# Ultra-Sensitive Targeted Haplotype Phasing Using TELL-Seq™

Veronika Mikhaylova[1], Long Pham[1], Peter Chang[1], Yu Xia[1], T. Christian Boles[2], Andrew Anfora[1], Ivan Garcia-Bassets[1], Yong Wang[3] , and Tom Chen[1]

[1] Universal Sequencing Technology, Carlsbad, CA (USA); [2] Sage Sciences, Beverly, MA (USA); [3] Universal Sequencing Technology, Canton, MA (USA)

UNIVERSAL SEQUENCING innovation for all

## 1. Abstract & Conclusions

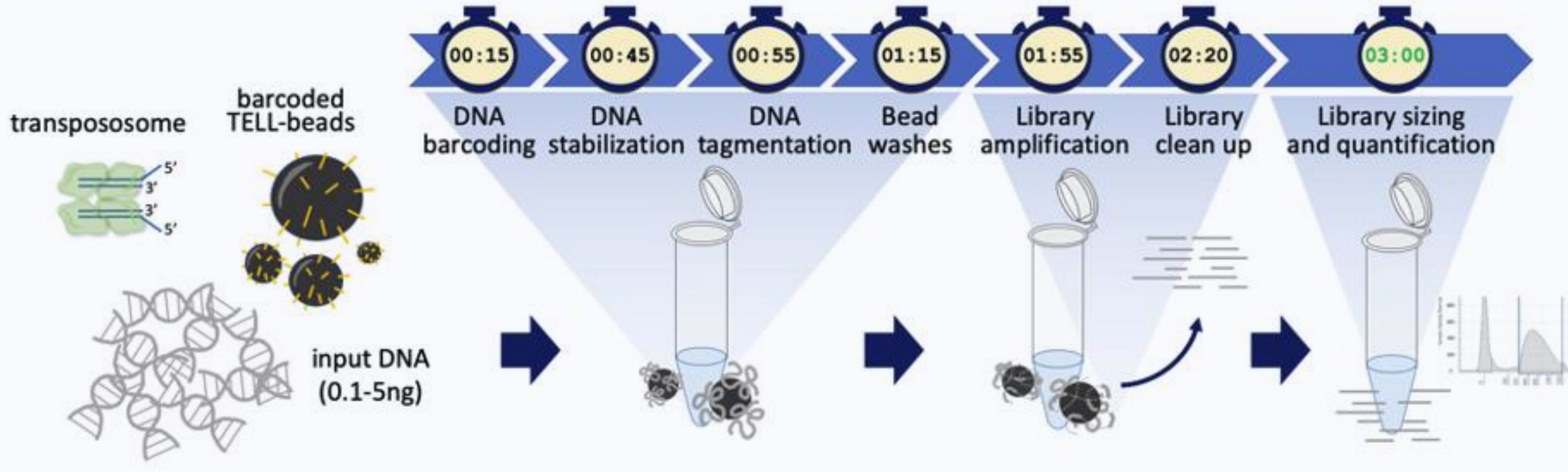
Humans have maternal and paternal sets of matched chromosomal copies that, combined, typically differ from the reference genome at 4 to 5 million sites. However, many of these sequence differences (*alternate alleles*) are present in one of the two copies (known as *heterozygous sites)*. Resolving whether neighboring alternate alleles belong to the same or different chromosomal copies (a process known as '*haplotype phasing*') is critical for many research and clinical applications. Highly accurate and low-cost conventional short-read next-generation sequencing (NGS) methods fail to phase heterozygous sites located more than a few hundred bases apart. Phasing must then rely on population allelic frequencies, parental genotypes, or long-read NGS methods with limited sequencing sensitivity or fidelity and at a higher cost. Alternative to conventional short-read NGS, linked-read approaches capture long-range information during NGS library preparation. Using linked reads, transposase enzyme-linked long-read sequencing (TELL-Seq™) can phase an entire genome (Chen et al., 2020).

