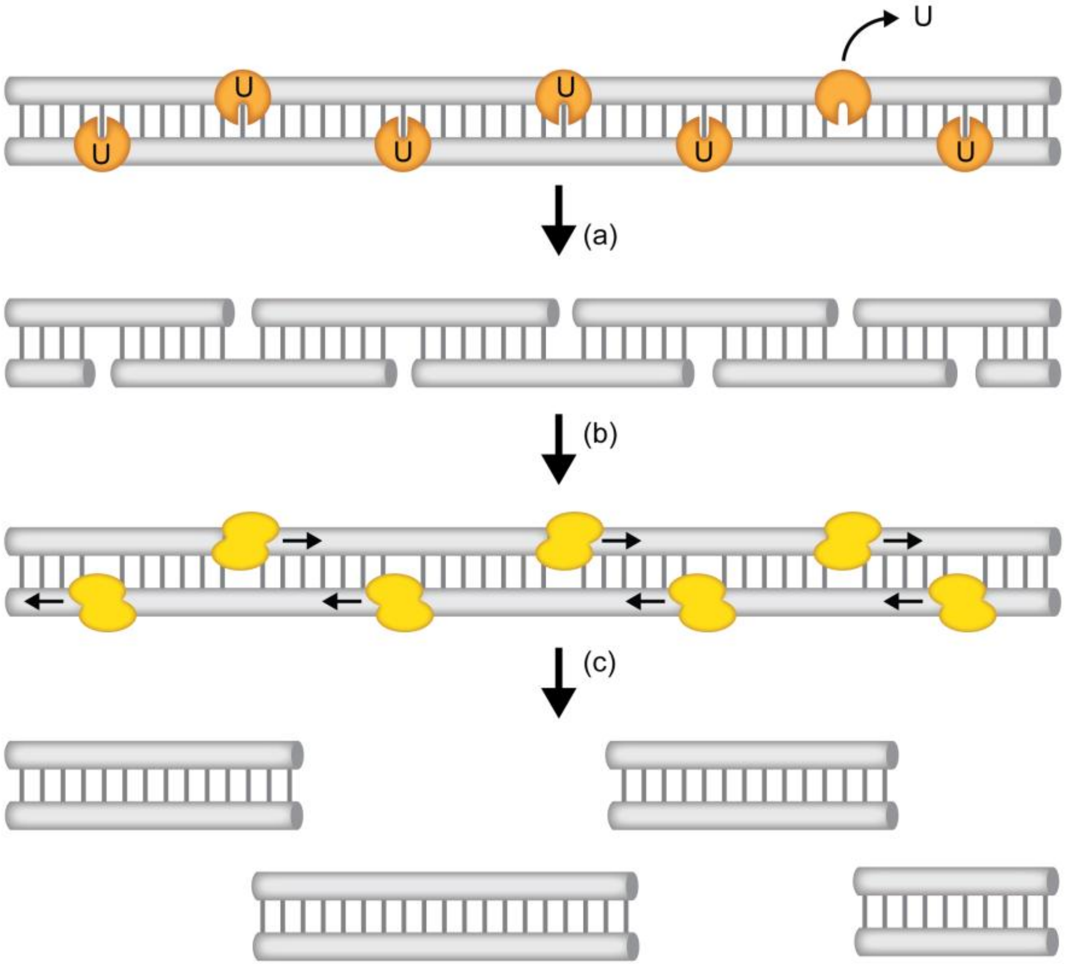
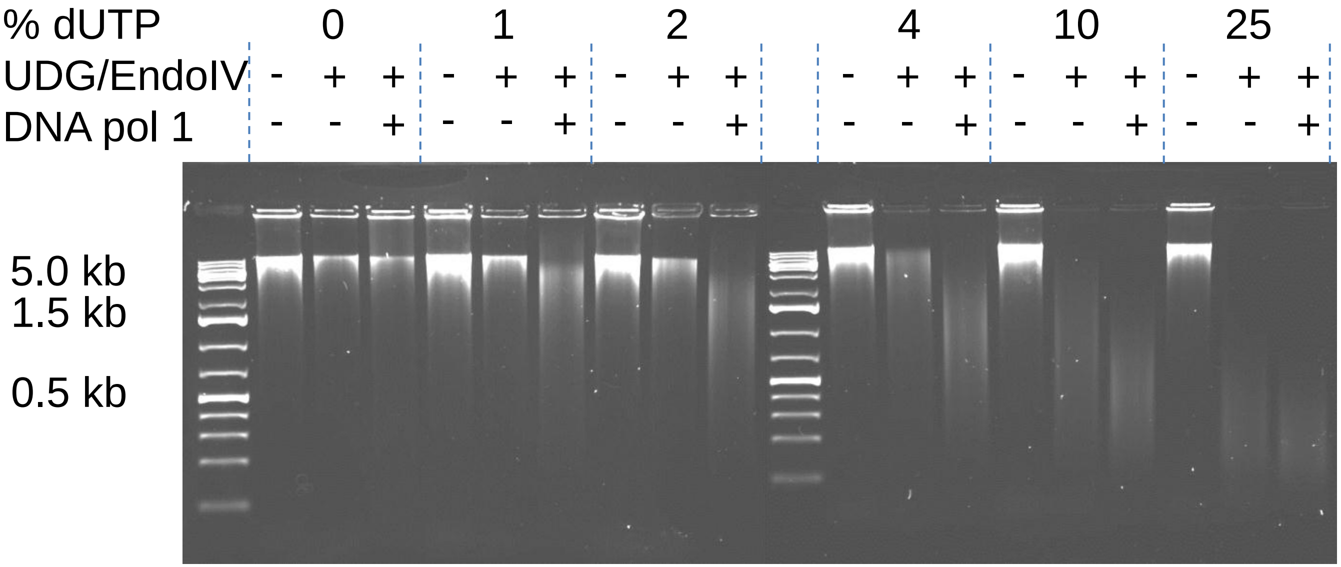
- **Long Fragment Read technology is a novel method for phasing genomes**
  - Supplementary Materials sections A and B, and Supplementary Methods
  - Figures 1 and 2, Supplementary Figure 2-4 and 6-10, Supplementary Tables 1-3
- **LFR phases up to 97% of heterozygous SNPs in a genome**
  - Supplementary Materials section B, and Supplementary Methods
  - Table 1, Supplementary Table 4
- **LFR can successfully sequence and haplotype from 10-20 cells**
  - Supplementary Materials, and Supplementary Methods
  - Table 1, Supplementary Table 4
- **LFR phases *de-novo* mutations**
  - Supplementary Table 10
- **LFR has a low phasing error rate**
  - Supplementary Table 8
- **LFR reduces the number of false positive SNVs incorporated in a genome**
  - Supplementary Table 11
- **LFR can “recall” regions of the genome not called due to low coverage**
  - Supplementary Materials section C
  - Supplementary Figure 12
- **Individual genomes contain many genes with inactiviting variations in both alleles**
  - Supplementary Methods
  - Table 2
- **6 examples of allelic expression difference linked to transcription factor binding site disruptions through LFR contigs**
  - Supplementary Materials section D and Supplementary Methods
  - Supplementary Table 12
- **Regions of low heterozygosity (RLHs) in non-African populations and their effects on phasing performance**
  - Supplementary Materials section E
  - Supplementary Figure 11 and Supplementary Tables 5-7
- **Highly divergent haplotypes in African and non-African genomes discovered by LFR**
  - Supplementary Materials section F
  - Supplementary Table 13

