Establishment of an eHAP1 human haploid cell line hybrid reference genome assembled from short and long reads

WilliamD. Law, René L. Warren, Andrew S. McCallion

PII: S0888-7543(19)30906-1

DOI: https://doi.org/10.1016/j.ygeno.2020.01.009

Reference: YGENO 9447

To appear in: *Genomics*

Received date: 19 November 2019

Revised date: 13 January 2020

Accepted date: 15 January 2020

Please cite this article as: W. Law, R.L. Warren and A.S. McCallion, Establishment of an eHAP1 human haploid cell line hybrid reference genome assembled from short and long reads, *Genomics* (2019), https://doi.org/10.1016/j.ygeno.2020.01.009

Establishment of an eHAP1 Human Haploid Cell Line Hybrid Reference Genome Assembled from Short and Long Reads

William D. Law[1], René L. Warren[2], and Andrew S. McCallion*[1,3,4]

1. McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

2. Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 4S6, Canada.

3. Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

4. Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

*To whom correspondence should be addressed

Electronic address: William Law - wlaw4@jhmi.edu, René Warren - rwarren@bcgsc.ca, Andrew McCallion - andy@jhmi.edu.

Abstract:

Haploid cell lines are a valuable research tool with broad applicability for genetic assays. As such the fully haploid human cell line, eHAP1, has been used in a wide array of studies. However, the absence of a corresponding reference genome sequence for this cell line has limited the potential for more widespread applications to experiments dependent on available sequence, like capture-clone methodologies. We generated ~15x coverage Nanopore long reads from ten GridION flowcells and utilized this data to assemble a *de novo* draft genome using minimap and miniasm and subsequently polished using Racon. This assembly was further polished using previously generated, low-coverage, Illumina short reads with Pilon and ntEdit. This resulted in a hybrid eHAP1 assembly with >90% complete BUSCO scores. We further assessed the eHAP1 long read data for structural variants using Sniffles and identify a variety of rearrangements, including a previously established Philadelphia translocation. Finally, we demonstrate how some of these variants overlap open chromatin regions, potentially impacting regulatory regions. By integrating both long and short reads, we generated a high-quality reference assembly for eHAP1 cells. The union of long and short reads demonstrates the utility in combining sequencing platforms to generate a high-quality reference genome *de novo* solely from low coverage data. We expect the resulting eHAP1 genome assembly to provide a useful resource to enable novel experimental applications in this important model cell line.

Introduction:

The vast majority of eukaryotic cells are diploid and many cellular models used experimentally are either diploid or polyploid. The presence of additional alleles, while evolutionarily beneficial, can pose challenges to genetic assays assessing loss of function mutations. This can occur through masking effects of a recessive mutation, or complicating experiments because of the

necessity of retargeting unmodified alleles. To alleviate these challenges, haploid cell lines have been developed from a variety of species including medaka[1], rat[2], mouse[3, 4], and monkey[5]. In humans, a near haploid cell line containing the Philadelphia translocation, KBM-7, spontaneously arose from a subculture of a human leukemia tumor[6], although it remained diploid for chromosomes 8 and a portion of 15. Further work with these cells, in an unsuccessful attempt to induce pluripotency, resulted in a new cell line, termed HAP1; this line grew adherently and had lost a copy of chromosome 8[7]. Karyotyping of these cells also revealed loss of the Y chromosome. Finally, through the use of CRISPR/Cas9, HAP1 cells were genetically engineered to delete the diploid portion of chromosome 15, resulting in a fully haploid cell line termed eHAP1[8]. **In additional to these immortalized cell lines, haploid stem cell lines have been derived which have retain the capacity to differentiate[9-11].**

The HAP1 and eHAP1 cells have been used in a variety of experiments including drug screens[12, 13], host-virus interactions[7, 14-16], and genetic screens[17-19]. Despite the wide utility of these cells, only low coverage Illumina short read sequencing data has been generated for eHAP1 cells[8], resulting in challenges to producing a reference genome specific to this cell line. The generation of a more contiguous and generally higher quality reference genome would enable additional experimental uses such as sequence target capture, where knowledge of the underlying variants is critical. To this end, we employed the Oxford Nanopore Technologies Ltd. (ONT) GridION sequencing technologies to leverage the ability of long reads to uncover structural variants, which are difficult to detect using short reads alone. Additionally, combining long, but error prone read information with short, but highly accurate reads yields a more complete genome assembly than can be achieved using either technology individually[20-22].

