Chapter 4: Large-scale sequencing of worldwide human populations

4.1 Introduction

Access to data from many different and diverse human populations is clearly key to efforts to understand human genetic diversity and population history. A number of different datasets have been produced and served as useful resources for the human genetics research community. The HapMap project was an early effort that provided array genotypes from populations of mainly European, East Asian and African ancestry (International HapMap Consortium 2005). The Human Genome Diversity Project (HGDP) and the CEPH foundation established a collection of cell lines from a large number of diverse human populations (Cann, et al. 2002), and array genotyping of this collection has resulted in widely used datasets (Jakobsson, et al. 2008; Li, et al. 2008; Patterson, et al. 2012). The POPRES resource released genotype array data on close to 6,000 individuals (Nelson, et al. 2008), with particularly good geographical coverage of Europe; however, its wider usage has been limited as the data are available only under managed access rather than open access. The 1000 Genomes Project released data based on a combination of low-coverage whole-genome and high-coverage exome sequencing data from worldwide human populations, initially from seven populations (1000 Genomes Project Consortium 2010), then 14 populations (1000 Genomes Project Consortium 2012) and finally 26 populations (1000 Genomes Project Consortium 2015). The aggregation of data generated in many different studies, particularly large medical genetics studies, has led to the establishment of resources based on data from tens of thousands of individuals, such as the Exome Aggregation Consortium (ExAC) (Lek, et al. 2016) and the Haplotype Reference Consortium (McCarthy, et al. 2016). While the donor consents for the constituent studies typically preclude open access to the individual genotypes, these resources enable accurate assessments of allele frequencies and/or high-quality genotype phasing and imputation. Recently, projects with sampling strategies focused explicitly on population history have provided high-coverage whole-genome sequencing data from very large numbers of diverse populations, but limited to typically 2-4 individuals from each of them; the Simons Genome Diversity Project (SGDP) with 300 individuals from 142 populations (Mallick, et al. 2016), and the Estonian Biocentre Human Genome Diversity Panel (EGDP) with 483 individuals from 148 populations (Pagani, et al. 2016).

A number of different factors influence the utility of a given dataset or resource to different areas of study within human genetics. The primary utility of population-genetic datasets to medical genetics studies is providing information on the frequency of particular alleles in populations of broadly-defined ancestry and in constituting a resource for phasing and imputation, such that the sample size of the dataset is key. For population history studies, the diversity of the sampled population is very important, as different populations might provide different information into

history.

The HGDP-CEPH collection, mentioned above, has several attractive features as a resource for human population genetics. It contains individuals from about 52 different populations, with typically around 20 individuals from each of these. While projects such as HapMap and the 1000 Genomes Project sampled individuals mainly from major continental populations (sampling criteria included being non-vulnerable and relevant to medical-genetic studies), the HGDP sampling encompasses many smaller populations of particular anthropological, linguistic, historical or genetic interest. As a few examples, from Africa, it contains the Khoe-San, believed to represent the earliest branching modern human population, and central African rain forest hunter-gatherers. From Europe, it contains the Basque population, one of the few European groups to speak a language that is not part of the Indo-European family, and the isolated Orkney islanders. From the Middle East and South Asia, it contains the Druze ethno-religious group and the isolated Kalash group from the western Himalayas. From East Asia, it contains several minority ethnic groups from China, including the Turkic speaking Uyghurs from western China. From Oceania, it importantly contains Papua New Guineans. From the Americas, it contains several Native American groups without European admixture, which otherwise is ubiquitous in the majority populations. As the resource consists of cell lines, an unlimited amount of DNA can be obtained. Additionally, while data obtained for population history studies sometimes has restrictions on how it can be used or shared, the policies of the HGDP-CEPH resource is such that all data can be analysed without restrictions and distributed openly. To date, array genotype and other data generated from this collection have been used in hundreds of population genetics studies.

Whole-genome, high-coverage sequencing of most of this panel was undertaken at the Wellcome Trust Sanger Institute. This project is still ongoing at the time of writing, but in this chapter I describe some initial work on this sequencing data, focusing on the more technical aspects rather than population genetics analyses.

