Human Pangenome Reference Consortium
Publication Summary

NIH_NHGRI(Total **52** published manuscripts and preprints to-date)

2022

- **HP-1#** Wang, Ting, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A. Lawson, Julian K. Lucas, Adam M. Phillippy, Alice B. Popejoy et al. "The Human Pangenome Project: a global resource to map genomic diversity." Nature 604, no. 7906 (2022): 437-446.

  Abstract: The human reference genome is the most widely used resource in human genetics and is due for a major update. Its current structure is a linear composite of merged haplotypes from more than 20 people, with a single individual comprising most of the sequence. It contains biases and errors within a framework that does not represent global human genomic variation. A high-quality reference with global representation of common variants, including single-nucleotide variants, structural variants and functional elements, is needed. The Human Pangenome Reference Consortium aims to create a more sophisticated and complete human reference genome with a graph-based, telomere-to-telomere representation of global genomic diversity. Here we leverage innovations in technology, study design and global partnerships with the goal of constructing the highest-possible quality human pangenome reference. Our goal is to improve data representation and streamline analyses to enable routine assembly of complete diploid genomes. With attention to ethical frameworks, the human pangenome reference will contain a more accurate and diverse representation of global genomic variation, improve gene–disease association studies across populations, expand the scope of genomics research to the most repetitive and polymorphic regions of the genome, and serve as the ultimate genetic resource for future biomedical research and precision medicine.

- **HP-2#** Jarvis, Erich D., Giulio Formenti, Arang Rhie, Andrea Guarracino, Chentao Yang, Jonathan Wood, Alan Tracey et al. "Automated assembly of high-quality diploid human reference genomes." bioRxiv (2022). (accepted, Nature)

  Abstract: The current human reference genome, GRCh38, represents over 20 years of effort to generate a high-quality assembly, which has greatly benefited society1, 2. However, it still has many gaps and errors, and does not represent a biological human genome since it is a blend of multiple individuals3, 4. Recently, a high-quality telomere-to-telomere reference genome, CHM13, was generated with the latest long-read technologies, but it was derived from a hydatidiform mole cell line with a duplicate genome, and is thus nearly homozygous5. To address these limitations, the Human Pangenome Reference Consortium (HPRC) recently formed with the goal of creating a collection of high-quality, cost-effective, diploid genome assemblies for a pangenome reference that represents human genetic diversity6. Here, in our first scientific report, we determined which combination of current genome sequencing and automated assembly approaches yields the most complete, accurate, and cost-effective diploid genome assemblies with minimal manual curation. Approaches that used highly accurate long reads and parent-child data to sort haplotypes during assembly outperformed those that did not. Developing a combination of all the top performing methods, we generated our first high- quality diploid reference assembly, containing only ~4 gaps (range 0-12) per chromosome, most within + 1% of CHM13's length. Nearly 1/4th of protein coding genes have

synonymous amino acid changes between haplotypes, and centromeric regions showed the highest density of variation. Our findings serve as a foundation for assembling near-complete diploid human genomes at the scale required for constructing a human pangenome reference that captures all genetic variation from single nucleotides to large structural rearrangements.

- **HP-3#** Liao, Wen-Wei, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu et al. "A draft human pangenome reference." bioRxiv (2022). (In revision, Nature)

  Abstract: The Human Pangenome Reference Consortium (HPRC) presents a first draft human pangenome reference. The pangenome contains 47 phased, diploid assemblies from a cohort of genetically diverse individuals. These assemblies cover more than 99% of the expected sequence and are more than 99% accurate at the structural and base-pair levels. Based on alignments of the assemblies, we generated a draft pangenome that captures known variants and haplotypes, reveals novel alleles at structurally complex loci, and adds 119 million base pairs of euchromatic polymorphic sequence and 1,529 gene duplications relative to the existing reference, GRCh38. Roughly 90 million of the additional base pairs derive from structural variation. Using our draft pangenome to analyze short-read data reduces errors when discovering small variants by 34% and boosts the detected structural variants per haplotype by 104% compared to GRCh38-based workflows, and by 34% compared to using previous diversity sets of genome assemblies.

- **HP-4#** Porubsky, David, Mitchell R. Vollger, William T. Harvey, Allison N. Rozanski, Peter Ebert, Glenn Hickey, Patrick Hasenfeld et al. "Gaps and complex structurally variant loci in phased genome assemblies." bioRxiv (2022). (in review, Nature)

  Abstract: There has been tremendous progress in the production of phased genome assemblies by combining long-read data with parental information or linking read data. Nevertheless, a typical phased genome assembly generated by trio-hifiasm still generates more than ~140 gaps. We perform a detailed analysis of gaps, assembly breaks, and misorientations from 77 phased and assembled human genomes (154 unique haplotypes). We find that trio-based approaches using HiFi are the current gold standard although chromosome-wide phasing accuracy is comparable when using Strand-seq instead of parental data. We find two-thirds of defined contig ends cluster near the largest and most identical repeats [including segmental duplications (35.4%) or satellite DNA (22.3%) or to regions enriched in GA/AT rich DNA (27.4%)]. As a result, 1513 protein-coding genes overlap assembly gaps in at least one haplotype and 231 are recurrently disrupted or missing from five or more haplotypes. In addition, we estimate that 6-7 Mbp of DNA are incorrectly orientated per haplotype irrespective of whether trio-free or trio-based approaches are employed. 81% of such misorientations correspond to bona fide large inversion polymorphisms in the human species, most of which are flanked by large identical segmental duplications. In addition, we also identify large-scale alignment discontinuities consistent with an 11.9 Mbp deletion and 161.4 Mbp of insertion per human haploid genome. While 99% of this variation corresponds to satellite DNA, we identify 230 regions of the euchromatic DNA with frequent expansions and contractions, nearly half of which overlap with 197 protein-coding genes. Although not completely resolved, these regions include copy number polymorphic and biomedically relevant genic regions where complete resolution and a pangenome representation will be most useful, yet most challenging, to realize.

- **HP-5#** Vollger, Mitchell R., William S. DeWitt, Philip C. Dishuck, William T. Harvey, Xavi Guitart, Michael E. Goldberg, Allison Rozanski et al. "Increased mutation rate and interlocus gene conversion within human segmental duplications." bioRxiv (2022). (in review, Nature)

Abstract: Single-nucleotide variants (SNVs) within segmental duplications (SDs) have not been systematically assessed because of the difficulty in mapping short-read sequence data to virtually identical repetitive sequences. Using 102 phased human haplotypes, we constructed 1:1 unambiguous alignments spanning high-identity SDs and compared the pattern of SNVs between unique and SD regions. We find that human SNVs are elevated 60% in SDs compared to unique regions. We estimate that at least 23% of this increase is due to interlocus gene conversion (IGC) with >7 Mbp of SD sequence converted on average per human haplotype. We develop a genome-wide map of IGC donors and acceptors, including 498 acceptor and 454 donor hotspots affecting the exons of ~800 protein-coding genes. The latter includes 171 genes that have “relocated” on average 1.61 Mbp in a subset of human haplotypes. Using a coalescent framework, we show that SD regions are evolutionarily older when compared to unique sequences with most of this signal originating from putative IGC loci. SNVs within SDs, however, also exhibit a distinct mutational spectrum where there is a 27.1% increase in transversions that convert cytosine to guanine or the reverse across all triplet contexts. In addition, we observe a 7.6% reduction in the frequency of CpG associated mutations when compared to unique DNA. We hypothesize that these distinct mutational properties help to maintain an overall higher GC content of SD DNA when compared to unique DNA, and we show that these GC-favoring mutational events are likely driven by GC-biased conversion between paralogous sequences.

- **HP-6#** Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger et al. "The complete sequence of a human genome." *Science* 376, no. 6588 (2022): 44-53.