Results:

Generating a hybrid reference genome assembly of eHAP1 cells.

Three independent replicates of high molecular weight genomic DNA was isolated from eHAP1s and prepared for Nanopore sequencing (Methods). We generated a total of ten individual libraries yielding 5 million reads and 48.1 Gb of sequence with an N50 of 31 Kb (Additional Table 1)[23]. The combined reads were aligned against the hg19 genome demonstrating an average coverage of ~15.6x (Fig 1a; Additional Figure 1).

Next, the reads were aligned against themselves to generate a *de novo* reference assembly. Subsequently, error correction was performed using five successive iterations of minimap2[24], miniasm[25], and Racon[26] (Fig 1b). After each round we calculated the BUSCO (Benchmarking universal single-copy orthologs) scores[27, 28] to evaluate the quality of the consensus assembly. We observed a 5.5% increase in complete BUSCOs from the first (3,369 complete BUSCOs) to the last (3,571 complete BUSCOs) round of error correction; however, after five rounds we observed no appreciable improvement using long reads alone.

To further improve the quality of the reference genome, we incorporated previously generated Illumina short reads[8] utilizing Pilon[29] to further polish the assembly (Fig 1c). While eHAP1 cells are a direct derivative of the E9 clone (~6x whole genome sequencing (WGS) coverage), we also chose to use short reads from a parallel eHAP1 clone (A11; ~6x WGS coverage) to obtain ~12x total coverage. To assess the quality of the reference genome, we again used BUSCO scores and saw a large improvement with just one round of Pilon polishing. The number of complete BUSCOs increased markedly from 3,571 to 5,204 (31.4%) and the number of fragmented or missing decreased by 1,633. Using the polished output from Pilon, we repeated the short read polishing two additional times and observed moderate improvements in BUSCO scores. Finally, due to the low Illumina sequencing coverage, we employed an

additional polishing step utilizing ntEdit[30], which functions well in low sequence coverage situations. We observed a slight improvement recovering an additional 81 complete BUSCOs, ultimately obtaining 90% BUSCO completeness with ~5% listed as fragmented or missing, respectively. Overall, we were able to generate a high quality *de novo* reference genome using a combination of low coverage Nanopore long reads and Illumina short reads.

Initially, we did not include the ~20x WGS coverage from the parental HAP1 cells to avoid any potential diploid regions or eHAP1 specific sequence variants. However, to assess if additional sequencing depth could further improve the reference genome, by resolving indels causing frameshift errors, we polished the reference genome using the parental HAP1 data. Starting with the eHAP1 reference assembly polished by 5x Racon and 3x Pilon, we used ntEdit to polish the assembly using the eHAP1 data first, followed by a second round using the HAP1 data, or the reverse order. In either case, we saw virtually no improvement in BUSCO scores, nine additional complete BUSCOs, compared with using the eHAP1 data alone (Fig 1d). We also utilized ABySS[31] to assess the assembly contiguity statistics of the three polished references (Fig 1e), and observe similar statistics across the three. To prevent potentially confounding variants from the parental cell line data, we focused on the polished reference assembly using eHAP1 reads only. By combining long and short reads, we were able to assemble a high-quality reference genome for the eHAP1 cells.

Long reads identify multiple structural variants

We next sought to identify structural variants (SVs) present in the eHAP1 cells. To do this, we used the structural variant caller Sniffles[32]. We first aligned the Nanopore reads to the hg19 reference genome using NGMLR[32] and passed the output into Sniffles to identify structural variants. Using the default threshold of a minimum of ten reads supporting the SV, we identified

11,451 SVs (Fig 2a; Additional Table 2); however, due to the lower coverage, we reduced the minimum number of reads to five, which yields 18,295 SVs (Additional Table 3). Despite this, the SV types between the two sets are very similar with a vast majority of SV subtypes identified as either deletions (Fig 2b) or insertions (Fig 2c). Additionally, using the five read threshold, many of the deletions (6,268/8,665; ~72.3%) or insertions (5,187/7,683; ~67.5%) are small, less than 250 base pairs (bp). Critically, ONT reads are known to have biases in deletions potentially due to difficulty in basecalling [32, 33]. We find 1,306 (15.1%) deletions detected by Sniffles contain homopolymeric runs of at least 20 bp and an additional 954 (11%) deletions overlapping dinucleotide repeats of at least 10 bp. This implies these detected deletions may be a technical artifact, rather than genuine rearrangements.