4.2 Sequencing of the HGDP-CEPH panel

The full HGDP-CEPH panel contains over 1000 samples; however, some of these are close relatives, duplicates, have ancestry deviations or other complications, such that a core set of 952 unrelated samples has been established (Rosenberg 2006). Out of these, approximately 135 had been sequenced to high-coverage (here meaning at least approximately 30x) as part of another project, the Simons Genome Diversity Project (Mallick, et al. 2016). These were sequenced mostly as PCR-free libraries on the Illumina HiSeq2000 machines with paired-end 100bp reads. In addition, 10 samples had been sequenced in an earlier study (Meyer, et al. 2012). The remaining samples were sequenced at the Wellcome Trust Sanger Institute (WTSI). An initial batch of 178 samples was sequenced as PCR-based libraries (PCR-free libraries were not available for large-scale production sequencing at WTSI at the time). The remaining approximately 650 samples were sequenced as PCR-free libraries. These were all sequenced on the Illumina HiSeqX

machines with paired-end 150 bp reads. A number of samples were deliberately sequenced more than once across technologies (SGDP versus WTSI versus Meyer, PCR libraries versus PCR-free libraries), to allow for assessments of reproducibility and batch effects. All reads were mapped to the GRCh38 reference genome.

4.3 Characterization of cell line chromosomal abnormalities

The HGDP-CEPH DNA samples derive from lymphoblastoid cell lines (LCLs) collected and established by a number of different laboratories (Cann, et al. 2002). Culturing of cell lines always comes with some risk of new mutations being introduced. In particular, large-scale chromosomal changes, e.g. loss, gain or rearrangements of large segments or entire chromosomes can occur, often without impact on the viability of the affected cells (Shirley, et al. 2012). As the lines are maintained as populations of cells, a de-novo variant might be present in just some subset of the cells in this population, or in all of them. When sequencing DNA from cell lines, these abnormalities could interfere with the accurate determination of the germ-line genome sequence of the donor.

To determine the extent of chromosomal abnormalities in the HGDP-CEPH cell lines, I analysed patterns in the depth of sequencing reads mapped against the reference genome. The excess, or depletion, of reads relative to the genome-wide average should be proportional to the frequency of the chromosomal abnormality variant in the cell population, reaching up to 1.5 in the case of gains or down to 0.5 in the case of losses, if all cells carry the change. Measuring the total number of reads mapping to a chromosome overall will likely be enough to identify high-frequency, whole-chromosome events, but to allow for lower-frequency and partial events I plotted coverage along the length of each chromosome and manually inspected these for deviations (Figure 4.1). This identified approximately 50 events across samples, though the majority constituted only very slight deviations from the genome-wide average. Most events affect whole chromosomes or large segments, e.g. half, of chromosomes, but there are also smaller events on the scales of tens of megabases. Chromosomes 9 and 12 were subject to a larger number of whole-chromosome gains than other chromosomes. There appeared to be no correlation between ancestry of the donor and the rate of chromosomal abnormalities.

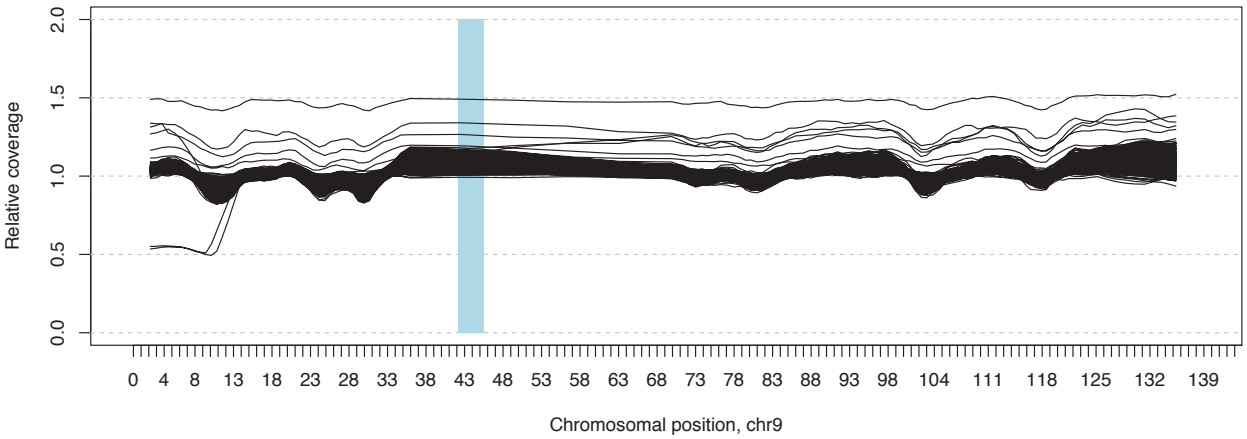

Figure 4.1: Identification of chromosomal abnormalities in the HGDP-CEPH cell lines from high-coverage whole-genome sequencing data: an example from chromosome 9. *Each black line represents a single sequenced individual, displaying the average coverage relative to the genome-wide median in a rolling window across the chromosome. The blue rectangle indicates the location of the centromere. Several samples containing copy-number gains of the whole chromosome, in varying proportions of the cells in the sequenced cell population, are visible. Additionally, two samples containing a ∼10 Mb deletion at the beginning of the chromosome, seemingly carried by close to all cells in the populations, are visible.*