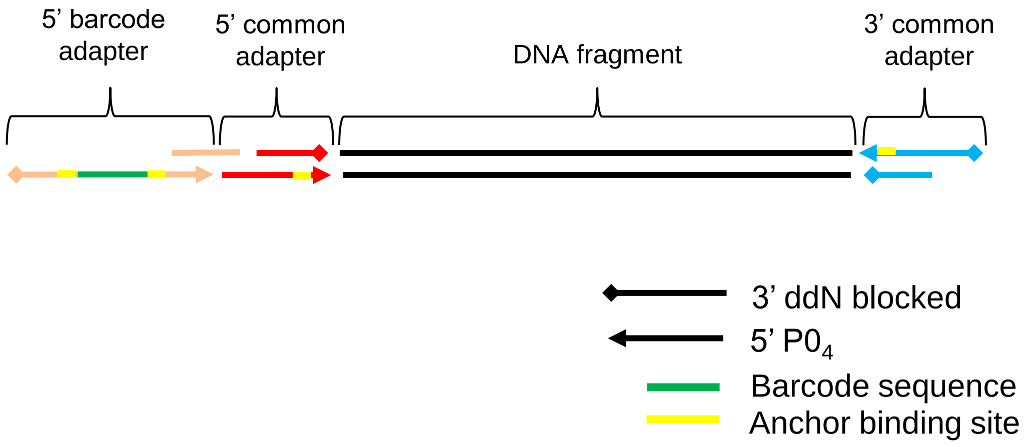
Supplementary Figure 1. Overview of main results and corresponding supporting figures, tables, and descriptions.