  Abstract: Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

- **HP-7#** Vollger, Mitchell R., Xavi Guitart, Philip C. Dishuck, Ludovica Mercuri, William T. Harvey, Ariel Gershman, Mark Diekhans et al. "Segmental duplications and their variation in a complete human genome." Science 376, no. 6588 (2022): eabj6965.

  Abstract: Despite their importance in disease and evolution, highly identical segmental duplications (SDs) are among the last regions of the human reference genome (GRCh38) to be fully sequenced. Using a complete telomere-to-telomere human genome (T2T-CHM13), we present a comprehensive view of human SD organization. SDs account for nearly one-third of the additional sequence, increasing the genome-wide estimate from 5.4 to 7.0% [218 million base pairs (Mbp)]. An analysis of 268 human genomes shows that 91% of the previously unresolved T2T-CHM13 SD sequence (68.3 Mbp) better represents human copy number variation. Comparing long-read assemblies from human (n = 12) and nonhuman primate (n = 5) genomes, we systematically reconstruct the evolution and structural haplotype diversity of biomedically relevant and duplicated genes. This analysis reveals patterns of structural heterozygosity and evolutionary differences in SD organization between humans and other primates.

- **HP-8#** Aganezov, Sergey, Stephanie M. Yan, Daniela C. Soto, Melanie Kirsche, Samantha Zarate, Pavel Avdeyev, Dylan J. Taylor et al. "A complete reference genome improves analysis of human genetic variation." Science 376, no. 6588 (2022): eabl3533.

Abstract: Compared to its predecessors, the Telomere-to-Telomere CHM13 genome adds nearly 200 million base pairs of sequence, corrects thousands of structural errors, and unlocks the most complex regions of the human genome for clinical and functional study. We show how this reference universally improves read mapping and variant calling for 3202 and 17 globally diverse samples sequenced with short and long reads, respectively. We identify hundreds of thousands of variants per sample in previously unresolved regions, showcasing the promise of the T2T-CHM13 reference for evolutionary and biomedical discovery. Simultaneously, this reference eliminates tens of thousands of spurious variants per sample, including reduction of false positives in 269 medically relevant genes by up to a factor of 12. Because of these improvements in variant discovery coupled with population and functional genomic resources, T2T-CHM13 is positioned to replace GRCh38 as the prevailing reference for human genetics.

- **HP-9#** Gershman, Ariel, Michael EG Sauria, Xavi Guitart, Mitchell R. Vollger, Paul W. Hook, Savannah J. Hoyt, Miten Jain et al. "Epigenetic patterns in a complete human genome." *Science* 376, no. 6588 (2022): eabj5089.

  Abstract: The completion of a telomere-to-telomere human reference genome, T2T-CHM13, has resolved complex regions of the genome, including repetitive and homologous regions. Here, we present a high-resolution epigenetic study of previously unresolved sequences, representing entire acrocentric chromosome short arms, gene family expansions, and a diverse collection of repeat classes. This resource precisely maps CpG methylation (32.28 million CpGs), DNA accessibility, and short-read datasets (166,058 previously unresolved chromatin immunoprecipitation sequencing peaks) to provide evidence of activity across previously unidentified or corrected genes and reveals clinically relevant paralog-specific regulation. Probing CpG methylation across human centromeres from six diverse individuals generated an estimate of variability in kinetochore localization. This analysis provides a framework with which to investigate the most elusive regions of the human genome, granting insights into epigenetic regulation.

- **HP-10#** Hoyt, Savannah J., Jessica M. Storer, Gabrielle A. Hartley, Patrick GS Grady, Ariel Gershman, Leonardo G. de Lima, Charles Limouse et al. "From telomere to telomere: The transcriptional and epigenetic state of human repeat elements." *Science* 376, no. 6588 (2022): eabk3112.

  Abstract: Mobile elements and repetitive genomic regions are sources of lineage-specific genomic innovation and uniquely fingerprint individual genomes. Comprehensive analyses of such repeat elements, including those found in more complex regions of the genome, require a complete, linear genome assembly. We present a de novo repeat discovery and annotation of the T2T-CHM13 human reference genome. We identified previously unknown satellite arrays, expanded the catalog of variants and families for repeats and mobile elements, characterized classes of complex composite repeats, and located retroelement transduction events. We detected nascent transcription and delineated CpG methylation profiles to define the structure of transcriptionally active retroelements in humans, including those in centromeres. These data expand our insight into the diversity, distribution, and evolution of repetitive regions that have shaped the human genome.

- **HP-11#** Altemose, Nicolas, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt et al. "Complete genomic and epigenetic maps of human centromeres." Science 376, no. 6588 (2022): eabl4178.

  Abstract: Existing human genome assemblies have almost entirely excluded repetitive sequences within and near centromeres, limiting our understanding of their organization, evolution, and functions, which include facilitating proper chromosome segregation. Now, a complete, telomere-to-telomere human genome assembly (T2T-CHM13) has enabled us to comprehensively characterize pericentromeric and centromeric repeats, which constitute 6.2% of the genome (189.9 megabases). Detailed maps of these regions revealed multi-megabase

structural rearrangements, including in active centromeric repeat arrays. Analysis of centromere-associated sequences uncovered a strong relationship between the position of the centromere and the evolution of the surrounding DNA through layered repeat expansions. Furthermore, comparisons of chromosome X centromeres across a diverse panel of individuals illuminated high degrees of structural, epigenetic, and sequence variation in these complex and rapidly evolving regions.

- **HP-12#** Mc Cartney, Ann M., Kishwar Shafin, Michael Alonge, Andrey V. Bzikadze, Giulio Formenti, Arkarachai Fungtammasan, Kerstin Howe et al. "Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies." Nature Methods (2022): 1-9.

  Abstract: Advances in long-read sequencing technologies and genome assembly methods have enabled the recent completion of the first telomere-to-telomere human genome assembly, which resolves complex segmental duplications and large tandem repeats, including centromeric satellite arrays in a complete hydatidiform mole (CHM13). Although derived from highly accurate sequences, evaluation revealed evidence of small errors and structural misassemblies in the initial draft assembly. To correct these errors, we designed a new repeat-aware polishing strategy that made accurate assembly corrections in large repeats without overcorrection, ultimately fixing 51% of the existing errors and improving the assembly quality value from 70.2 to 73.9 measured from PacBio high-fidelity and Illumina k-mers. By comparing our results to standard automated polishing tools, we outline common polishing errors and offer practical suggestions for genome projects with limited resources. We also show how sequencing biases in both high-fidelity and Oxford Nanopore Technologies reads cause signature assembly errors that can be corrected with a diverse panel of sequencing technologies.

- **HP-13#** Deorowicz, Sebastian, Agnieszka Danek, and Heng Li. "AGC: Compact representation of assembled genomes." bioRxiv (2022).

  Abstract: High-quality sequence assembly is the ultimate representation of complete genetic information of an individual. Several ongoing pangenome projects are producing collections of high-quality assemblies of various species. Here, we show how to represent the sequenced genomes in 2–3 orders of magnitude smaller space, allowing easy and fast extraction of any contig or its part.

- **HP-14#** Cheng, Haoyu, Erich D. Jarvis, Olivier Fedrigo, Klaus-Peter Koepfli, Lara Urban, Neil J. Gemmell, and Heng Li. "Haplotype-resolved assembly of diploid genomes without parental data." *Nature Biotechnology* (2022): 1-4.

  Abstract: Routine haplotype-resolved genome assembly from single samples remains an unresolved problem. Here we describe an algorithm that combines PacBio HiFi reads and Hi-C chromatin interaction data to produce a haplotype-resolved assembly without the sequencing of parents. Applied to human and other vertebrate samples, our algorithm consistently outperforms existing single-sample assembly pipelines and generates assemblies of similar quality to the best pedigree-based assemblies.