Chromosomes X and Y have sex-dependent expectations for the sequencing coverage relative to that of the autosomes. Furthermore, in contrast to the autosomes where any larger-scale chromosomal abnormalities in-vivo are invariable associated with developmental disorders, copy-number deviations on the X and Y often do not have major phenotypic effects (other than affecting fertility). A large population sample will therefore contain healthy individuals with such deviations with a non-trivial probability; for example, a XXY configuration (diagnosed as Klinefelter syndrome) occurs at a rate of approximately 1 in 1000 males (and one example was detected in the 1000 Genomes Project (1000 Genomes Project Consortium 2015)). Analysing coverage on the X and Y chromosomes in the HGDP samples revealed a number of individuals with deviating copy numbers (Figure 4.2). Two males display substantially reduced coverage on the Y chromosome, likely reflecting loss of Y in the cell line, but potentially also reflecting actual loss of Y in the healthy donor, a phenomenon known to occur in certain tissues, particularly in older men (Forsberg 2017). 10 or so females display varying degrees of loss of X, likely on the cell line level. One male displays coverage consistent with a XXY configuration, thereby likely representing a naturally occurring example of this karyotype.

In addition to these copy number variations, one sample (HGDP01097, from the Chinese Tujia population) was found to be completely homozygous across the entire length of chromosome 1 (confirmed using array genotype data from this sample (Li, et al. 2008)), but the copy number of the chromosome is normal. This is likely a cell line artefact; however, the phenomenon does occur at low frequencies in-vivo owing to a process known as uniparental disomy. While this is typically associated with genetic disorders due to increased risk for a homozygous state at recessive disease variants (King, et al. 2014), it has been described in healthy individuals as well (Field, et al. 1998).

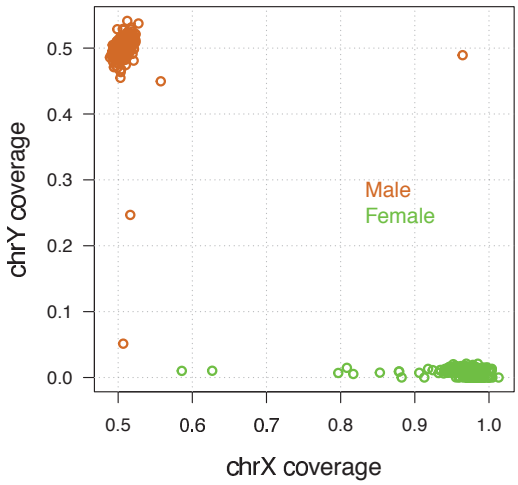

Figure 4.2: Sex chromosome sequencing coverage in the HGDP samples. *Depth of coverage on the X and the Y chromosomes relative to the genome-wide median are plotted against each other for each individual sample. In males, the X and Y chromosomes are haploid, and coverage is therefore expected to be half of the genome-wide median.*

4.4 Effects of chromosomal abnormalities on sequencing and genotyping

The HGDP samples have been genotyped on arrays in the past (Jakobsson, et al. 2008; Li, et al. 2008; Patterson, et al. 2012), data that have been used in hundreds of studies. An unequal copy number of the two chromosomes could conceivably interfere with accurate array genotyping, e.g. leading to some heterozygous genotypes incorrectly being called as homozygous for the allele carried by the chromosome in higher abundance. The genotyping of homozygous sites should not be affected. I investigated the effects in the most heavily used dataset, generated with the Illumina 650K array (Li, et al. 2008). Coverage per chromosome in the whole-genome sequencing data for a given sample is associated with reduced heterozygosity in the array genotypes (Figure 4.3). This thus shows that these cell line abnormalities have affected the array genotypes. It also shows that the abnormalities themselves have been present in the cell lines for a long time, and have not arisen recently or just locally in a subset of cells that was used to extract DNA for this particular whole-genome sequencing project.