In spite of the potential false positives, once the SVs overlapping repeats are removed, we identify 271 deletions and 386 insertions that specifically overlap an open chromatin region[34, 35] (OCR; Fig 2d) identified by ENCODE in any assessed cell line. There are an additional 1,334 deletions and 1,417 insertions that do not overlap an OCR but do impact a transcription factor binding site (TFBS) [34, 36]. Finally, there were 411 and 684 regions that disrupt both an OCR and TFBS for insertions and deletions, respectively. Collectively, we find a large number of SVs impacting putative regulatory regions that could impact experimental design or interpretations.

Other types of SVs identified include translocations. As the parental KBM-7 cell line contained the Philadelphia chromosome that was retained throughout subcloning, we anticipated identifying a translocation between chromosomes 9 and 22. Sniffles correctly identified the translocation (chr9: 133,681,711 - chr22:23,632,359; hg19) directly within in the *BCR* and *ABL* genes. In all, we detect 250 translocations, of which 60 are classified as precise, indicating confidence of the exact breakpoint position at the nucleotide level. Interestingly, Sniffles detects 31 translocations involving the Y chromosome, despite karyotyping data suggesting it was lost

between the KBM-7[6] to HAP1[7]. Our polished assembly potentially suggests a translocation of the Y chromosome onto the X chromosome, but the alignment quality to the Y is moderate (~60%). While Sniffles suggests the Y chromosome may have been broken and scattered throughout the genome, other data indicates it may have been lost entirely, and additional experiments are necessary to distinguish these possibilities.

Finally, we generated an assembly consistency (Jupiter) plot[37] showing the polished eHAP1 reference genome scaffolds against GRCh38 (Fig 2e). From it, we clearly identify the Philadelphia translocation and additionally identify other SVs that corroborate Sniffles' findings in the ONT reads. Utilizing long reads exclusively, we were able to discover a variety of SVs, many of which are insertions or deletions potentially impacting OCRs, as well as larger translocations.

Discussion:

We employed ONT long read sequencing to improve the reference genome quality and identify SVs of an important human cell line, eHAP1. By utilizing previously published Illumina short read data with low coverage (~12x) and combining it with our long read data, we were able to generate a high-quality hybrid genome assembly with complete BUSCO scores of 90%. This required the use of a variety of polishing tools including Racon[26], Pilon[29], and ntEdit[30]. While we observed the greatest improvement through one round of Pilon utilizing short reads, it is important to note it required the greatest amount of computational resources per round: 3-4 days, 48 processors, and 384 Gb of RAM. In comparison, ntEdit required 36 minutes, 48 processors, and 22.2 Gb of RAM. While ntEdit benefited from a Pilon polished reference, we did not see any appreciable resource reduction between sequential Pilon rounds. Additionally, we did not need to employ ntEdit prior to Pilon polishing, but it may be beneficial in situations where

computational resources are limited. Regardless of the computational requirements, Pilon produced the largest increase in BUSCO scores despite the relatively low sequencing coverage of Illumina data. Providing additional accurate short reads, using either deeper coverage Illumina sequencing or linked reads, would likely improve base pair accuracy, scaffolding, and contiguity of the reference generated here; however, utilizing the deeper (~20x) coverage of the parental HAP1 cell line[8] had little impact on the final quality of the reference genome generated, as assessed by BUSCO analysis.

One of the greatest advantages of long reads is the ability to easily detect structural rearrangements. We were able to use the Nanopore data aligned against the human reference genome to identify over 18,000 SVs, a majority of which were small insertions or deletions. It is important to note that Nanopore reads are prone to over-calling deletions residing in repetitive regions of the genome. While a portion of these deletions may not be validated, Sniffles was able to find highly confident insertions and deletions, some of which reside within OCR and within TFBS. This type of information would be useful for experiments interested in using this cell line to assess regulatory regions[38] or in cases where the eHAP1 cells are used as primary genomic DNA isolation for capture-clone experiments[39, 40]. Additionally, Sniffles detected the presence of the Philadelphia chromosome translocation with high confidence using the long reads exclusively.