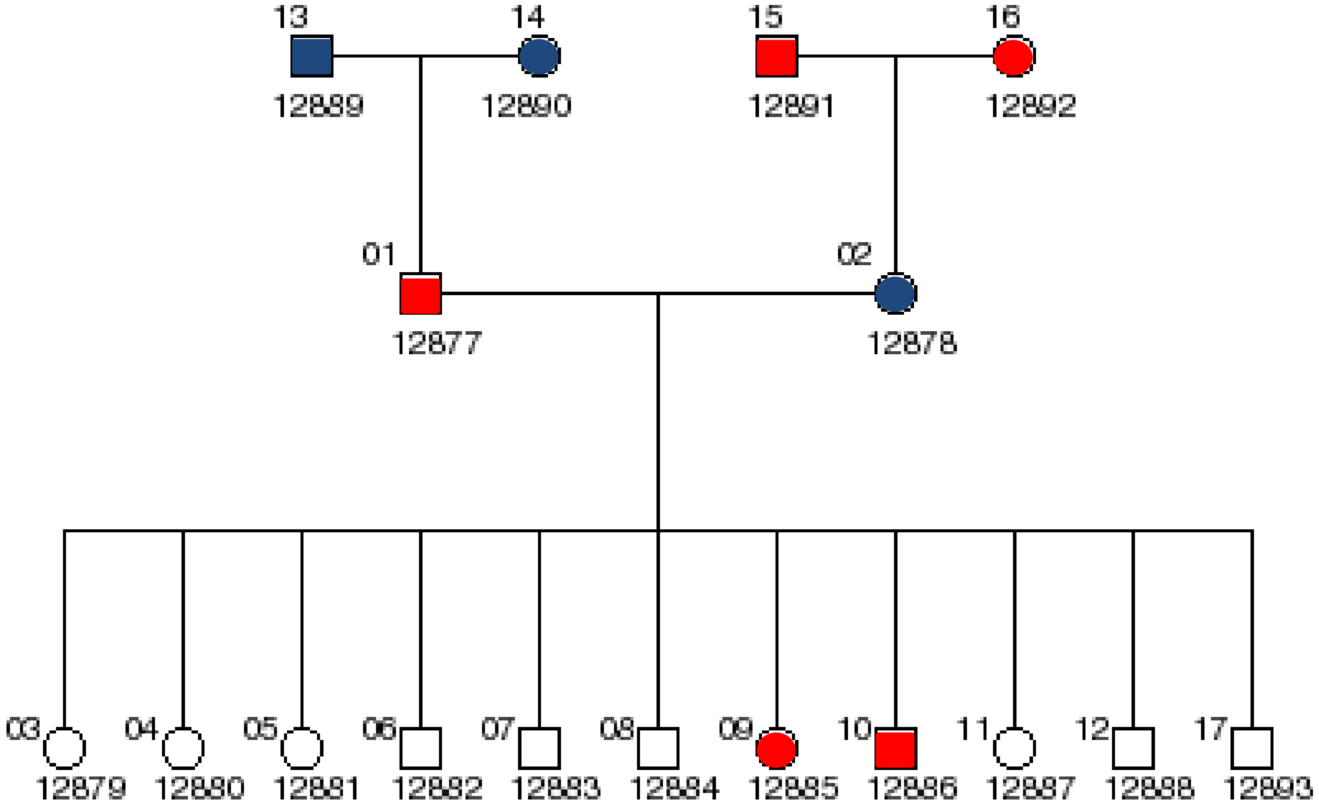
Supplementary Figure 2. Controlled Random Enzymatic (CoRE) fragmentation. CoRE fragmenting begins with MDA products containing the uracil base incorporated at a specific frequency. (a) The MDA product is treated with uracil DNA glycosylase and endonuclease IV to introduce 1 base pair gaps, (b) DNA polymerase I extends from each gap, (c) after DNA polymerase I extension the MDA is resolved into 300-1,500 base pair blunt end fragments.


% dUTP
0
1
2
4
10
25
UDG/EndoIV - + + - + + - + + - + + - + + - + +
DNA pol 1 - - + - - + - - + - - + - - + - - +
5.0 kb
1.5 kb
0.5 kb

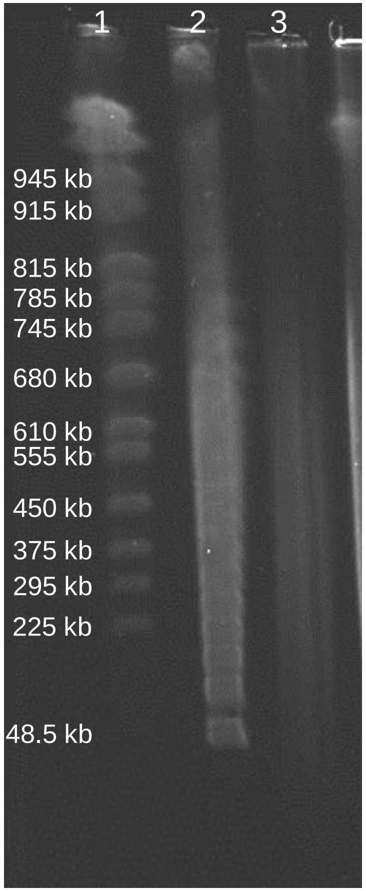
Supplementary Figure 3. Demonstration of DNA fragmenting by CoRE. MDA material of varying degrees of uracil incorporation was fragmented by removal of uracil with uracil DNA glycosylase (UDG) and Endonuclease IV (EndoIV) followed by nick translation with DNA polymerase 1.