Here, we sought to determine whether TELL-Seq can also phase discrete DNA fragments in targeted experiments:

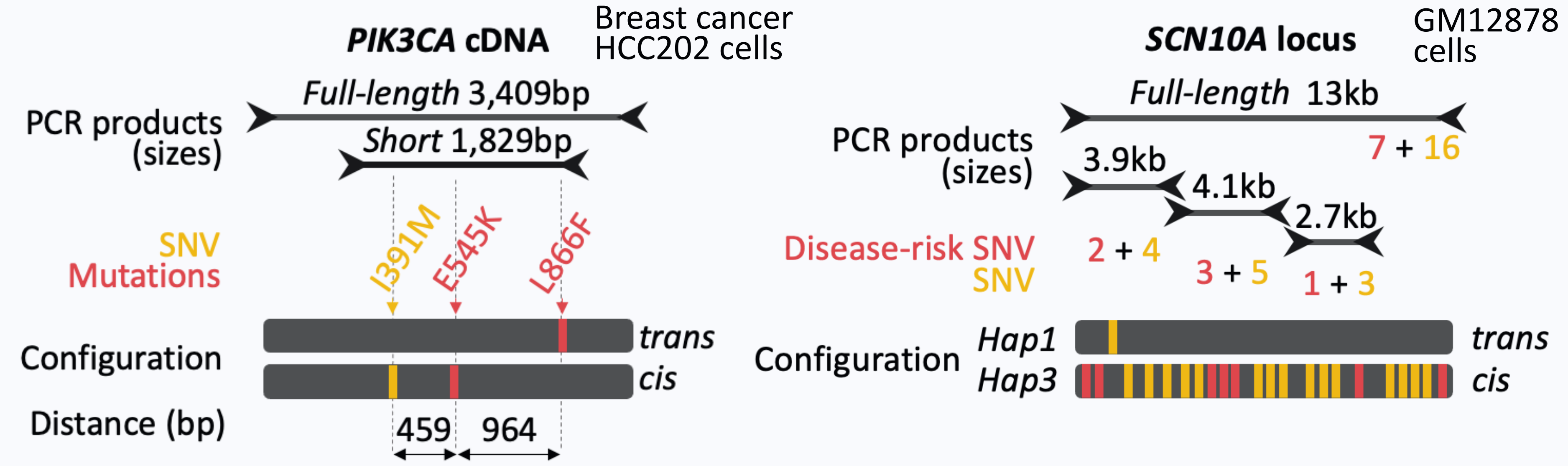
- First, we show the phasing of two double somatic mutations in 1.8 and 3.4kb PCR fragments of amplified *PIK3CA* cDNA. These mutations are relevant for cancer prognosis and treatment (Vasan et al.., 2019).
- Second, we show the phasing of four to seven polymorphic sites in 2.7, 3.9, 4.1, and 13kb PCR fragments of amplified genomic (g)DNA containing a region of the *SCN10A* locus. These sites have been associated with life-threatening heart arrhythmia (Pinsach-Abuin et al., 2020).
- Third, we show the phasing of the entire *BRCA1* and *BRCA2* loci in the Ashkenazi Trio (family trio). These genes have been extensively associated with breast cancer. In this case, we used CRISPR-Cas9 and pulse-field electrophoresis (HLS-CATCH™ system developed by Sage Science) to excise and isolate the two almost 200kb DNA fragments (high-molecular weight or HMW DNA) prior to TELL-Seq processing.

In addition, we developed a computational pipeline that outputs the sets of phased heterozygous sites aligned to the reference genome. Together, our results demonstrate high phasing accuracy combining conventional NGS data collection with TELL-Seq using DNA fragments in the range of 2.7 kb and up to 200 kb (PCR products amplified from cDNA or gDNA) and HMW DNA excised from gDNA.