While eHAP1 cells are a useful tool, they may not be the most biologically relevant. Use of haploid stem cells allows for similar advantages in genetic modifications compared to eHAP1 cells, with the added benefit to differentiate the modified cells into a more appropriate cell type. Human haploid stem cells have been successfully established and differentiated into a plethora of cell types[10]. The major drawback of using stem cells is the time and difficulty in maintaining the cells, often requiring the use of a feeder cell layer[10, 11]. Additional time and resources are necessary for differentiation, which may

become prohibitive[9]. The decision to use an immortalized cell line or a stem cell line is dependent on the question and should be carefully considered to minimize costs while fully addressing the hypothesis.

In summary, we applied Nanopore long read sequencing technology to an important human haploid cellular model. Utilizing a combination of long and short reads, we were able to generate a high-quality reference genome and demonstrate the utility of a hybrid assembly despite comparatively low sequencing coverage. We anticipate this work will enable novel applications of eHAP1 cells, such as capture sequencing experiments and targeted CRISPR screens, to be conducted in an accelerated time frame.

Methods:

eHAP1 cell culture: eHAP1 cells were purchased from Horizon Discovery (SKU: c669). The cells were cultured using the following growth media: 445 mL IMDM media (Gibco: 12440-053), 50 mL FBS, and 5 mL 100x Pen/Strep. Cells were passaged every 2-3 days at a ratio of 1:5. The cells were rapidly expanded post purchase to reduce the number of passages and possible ploidy changes, prior to genomic DNA isolation.

Genomic DNA isolation, library prep, and sequencing: Genomic DNA was harvested from 5 million cells using the Circulomics Nanobind CBB Big DNA kit (Part #NB-900-001-01). The DNA was extracted following the included handbook (v1.7) protocol for "Cultured Mammalian Cells – HMW" with minor modifications. Specifically, cells were vortexed intensively (1 second pulses, 10x pulses), the final DNA was pipetted 10 times through a p200 tip, and immediately prior to library preparation, the DNA was run through a 28G needle five times. This was done to help the DNA into solution with minimal effect on length.

The genomic DNA was prepared using the Nanopore Ligation Sequencing Kit (SQK-LSK109) following the manufacturer's protocol (GDE_9063_v109_revD_23May2018). An initial starting amount of 1 µg of genomic DNA was used, and after library preparation, a final amount of 250-400 ng was obtained. Regardless of the final mass of DNA obtained, the entire library preparation was subjected to R9.4 flowcells (FLO-MIN106) in a 1:1 library preparation:flowcell ratio. Basecalling was performed using guppy v2.3.7. Run statistics were calculated using NanoPlot (-t 12) [23].

De novo genome assembly and polishing: The Nanopore long read fastq files from all 10 replicates were combined and aligned against each other using minimap2 (v2.16-r922)[24] with the -x ava-ont and -t 24 flags. A layout was generated using miniasm (v0.3-r179)[25] with the -t 24 flag, and the resulting .gfa file was converted into a .fasta format using awk. Then, the original Nanopore reads were aligned against the .fasta file using minimap2 (-t 24), and the resulting pairwise mapping format (.paf) file was combined with the .fasta and the combined .fastq files to be polished using Racon (v1.3.3)[26] (-t 24). The output, a .fasta polished formatted file, was passed back to minimap2, and the reads were aligned a second time. This process was repeated a total of five times.

After five rounds, the resulting polished .fasta file was used as a reference to map the previously generated[8] Illumina short reads using minimap2 (-ax sr). Data from both clone E9 and A11 were mapped independently, and the resulting .sam files were converted to sorted and indexed .bam files using samtools (v1.9)[41]. Finally, both .bam files were used to polish the .fasta file using Pilon[29] (-Xmx700G; v1.22). The resulting Pilon polished .fasta format was used to re-map the short reads using minimap2. This process was performed a total of three times.

The resulting Pilon-polished assembly was used as input for ntEdit (v1.2.2)[30]. Briefly, we ran ntEdit iteratively 3 times (-k 50-40 step 5, -i 5 -d 5 -m 1 -t 48) each with k=50, k=45 and k=40 kmer Bloom filters derived from running ntHits (v0.0.1 --outbloom --solid -b 36 -k 50-40 step 5 -t

48) on the combined Illumina short read data. Run time and memory usage was benchmarked on a CentOS 7 system with 128 Intel(R) Xeon(R) E7-8867 v3 CPUs @ 2.50GHz.