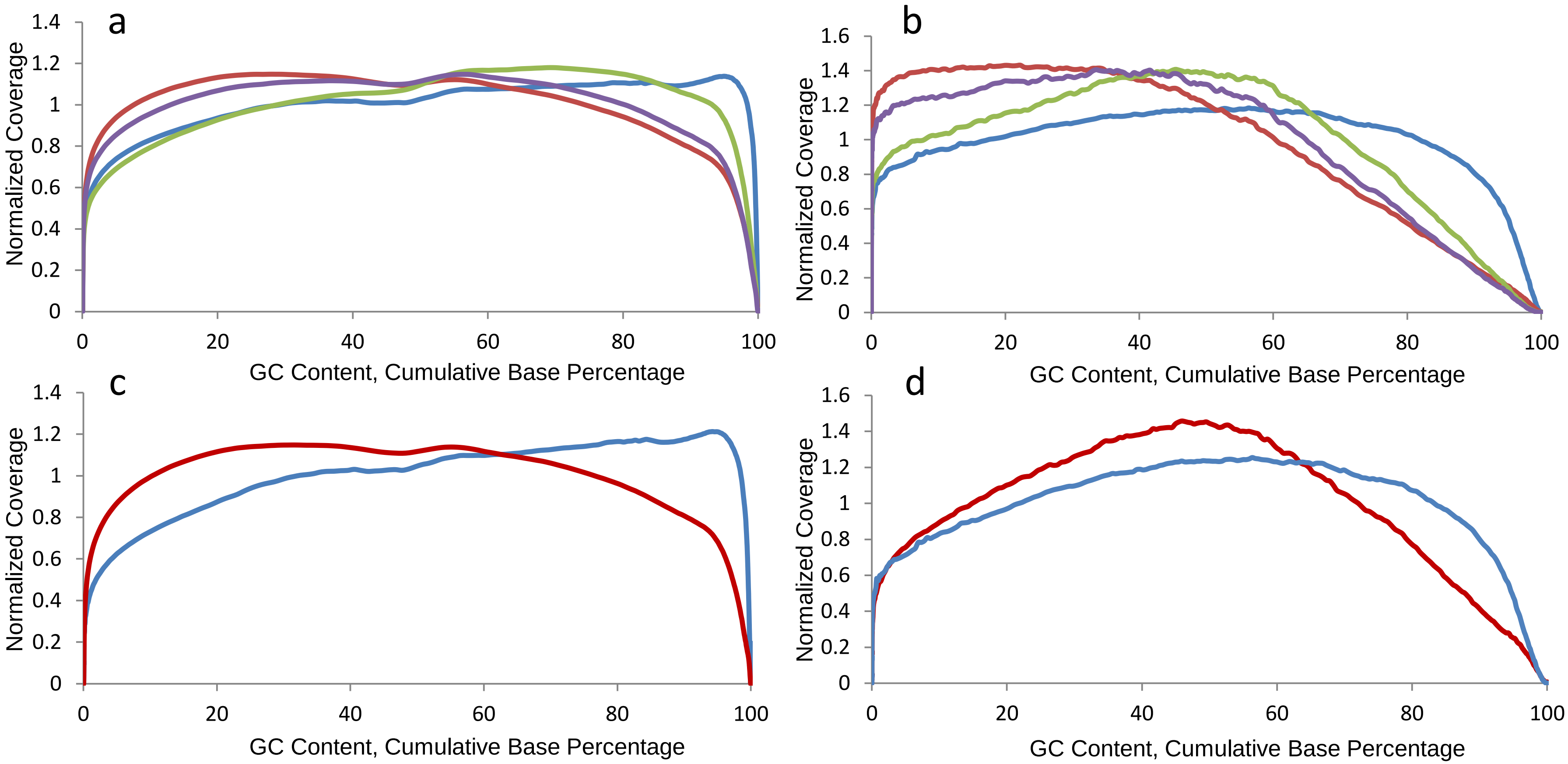
Supplementary Figure 4. LFR barcode adapters. LFR adapters are composed of a unique 5’ barcode adapter, a common 5’ adapter, and a common 3’ adapter. The common adapters are both designed with 3’ dideoxy nucleotides that are unable to ligate to the 3’ fragment, this eliminates adapter dimer formation. After ligation, the block portion of the adapter is removed and replaced with an unblocked oligo. The remaining nick is resolved by subsequent nick translation with *Taq* polymerase and ligation with T4 ligase.



