de novo assembly with the updated hifiasm*)
2024 Talk, Computational Genomics Summer Institute (CGSI 2024) (*Title: Scalable Telomere-to-Telomere Assembly for Complex Genomes with HiFiasm*)
2024 Talk, International Plant and Animal Genome Conference (PAG 31) (*Title: Scalable Telomere-to-Telomere Assembly for Complex Genomes with HiFiasm*)
2023 Talk, Biodiversity Genomics Academy 2023 (BGA23) (*Title: Scalable telomere-to-telomere assembly with hifiasm*)
2023 Talk, Human Pangenome Reference Consortium Annual Meeting 2023 (*Title: Scalable telomere-to-telomere assembly for diploid, polyploid and cancer genomes with double graph*)
2023 Poster, Genome Informatics 2023 (*Title: Scalable telomere-to-telomere assembly for diploid, polyploid and cancer genomes with double graph*)
2022 Talk, Biological Data Science 2022 (*Title: An integrated algorithm for robust and cost-effective telomere-to-telomere genome assembly*)
2021 Talk, Genome Informatics 2021 (*Title: Robust haplotype-resolved assembly of diploid individuals without parental data*)
2021 Talk, Human Pangenome Reference Consortium Annual Meeting 2021 (*Title: Robust haplotype-resolved assembly of diploid individuals without parental data*)
2020 Talk, Genome Informatics 2020 (*Title: Haplotype-resolved de novo assembly with phased assembly graphs*)
2020 – Member of the Human Pangenome Reference Consortium (HPRC)
2020 – Member of the Telomere-to-Telomere consortium (T2T)
2019 Poster, Genome Informatics 2019 (*Title: Haplotype-resolved de novo assembly with accurate long reads*)

AWARDS AND HONORS

2023 National Human Genome Research Institute - Pathway to Independence Award (K99/R00)

CONTRIBUTIONS TO SCIENCE

1. **Haplotype-resolved *de novo* assembly for accurate long reads.** The advance of sequencing technologies makes it possible to produce high-quality reads that are both long and accurate. However, most of the existing *de novo* assembly algorithms could not take full advantage of the power of long accurate reads, as they were originally designed for long error-prone reads. To this end, I developed a haplotype-resolved assembly algorithm hifiasm, which introduces a haplotype-aware error correction strategy to faithfully reconstruct different haplotypes and resolve repeats. Hifiasm also consists of the graph-binning strategy to significantly improve the quality of trio-binning assemblies by utilizing the topological information of the assembly graph. The benchmark conducted by the Human Pangenome Reference Consortium (HPRC) showed that hifiasm outperformed all other algorithms by a large margin, enabling it to be the assembler of choice by HPRC. Hifiasm has also been widely used by many other large-scale sequencing projects, such as the Genome in a Bottle (GIAB), the Vertebrate Genomes Project (VGP), as well as the Darwin Tree of Life project (DToL). In addition, I contributed to the development of hifiasm-meta, a specific version of hifiasm that is designed for metagenome assemblies.

a) **Cheng H**, Concepcion GT, Feng X, Zhang H, Li H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. ***Nat Methods***, 18(2):170-175. PMCID: PMC7961889.
b) Feng X, **Cheng H**, Portik D. and Li H. (2022) Metagenome assembly of high-fidelity long reads with hifiasm-meta. ***Nat Methods***, 19(6):671-674. PMCID: PMC9343089.
c) Larivière D, Abueg L, Brajuka N, Gallardo-Alba C, Grüning B, Ko BJ, Ostrovsky A, Palmada-Flores M, Pickett BD, Rabbani K, Antunes A, Balacco JR, Chaisson MJP, **Cheng H**, Collins J, Couture M, Denisova A, Fedrigo O, Gallo GR, Giani AM, Gooder GM, Horan K, Jain N, Johnson C, Kim H, Lee C, Marques-Bonet T, O'Toole B, Rhie A, Secomandi S, Sozzoni M, Tilley T, Uliano-Silva M, van den Beek M, Williams RW, Waterhouse RM, Phillippy AM, Jarvis ED, Schatz MC, Nekrutenko A, Formenti G. (2024) Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. ***Nat Biotechnol***. 42(3):367-370. PMID: 38278971.

2. **Near telomere-to-telomere hybrid assembly.** The advent of accurate PacBio High-Fidelity (HiFi) long reads enables the haplotype-resolved assembly to become a routine procedure for large genomes. However, HiFi reads, while precise, often fall short in length to resolve long exact repeats, leading to fragmented segments in repeat-dense areas, such as centromeres. Building upon my earlier HiFi-only hifiasm algorithm, I developed a hybrid assembly approach, termed hifiasm (UL). This incorporates the considerably longer, albeit less accurate, ultra-long ONT reads. The advantages of hifiasm (UL) stem mainly from the novel double graph framework, which integrates assembly graphs at different scales, maximizing the capabilities of both HiFi and ultra-long reads. In addition, I developed a hybrid algorithm that combines the local phasing information of long reads with the long-range phasing information of Hi-C reads to achieve single-sample fully-phased assembly for diploid individuals. The Human Pangenome Reference Consortium (HPRC) is assembling 150 telomere-to-telomere human genomes utilizing these algorithms due to their demonstrated superior performance.