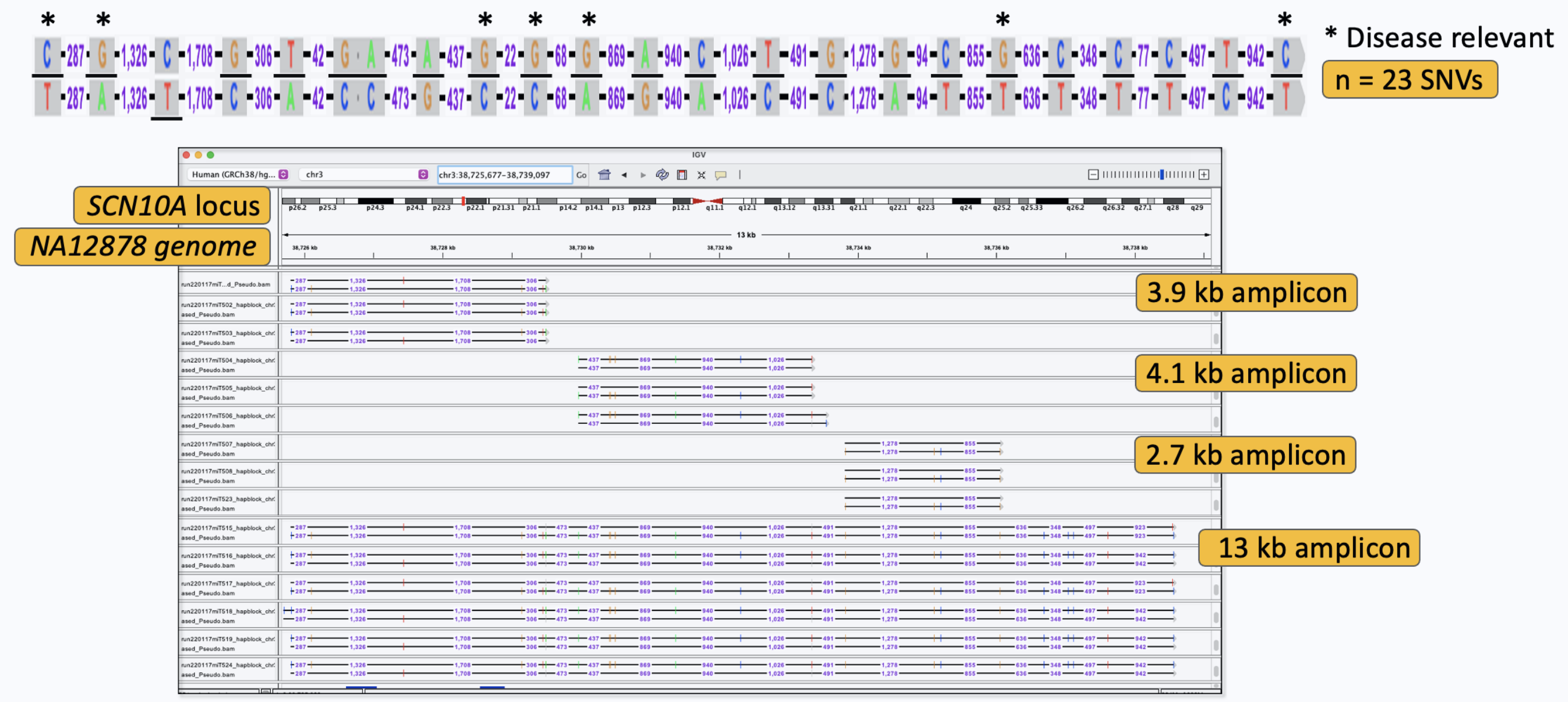
**TELL-seq workflow**



## 2. Results: Phasing of PCR products (2.7-13 kb amplicons)



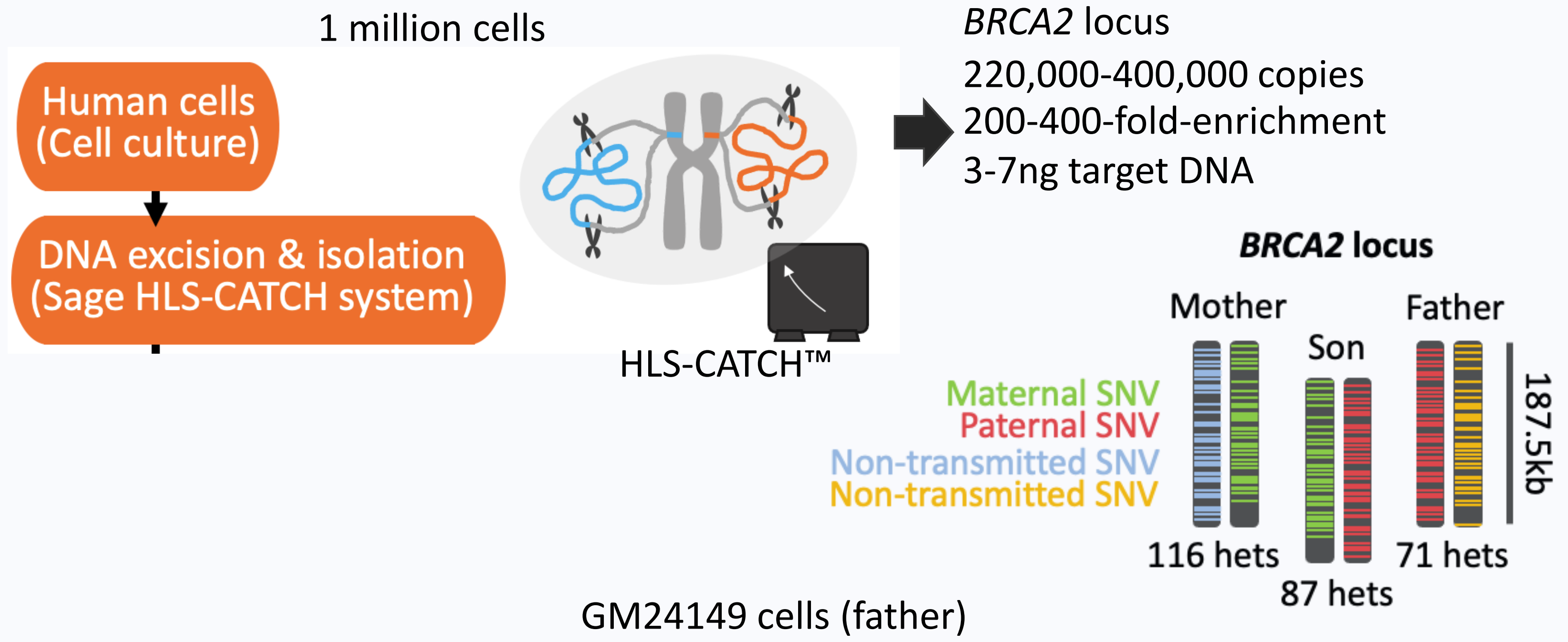
GM12878 cells



**Summary table**

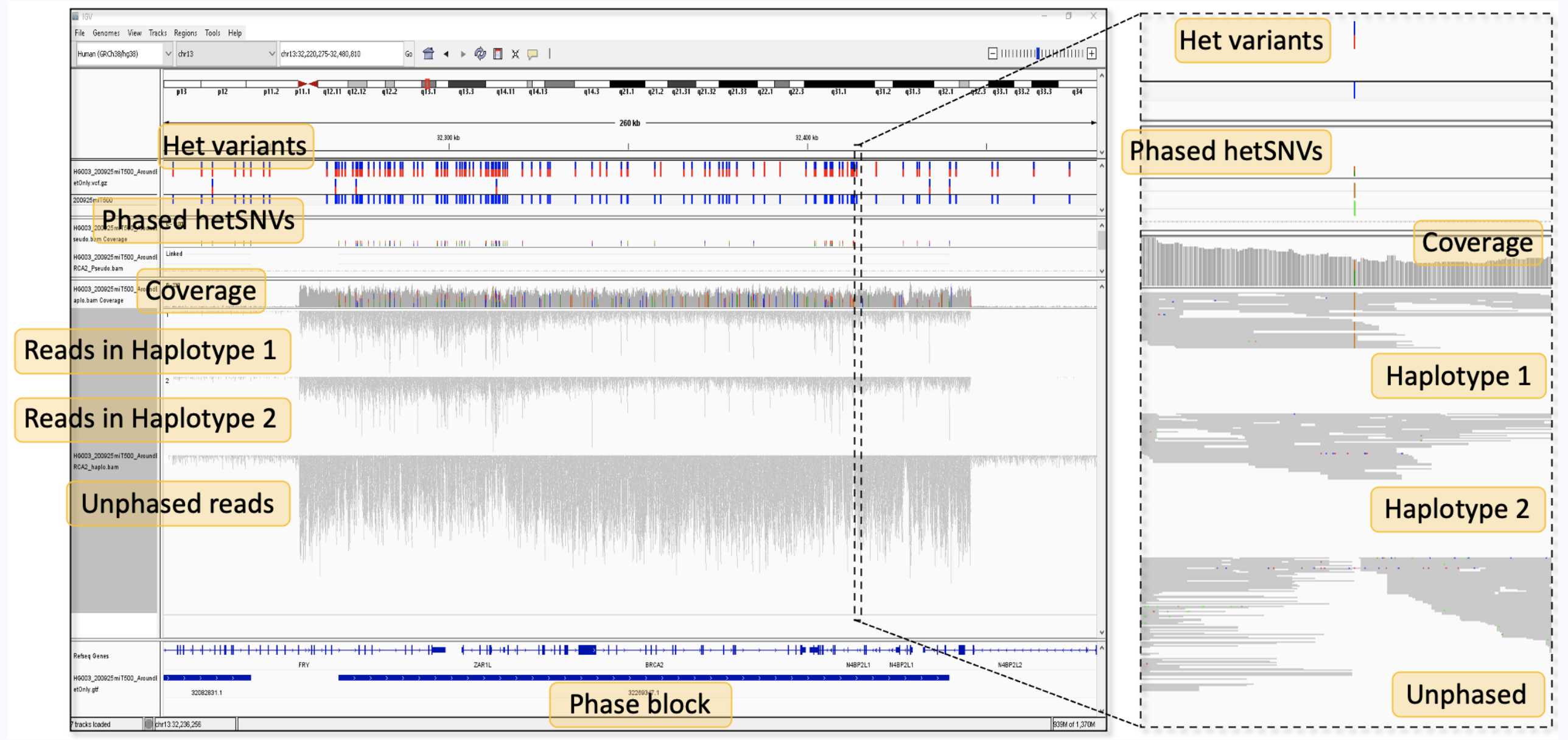
| | *PIK3CA* amplicons | | | | | | | *SCN10A* amplicons | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sample characteristics** | C#1 | C#2 | C#3 | C#4 | C#5 | C#6 | C#7 | C#8 | C#9 | C#10 | C#11 | C#12 | C#13 | C#14 |
| Priming method for cDNA (R6 vs dT) | R6 | R6 | R6 | R6 | dT | dT | dT | -- | -- | -- | -- | -- | -- | -- |
| DNA source (genomic, gen.) | -- | -- | -- | -- | -- | -- | -- | gDNA | gDNA | gDNA | gDNA | gDNA | gDNA | gDNA |