- **HP-15#** Eizenga, Jordan M., and Benedict Paten. "Improving the time and space complexity of the WFA algorithm and generalizing its scoring." bioRxiv (2022).

  Abstract: Modern genomic sequencing data is trending toward longer sequences with higher accuracy. Many analyses using these data will center on alignments, but classical exact alignment algorithms are infeasible for long sequences. The recently proposed WFA algorithm demonstrated how to perform exact alignment for long, similar sequences in O(sN) time and O(s2) memory, where s is a score that is low for similar sequences (Marco-Sola et al., 2021). However, this algorithm still has infeasible memory requirements for longer sequences. Also, it

uses an alternate scoring system that is unfamiliar to many bioinformaticians. We describe variants of WFA that improve its asymptotic memory use from O(s2) to O(s3/2) and its asymptotic run time from O(sN) to O(s2 + N). We expect the reduction in memory use to be particularly impactful, as it makes it practical to perform highly multithreaded megabase-scale exact alignments in common compute environments. In addition, we show how to fold WFA's alternate scoring into the broader literature on alignment scores.

- **HP-16#** Marco-Sola, Santiago, Jordan M. Eizenga, Andrea Guarracino, Benedict Paten, Erik Garrison, and Miquel Moreto. "Optimal gap-affine alignment in O (s) space." bioRxiv (2022).

  Abstract: Pairwise sequence alignment remains a fundamental problem in computational biology and bioinformatics. Recent advances in genomics and sequencing technologies demand faster and scalable algorithms that can cope with the ever-increasing sequence lengths. Classical pairwise alignment algorithms based on dynamic programming are strongly limited by quadratic requirements in time and memory. The recently proposed wavefront alignment algorithm (WFA) introduced an efficient algorithm to perform exact gap-affine alignment in O(ns) time, where s is the optimal score and n is the sequence length. Notwithstanding these bounds, WFA's O(s2) memory requirements become computationally impractical for genome-scale alignments, leading to a need for further improvement. In this paper, we present the bidirectional WFA algorithm (BiWFA), the first gap-affine algorithm capable of computing optimal alignments in O(s) memory while retaining WFA's time complexity of O(ns). As a result, this work improves the lowest known memory bound O(n) to compute gap-affine alignments. In practice, our implementation never requires more than a few hundred MBs aligning noisy Oxford Nanopore Technologies reads up to 1 Mbp long while maintaining competitive execution times

- **HP-17#** Zhang, Haowen, Shiqi Wu, Srinivas Aluru, and Heng Li. "Fast sequence to graph alignment using the graph wavefront algorithm." arXiv preprint arXiv:2206.13574 (2022).

  Abstract: A pan-genome graph represents a collection of genomes and encodes sequence variations between them. It is a powerful data structure for studying multiple similar genomes. Sequence-to-graph alignment is an essential step for the construction and the analysis of pan-genome graphs. However, existing algorithms incur runtime proportional to the product of sequence length and graph size, making them inefficient for aligning long sequences against large graphs. Results: We propose the graph wavefront alignment algorithm (Gwfa), a new method for aligning a sequence to a sequence graph. Although the worst-case time complexity of Gwfa is the same as the existing algorithms, it is designed to run faster for closely matching sequences, and its runtime in practice often increases only moderately with the edit distance of the optimal alignment. On four real datasets, Gwfa is up to four orders of magnitude faster than other exact sequence-to-graph alignment algorithms. We also propose a graph pruning heuristic on top of Gwfa, which can achieve an additional ~10-fold speedup on large graphs. Availability: Gwfa code is accessible at this https URL.

- **HP-18#** Bailey, Andrew D., Jason Talkish, Hongxu Ding, Haller Igel, Alejandra Duran, Shreya Mantripragada, Benedict Paten, and Manuel Ares. "Concerted modification of nucleotides at functional centers of the ribosome revealed by single-molecule RNA modification profiling." Elife 11 (2022): e76562.

  Abstract: Nucleotides in RNA and DNA are chemically modified by numerous enzymes that alter their function.

Eukaryotic ribosomal RNA (rRNA) is modified at more than 100 locations, particularly at highly conserved and functionally important nucleotides. During ribosome biogenesis, modifications are added at various stages of assembly. The existence of differently modified classes of ribosomes in normal cells is unknown because no method exists to simultaneously evaluate the modification status at all sites within a single rRNA molecule. Using a combination of yeast genetics and nanopore direct RNA sequencing, we developed a reliable method to track the modification status of single rRNA molecules at 37 sites in 18 S rRNA and 73 sites in 25 S rRNA. We use our method to characterize patterns of modification heterogeneity and identify concerted modification of nucleotides found near functional centers of the ribosome. Distinct, undermodified subpopulations of rRNAs accumulate upon loss of Dbp3 or Prp43 RNA helicases, suggesting overlapping roles in ribosome biogenesis. Modification profiles are surprisingly resistant to change in response to many genetic and acute environmental conditions that affect translation, ribosome biogenesis, and pre-mRNA splicing. The ability to capture single-molecule RNA modification profiles provides new insights into the roles of nucleotide modifications in RNA function.

- **HP-19#** Markello, Charles, Charles Huang, Alex Rodriguez, Andrew Carroll, Pi-Chuan Chang, Jordan Eizenga, Thomas Markello, David Haussler, and Benedict Paten. "A complete pedigree-based graph workflow for rare candidate variant analysis." Genome Research 32, no. 5 (2022): 893-903.

  Abstract: Methods that use a linear genome reference for genome sequencing data analysis are reference-biased. In the field of clinical genetics for rare diseases, a resulting reduction in genotyping accuracy in some regions has likely prevented the resolution of some cases. Pangenome graphs embed population variation into a reference structure. Although pangenome graphs have helped to reduce reference mapping bias, further performance improvements are possible. We introduce VG-Pedigree, a pedigree-aware workflow based on the pangenome-mapping tool of Giraffe and the variant calling tool DeepTrio using a specially trained model for Giraffe-based alignments. We demonstrate mapping and variant calling improvements in both single-nucleotide variants (SNVs) and insertion and deletion (indel) variants over those produced by alignments created using BWA-MEM to a linear-reference and Giraffe mapping to a pangenome graph containing data from the 1000 Genomes Project. We have also adapted and upgraded deleterious-variant (DV) detecting methods and programs into a streamlined workflow. We used these workflows in combination to detect small lists of candidate DVs among 15 family quartets and quintets of the Undiagnosed Diseases Program (UDP). All candidate DVs that were previously diagnosed using the Mendelian models covered by the previously published methods were recapitulated by these workflows. The results of these experiments indicate that a slightly greater absolute count of DVs are detected in the proband population than in their matched unaffected siblings.

- **HP-20#** Formenti, Giulio, Arang Rhie, Brian P. Walenz, Françoise Thibaud-Nissen, Kishwar Shafin, Sergey Koren, Eugene W. Myers, Erich D. Jarvis, and Adam M. Phillippy. "Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation." Nature Methods (2022): 1-9.

  Abstract: Variant calling has been widely used for genotyping and for improving the consensus accuracy of long-read assemblies. Variant calls are commonly hard-filtered with user-defined cutoffs. However, it is impossible to define a single set of optimal cutoffs, as the calls heavily depend on the quality of the reads, the variant caller of choice and the quality of the unpolished assembly. Here, we introduce Merfin, a k-mer based variant-filtering algorithm for improved accuracy in genotyping and genome assembly polishing. Merfin evaluates each variant based on the expected k-mer multiplicity in the reads, independently of the quality of the read alignment and variant caller's internal score. Merfin increased the precision of genotyped calls in several benchmarks, improved consensus accuracy and reduced frameshift errors when applied to human and nonhuman assemblies built from Pacific Biosciences HiFi and continuous long reads or Oxford Nanopore reads, including the first complete human genome. Moreover, we introduce assembly quality and completeness metrics that account for the expected genomic copy numbers.