a) **Cheng H**, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, Li H. (2022) Haplotype-resolved assembly of diploid genomes without parental data. ***Nat Biotechnol***, 21(6):967-970. PMCID: PMC11214949.
b) **Cheng H**, Asri M, Lucas J, Koren S, Li H. (2024) Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. ***Nat Methods***, 21(6):967-970. PMCID: PMC11214949.
c) **Cheng H**, Qu H, McKenzie S, Lawrence KR, Windsor R, Vella M, Park PJ, Li H. (2025) Efficient near telomere-to-telomere assembly of Nanopore Simplex reads. ***bioRxiv***, 2025.04.14.648685.

3. **String algorithms for short-read mapping.** In addition to genome assembly algorithms, I proposed several algorithms for fundamental problems of short-read mapping. (a) I developed a compact bit-vector algorithm that is able to simultaneously calculate the similarity of multiple reads to the reference genome. (b) I proposed an efficient algorithm to reduce unnecessary operations and improve the data locality for the locating step of the FM-index. When querying short patterns via the FM-index, my algorithm could significantly reduce the query time. (c) By utilizing the massively parallel computing capability of GPUs, I designed a parallel short-read mapping algorithm that is nearly 10 times faster than the corresponding CPU-based algorithms. (d) I contributed to the development of a pipeline that is designed specifically to preprocess and align chromatin profiles.

a) **Cheng H**, Jiang H, Yang J, Xu Y, Shang Y. (2015) BitMapper: an efficient all-mapper based on bit-vector computing. ***BMC Bioinformatics***, 16:192. PMCID: PMC4462005.
b) **Cheng H**, Wu M, Xu Y. (2018) FMtree: a fast locating algorithm of FM-indexes for genomic data. ***Bioinformatics***, 34(3):416-424. Not based on US Government funded research.
c) **Cheng H**, Zhang Y, Xu Y. (2018) BitMapper2: A GPU-Accelerated All-Mapper Based on the Sparse q-Gram Index. ***IEEE/ACM Trans Comput Biol Bioinform***, 16(3):886-897. Not based on US Government funded research.
d) Zhang H, Song L, Wang X, **Cheng H**, Wang C, Meyer CA, Liu T, Tang M, Aluru S, Yue F, Liu XS, Li H. (2021) Fast alignment and preprocessing of chromatin profiles with Chromap. ***Nat Commun***, 12(1):6566. PMCID: PMC8589834.

4. **Large-scale sequencing projects.** As a member of the Telomere-to-Telomere (T2T) Consortium and the Human Pangenome Reference Consortium (HPRC), I contributed to the creation of the first complete human genome and the first human pangenome reference. In particular, I took on important roles in the data analysis of the HPRC. During the benchmark of the HPRC, my hifiasm genome assembler produced the best haplotype-resolved assemblies among 23 submitted tools. As a result, I worked closely with the whole HPRC consortium to assemble 47 human genomes with hifiasm, and build the human pangenome reference on top of the assemblies of these genomes. I also collaborated with the Genome in a Bottle (GIAB) team to build a curated benchmark including 273 challenging medically-relevant genes, which can only be resolved through the haplotype-resolved assemblies produced by hifiasm.

    a) Wagner J, Olson ND, Harris L, McDaniel J, **Cheng H**, Fungtammasan A, Hwang YC, Gupta R, Wenger AM, Rowell WJ, Khan ZM, Farek J, Zhu Y, Pisupati A, Mahmoud M, Xiao C, Yoo B, Sahraeian SME, Miller DE, Jáspez D, Lorenzo-Salazar JM, Muñoz-Barrera A, Rubio-Rodríguez LA, Flores C, Narzisi G, Evani US, Clarke WE, Lee J, Mason CE, Lincoln SE, Miga KH, Ebbert MTW, Shumate A, Li H, Chin CS, Zook JM, Sedlazeck FJ. (2022) Curated variation benchmarks for challenging medically relevant autosomal genes. ***Nat Biotechnol***. 40(5):672-680. PMCID: PMC9117392. (with the Genome in a Bottle team)
    b) The Telomere-to-Telomere Consortium. (2022) The complete sequence of a human genome. ***Science***, 376(6588):44-53. PMCID: PMC9186530. (with the Telomere-to-Telomere Consortium)
    c) The Human Pangenome Reference Consortium. (2022) Semi-automated assembly of high-quality diploid human reference genomes. ***Nature***, 611(7936):519-531. PMCID: PMC9668749. (with the Human Pangenome Reference Consortium)
    d) The Human Pangenome Reference Consortium. (2023) A draft human pangenome reference. ***Nature***, 617(7960):312-324. PMCID: PMC10172123. (with the Human Pangenome Reference Consortium)

COMPLETE LIST OF PUBLISHED WORK:
https://scholar.google.com/citations?user=Vff5EiwAAAAJ&hl=en&oi=sra