I next examined the effects that chromosomal copy number abnormalities might have on the calling of variants from whole-genome sequencing data. While changes affecting large chromosomal regions would have obvious confounding effects on the identification of germ line copy number variation in the donors, it's less clear how it might affect the identification and genotyping of small nucleotide variants. To address this, I performed a down-sampling experiment using the necessarily haploid X chromosomes in male samples. By pooling reads from two different male X chromosomes at different relative abundance and calling variants across these reads, treating them as coming from a single sample, the sensitivity in calling heterozygote variants can be assessed. The calling of homozygous variants should not be affected by changes in chromosomal copy numbers. The results show that as the copy number of the second chromosome decreases, the ability

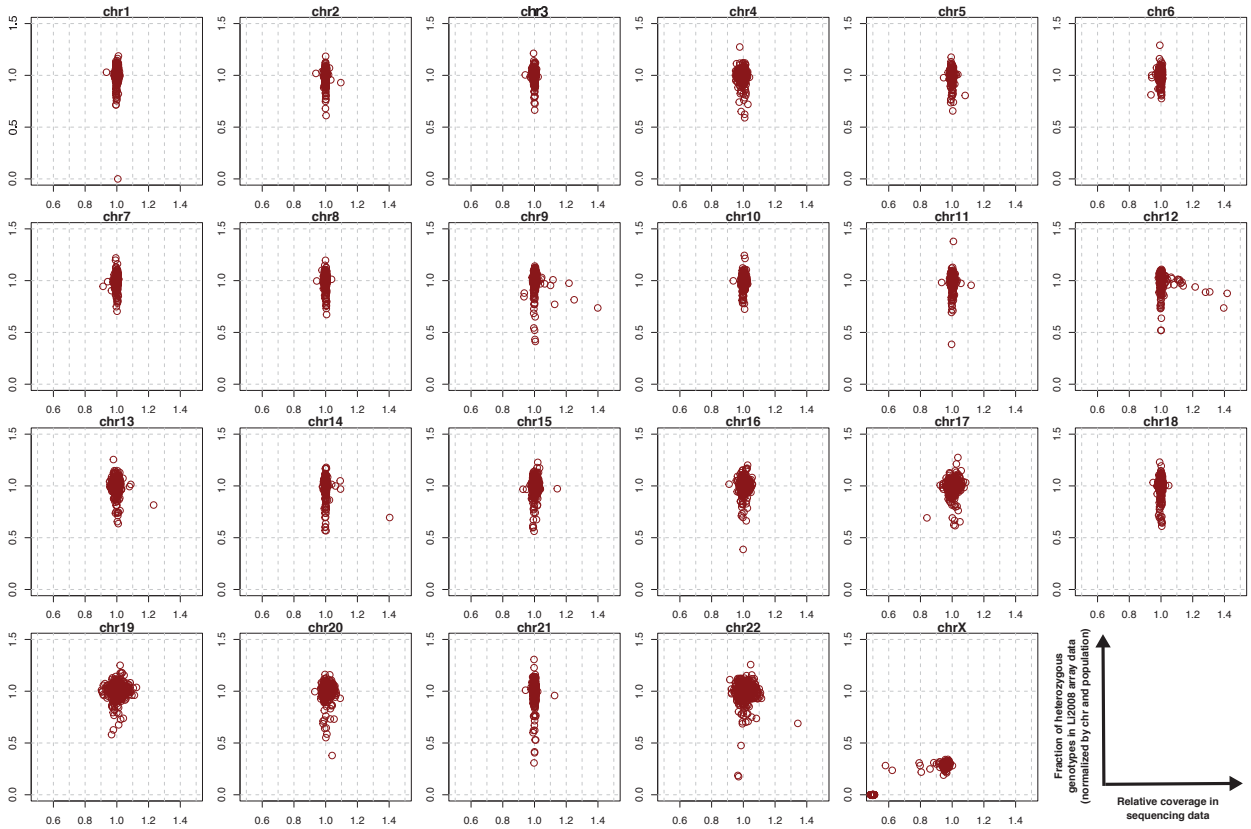

Figure 4.3: The effects of cell line chromosomal copy number abnormalities on array genotyping in the HGDP panel. *For each chromosome separately, sequencing coverage in the whole-genome sequencing data (normalized by the genome-wide median) is plotted against the fraction of heterozygous genotypes in the array data (normalized by chromosome and by population) for each individual sample. Deviations from a balanced chromosomal copy number is associated with reductions in heterozygosity; see for example chromosomes 9 and 12.*

to call heterozygote genotypes decreases quite rapidly (Figure 4.4). However, the effect is not as strong with increases in the copy number of the second chromosome. This is expected, as a variant caller is more likely to incorrectly call a heterozygous site as homozygous if the read count for the less frequent allele is low. Also as expected, the negative effects were more pronounced with lower levels of overall coverage.

In summary, the cell line artefacts affect genotypes called both from genotype arrays and from whole-genome sequencing data. The latter are affected more by chromosomal losses than by gains. On the basis of this, we decided to exclude nine samples that carried large, high frequency losses (as well as one sample that had gains on four separate chromosomes, and the sample with a homozygous chromosome 1) from most downstream analyses.