- **HP-21#** Sirén, Jouni, and Benedict Paten. "GBZ File Format for Pangenome Graphs." bioRxiv (2022), under revision, Bioinformatics

  Abstract: *Motivation* Pangenome graphs representing aligned genome assemblies are being shared in the text-based Graphical Fragment Assembly format. As the number of assemblies grows, there is a need for a file format that can store the highly repetitive data space-efficiently. *Results* We propose the GBZ file format based on data structures used in the Giraffe short read aligner. The format provides good compression, and the files can be efficiently loaded into in-memory data structures. We provide compression and decompression tools and libraries for using GBZ graphs, and we show that they can be efficiently used on a variety of systems. *Availability* C++ and Rust implementations are available at https://github.com/jltsiren/gbwtgraph and https://github.com/jltsiren/gbwt-rs, respectively.

- **HP-22#** Goenka, Sneha D., John E. Gorzynski, Kishwar Shafin, Dianna G. Fisk, Trevor Pesout, Tanner D. Jensen, Jean Monlong et al. "Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing." Nature Biotechnology (2022): 1-7.

  Abstract: Whole-genome sequencing (WGS) can identify variants that cause genetic disease, but the time required for sequencing and analysis has been a barrier to its use in acutely ill patients. In the present study, we develop an approach for ultra-rapid nanopore WGS that combines an optimized sample preparation protocol, distributing sequencing over 48 flow cells, near real-time base calling and alignment, accelerated variant calling and fast variant filtration for efficient manual review. Application to two example clinical cases identified a candidate variant in <8 h from sample preparation to variant identification. We show that this framework provides accurate variant calls and efficient prioritization, and accelerates diagnostic clinical genome sequencing twofold compared with previous approaches.

- **HP-23#** Sibbesen, Jonas A., Jordan M. Eizenga, Adam M. Novak, Jouni Sirén, Xian Chang, Erik Garrison, and Benedict Paten. "Haplotype-aware pantranscriptome analyses using spliced pangenome graphs." BioRxiv (2022): 2021-03, under revision, Nature Methods

  Abstract: Pangenomics is emerging as a powerful computational paradigm in bioinformatics. This field uses population-level genome reference structures, typically consisting of a sequence graph, to mitigate reference bias and facilitate analyses that were challenging with previous reference-based methods. In this work, we extend these methods into transcriptomics to analyze sequencing data using the pantranscriptome: a population-level transcriptomic reference. Our novel toolchain can construct spliced pangenome graphs, map RNA-seq data to these graphs, and perform haplotype-aware expression quantification of transcripts in a pantranscriptome. This workflow improves accuracy over state-of-the-art RNA-seq mapping methods, and it can efficiently quantify haplotype-specific transcript expression without needing to characterize a sample's haplotypes beforehand.

- **HP-24#** Wagner, Justin, Nathan D. Olson, Lindsay Harris, Jennifer McDaniel, Haoyu Cheng, Arkarachai Fungtammasan, Yih-Chii Hwang et al. "Curated variation benchmarks for challenging medically relevant autosomal genes." Nature biotechnology 40, no. 5 (2022): 672-680.

  Abstract: The repetitive nature and complexity of some medically relevant genes poses a challenge for their accurate analysis in a clinical setting. The Genome in a Bottle Consortium has provided variant benchmark sets, but these exclude nearly 400 medically relevant genes due to their repetitiveness or polymorphic complexity. Here, we characterize 273 of these 395 challenging autosomal genes using a haplotype-resolved whole-genome assembly. This curated benchmark reports over 17,000 single-nucleotide variations, 3,600 insertions and

deletions and 200 structural variations each for human genome reference GRCh37 and GRCh38 across HG002. We show that false duplications in either GRCh37 or GRCh38 result in reference-specific, missed variants for short- and long-read technologies in medically relevant genes, including CBS, CRYAA and KCNE1. When masking these false duplications, variant recall can improve from 8% to 100%. Forming benchmarks from a haplotype-resolved whole-genome assembly may become a prototype for future benchmarks covering the whole genome.

- **HP-25#** Porubsky, David, Wolfram Höps, Hufsah Ashraf, PingHsun Hsieh, Bernardo Rodriguez-Martin, Feyza Yilmaz, Jana Ebler et al. "Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders." Cell 185, no. 11 (2022): 1986-2005.

  Abstract: Unlike copy number variants (CNVs), inversions remain an underexplored genetic variation class. By integrating multiple genomic technologies, we discover 729 inversions in 41 human genomes. Approximately 85% of inversions <2 kbp form by twin-priming during L1 retrotransposition; 80% of the larger inversions are balanced and affect twice as many nucleotides as CNVs. Balanced inversions show an excess of common variants, and 72% are flanked by segmental duplications (SDs) or retrotransposons. Since flanking repeats promote non-allelic homologous recombination, we developed complementary approaches to identify recurrent inversion formation. We describe 40 recurrent inversions encompassing 0.6% of the genome, showing inversion rates up to 2.7 × 10−4 per locus per generation. Recurrent inversions exhibit a sex-chromosomal bias and co-localize with genomic disorder critical regions. We propose that inversion recurrence results in an elevated number of heterozygous carriers and structural SD diversity, which increases mutability in the population and predisposes specific haplotypes to disease-causing CNVs.

- **HP-26#** Benet Pagès, Anna, Kate R. Rosenbloom, Luis R. Nassar, Christopher M. Lee, Brian J. Raney, Hiram Clawson, Daniel Schmelter et al. "Variant interpretation: UCSC Genome Browser recommended track sets." Human Mutation 43, no. 8 (2022): 998-1011.

  Abstract: The UCSC Genome Browser has been an important tool for genomics and clinical genetics since the sequence of the human genome was first released in 2000. As it has grown in scope to display more types of data it has also grown more complicated. The data, which are dispersed at many locations worldwide, are collected into one view on the Browser, where the graphical interface presents the data in one location. This supports the expertise of the researcher to interpret variants in the genome. Because the analysis of single nucleotide variants and copy number variants require interpretation of data at very different genomic scales, different data resources are required. We present here several Recommended Track Sets designed to facilitate the interpretation of variants in the clinic, offering quick access to datasets relevant to the appropriate scale.

- **HP-27#** Li, Daofeng, Deepak Purushotham, Jessica K. Harrison, Silas Hsu, Xiaoyu Zhuo, Changxu Fan, Shane Liu et al. "WashU Epigenome Browser update 2022." Nucleic Acids Research (2022).

  Abstract: WashU Epigenome Browser (https://epigenomegateway.wustl.edu/browser/) is a web-based genomic data exploration tool that provides visualization, integration, and analysis of epigenomic datasets. The newly renovated user interface and functions have enabled researchers to engage with the browser and genomic data more efficiently and effectively since 2018. Here, we introduce a new integrated panel design in the browser that

allows users to interact with 1D (genomic features), 2D (such as Hi-C), 3D (genome structure), and 4D (time series) data in a single web page. The browser can display three-dimensional chromatin structures with the 3D viewer module. The 4D tracks, called 'Dynamic' tracks, animatedly display time-series data, allowing for a more striking visual impact to identify the gene or genomic region candidates as a function of time. Genomic data, such as annotation features, numerical values, and chromatin interaction data can all be viewed in the dynamic track mode. Imaging data from microscopy experiments can also be displayed in the browser. In addition to software development, we continue to service and expand the data hubs we host for large consortia including 4DN, Roadmap Epigenomics, TaRGET and ENCODE, among others. Our growing user/developer community developed additional track types as plugins, such as qBed and dynseq tracks, which extend the utility of the browser. The browser serves as a foundation for additional genomics platforms including the WashU Virus Genome Browser (for COVID-19 research) and the Comparative Genome Browser. The WashU Epigenome Browser can also be accessed freely through Amazon Web Services at https://epigenomegateway.org/.