After each round of polishing, regardless of the program, BUSCO (v3.1.0)[27, 28] scores were assessed. The program was run in --mode genome, with --cpu 24 and --blast_single_core. The files were compared against the euarchontoglires_odb9 lineage. ABySS (v2.1.0) [31] statistics were calculated using the abyss-fac function.

Structural variant detection: Structural variants were detected using Sniffles[32] after alignment of the long reads against the hg19 reference genome using NGMLR[32] (-t 24, -x ont). The resulting .sam file was converted into a sorted .bam using samtools[41] and passed onto Sniffles in either default mode (-s 10) or five read minimum (-s 5). The resulting .vcf file was parsed into SV types using grep, and figures were made using ggplot2[42].

The overlap with ENCODE DNaseI [34, 35] and TF ChIP [34, 36] datasets was performed using the UCSC Table Browser[43]. A .bed file was made from structural variants detected (five read minimum) using the left-most coordinate and adding one basepair. The .bed file was uploaded to the UCSC Table Browser and intersected, with 100% overlap, with the "wgEncodeRegDnaseClustered V3"[34, 35] or "encRegTfbsClustered"[34, 36]. The resulting regions were filtered into insertions and deletions using Sniffles SVTYPE information, and the sequences were further filtered for repeats. For homopolymeric repeats, insertions or deletions containing 20 identical basepairs in a row and dinucleotide repeats of 10 pairs of any two basepairs consecutively were removed.

Jupiter plot generation: An assembly consistency (Jupiter v1.0) plot[37] of the polished reference eHAP1 genome was generated. As part of the Jupiter plot pipeline, scaftigs from the largest eHAP1 scaffolds, consisting of 75% (NG75) of the genome, were aligned to GRCh38 with minimap2 (v2.17-r941) and plotted with Circos (v0.69-6_1)[44].

Data Accession: The raw Nanopore reads generated in this study are available at **BioProject (PRJNA580215)**. The previously generated[8] Illumina short reads for clone A11 (SRR1518295) and E9 (SRR1518293) are available from the NCBI Sequence Read Archive. The final polished .fasta formatted eHAP1 reference **genomes are available at GenBank (using eHAP1 data only [WUWL00000000], using eHAP1 then HAP1 data [WUWK00000000], or using HAP1 then eHAP1 data [WUWJ00000000]).** The eHAP1 cell line may be purchased from Horizon Discovery. All custom programs and intermediate files are available upon request.

Abbreviations: SV: Structural Variant; BUSCO: Benchmarking Universal Single-Copy Orthologs; bp: base pairs; WGS: Whole Genome Sequencing; OCR: Open chromatin region; ONT: Oxford Nanopore Technologies Ltd.; TFBS: Transcription Factor Binding Site

Declaration of Competing Interest: The authors declare that they have no competing financial interests or personal relationships that could influence the work reported in this paper.

Acknowledgments: This work was supported from the NIH (MH106522) and the National Institutes of Health [2R01HG007182-04A1]. This work was also supported through internal funding from the Johns Hopkins University School of Medicine as part of the Core Coins Program. We acknowledge assistance for Nanopore sequencing from the Genetic Resources Core Facility High-Throughput Sequencing Core. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.

We thank David Mohr for providing guidance and computational resources, Jeffrey Burke from Circulomics for assistance in genomic DNA extraction, and Paul W. Hook and Sarah A. McClymont for critical reading of the manuscript. Additional computational resources were provided by the Maryland Advanced Research Computing Center (MARCC).

Figure Legends

Figure 1: *Polishing the eHAP1 reference genome*. (**A**) Histogram of the combined ten ONT GridION flow cells coverage relative to the human genome (hg19). Greater than 30 reads were collapsed into a single bin, and the red line indicates the average mean coverage. BUSCO [27, 28] scores were calculated after five rounds of Racon[26] polishing (**B**) and three rounds of Pilon[29] (**C**). Left indicates the number of rounds of each program, and bars display BUSCO notation. (**D**) ntEdit[30] was performed using the eHAP1 short reads on the 5x Racon/3x Pilon (eHAP1 Only) polished assembly and using eHAP1 then HAP1 (eHAP1-HAP1) or HAP1 then eHAP1 (HAP1-eHAP1) short reads. BUSCO sores were calculated after each round. (**E**) ABySS [31] contiguity statistics were calculated for the three ntEdit polished assemblies.