| PCR product size (kb) | 3.4 | 3.4 | 1.8 | 1.8 | 3.4 | 3.4 | 3.4 | 13 | 2.7 | 2.7 | 3.9 | 3.9 | 4.1 | 4.1 |
| *Target* input amount (pg) | 20 | 20 | 20 | 5 | 20 | 5 | 0.4 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| *Target* fraction over total DNA | 1/4 | -- | 1/4 | 1/14 | 1/5 | 1/20 | 1/250 | 1/4 | 1/4 | -- | 1/4 | -- | 1/4 | -- |
| *E.coli* gDNA amount (pg) | 60 | -- | 60 | 65 | -- | -- | -- | 60 | 60 | -- | 60 | -- | 60 | -- |
| Lambda phage gDNA amount (pg) | -- | -- | -- | -- | 80 | 95 | 99.6 | -- | -- | -- | -- | -- | -- | -- |
| Other human fragments (pg) | -- | 60 | -- | -- | -- | -- | -- | -- | -- | 60 | -- | 60 | -- | 60 |
| **Data analysis** | | | | | | | | | | | | | | |
| Technical replicates | 6 | 6 | 3 | 3 | 3 | 2 | 2 | 6 | 3 | 6 | 3 | 6 | 3 | 6 |
| Average on-target coverage (x) | 213 | 257 | 3,574 | 834 | 423 | 73 | 58 | 480 | 307 | 728 | 301 | 397 | 58.8 | 124 |
| Average on-target, dedup reads (number) | 6.0k | 7.2k | 5.5k | 12.8k | 12.4k | 2.1k | 1.6k | 57.8k | 6.8k | 15.9k | 9.9k | 12.6k | 2.1k | 4.2k |
| Expected hets per test | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 23 | 4 | 4 | 6 | 6 | 8 | 8 |
| Tests w/all expected hets recalled correctly | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Tests w/all expected hets phased correctly | 100% | 100% | 100% | 100% | 66% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