- **HP-28#** Nair, Surag, Arjun Barrett, Daofeng Li, Brian J. Raney, Brian T. Lee, Peter Kerpedjiev, Vivekanandan Ramalingam et al. "The dynseq genome browser track enables visualization of context-specific, dynamic DNA sequence features at single nucleotide resolution." bioRxiv (2022).

  Abstract: We introduce the dynseq genome browser track, which displays DNA nucleotide characters scaled by user-specified, base-resolution scores provided in the BigWig file format. The dynseq track enables visualization of context-specific, informative genomic sequence features. We demonstrate its utility in three popular genome browsers for interpreting cis-regulatory sequence syntax and regulatory variant interpretation by visualizing nucleotide importance scores derived from machine learning models of regulatory DNA trained on protein-DNA binding and chromatin accessibility experiments.

- **HP-29#** Kokot, Marek, Adam Gudyś, Heng Li, and Sebastian Deorowicz. "CoLoRd: Compressing long reads." Nature Methods 19, no. 4 (2022): 441-444.

  Abstract: The cost of maintaining exabytes of data produced by sequencing experiments every year has become a major issue in today's genomic research. In spite of the increasing popularity of third-generation sequencing, the existing algorithms for compressing long reads exhibit a minor advantage over the general-purpose gzip. We present CoLoRd, an algorithm able to reduce the size of third-generation sequencing data by an order of magnitude without affecting the accuracy of downstream analyses.

- **HP-30#** Fulton, Lucinda, Ting Wang, Robert Fulton, Tina Lindsay, and Sarah Cody. "eP348: Human Pangenome Reference Consortium Coordinating Center." Genetics in Medicine 24, no. 3 (2022): S218-S219.

  Abstract: The Human Pangenome Reference Consortium (HPRC) aims to create a diverse pangenome reference comprised of many genomes. Development efforts focus on developing tools needed by the research and clinical communities to improve clinical outcomes by better utilizing genomic diversity. As part of our mission, we aim to understand the needs in the clinical space related to the utility of the pangenome reference and understand any barriers to its adoption, whether they be technical or non-technical.

- **HP-31#** Guarracino, Andrea, Silvia Buonaiuto, Tamara Potapova, Arang Rhie, Sergey Koren, Boris Rubinstein, Christian Fischer et al. "Recombination between heterologous human acrocentric chromosomes." bioRxiv (2022).

  Abstract: The short arms of the human acrocentric chromosomes 13, 14, 15, 21, and 22 share large homologous regions, including the ribosomal DNA repeats and extended segmental duplications (1,2). While the complete assembly of these regions in the Telomere-to-Telomere consortium's (T2T) CHM13 provided a model of their

homology (3), it remained unclear if these patterns were ancestral or maintained by ongoing recombination exchange. Here, we use pangenomic resources and methods to show that specific pseudo-homologous regions of the acrocentrics recombine. Considering an all-to-all comparison of the high-quality human pangenome from the Human Pangenome Reference Consortium (HPRC) (4), we find that contigs from all of the acrocentric short arms form a community similar to those formed by single chromosomes or the sex chromosome pair. A variation graph (5) constructed from centromere-spanning acrocentric contigs indicates the presence of pseudo-homologous regions where most contigs appear as recombinants of heterologous CHM13 acrocentrics. A diploid T2T assembly of a target sample cross-validates these patterns (6). On chromosomes 13, 14, 21, and 22, we observe lower levels of linkage disequilibrium in pseudo-homologous regions than in their short and long arms, indicating higher rates of recombination and/or a larger effective population size (7). The ubiquity of signals of heterologous recombination seen in the HPRC draft pangenome's acrocentric assemblies suggests that this phenomenon is a basic feature of human cellular biology, providing sequence and population-based confirmation of hypotheses first developed cytogenetically fifty years ago (8).

2021

- **HP-32#** Sirén, Jouni, Jean Monlong, Xian Chang, Adam M. Novak, Jordan M. Eizenga, Charles Markello, Jonas A. Sibbesen et al. "Pangenomics enables genotyping of known structural variants in 5202 diverse genomes." Science 374, no. 6574 (2021): abg8871

  Abstract: We introduce Giraffe, a pangenome short-read mapper that can efficiently map to a collection of haplotypes threaded through a sequence graph. Giraffe maps sequencing reads to thousands of human genomes at a speed comparable to that of standard methods mapping to a single reference genome. The increased mapping accuracy enables downstream improvements in genome-wide genotyping pipelines for both small variants and larger structural variants. We used Giraffe to genotype 167,000 structural variants, discovered in long-read studies, in 5202 diverse human genomes that were sequenced using short reads. We conclude that pangenomics facilitates a more comprehensive characterization of variation and, as a result, has the potential to improve many genomic analyses.

- **HP-33#** Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm." *Nature Methods* 18, no. 2 (2021): 170-175.

  Abstract: Haplotype-resolved de novo assembly is the ultimate solution to the study of sequence variations in a genome. However, existing algorithms either collapse heterozygous alleles into one consensus copy or fail to cleanly separate the haplotypes to produce high-quality phased assemblies. Here we describe hifiasm, a de novo assembler that takes advantage of long high-fidelity sequence reads to faithfully represent the haplotype information in a phased assembly graph. Unlike other graph-based assemblers that only aim to maintain the contiguity of one haplotype, hifiasm strives to preserve the contiguity of all haplotypes. This feature enables the development of a graph trio binning algorithm that greatly advances over standard trio binning. On three human and five nonhuman datasets, including California redwood with a ~30-Gb hexaploid genome, we show that hifiasm frequently delivers better assemblies than existing tools and consistently outperforms others on haplotype-resolved assembly.

- **HP-34#** Mao, Yafei, Claudia R. Catacchio, LaDeana W. Hillier, David Porubsky, Ruiyang Li, Arvis Sulovari, Jason D. Fernandes et al. "A high-quality bonobo genome refines the analysis of hominid evolution." Nature 594, no. 7861 (2021): 77-81.

Abstract: The divergence of chimpanzee and bonobo provides one of the few examples of recent hominid speciation1,2. Here we describe a fully annotated, high-quality bonobo genome assembly, which was constructed without guidance from reference genomes by applying a multiplatform genomics approach. We generate a bonobo genome assembly in which more than 98% of genes are completely annotated and 99% of the gaps are closed, including the resolution of about half of the segmental duplications and almost all of the full-length mobile elements. We compare the bonobo genome to those of other great apes1,3,4,5 and identify more than 5,569 fixed structural variants that specifically distinguish the bonobo and chimpanzee lineages. We focus on genes that have been lost, changed in structure or expanded in the last few million years of bonobo evolution. We produce a high-resolution map of incomplete lineage sorting and estimate that around 5.1% of the human genome is genetically closer to chimpanzee or bonobo and that more than 36.5% of the genome shows incomplete lineage sorting if we consider a deeper phylogeny including gorilla and orangutan. We also show that 26% of the segments of incomplete lineage sorting between human and chimpanzee or human and bonobo are non-randomly distributed and that genes within these clustered segments show significant excess of amino acid replacement compared to the rest of the genome.

- **HP-35#** Eizenga, Jordan M., Adam M. Novak, Emily Kobayashi, Flavia Villani, Cecilia Cisar, Simon Heumos, Glenn Hickey, Vincenza Colonna, Benedict Paten, and Erik Garrison. "Efficient dynamic variation graphs." Bioinformatics 36, no. 21 (2021): 5139-5144.

