Supplemental Materials

Dense and accurate whole-chromosome haplotyping of individual genomes

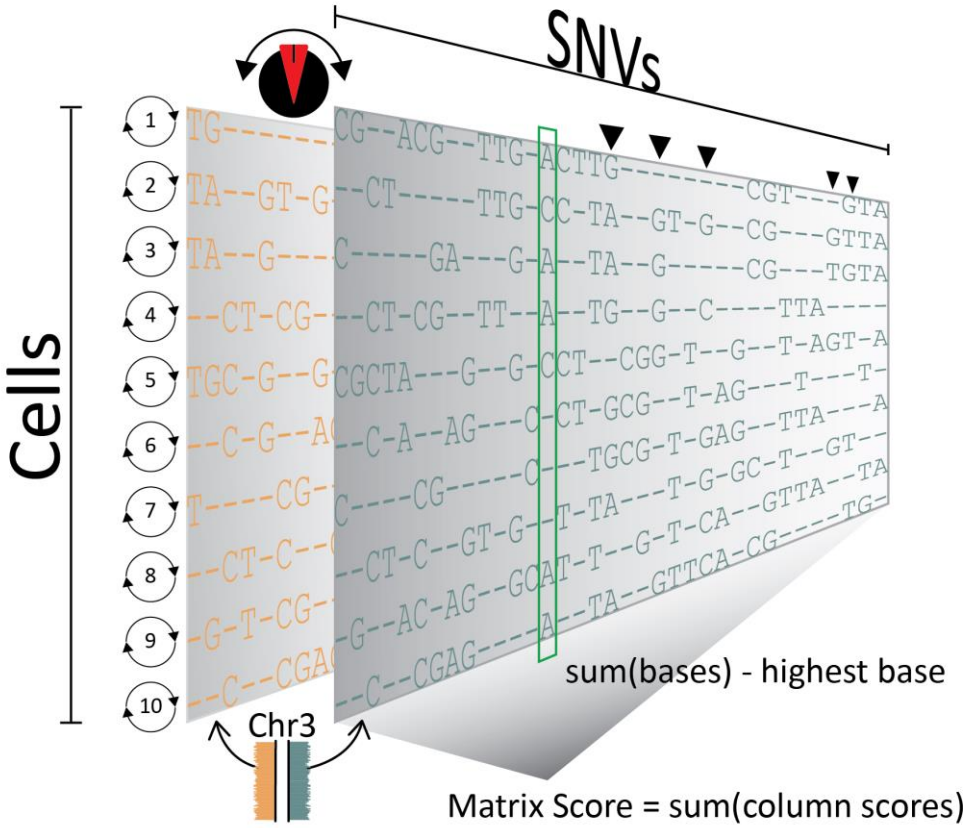
David Porubsky[1*], Shilpa Garg[2,3,4*], Ashley D. Sanders[5*], Victor Guryev[1], Peter M. Lansdorp[1,6,7], Tobias Marschall[2,3]

1. *European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, 9713 AV Groningen, The Netherlands*
2. *Center for Bioinformatics, Saarland University, Saarbrücken, Germany*
3. *Max Planck Institute for Informatics, Saarbrücken, Germany*
4. *Graduate School of Computer Science, Saarland University, Saarbrücken, Germany*
5. *European Molecular Biology Laboratory (EMBL), Heidelberg, Germany*
6. *Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada*
7. *Department of Medical Genetics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada*