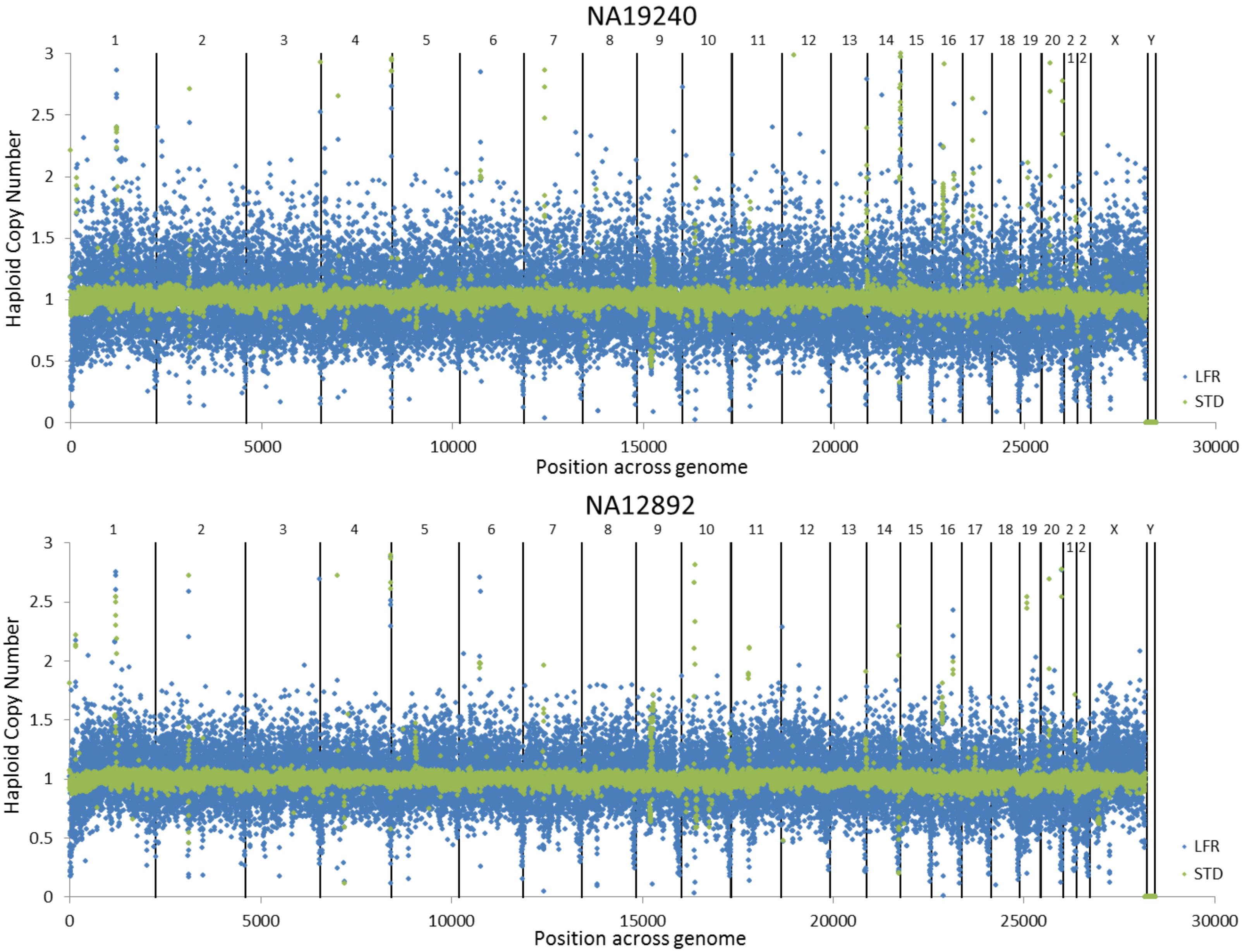
Supplementary Figure 5. CEPH Pedigree 1463. Colored individuals were used in the various analyses performed in this paper. Individuals in red were analyzed using LFR and their complete genome sequences were part of a public release of data by Complete Genomics. The complete genome sequences of individuals in blue were provided as part of the Complete Genomics' data release.



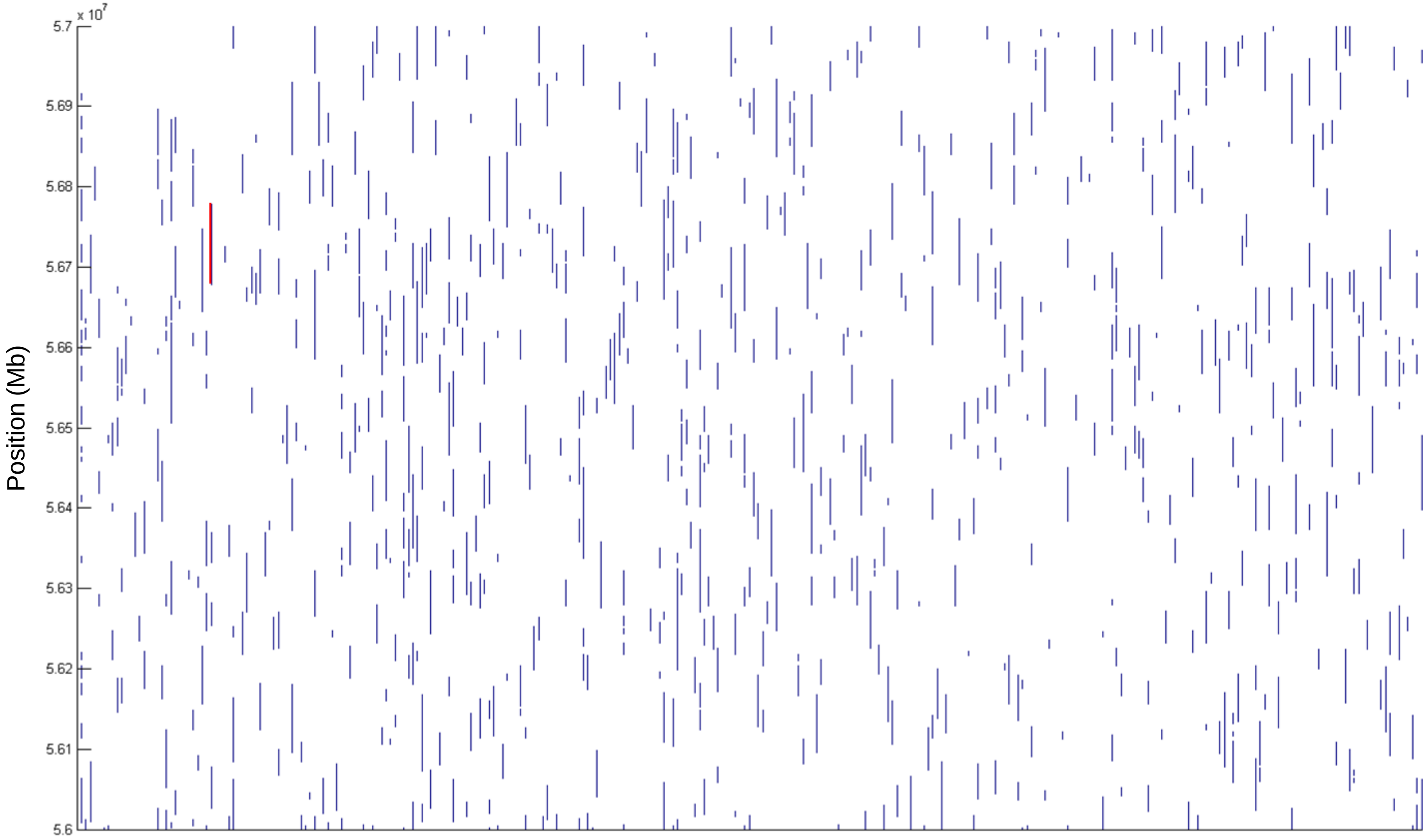
Supplementary Figure 6. NA19240 genomic DNA length. Genomic DNA, isolated from NA19240 cells using a dialysis purification method, was analyzed by pulse-field electrophoresis on a BioRad CHEF-DR II (see Methods). Lane one contains 500 ng of Yeast Chromosome PFG Marker, lane 2 contains 500 ng of Lambda Ladder PFG Marker, and lane 3 contains 200 ng of NA19240 genomic DNA. The majority of NA19240 DNA is longer than 225 kb.