  Abstract: Pangenomics is a growing field within computational genomics. Many pangenomic analyses use bidirected sequence graphs as their core data model. However, implementing and correctly using this data model can be difficult, and the scale of pangenomic datasets can be challenging to work at. These challenges have impeded progress in this field. Here, we present a stack of two C++ libraries, libbdsg and libhandlegraph, which use a simple, field-proven interface, designed to expose elementary features of these graphs while preventing common graph manipulation mistakes. The libraries also provide a Python binding. Using a diverse collection of pangenome graphs, we demonstrate that these tools allow for efficient construction and manipulation of large genome graphs with dense variation. For instance, the speed and memory usage are up to an order of magnitude better than the prior graph implementation in the VG toolkit, which has now transitioned to using libbdsg's implementations.

- **HP-36#** Porubsky, David, Peter Ebert, Peter A. Audano, Mitchell R. Vollger, William T. Harvey, Pierre Marijon, Jana Ebler et al. "Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads." Nature biotechnology 39, no. 3 (2021): 302-308.

  Abstract: Human genomes are typically assembled as consensus sequences that lack information on parental haplotypes. Here we describe a reference-free workflow for diploid de novo genome assembly that combines the chromosome-wide phasing and scaffolding capabilities of single-cell strand sequencing1,2 with continuous long-read or high-fidelity3 sequencing data. Employing this strategy, we produced a completely phased de novo genome assembly for each haplotype of an individual of Puerto Rican descent (HG00733) in the absence of parental data. The assemblies are accurate (quality value > 40) and highly contiguous (contig N50 > 23 Mbp) with low switch error rates (0.17%), providing fully phased single-nucleotide variants, indels and structural variants. A comparison of Oxford Nanopore Technologies and Pacific Biosciences phased assemblies identified 154 regions that are preferential sites of contig breaks, irrespective of sequencing technology or phasing algorithms.

- **HP-37#** Ding, Hongxu, Ioannis Anastopoulos, Andrew D. Bailey, Joshua Stuart, and Benedict Paten. "Towards inferring nanopore sequencing ionic currents from nucleotide chemical structures." Nature communications 12, no. 1 (2021): 1-9.

Abstract: The characteristic ionic currents of nucleotide kmers are commonly used in analyzing nanopore sequencing readouts. We present a graph convolutional network-based deep learning framework for predicting kmer characteristic ionic currents from corresponding chemical structures. We show such a framework can generalize the chemical information of the 5-methyl group from thymine to cytosine by correctly predicting 5-methylcytosine-containing DNA 6mers, thus shedding light on the de novo detection of nucleotide modifications.

- **HP-38#** Shafin, Kishwar, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid et al. "Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads." Nature methods 18, no. 11 (2021): 1322-1332.

  Abstract: Long-read sequencing has the potential to transform variant detection by reaching currently difficult-to-map regions and routinely linking together adjacent variations to enable read-based phasing. Third-generation nanopore sequence data have demonstrated a long read length, but current interpretation methods for their novel pore-based signal have unique error profiles, making accurate analysis challenging. Here, we introduce a haplotype-aware variant calling pipeline, PEPPER-Margin-DeepVariant, that produces state-of-the-art variant calling results with nanopore data. We show that our nanopore-based method outperforms the short-read-based single-nucleotide-variant identification method at the whole-genome scale and produces high-quality single-nucleotide variants in segmental duplications and low-mappability regions where short-read-based genotyping fails. We show that our pipeline can provide highly contiguous phase blocks across the genome with nanopore reads, contiguously spanning between 85% and 92% of annotated genes across six samples. We also extend PEPPER-Margin-DeepVariant to PacBio HiFi data, providing an efficient solution with superior performance over the current WhatsHap-DeepVariant standard. Finally, we demonstrate de novo assembly polishing methods that use nanopore and PacBio HiFi reads to produce diploid assemblies with high accuracy (Q35+ nanopore-polished and Q40+ PacBio HiFi-polished).

- **HP-39#** Miga, Karen H., and Ting Wang. "The need for a human pangenome reference sequence." Annual review of genomics and human genetics 22 (2021): 81.

  Abstract: The reference human genome sequence is inarguably the most important and widely used resource in the fields of human genetics and genomics. It has transformed the conduct of biomedical sciences and brought invaluable benefits to the understanding and improvement of human health. However, the commonly used reference sequence has profound limitations, because across much of its span, it represents the sequence of just one human haplotype. This single, monoploid reference structure presents a critical barrier to representing the broad genomic diversity in the human population. In this review, we discuss the modernization of the reference human genome sequence to a more complete reference of human genomic diversity, known as a human pangenome.

- **HP-40#** Howe, Kevin L., Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean et al. "Ensembl 2021." Nucleic acids research 49, no. D1 (2021): D884-D891.

  Abstract: The Ensembl project (https://www.ensembl.org) annotates genomes and disseminates genomic data for vertebrate species. We create detailed and comprehensive annotation of gene structures, regulatory elements and variants, and enable comparative genomics by inferring the evolutionary history of genes and genomes. Our integrated genomic data are made available in a variety of ways, including genome browsers, search interfaces, specialist tools such as the Ensembl Variant Effect Predictor, download files and programmatic interfaces. Here, we present recent Ensembl developments including two new website portals. Ensembl Rapid Release (http://rapid.ensembl.org) is designed to provide core tools and services for genomes as soon as possible and has been deployed to support large biodiversity sequencing projects. Our SARS-CoV-2 genome browser

(https://covid-19.ensembl.org) integrates our own annotation with publicly available genomic data from numerous sources to facilitate the use of genomics in the international scientific response to the COVID-19 pandemic. We also report on other updates to our annotation resources, tools and services. All Ensembl data and software are freely available without restriction.

- **HP-41#** Logsdon, Glennis A., Mitchell R. Vollger, PingHsun Hsieh, Yafei Mao, Mikhail A. Liskovykh, Sergey Koren, Sergey Nurk et al. "The structure, function and evolution of a complete human chromosome 8." Nature 593, no. 7857 (2021): 101-107.

  Abstract: The complete assembly of each human chromosome is essential for understanding human biology and evolution1,2. Here we use complementary long-read sequencing technologies to complete the linear assembly of human chromosome 8. Our assembly resolves the sequence of five previously long-standing gaps, including a 2.08-Mb centromeric α-satellite array, a 644-kb copy number polymorphism in the β-defensin gene cluster that is important for disease risk, and an 863-kb variable number tandem repeat at chromosome 8q21.2 that can function as a neocentromere. We show that the centromeric α-satellite array is generally methylated except for a 73-kb hypomethylated region of diverse higher-order α-satellites enriched with CENP-A nucleosomes, consistent with the location of the kinetochore. In addition, we confirm the overall organization and methylation pattern of the centromere in a diploid human genome. Using a dual long-read sequencing approach, we complete high-quality draft assemblies of the orthologous centromere from chromosome 8 in chimpanzee, orangutan and macaque to reconstruct its evolutionary history. Comparative and phylogenetic analyses show that the higher-order α-satellite structure evolved in the great ape ancestor with a layered symmetry, in which more ancient higher-order repeats locate peripherally to monomeric α-satellites. We estimate that the mutation rate of centromeric satellite DNA is accelerated by more than 2.2-fold compared to the unique portions of the genome, and this acceleration extends into the flanking sequence.

- **HP-42#** Lu, Tsung-Yu, and Mark JP Chaisson. "Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs." Nature communications 12, no. 1 (2021): 1-12.