Supplemental Figures 1-3

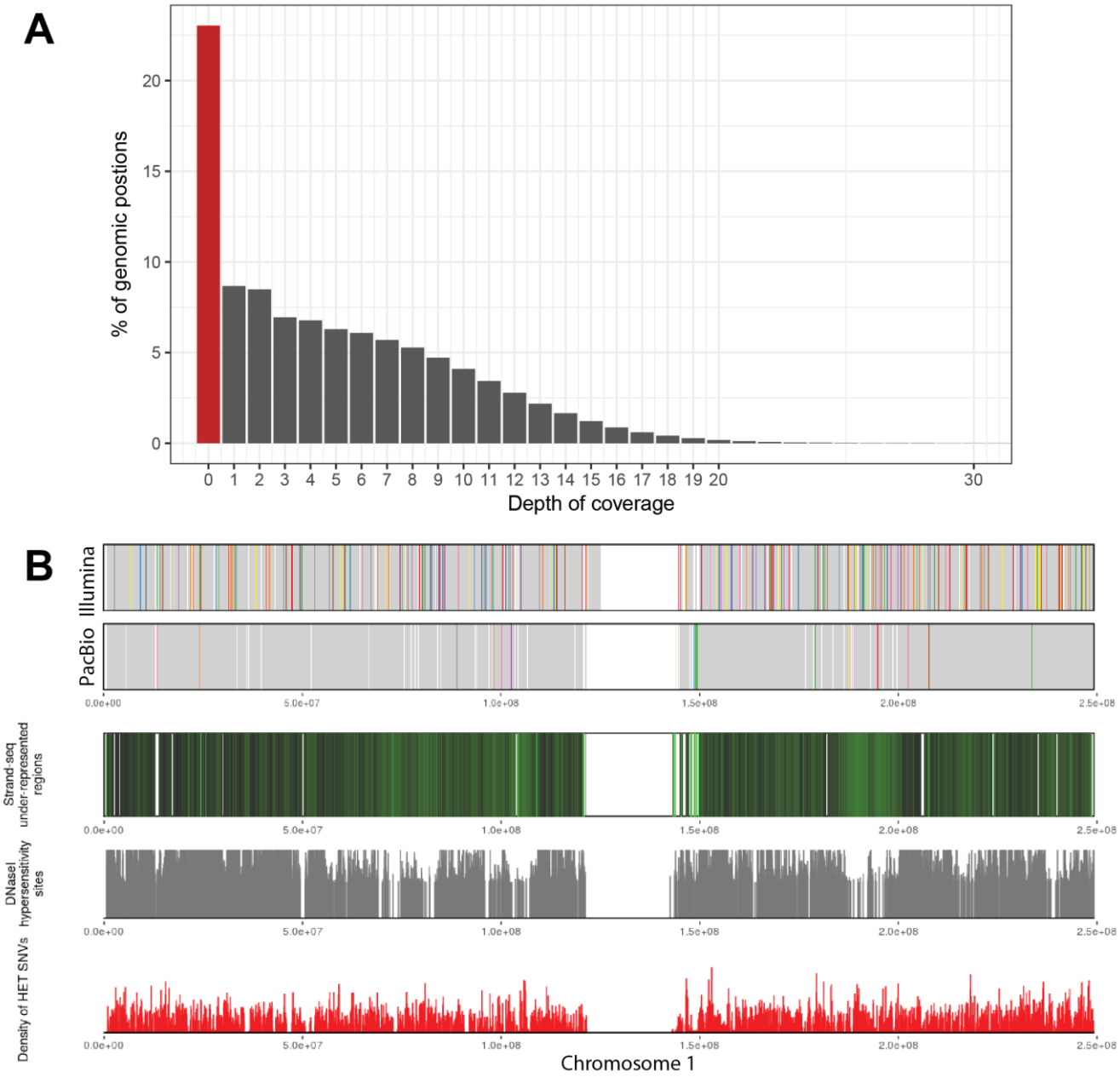
Supplemental Tables 1-3

Supplemental Figures 1-3



Supplemental Fig. S1: StrandPhaseR phasing algorithm.

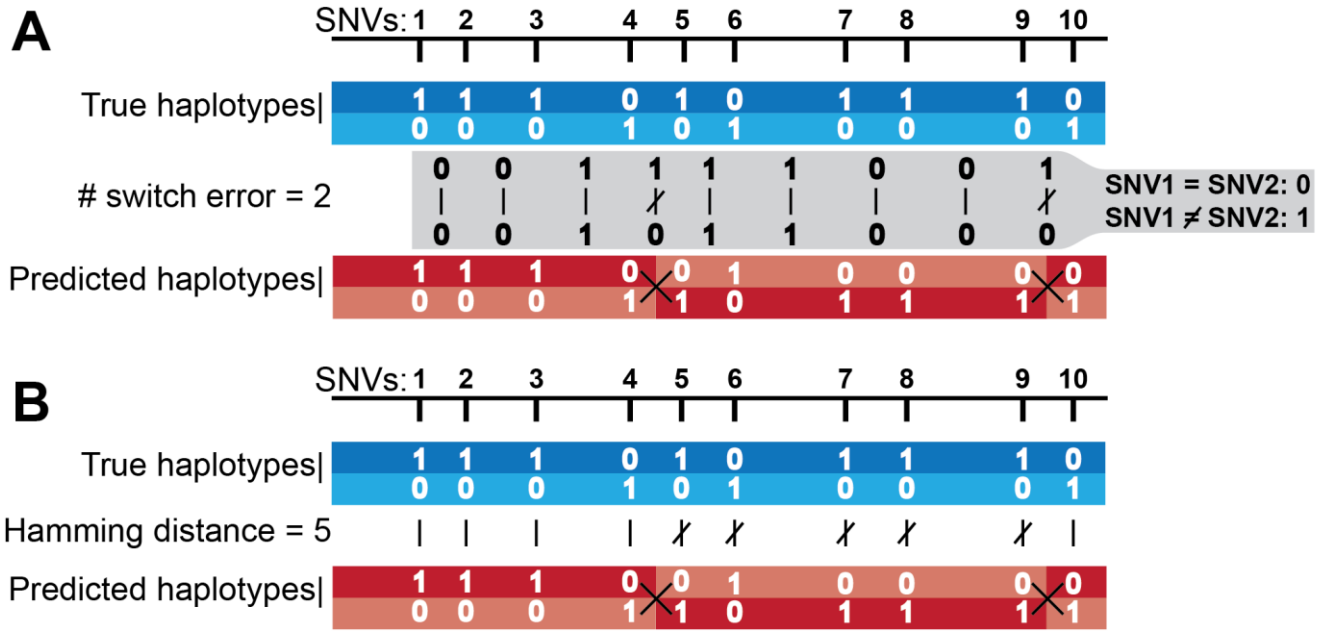
Two parallel matrices are shown. Rows represent single cell Strand-seq libraries and columns represent variants covered in those cells. Initially, one matrix stores all alleles (SNVs) covered by Watson reads and the second matrix stores all alleles covered by Crick reads (e.g. of Chromosome 3 - Chr3) separately for every single cell. Switch button at the top of the figure illustrates swap of alleles in every row between two matrices. Matrix score is calculated for each iteration to minimize the level of disagreement seen across the columns and determine the consensus haplotype.