Figure 2: *Structural variant analysis of the eHAP1 cell line*. (**A**) A breakdown of the types of structural variants (SV) identified by Sniffles[32] from the eHAP1 cell line. The X-axis refers to the minimum number of reads required to support a SV. Visualization using IGV genome browser[45] of a deletion (**B**) or insertion (**C**) DNaseI hypersensitivity sites (DNaseI) [34, 35] and transcription factor binding sites (TF ChIP) [34, 36]. The numbers inside the SV indicate the size in base pairs. (**D**) Venn diagram illustrating the number of SVs overlapping TF ChIP or DNaseI sites. (**E**) A Jupiter plot[37] of the eHAP1 only polished assembly against the human genome (GRCh38). GRCh38 chromosomes are displayed incrementally from 1 (bottom, red) to Y (top, fuchsia) on the left while scaffolds (grey with black outlines) are displayed on the right side of the rim. The highlighted lines indicate potential translocations. The black lines indicate potential translocations not found using Sniffles. The green lines indicate potential chromosomal

translocations where Sniffles also indicates a translocation between the two chromosomes. The red lines indicate the Philadelphia translocations, identified here and by Sniffles.

Additional Figure 1: *Coverage per individual flow cell*. Histogram of the coverage plots for each of the ten replicates. The blue line and number indicates the mean coverage for a flow cell. Coverages above 20 were collapsed into a single bin (20+).

Additional Table 1: *NanoPlot statistics per individual flow cell*. NanoPlot[23] statistics were computed for each flow cell. For appropriate statistics, the mean was calculated across the ten replicates (column L). NanoPlot was also run on the combined flow cells (column M).

Additional Table 2: *Structural variants detected by Sniffles (Default; ten reads).* Sniffles was run in default mode, ten reads minimum supporting SV calls, using the long reads generated from eHAP1 cells. Column information is indicated in the header, and additional information can be found on the Sniffles wiki: https://github.com/fritzsedlazeck/Sniffles/wiki/Output.

Additional Table 3: *Structural variants detected by Sniffles (Default; five reads).* Sniffles was run in using five read minimum supporting SV calls, using the long reads generated from eHAP1 cells. Column information is indicated in the header, and additional information can be found on the Sniffles wiki: https://github.com/fritzsedlazeck/Sniffles/wiki/Output.

1. Yi, M., N. Hong, and Y. Hong, *Generation of medaka fish haploid embryonic stem cells.* Science, 2009. **326**(5951): p. 430-3.
2. Li, W., et al., *Genetic modification and screening in rat using haploid embryonic stem cells.* Cell Stem Cell, 2014. **14**(3): p. 404-14.
3. Leeb, M. and A. Wutz, *Derivation of haploid embryonic stem cells from mouse embryos.* Nature, 2011. **479**(7371): p. 131-4.
4. Elling, U., et al., *Forward and reverse genetics through derivation of haploid mouse embryonic stem cells.* Cell Stem Cell, 2011. **9**(6): p. 563-74.
5. Yang, H., et al., *Generation of haploid embryonic stem cells from Macaca fascicularis monkey parthenotes.* Cell Res, 2013. **23**(10): p. 1187-200.
6. Kotecki, M., P.S. Reddy, and B.H. Cochran, *Isolation and characterization of a near-haploid human cell line.* Exp Cell Res, 1999. **252**(2): p. 273-80.
7. Carette, J.E., et al., *Ebola virus entry requires the cholesterol transporter Niemann-Pick C1.* Nature, 2011. **477**(7364): p. 340-3.
8. Essletzbichler, P., et al., *Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line.* Genome Res, 2014. **24**(12): p. 2059-65.
9. Wang, H., et al., *Genetic screening and multipotency in rhesus monkey haploid neural progenitor cells.* Development, 2018. **145**(11).
10. Sagi, I., et al., *Derivation and differentiation of haploid human embryonic stem cells.* Nature, 2016. **532**(7597): p. 107-11.
11. Peng, K., et al., *Derivation of Haploid Trophoblast Stem Cells via Conversion In Vitro.* iScience, 2019. **11**: p. 508-518.
12. Birsoy, K., et al., *MCT1-mediated transport of a toxic molecule is an effective strategy for targeting glycolytic tumors.* Nat Genet, 2013. **45**(1): p. 104-8.
13. Gerhards, N.M., et al., *Haploid genetic screens identify genetic vulnerabilities to microtubule-targeting agents.* Mol Oncol, 2018. **12**(6): p. 953-971.
14. Jae, L.T., et al., *Deciphering the glycosylome of dystroglycanopathies using haploid screens for lassa virus entry.* Science, 2013. **340**(6131): p. 479-83.
15. Carette, J.E., et al., *Haploid genetic screens in human cells identify host factors used by pathogens.* Science, 2009. **326**(5957): p. 1231-5.
16. Pillay, S., et al., *An essential receptor for adeno-associated virus infection.* Nature, 2016. **530**(7588): p. 108-12.
17. Lenk, G.M., et al., *CRISPR knockout screen implicates three genes in lysosome function.* Sci Rep, 2019. **9**(1): p. 9609.
18. Blomen, V.A., et al., *Gene essentiality and synthetic lethality in haploid human cells.* Science, 2015. **350**(6264): p. 1092-6.
19. Rong, Y., et al., *Genome-Wide Screening of Genes Required for Glycosylphosphatidylinositol Biosynthesis.* PLoS One, 2015. **10**(9): p. e0138553.
20. Tan, M.H., et al., *Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly.* Gigascience, 2018. **7**(3): p. 1-6.