  Abstract: Variable number tandem repeats (VNTRs) are composed of consecutive repetitive DNA with hypervariable repeat count and composition. They include protein coding sequences and associations with clinical disorders. It has been difficult to incorporate VNTR analysis in disease studies that use short-read sequencing because the traditional approach of mapping to the human reference is less effective for repetitive and divergent sequences. In this work, we solve VNTR mapping for short reads with a repeat-pangenome graph (RPGG), a data structure that encodes both the population diversity and repeat structure of VNTR loci from multiple haplotype-resolved assemblies. We develop software to build a RPGG, and use the RPGG to estimate VNTR composition with short reads. We use this to discover VNTRs with length stratified by continental population, and expression quantitative trait loci, indicating that RPGG analysis of VNTRs will be critical for future studies of diversity and disease.

- **HP-43#** Zhuo, Xiaoyu, Alan Y. Du, Erica C. Pehrsson, Daofeng Li, and Ting Wang. "Epigenomic differences in the human and chimpanzee genomes are associated with structural variation." Genome research 31, no. 2 (2021): 279-290.

  Abstract: Structural variation (SV), including insertions and deletions (indels), is a primary mechanism of genome evolution. However, the mechanism by which SV contributes to epigenome evolution is poorly understood. In this study, we characterized the association between lineage-specific indels and epigenome differences between human and chimpanzee to investigate how SVs might have shaped the epigenetic landscape. By intersecting medium-to-large human–chimpanzee indels (20 bp–50 kb) with putative promoters and enhancers in cranial

neural crest cells (CNCCs) and repressed regions in induced pluripotent cells (iPSCs), we found that 12% of indels overlap putative regulatory and repressed regions (RRRs), and 15% of these indels are associated with lineage-biased RRRs. Indel-associated putative enhancer and repressive regions are approximately 1.3 times and approximately three times as likely to be lineage-biased, respectively, as those not associated with indels. We found a twofold enrichment of medium-sized indels (20–50 bp) in CpG island (CGI)–containing promoters than expected by chance. Lastly, from human-specific transposable element insertions, we identified putative regulatory elements, including NR2F1-bound putative CNCC enhancers derived from SVAs and putative iPSC promoters derived from LTR5s. Our results show that different types of indels are associated with specific epigenomic diversity between human and chimpanzee.

2020

- **HP-44#** Warren, Wesley C., R. Alan Harris, Marina Haukness, Ian T. Fiddes, Shwetha C. Murali, Jason Fernandes, Philip C. Dishuck et al. "Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility." Science 370, no. 6523 (2020): eabc6617.

  Abstract: The rhesus macaque (Macaca mulatta) is the most widely studied nonhuman primate (NHP) in biomedical research. We present an updated reference genome assembly (Mmul_10, contig N50 = 46 Mbp) that increases the sequence contiguity 120-fold and annotate it using 6.5 million full-length transcripts, thus improving our understanding of gene content, isoform diversity, and repeat organization. With the improved assembly of segmental duplications, we discovered new lineage-specific genes and expanded gene families that are potentially informative in studies of evolution and disease susceptibility. Whole-genome sequencing (WGS) data from 853 rhesus macaques identified 85.7 million single-nucleotide variants (SNVs) and 10.5 million indel variants, including potentially damaging variants in genes associated with human autism and developmental delay, providing a framework for developing noninvasive NHP models of human disease.

- **HP-45#** Li, Heng, Xiaowen Feng, and Chong Chu. "The design and construction of reference pangenome graphs with minigraph." Genome biology 21, no. 1 (2020): 1-19.

  Abstract: The recent advances in sequencing technologies enable the assembly of individual genomes to the quality of the reference genome. How to integrate multiple genomes from the same species and make the integrated representation accessible to biologists remains an open challenge. Here, we propose a graph-based data model and associated formats to represent multiple genomes while preserving the coordinate of the linear reference genome. We implement our ideas in the minigraph toolkit and demonstrate that we can efficiently construct a pangenome graph and compactly encode tens of thousands of structural variants missing from the current reference genome.

- **HP-46#** Armstrong, Joel, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang et al. "Progressive Cactus is a multiple-genome aligner for the thousand-genome era." Nature 587, no. 7833 (2020): 246-251.

  Abstract: New genome assemblies have been arriving at a rapidly increasing pace, thanks to decreases in sequencing costs and improvements in third-generation sequencing technologies1,2,3. For example, the number of vertebrate genome assemblies currently in the NCBI (National Center for Biotechnology Information) database4 increased by more than 50% to 1,485 assemblies in the year from July 2018 to July 2019. In addition to this influx of assemblies from different species, new human de novo assemblies5 are being produced, which enable the analysis of not only small polymorphisms, but also complex, large-scale structural differences between human individuals and haplotypes. This coming era and its unprecedented amount of data offer the opportunity to

uncover many insights into genome evolution but also present challenges in how to adapt current analysis methods to meet the increased scale. Cactus6, a reference-free multiple genome alignment program, has been shown to be highly accurate, but the existing implementation scales poorly with increasing numbers of genomes, and struggles in regions of highly duplicated sequences. Here we describe progressive extensions to Cactus to create Progressive Cactus, which enables the reference-free alignment of tens to thousands of large vertebrate genomes while maintaining high alignment quality. We describe results from an alignment of more than 600 amniote genomes, which is to our knowledge the largest multiple vertebrate genome alignment created so far.

- **HP-47#** Goenka, Sneha D., John E. Gorzynski, Kishwar Shafin, Dianna G. Fisk, Trevor Pesout, Tanner D. Jensen, Jean Monlong et al. "Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting", New England Journal Medicine. 2022 Feb 17;386(7):700-702.

  Abstract: Whole-genome sequencing (WGS) can identify variants that cause genetic disease, but the time required for sequencing and analysis has been a barrier to its use in acutely ill patients. In the present study, we develop an approach for ultra-rapid nanopore WGS that combines an optimized sample preparation protocol, distributing sequencing over 48 flow cells, near real-time base calling and alignment, accelerated variant calling and fast variant filtration for efficient manual review. Application to two example clinical cases identified a candidate variant in <8 h from sample preparation to variant identification. We show that this framework provides accurate variant calls and efficient prioritization, and accelerates diagnostic clinical genome sequencing twofold compared with previous approaches.

- **HP-48#** Shafin, Kishwar, Trevor Pesout, Ryan Lorig-Roach, Marina Haukness, Hugh E. Olsen, Colleen Bosworth, Joel Armstrong et al. "Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes." *Nature biotechnology* 38, no. 9 (2020): 1044-1053.

  Abstract: De novo assembly of a human genome using nanopore long-read sequences has been reported, but it used more than 150,000 CPU hours and weeks of wall-clock time. To enable rapid human genome assembly, we present Shasta, a de novo long-read assembler, and polishing algorithms named MarginPolish and HELEN. Using a single PromethION nanopore sequencer and our toolkit, we assembled 11 highly contiguous human genomes de novo in 9 d. We achieved roughly 63× coverage, 42-kb read N50 values and 6.5× coverage in reads >100 kb using three flow cells per sample. Shasta produced a complete haploid human genome assembly in under 6 h on a single commercial compute node. MarginPolish and HELEN polished haploid assemblies to more than 99.9% identity (Phred quality score QV = 30) with nanopore reads alone. Addition of proximity-ligation sequencing enabled near chromosome-level scaffolds for all 11 genomes. We compare our assembly performance to existing methods for diploid, haploid and trio-binned human samples and report superior accuracy and speed.

- **HP-49#** Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks et al. "Telomere-to-telomere assembly of a complete human X chromosome." Nature 585, no. 7823 (2020): 79-84.