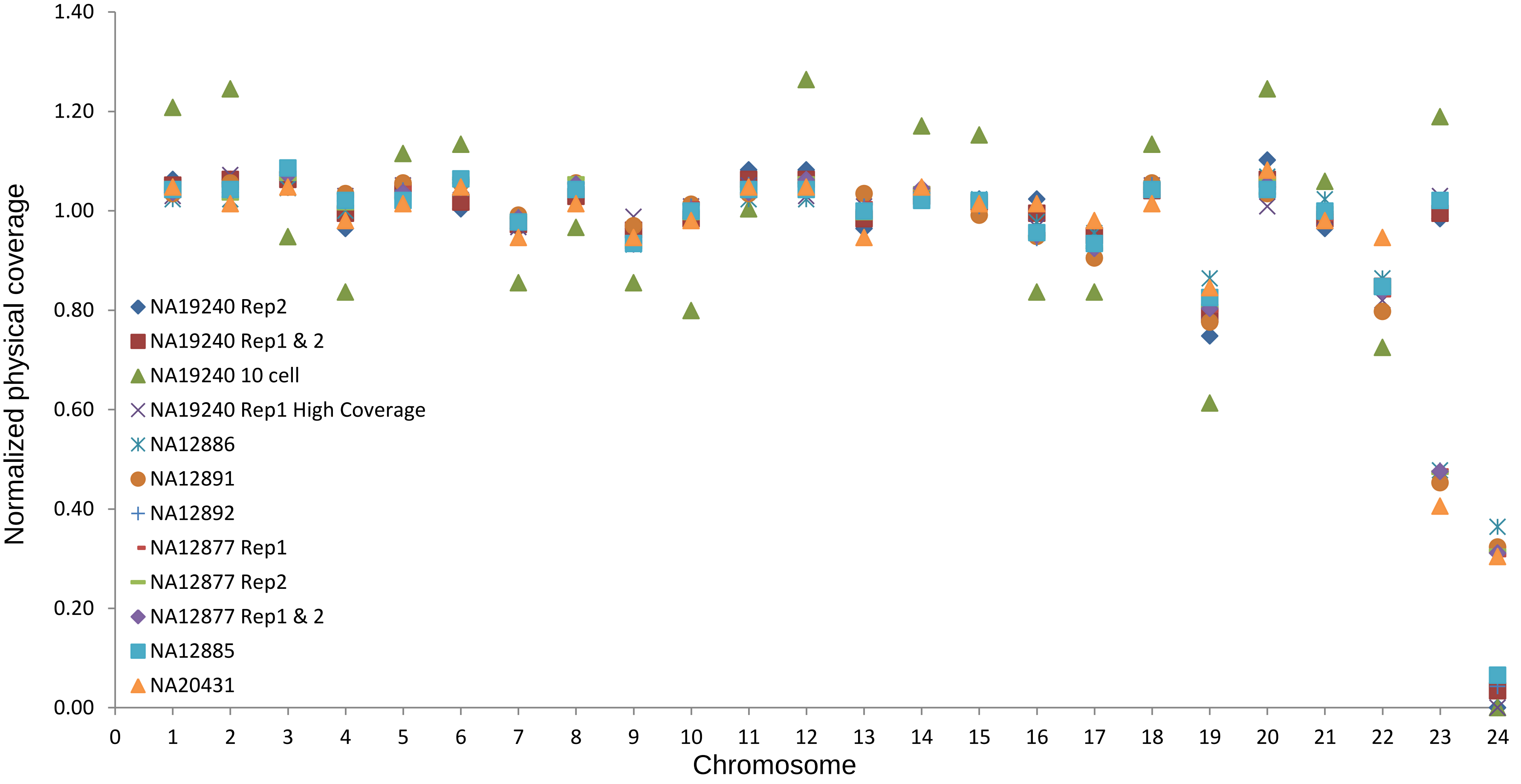
Supplementary Figure 7. Cumulative GC coverage plots. Cumulative coverage of GC was plotted for LFR and standard libraries to compare GC bias differences. For sample NA19240 (a and b), three LFR libraries (Replicate 1, red, Replicate 2, green, and 10 cell, purple) and one standard library were analyzed (blue). Coverage plotted for the entire genome (a) and the coding regions only (b). The same was plotted for NA12892 (c and d). One LFR library (red) and one standard library (blue) are plotted for both the entire genome (c) and the coding only portions (d). In all LFR libraries a loss of coverage in high GC regions is evident, this is more pronounced in coding regions (b and d) which contain a higher proportion of GC rich regions.