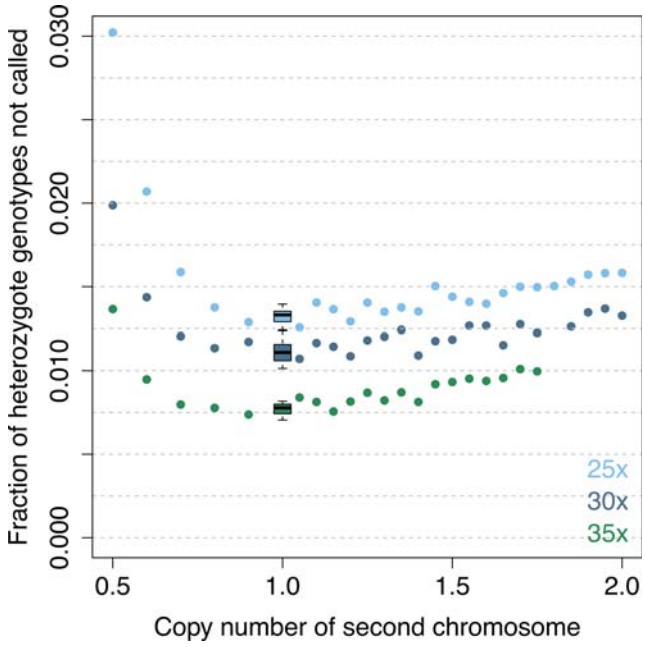

Figure 4.4: The effects of chromosomal copy number abnormalities on the ability to call heterozygote genotypes in high-coverage whole-genome sequencing data. *Data from the X chromosome of two male samples (both from the South Asian Pathan population) were pooled together in different relative abundance, and variants were called across these reads, treated as a single sample, using GATK HaplotypeCaller + GenotypeGVCFs. The horizontal axis shows the copy number of the second chromosome relative to the first chromosome. The vertical axis shows the fraction of heterozygote genotypes that were not called, relative to what was called with a balanced read count from the two chromosomes. The experiment was performed at three different levels of background genome-wide coverage; 25x, 30x and 35x. Twenty replicate down-sampling experiments were performed for the case with a balanced read count, the results of which are displayed with a boxplot.*

4.5 Variant call properties and filtering strategies

There is a myriad of strategies for the calling and filtering of genotypes from whole-genome sequencing data, and strategies also depend on the objective of the given study. I performed various analyses on a preliminary callset of the HGDP samples, produced at the Wellcome Trust Sanger Institute using the GATK HaplotypeCaller software (McKenna, et al. 2010), to try to understand the properties of the variant calls, discover technical issues that need to be addressed and identify an appropriate filtering strategy for this data.

As the sequencing data for this project were produced by two different institutes using a combination of PCR-based and PCR-free libraries, as well as different sequencing platforms, some technical batch effects might be expected, and these might affect downstream ancestry analyses. PCA analyses reveal that there are indeed batch effects, with the strongest one being between WTSI and SGDP datasets, but also some between PCR-based and PCR-free libraries (Figure 4.5). This is also seen in direct *D*-statistic tests of allele frequency correlations between single individuals, for which there is a very strong expectation of a value of zero; for example: D(Chimp,Biaka$_{\text{SGDP.PCRfree}}$;Papuan$_{\text{WTSI.PCRfree}}$,Papuan$_{\text{SGDP.PCRfree}}$) = 0.0194, Z = 3.771. The extent of the batch effect revealed by PCA is reduced by applying increasingly stringent filter thresholds as determined by the GATK VQSR (Variant Quality Score Recalibration) filtering engine, as well as by restricting to regions of the genome with good mapping properties for short reads (Figure 4.5). With the 1000 Genomes "strict mask", which covers ∼78% of the reference genome,

the batch effect is not discernible. The same is true in D-statistic tests, e.g. the above test statistic reduces to $D = 0.0064$, $Z = 1.334$ when applying the strict mask (though it cannot be ruled out that there is still a subtle but non-significant effect remaining). These results imply that the batch effect is not primarily driven by genotype differences at genuine and non-problematic sites, but rather by sites in difficult parts of the genome.