  Abstract: After two decades of improvements, the current human reference genome (GRCh38) is the most accurate and complete vertebrate genome ever produced. However, no single chromosome has been finished end to end, and hundreds of unresolved gaps persist1,2. Here we present a human genome assembly that surpasses the continuity of GRCh382, along with a gapless, telomere-to-telomere assembly of a human chromosome. This was enabled by high-coverage, ultra-long-read nanopore sequencing of the complete hydatidiform mole CHM13 genome, combined with complementary technologies for quality improvement and validation. Focusing our efforts on the human X chromosome3, we reconstructed the centromeric satellite DNA

array (approximately 3.1 Mb) and closed the 29 remaining gaps in the current reference, including new sequences from the human pseudoautosomal regions and from cancer-testis ampliconic gene families (CT-X and GAGE). These sequences will be integrated into future human reference genome releases. In addition, the complete chromosome X, combined with the ultra-long nanopore data, allowed us to map methylation patterns across complex tandem repeats and satellite arrays. Our results demonstrate that finishing the entire human genome is now within reach, and the data presented here will facilitate ongoing efforts to complete the other human chromosomes.

- **HP-50#** Nurk, Sergey, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. "HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads." Genome research 30, no. 9 (2020): 1291-1305.

  Abstract: Complete and accurate genome assemblies form the basis of most downstream genomic analyses and are of critical importance. Recent genome assembly projects have relied on a combination of noisy long-read sequencing and accurate short-read sequencing, with the former offering greater assembly continuity and the latter providing higher consensus accuracy. The recently introduced Pacific Biosciences (PacBio) HiFi sequencing technology bridges this divide by delivering long reads (>10 kbp) with high per-base accuracy (>99.9%). Here we present HiCanu, a modification of the Canu assembler designed to leverage the full potential of HiFi reads via homopolymer compression, overlap-based error correction, and aggressive false overlap filtering. We benchmark HiCanu with a focus on the recovery of haplotype diversity, major histocompatibility complex (MHC) variants, satellite DNAs, and segmental duplications. For diploid human genomes sequenced to 30× HiFi coverage, HiCanu achieved superior accuracy and allele recovery compared to the current state of the art. On the effectively haploid CHM13 human cell line, HiCanu achieved an NG50 contig size of 77 Mbp with a per-base consensus accuracy of 99.999% (QV50), surpassing recent assemblies of high-coverage, ultralong Oxford Nanopore Technologies (ONT) reads in terms of both accuracy and continuity. This HiCanu assembly correctly resolves 337 out of 341 validation BACs sampled from known segmental duplications and provides the first preliminary assemblies of nine complete human centromeric regions. Although gaps and errors still remain within the most challenging regions of the genome, these results represent a significant advance toward the complete assembly of human genomes.

- **HP-51#** Miga, Karen H. "Centromere studies in the era of ‘telomere-to-telomere’genomics." *Experimental cell research* 394, no. 2 (2020): 112127.

  Abstract: We are entering into an exciting era of genomics where truly complete, high-quality assemblies of human chromosomes are available end-to-end, or from ‘telomere-to-telomere’ (T2T). This technological advance offers a new opportunity to include endogenous human centromeric regions in high-resolution, sequence-based studies. These emerging reference maps are expec ted to reveal a new functional landscape in the human genome, where centromere proteins, transcriptional regulation, and spatial organization can be examined with base-level resolution across different stages of development and disease. Such studies will depend on innovative assembly methods of extremely long tandem repeats (ETRs), or satellite DNAs, paired with the development of new, orthogonal validation methods to ensure accuracy and completeness. This review reflects the progress in centromere genomics, credited by recent advancements in long-read sequencing and assembly methods. In doing so, I will discuss the challenges that remain and the promise for a new period of scientific discovery for satellite DNA biology and centromere function.

- **HP-52#** Chen, Xiaoying, Bo Zhang, Ting Wang, Azad Bonni, and Guoyan Zhao. "Robust principal component analysis for accurate outlier sample detection in RNA-Seq data." BMC bioinformatics 21, no. 1 (2020): 1-20.

  Abstract: High throughput RNA sequencing is a powerful approach to study gene expression. Due to the complex multiple-steps protocols in data acquisition, extreme deviation of a sample from samples of the same treatment group may occur due to technical variation or true biological differences. The high-dimensionality of the data with few biological replicates make it challenging to accurately detect those samples, and this issue is not well studied in the literature currently. Robust statistics is a family of theories and techniques aim to detect the outliers by first fitting the majority of the data and then flagging data points that deviate from it. Robust statistics have been widely used in multivariate data analysis for outlier detection in chemometrics and engineering. Here we apply robust statistics on RNA-seq data analysis. We report the use of two robust principal component analysis (rPCA) methods, PcaHubert and PcaGrid, to detect outlier samples in multiple simulated and real biological RNA-seq data sets with positive control outlier samples. PcaGrid achieved 100% sensitivity and 100% specificity in all the tests using positive control outliers with varying degrees of divergence. We applied rPCA methods and classical principal component analysis (cPCA) on an RNA-Seq data set profiling gene expression of the external granule layer in the cerebellum of control and conditional SnoN knockout mice. Both rPCA methods detected the same two outlier samples but cPCA failed to detect any. We performed differentially expressed gene detection before and after outlier removal as well as with and without batch effect modeling. We validated gene expression changes using quantitative reverse transcription PCR and used the result as reference to compare the performance of eight different data analysis strategies. Removing outliers without batch effect modeling performed the best in term of detecting biologically relevant differentially expressed genes. rPCA implemented in the PcaGrid function is an accurate and objective method to detect outlier samples. It is well suited for high-dimensional data with small sample sizes like RNA-seq data. Outlier removal can significantly improve the performance of differential gene detection and downstream functional analysis.

- **HP-53#** Eimear E. Kenny, Robert Cook-Deegan, Barbara Koenig, Alice Popejoy, Nanibaa' Garrison, Erich Jarvis, Evan E. Eichler, Ira Hall, Benedict Paten, Adam Felsenfeld, Heng Li, Matthew Mitchell, Heather Lawson, Ting Wang, David Haussler, Karen H. Miga, Human Pangenome Reference Consortium. Policy Perspective: Population sampling and representation for the Human Pangenome Reference Consortium *(in preparation)*

  Abstract: The Human Pangenome Reference will represent the genomic diversity of 350 finished (T2T) diploid genomes. Participants in this study will be selected to represent global genetic and genomic diversity. In the first phase of the project, the HPRC will focus on existing cell lines established from the 1000 Genomes Consortium, which offers a deep catalog of human variation from 26 populations with compatible consent for unrestricted ("open") access data release. Cell lines from the 1000 Genome Collection, available in the NHGRI Biorepository at Coriell, are prioritized based on genetic and geographic diversity, the use of parental data for haplotype phasing method development, and limited time in cell culture. Nevertheless, the use of 1000 Genome Consortium data offers limited geographical sampling and is insufficient on its own to support the ambitious sampling and genetic representation goals of the Human Pangenome Project. New participant recruitment will require new domestic and international partnerships. Such international genetic and genomic research will involve broader ethical, social, and political considerations and strategic partnerships with organizations, like GA4GH. We will conduct an analytical assessment of samples to identify those individuals who will extend, as much as possible, the variation represented in the Human Pangenome Reference. Further, to facilitate cell-line-based resources to broadly advance biomedical research we will establish lymphoblastoid cell lines (LCLs) representing each reference line and work directly with Coriell to ensure consistency, quality, and universal availability of these lines for future studies. The goal to represent genetic information across humanity is positioned at the intersection of genetics and society. We anticipate a multitude of complex ethical, legal, and social implications (ELSI) that must be thoroughly addressed by scholars with expertise in research ethics, law, social sciences, demography, and

population genetics. Our effort will benefit from the existing legislative framework for the ethical collection of samples for open data sharing and the prevention of genetic discrimination. However, we understand that we will be faced with an additional need for careful policy and ethical oversight about concerns—both anticipated and unanticipated—related to consent procedures, ethical and scientific selection of study samples, engagement of communities, and definition(s) of diversity to be used for selecting and reporting on genetic diversity.