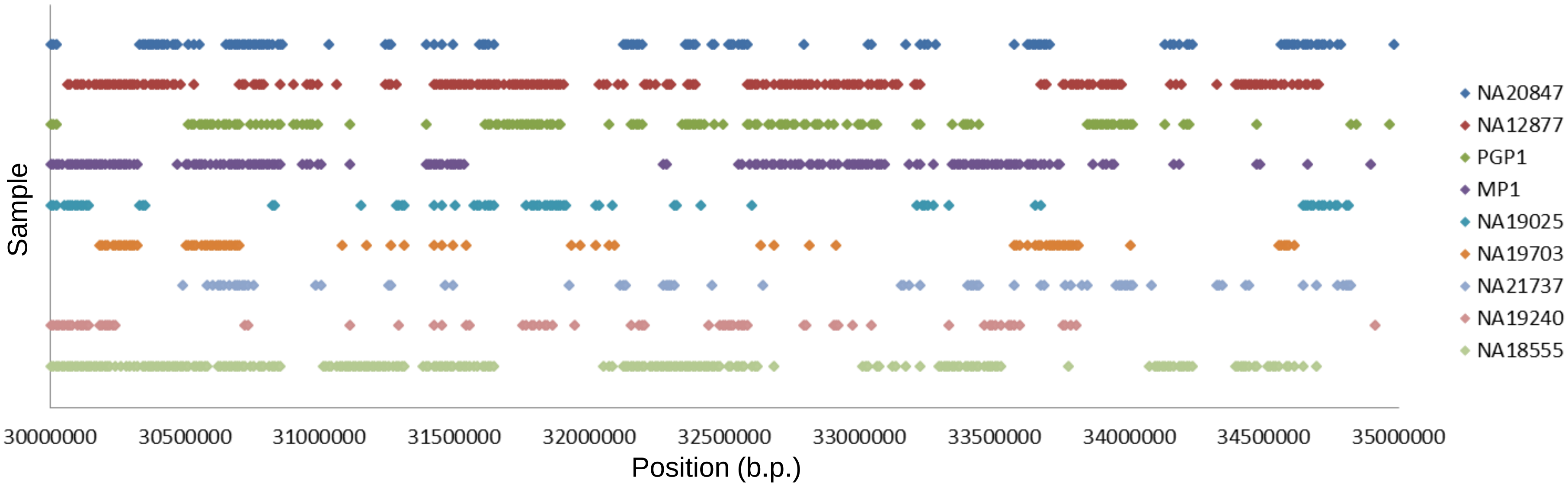
Supplementary Figure 8. 100kb variability plots. Coverage was binned in consecutive 100kb windows, normalized to the total coverage for each library and plotted by position across the genome. Both LFR (blue) and standard (green) libraries were plotted for samples NA19240 and NA12892. LFR introduces approximately 2 fold more coverage variability.