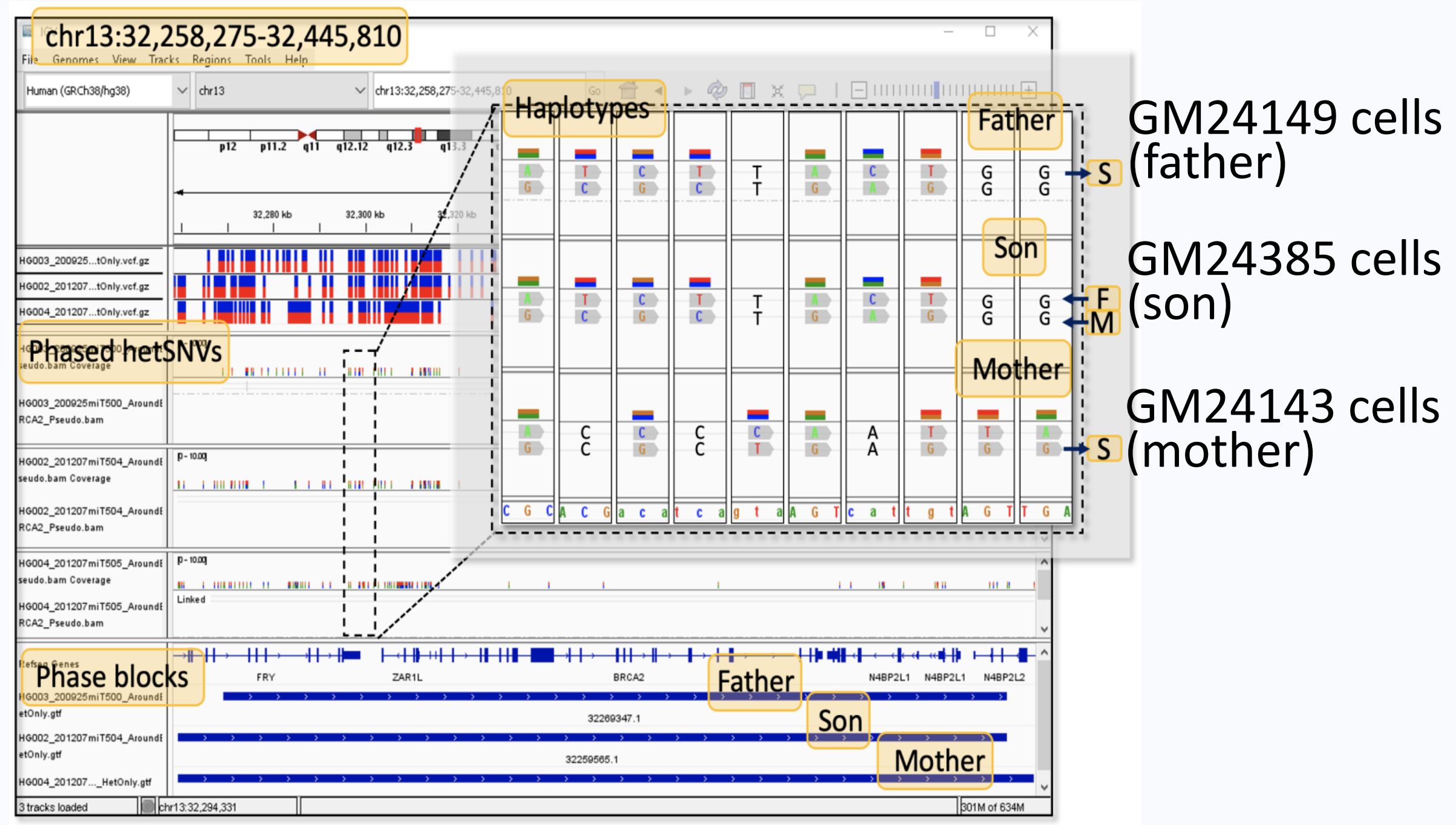
## 3. Results: Phasing of excised HMW DNA (187.5kb)