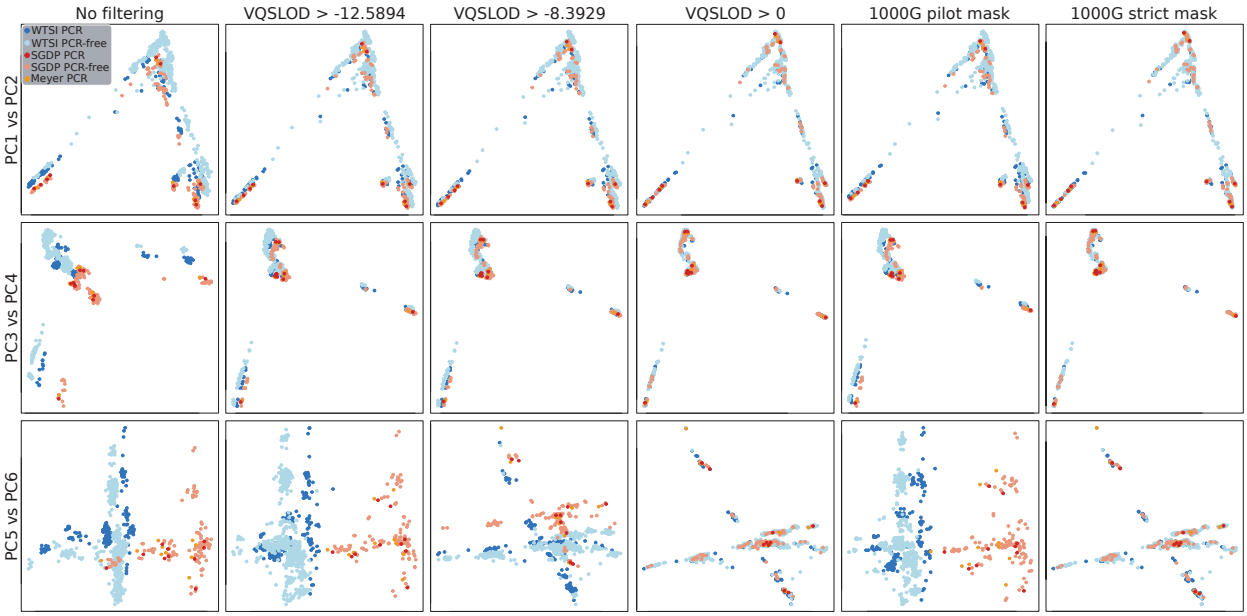

Figure 4.5: Technical batch effects in HGDP genotype calls and the effects of different filters. *Principal component analyses were performed on whole-genome SNP genotype calls, restricted to approximately 1 million variants with a minor allele frequency of at least 5% and pruned for LD. Without filtering (the left-most panel), a clear batch effect is visible between the sets of libraries sequenced with different technologies. This effect is reduced by increasingly stringent VQSR filtering thresholds, as well as by restricting to parts of the genome covered by mappability masks from the 1000 Genomes Project. With more stringent filters (i.e. VSQLOD $>$ 0 and 1000G strict mask), in terms of ancestry, PC1 separates Africans from non-Africans, PC2 separates western non-Africans from eastern non-Africans, PC3 separates out Oceanians, PC4 separates out Native Americans, PC5 separates South Asians from Europeans and Middle Easterners and PC6 separates different African populations. With less stringent filtering, PC5 corresponds to the WTSI vs SGDP batch effect while PC6 separates South Asians from Europeans and Middle Easterners.*

Properties of the subset of SNPs that strongly differentiate between the sample sets might provide some insights into the causes of the batch effect. SNPs that are associated with higher alternative allele frequencies in the WTSI libraries also tend to have higher coverage and lower rates of missing genotypes across individuals in these compared to SGDP libraries (Figure 4.6). Likewise, SNPs that are associated with higher alternative allele frequencies in SGDP libraries tend to have higher coverage in these compared to WTSI libraries (though no clear difference in rates of missing genotypes). The latter SNPs have slightly lower GC content in the 100bp windows surrounding them, but this is not a large difference. Overall, these trends suggest that there are sets of sites in the genome which are differentially accessible to reads from these two sources, in each case leading to reduced coverage and ability to call alternative alleles. One underlying reason for this might be that the WTSI reads are 150 bp long, while the SGDP reads are only 100 bp, likely leading to some difference in mappability in repetitive parts of the genome.

The patterns of missing genotypes across individuals are non-random. There are some differences in the overall rate of missing genotypes across the called variants, with slightly lower rates in the