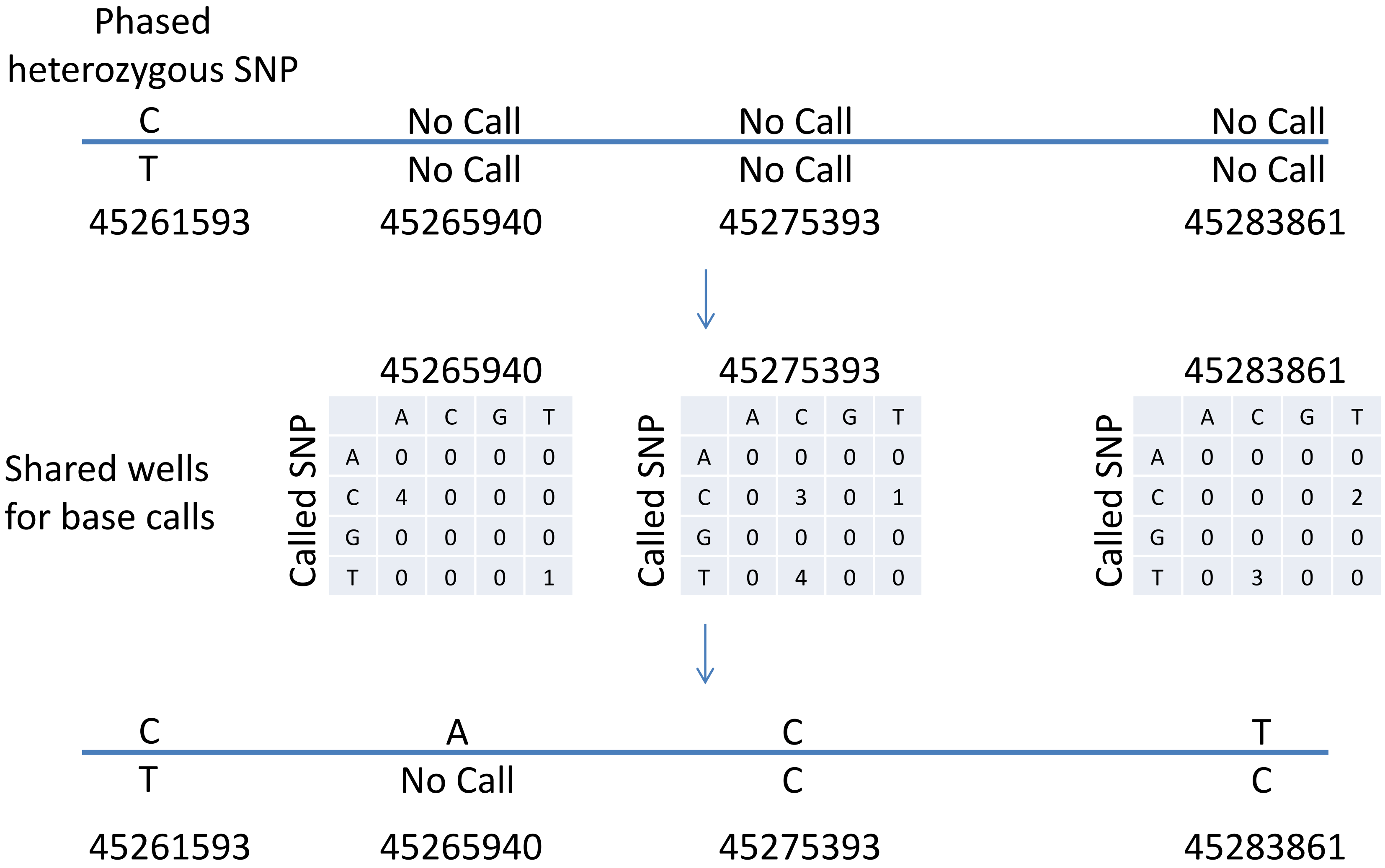
Supplementary Figure 9. Fragment coverage per well. Coverage of single molecule fragments from LFR library NA19240 replicate 1 were mapped across a 1 Mb region of chromosome 15 for 300 independent wells. Blue dashes represent contiguous DNA molecules. Each column of dashes represents and individual well. The red molecule towards the top left of the figure is 100kb in length.



Supplementary Figure 10. Physical chromosome coverage. In general LFR fragments were evenly distributed across all chromosomes. The 10 cell NA19240 library shows more variability, this is likely do to the cells being in different phases of the cell cycle at the time of isolation. Chromosomes 19 and 22 are consistently underrepresented in all libraries. This is likely because these chromosomes have the highest GC content in the genome and are amplified less efficiently by MDA. Chromosome number 23 is the X chromosome and 24 is the Y and have the expected half coverage for X in males and no coverage for Y in females.



Supplementary Figure 11. Contiguous stretches of low heterozygosity. Regions of low heterozygosity were plotted along a 5 Mb region of chromosome 18 for 8 ethnically diverse genomes. Each dot represents a 10 kb segment of the genome with less than 1.4 heterozygous SNPs within that particular sample.



Phased heterozygous SNP

45261593	45265940	45275393	45283861
C	No Call	No Call	No Call
T	No Call	No Call	No Call

Shared wells for base calls

45265940

Called SNP	A	C	G	T
A	0	0	0	0
C	4	0	0	0
G	0	0	0	0
T	0	0	0	1

45275393

Called SNP	A	C	G	T
A	0	0	0	0
C	0	3	0	1
G	0	0	0	0
T	0	4	0	0

45283861

Called SNP	A	C	G	T
A	0	0	0	0
C	0	0	0	2
G	0	0	0	0
T	0	3	0	0

45261593	45265940	45275393	45283861
C	A	C	T
T	No Call	C	C

Supplementary Figure 12. LFR recalling of no call positions. To demonstrate the potential of LFR to rescue no call positions three example positions were selected on chromosome18 that were uncalled by standard software. By phasing them with a C/T heterozygous SNP that is part of an LFR contig these positions can be partially or fully called. The distribution of shared wells (wells having at least one read for each of two bases in a pair; there are 16 pairs of bases for an assessed pair of loci) allows for the recalling of three N/N positions to A/N, C/C and T/C calls and defines C-A-C-T and T-N-C-C as haplotypes. Using well information allows LFR to accurately call an allele with as few as 2-3 reads if found in 2-3 expected wells, about 3-fold less than without having well information.