Supplemental Fig. S2: Underrepresented genomic regions in Strand-seq libraries.

A) To assess the distribution of coverage in merged Strand-seq libraries (N = 134). Libraries were merged using SAMtools 'merge' function, reads were filtered for mapping quality 10, and the coverage of the genome was examined using SAMtools 'mpileup' function. On the x axis of the bar plot the depth of coverage is shown, with the corresponding percentage of the genome covered at the given depth, shown on the y-axis. The red bar highlights the portion of the genome (~ 23%) that was never covered in any Strand-seq library. The uncovered fraction may be caused by problems in assigning high-quality read alignment to the reference assembly, as well as by the inaccessibility of certain parts of the genome during library preparation. **B**) Even at high numbers of Strand-seq libraries (N = 134) there are a number of genomic regions that could not be stitched together into a single haplotype. The black and green density track shows the genomic regions that are represented in the merged Strand-seq data for chromosome 1, where bright green highlights the underrepresented regions. We found the low density linkage information in Strand-seq data correlated more with the DNaseI hypersensitivity

sites reported for NA12878 (grey track; obtained from UCSC genome browser) than with the corresponding SNV density (red track). This suggests the regions underrepresented in Strand-seq data may be due to chromatin organisation of the genome (as the Strand-seq protocol utilizes an MNase-digestion fragmentation step).



Supplemental Fig. S3: Quality measures used to evaluate predicted haplotypes.

Hypothetical phasing of 10 single nucleotide variants (SNVs) along a defined chromosomal region is shown here. Each heterozygous SNV is represented in its two allelic forms (0 – reference allele, 1 – alternative allele). True (reference) haplotypes are distinguished in blue colors and predicted haplotypes in red. **A**) To count the number of switch errors (black crosses) between the true and predicted haplotypes, neighbouring pairs of SNVs are compared along each haplotype and recorded as a new binary string of 0's and 1's depending on whether the allele state changes (see gray box). A zero value is assigned if the given pair of SNVs have the same value, otherwise a value of 1 is assigned value 1. The absolute number of differences in the binary string generated for the true and predicted haplotypes is reported as the total number of switch errors. **B**) To calculate the Hamming distance, the absolute number of differences between reference and predicted haplotypes is calculated for all SNV positions.

Supplemental Tables 1-3

Chromosome1	Illumina data	PacBio data
Sequencing protocol	Paired-end (102bp)	
AVG length of mapped DNA fragments	433 bp	15,057 bp
AVG depth of coverage	49,5x	39,6x

Supplemental Table S1: Summary measures for PacBio and Illumina data (example Chromosome 1).

# Strand-seq libaries	134
Sequencing protocol	Paired-end
Read length	100bp
AVG genome coverage per library	2.97%
AVG depth of coverage per library	0.037
Genome coverage in merged libraries	77.02%
Depth of coverage in merged libraries	5.004
Depth of coverage in all WC regions	2.641

Supplemental Table S2: Summary measures for Strand-seq libraries.

Genome coverage is calculated as a percentage of genomic positions (excluding gaps in the genome) covered with at least one read. Depth of coverage is calculated as an overall number of bases sequenced per genomic position (excluding gaps in the genome).

	Covered variants (%)	Switch error (%)	Hamming error (%)
Illumina+StrandS	68.1	0.45	0.99
PacBio+StrandS	95.56	0.25	0.91
10xGen+StrandS	98.13	0.05	2.18

Supplemental Table S3: Summary of integrative whole-genome phasing.