Supplementary Materials

Section A: Optimal number of cells or amount of DNA needed per LFR library

The optimal input number of cells or amount of DNA needed to make a LFR library primarily depends on the total effective read coverage and the number of DNA aliquots. The optimal input maximizes the number of usable shared wells between two heterozygote loci. More input DNA increases the number of representative molecules for each genomic region but decreases the read coverage per each molecule (and makes them unusable when both loci are not read on the same molecule) and increases the number of unusable aliquots that comprise both alleles.

Based on a total effective read coverage of 60x and using 384 distinct pools of fragments we estimated that the optimal number of starting cells would be 10 if the DNA was denatured before dispensing in wells (equivalent to 20 cells of dsDNA; Supplementary Table 2).

It was also determined that when starting from denatured isolated DNA the optimal amount per library would be ~100 pg (Supplementary Table 2). This amount was selected to achieve more uniform genome coverage by minimizing stochastic sampling of fragments.

The following definitions, conditions and calculations are used in Supplementary Table2. A shared well is defined as a well that has DNA only from one haplotype at a given loci and at least one read for both alleles at loci 10kb apart assuming initial DNA fragments are 100kb in length. We limited calculations to our 384-aliquot process and evaluated six levels of cells/DNA, from 10 to 60 cells. We based our calculations on the 80x total mapped read coverage of the genome and estimated that such total read coverage will result in 60x effective read coverage due to loses of reads from improperly read barcodes or a biased excess of read coverage in a fraction of the genome.

For each input cell level we first calculated the effective coverage per allele by dividing the effective read coverage by the number of used cells multiplied by 2. For 10 cells that is 60/10x2=3x coverage. Then, we calculated the usable wells per haplotype by taking into account wells having more than one fragment with the same haplotype or mix of both haplotypes. We used this formula that approximates the expected number of usable wells per haplotype in a 384-well plate: Cells-(1.75x cells x (cells/384). For 60 cells the result is 60-1.75x60x(60/384)=60-16=44, ~70% of maximal number of 60. For 10 cells, the usable number is 9.5, close to maximal number of 10. Then, we calculated the percentage of bases read in each molecule by using binomial distribution on the effective coverage per allele: 1-0.999exp(1000xeffective coverage per allele). For 60X total effective coverage, for 10 cells the percentage of read bases is 1-0.999exp3000=95% and for 60 cells that is 1-0.999exp500=39%. At the end we calculated Shared wells as (usable wells) x (%bases read)exp2 x (0.9, frequency of not separating two 10kb apart loci on two 100kb fragments). Squaring "% bases read" defines fraction of cases where two heterozygous loci are read on the same molecule. Because the number of usable wells increases from 10 to 60 cells and the percentage of read bases is reduced from 95% to 39%, this defines 20 cells as the optimal number.

Section B: Genome sequencing and haplotyping from 100pg of DNA using only LFR libraries

All of the prior analyses in this paper incorporated heterozygous SNPs from both a standard and an LFR library. However, it should be possible to use only an LFR library given that full representation of the genome is expected as a result of starting with an amount of DNA

equivalent to that found in10-20 cells. We have demonstrated that our MDA provides a sufficiently uniform amplification and with high (80x) overall read coverage an LFR library alone allows for detection of up to 93% of heterozygous SNPs without any modifications to our standard library variation calling algorithms[32] (Supplementary Table 3). To demonstrate the potential of using only a LFR library, we phased NA19240 Replicate 1 as well as an additional 250 Gb of reads from the same library (500 Gb total). We did see 15% and 5% reductions, respectively, in the total number of SNPs phased (Table 1). This result is not surprising given that this library was made from 60 pg of DNA instead of the optimal amount of 200 pg (Supplementary Table 2) and given the previously mentioned GC bias incorporated during *in vitro* amplification by MDA. Another 285Gb LFR only library (NA12892) called and phased 90% of all variants phased using SNP loci from both standard and LFR libraries combined (Table 1). Despite the reduction in total SNPs phased, the N50 contig lengths in NA12892 and NA19240 were still greater than 500 kb and 1 Mb, respectively.

Section C: Improving base calling with LFR information

In addition to phasing and eliminating false-positive heterozygous SNVs, LFR can "rescue" "no-call" positions or verify other calls (*e.g.,* homozygous reference or homozygous variant) by assessing the well origin of the reads that support each base call. As a demonstration we found positions in the genome of NA19240 replicate one that were not called but were adjacent to a neighbouring phased heterozygous SNP. In these examples the position was able to be "recalled" as a phased heterozygous SNP do to the presence of shared wells between the neighbouring phased SNP and the no-call position (Supplementary Figure 10). While LFR may

not be able to rescue all no call positions, this simple demonstration highlights the usefulness of LFR in more accurate calling of all genomic positions to reduce no-calls.

Section D: TFBS disruption linked to differences in allelic expression

Long haplotypes that encompass both cis-regulatory regions and coding sequences are critical for understanding and predicting expression levels of each allele of a gene. By analysing 5.6 Gbs of non-exhaustive expression data from RNA sequencing of lymphocytes (Supplementary Methods) from NA20431 we identified a small number of genes that have significant differences in allele expression. In each of these genes 5 kb of the regulatory region upstream of the transcription start site and 1kb downstream were scanned for SNVs that significantly alter the binding sites of over 300 different transcription factors[38]. In 6 examples (Supplementary Table 13), 1-3 bases between the two alleles were found to differ in each gene causing a significant impact to one or more putative binding sites and potentially explaining the observed differential expression between alleles. While this is just one data set and it is not currently clear how large an impact these changes have on transcription factor binding, these results demonstrate that with large scale studies of this type[41], that become feasible using LFR haplotyping, the consequences of sequence changes to transcription factor binding sites may be elucidated.