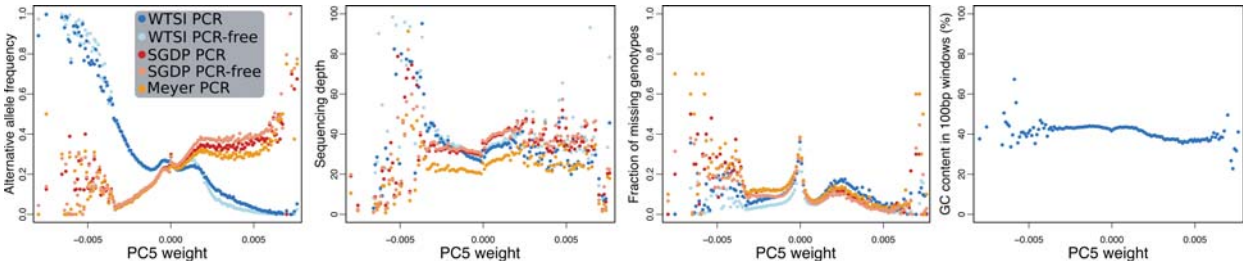

Figure 4.6: Properties of SNPs that separate between WTSI and SGDP sequenced libraries. *In PCA analyses of unfiltered SNP calls, PC5 separates WTSI and SGDP sequenced libraries (see Figure 4.5). The weights of each SNP for this component were extracted and plotted against various features of these SNPs.*

WTSI PCR-free than in other libraries (Figure 4.7A). Principal component analyses of the patterns of missing genotypes, rather than the genotypes themselves, also reveal a very strong clustering by technology rather than by ancestry, for both SNPs (Figure 4.7B) and indels (Figure 4.7C). These systematic differences in the ability to call genotypes might underlie some of the ancestry batch effects. Further to this, patterns of depth of coverage are also non-random (Figure 4.6), and the variant caller will output genotypes even at sites where there are just a few reads or an excessively high number of reads, resulting in low-confidence genotype calls that might be more susceptible to technical artefacts. Applying a site-level filter on the depth of sequencing coverage, setting any genotype to missing if the number of reads covering the site is lower than one third or higher than double the genome-wide average for the particular library (hereafter referred to as "DP3rd2x"), leads to some reduction in the strength of the batch effect in PCA (not shown). Similarly, the effect is partly reduced in *D*-statistic tests, i.e. the above test statistic reduces to $D = 0.0141$, $Z = 2.271$.

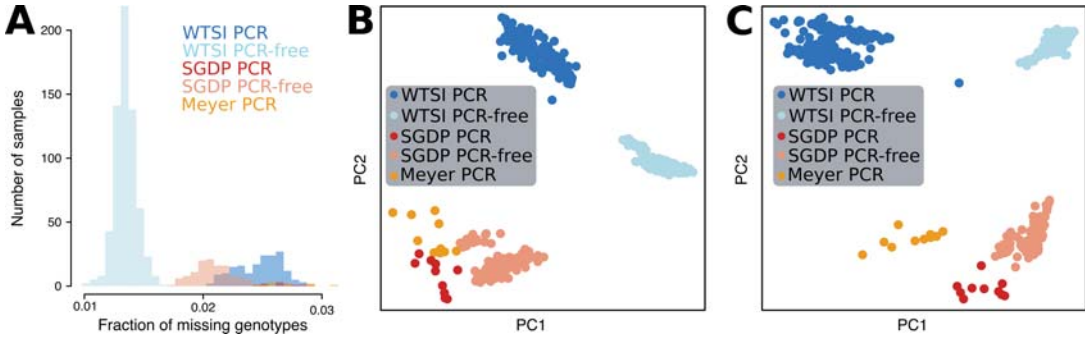

Figure 4.7: Missing genotypes in the HGDP callset. *A) Overall rates of missing genotypes at SNPS in the unfiltered callset for the different sets of libraries. B) A PCA on the patterns of missing genotypes at SNPs reveals very strong clustering by technology. C) A PCA on the patterns of missing genotypes at indels reveals very strong clustering by technology.*

The duplicate sequencing of a number of individuals with more than one technology allows for concordance assessments. Comparing WTSI PCR-free and SGDP PCR-free, representing the two dominant technology sets, there are an average of 263,779 discordant SNP genotypes between duplicate samples if no filtering is applied. This reduces to 133,225 at a VQSR LQSLOD threshold of >-12.5894, 103,325 at VQSLOD >-8.3929, 13,239 at VQSLOD >0, 159,837 with the DP3rd2x filter, 104,903 with the 1000 Genomes pilot mask, 6,669 with the 1000 Genomes strict mask

and 6,280 with the combination of the DP3rd2x filter and the latter mask. If restricting to the $\sim$650,000 sites that are present on the Illumina 650K array and therefore very likely to be real, and easily accessible, variants, there are an average of 117 discordant genotypes, which compares favourably to concordance observed between sequencing and array (which generally have very high quality) genotypes. Concordance at indel genotypes is substantially lower, considering that the total number of indel calls is approximately an order of magnitude lower than the number of SNPs calls, with e.g. an average of 197,929 discordant genotypes without filtering and 7,448 with the 1000 Genomes strict mask. Comparisons between duplicate samples in which at least one experiment was done with PCR-based libraries reveal substantially higher indel discordance (e.g. 28,855 discordant genotypes in the strict mask between a WTSI and a SGDP PCR duplicate sample), consistent with the notion that the PCR reaction introduces some level of indel error into the libraries. Overall, the observation that restricting to the strict mask leads to substantial reductions in duplicate discordance is consistent with the notion that most genotype errors are driven by mapping, alignment, coverage or other issues at problematic sites, rather than random sequencing errors or binomial sampling errors at well-behaved sites. Intuitively, the latter should be very infrequent for sequencing experiments with on the order of 30x coverage.