*BRCA2* locus
220,000-400,000 copies
200-400-fold-enrichment
3-7ng target DNA

GM24149 cells (father)





## 4. Material & Methods.

**DNA sources -** Human epithelial breast cancer HCC202 cells (ATCC®, Cat. #CRL-2316): RNA was extracted using the *Quick*-RNA Miniprep Kit$^{TM}$ (Zymo Research®, Cat. #R1054) and cDNA was generated using the SuperScript$^{TM}$ III First-Strand Synthesis System (ThermoFisher Scientific®, Cat. #18080051). Human lymphoblastoid GM12878 cells (Coriell Cell Repositories®, Cat. #GM12878): gDNA (NA12878) was extracted using a salting out method available on the UST website. *Escherichia coli* DH10B cells (New England Biolabs®, Cat. #FEREC0113): gDNA was extracted using the *Quick*-DNA Miniprep Kit (Zymo Research, Cat. #D3024). BstI P-digested lambda phage gDNA (TaKaRa®, Cat. #3402). Genome in a Bottle (GIAB) B lymphocytes for GM24385, GM24149, and GM24143: approximately 200kb HMW DNA segments containing the *BRCA2* gene were excised from one million cells using the HLS-CATCH system (Sage Sciences) with 2uM Cas9 and a pair of guide RNAs in each case (yield: 3-7 ng). **PCR amplification -** *PIK3CA* fragments were amplified with Phusion Hot Start II High-Fidelity polymerase (Thermo-Fisher Scientific, Cat. #F-565L) according to Vasan et al., (2019) and 2 ng of HCC202 cDNA. Cycling protocol: 98°C for 30 sec; 30 cycles of 98°C for 10 sec, 65°C for 20 sec, and 72°C for 1 min; final step at 72°C for 8 min. *SCN10A* fragments were amplified with Supreme NZYLong DNA Polymerase (NZYTech®, Cat. #MB331) according to Pinsach-Albuin et al., (2020) using NA12878 gDNA. Cycling protocol: 94°C for 5 min; 30 cycles of 94°C for 20 sec, 68°C for 30 sec, and 68°C for 14 min; final step at 68°C for 21 min. PCR cleaned up using ExoSap-IT (Thermo-Fisher Scientific , Cat. #78200.200.UL) and two rounds of 0.41x HighPrep$^{TM}$ PCR Clean-up beads (MagBio Genomics®, Cat. #AC-60050). **Library preparation & sequencing -** Libraries were generated using the TELL-Seq WGS Library Prep kit developed by UST, amplicon protocol available on the UST website, combining 20 pg of a single *PIK3CA* or *SCN10A* PCR product, 60 pg *E.coli* gDNA or 80 pg BstP I-digested lambda phage gDNA, and 3M TELL beads. When testing pools, we combined 20pg of each fragment (4) and 3M TELL beads without adding competitor DNA. We used 75,000 TELL beads for indexing (18-21 PCR cycles). (2×145 bp Illumina-compatible PE reads) and were sequenced on a MiSeq™ instrument (Illumina). For *BRCA2*, 100pg of CATCH'ed DNA was processed with 2M TELL beads: 2×150 PE reads. **Data analysis -** Sequencing output files were processed with Tell-read and Tell-Sort developed by UST. Tell-Sort incorporates BWA-MEM for read alignment, GATK-v4.0.3.0/HaplotypeCaller for variant calling (Broad Institute), and HapCUT2 (github/Vibansal) for phasing. Hg38 was used as reference for *SCN10A* and targeted *PIK3CA* exonic regions were used as reference for *PIK3CA*. The Integrative Genomics Viewer, IGV-v2.11.1 (Broad Institute) was used for data visualization assisted w/UST software available on the UST website.

## 5. References.

Bailey et al. *Cell*, 371-85 (2018); Chen et al. *Genome Res*., 30:898-909 (2020); Gorelick et al. *Nature*, 582:100-3 (2020); Klingstrom et al., *BioRxiv*, 10/1101/254276 (2018); Pinsach-Abuin et al. *Cell Rep*. 2:100250 (2021); The 1000 Genomes Project Consor., *Nature*: 526:68-74 (2015); Vasan et al. *Science*, 366:714-23 (2019); Zook et al. Sci. Data, 3:160025 (2016)