Section E: Origin and impact of regions of low heterozygosity in non-African populations

There are approximately two-fold more regions of low heterozygosity (RLHs, defined as genomic regions of 30 kb with less than 1.4 heterozygous SNPs per 10 kb, approximately 7 times lower than the median density) of 30 kb-3 Mb in the European pedigree samples than in NA19240 (Supplementary Table 4) clarifying a previously reported relative excess of

homozygotes in non-Africans[36,42] and further supported by an analysis of 52 complete genomes (Nicholas Schork, personal communication). These regions are barriers to phasing, resulting in a two-fold smaller N50 contig length. Over 90% of the contigs in European genomes end in these RLHs that vary between unrelated individuals (Supplementary Figure 9).

Approximately 3% of all heterozygous SNPs in non-African genomes (30-60% of all non-phased heterozygous SNPs) belong to these RLHs which cover a very large fraction (30-40%) of these genomes (Supplementary Table 4). In Chinese and European genomes, long RLHs cluster around 45 heterozygous SNPs per Mb (Supplementary Table 6; the genomic average is approximately 1000 per Mb outside RLHs) indicating they shared a common ancestor around 37,000-43,000 years ago (based on a mutation rate of 60-70 SNPs per 20-year generation[33,43]). This is probably due to a strong bottleneck at the time of or after the human exodus out of Africa and within a previously determined range from 10,000-65,000 years ago[44]. Furthermore, an excess of RLHs is observed on the X chromosome in European and Indian women (NA12885, NA12892, and NA20847) when compared to an African woman (NA19240) covering ~50% vs. 17% of this chromosome, respectively (Supplementary Table 4; 30% vs. 14% for the entire genome in these same individuals). This indicates an even stronger out-of-Africa bottleneck for the X chromosome. A possible explanation is that substantially fewer females left Africa and had offspring with multiple males.

These observations suggest that whole genome variation analyses, including haplotyping, in thousands of diverse genomes will provide a deep understanding of human population genetics and the impact of these extensive “inbred” regions, frequently comprising >100 homozygote

variants each, on human disease and other extreme phenotypes. In addition, it shows that about 2,000 RLHs >100 kb in length will be present in all non-African individuals (Supplementary Table 7). Populations with limited numbers of high-frequency haplotypes, as can result from recent bottlenecks or in-breeding[42], can also have long runs of identical heterozygous SNPs present in both parents, limiting use of parents for phasing or assigning shorter LFR contigs. Thus, population history and some reproduction patterns can make phasing challenging, as exhibited by the X chromosome of non-African woman. Regardless of these factors LFR phasing performance is approximately equivalent with up to 97% of heterozygous SNPs phased in both European and African individuals, a result that should translate across all populations. In addition to combining LFR with standard genotyping of one parent as described below (a strategy that will be more limited in some families, as discussed above), using initial DNA fragments longer than 300 kb, for example by entrapping cells or pre-purified DNA in gel blocks[45], would span ~95% of all RLHs (Supplementary Table 7) and haplotype most of the de-novo mutations that occur in these regions. This would not be feasible with current fosmid cloning strategies[11,13] which are limited to 40 kb fragments.

Section F: Highly divergent haplotypes present in African and non-African genomes

Haplotype analyses, enabled by large scale genotyping studies such as the HapMap project, have been immensely important to understanding population genetics. However, the resolution of the complete haplotypes of individuals has largely been intractable or prohibitively expensive. Highly accurate haplotypes, filtered of clustered false heterozygotes accumulated due to false mapping of repeated regions[43,44], will help understand many of the population phenomena found within individual genomes. As a demonstration, we scanned the LFR contigs of NA19240 for

regions of high divergence between the maternal and paternal copies. 7,000 10 kb regions containing >33 SNVs were identified; a threefold increase over the expected 10 SNVs. Assuming 0.1% standing variation and 0.15% base difference per 1 Myr (based on the 1% divergence of human and chimpanzee genomes evolving from a common ancestor ~6 Myr) our calculations suggest that ~50 Mb of these regions found in this African genome (~2.0% of "non-inbred" genome) may have been evolving separately for over 1.5 million years (Supplementary Table 12). This estimate is closer to 1Myr if the chimpanzee-human separation was less than 5Myr ago[46]. This whole genome analysis is in agreement with a recent study by Hammer *et al.*[47] on a few targeted genomic regions in African populations postulating a possible interbreeding of separate *Homo* species in Africa. Our analysis shows that 2.1% of European non- inbred genomes (Supplementary Table 12) also have similarly diverged sequences mostly at distinct genomic positions. The majority of these were likely introduced prior to the exodus of humans from Africa. As indicated by this preliminary analysis, reproduction patterns[44], retention of old haplotypes[46], and cross species mating[48] are some of the research areas that will benefit immensely from the 7 phased African and non-African genomes presented in this paper and the large scale human haplotyping studies enabled by LFR.