21. Zimin, A.V., et al., *Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm.* Genome Res, 2017. **27**(5): p. 787-792.
22. Austin, C.M., et al., *De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read.* Gigascience, 2017. **6**(8): p. 1-6.
23. De Coster, W., et al., *NanoPack: visualizing and processing long-read sequencing data.* Bioinformatics, 2018. **34**(15): p. 2666-2669.
24. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
25. Li, H., *Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.* Bioinformatics, 2016. **32**(14): p. 2103-10.
26. Vaser, R., et al., *Fast and accurate de novo genome assembly from long uncorrected reads.* Genome Res, 2017. **27**(5): p. 737-746.
27. Simao, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210-2.
28. Waterhouse, R.M., et al., *BUSCO applications from quality assessments to gene prediction and phylogenomics.* Mol Biol Evol, 2017.
29. Walker, B.J., et al., *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.* PLoS One, 2014. **9**(11): p. e112963.
30. Warren, R.L., et al., *ntEdit: scalable genome sequence polishing.* Bioinformatics, 2019.
31. Jackman, S.D., et al., *ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter.* Genome Res, 2017. **27**(5): p. 768-77.
32. Sedlazeck, F.J., et al., *Accurate detection of complex structural variations using single-molecule sequencing.* Nat Methods, 2018. **15**(6): p. 461-468.
33. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads.* Nat Biotechnol, 2018. **36**(4): p. 338-345.
34. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
35. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.
36. Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.* Genome Res, 2012. **22**(9): p. 1798-812.
37. Chu, J. *Jupiter Plot: A Circos-based tool to visualize genome assembly consistency*. 2018; 1.0:[Available from: https://doi.org/10.5281/zenodo.1241235.
38. Gasperini, M., et al., *CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions.* Am J Hum Genet, 2017. **101**(2): p. 192-205.
39. Ali, O.A., et al., *RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping.* Genetics, 2016. **202**(2): p. 389-400.
40. Shen, S.Q., et al., *Massively parallel cis-regulatory analysis in the mammalian central nervous system.* Genome Res, 2016. **26**(2): p. 238-55.
41. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
42. Wickham, H. and SpringerLink (Online service), *ggplot2 Elegant Graphics for Data Analysis*. 2016, Springer International Publishing : Imprint: Springer: Cham.
43. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.

44. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45.
45. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

Declaration of Competing Interest: The authors declare that they have no competing financial interests or personal relationships that could influence the work reported in this paper.

Generation of a reference genome for the human haploid cell line, eHAP1

Long read, but noisy Nanopore sequencing combined with short read, but accurate Illumina sequencing yields a polished reference genome

Structural variants, including the Philadelphia translocation, can be identified using exclusively long reads