In summary, even with high-coverage sequencing, technology differences lead to considerable batch effects which impact ancestry analyses. However, the strength of these effects can be reduced by filtering. Application of strict VQSR filtering removes most of the batch effects, however this comes at a price of excluding a fairly large number of variants (e.g. at VQSLOD > 0, less than 80% of variants are retained, though many of the excluded ones will not be real variants). There is also a conceptual issue related to VQSR and similar filtering strategies, which is that they only apply to polymorphic sites. Certain population-genetic methods, however, need genotype information also at monomorphic sites, especially methods that in one way or another perform inference on the process of mutation, e.g. the PSMC (Li and Durbin 2011), MSMC (Schiffels and Durbin 2014) and SMC++ (Terhorst, et al. 2017) methods, as well as site-frequency spectrum modelling methods such as fastsimcoal (Excoffier and Foll 2011) (or more generally, any divergence calculation, e.g. the basic divergence between two genomes or the heterozygosity of a single genome). Making use of filters that only apply to polymorphic sites would introduce a bias against such sites, potentially confounding these methods. The above analyses, however, also show that by restricting to the regions of the genome that are easily accessible to short read mapping and non-repetitive, most if not all of the batch effects disappear. This kind of site-level, rather than variant-level filtering, would not introduce a bias against polymorphic sites and the resulting data would therefore be suitable for any kind of population-genetic analysis. It does, however, come at the cost of excluding large parts of the genome – while 78% of the reference genome (which is what the 1000 Genomes strict mask covers) is probably enough for most population-history and demographic analyses, an ideal genomics resource should provide data on as large a fraction of the genome as possible, to also maximally enable e.g. functional, medical and selection studies. There is thus, as always, a

trade-off involved in these filtering decisions, and perhaps a sensible approach is to produce data for the whole genome and then restrict to suitable subsets depending on the particular type of analysis to be carried out.

4.6 Rare variant sharing patterns

One of the areas where sequencing of a large number of individuals per population has clear benefits, compared to array genotyping or sequencing of small numbers of individuals per population, is in the analysis of rare variants. There is an increasing interest in using rare variants to learn about population history (1000 Genomes Project Consortium 2012; Mathieson and McVean 2014; 1000 Genomes Project Consortium 2015; Field, et al. 2016; Schiffels, et al. 2016). As they typically result from recent mutations, their distribution across individuals might give insight into more recent population history than do common variants. An analysis of the sharing of doubletons, meaning variants that are observed exactly twice in this particular dataset, reveals an abundance of structure among the HGDP samples (Figure 4.8). This holds promise for the application of e.g. rare variant sharing asymmetry tests conceptually similar to the *D*-statistic, or more sophisticated, model-based approaches (Schiffels, et al. 2016).

On a general level, the dependence of the non-normalized doubleton counts on the background level of genetic diversity is evident in these results. Most strikingly, the number of variants shared between any African populations is greater than between many populations within non-African continental regions, even if the latter will typically be more closely related, reflecting the greater genetic diversity and therefore larger number of rare variant sharing opportunities in Africa. Most non-African populations share more doubletons with the San population than with other Africans, despite likely being less closely related to them, probably similarly reflecting the great genetic diversity of the San (Kim, et al. 2014). This demonstrates the need for appropriate normalization when using rare variant sharing counts to infer shared ancestry.

On a more detailed level, several known features of the history of particular populations are visible in the patterns of doubleton sharing. The South Asian Hazara population displays elevated sharing with East Asians, reflecting East Asian admixture described in this group. Specific South Asian and Middle Eastern individuals display elevated sharing with Africans, consistent with recent sub-Saharan admixture in these. The Uygurs from western China display elevated sharing with South Asians and Europeans, relative to other Chinese populations, reflecting known west Eurasian admixture in this group. The Melanesian population from Bougainville Island in Papua New Guinea display elevated sharing with East Asians (particularly the south-eastern groups of Cambodian and Dai) relative to the mainland Papuans, reflecting the $\sim$20% of their ancestry deriving from Southeast Asian admixture. The substructure within the mainland Papua New Guinean samples, described in Chapter 3, is clearly visible.

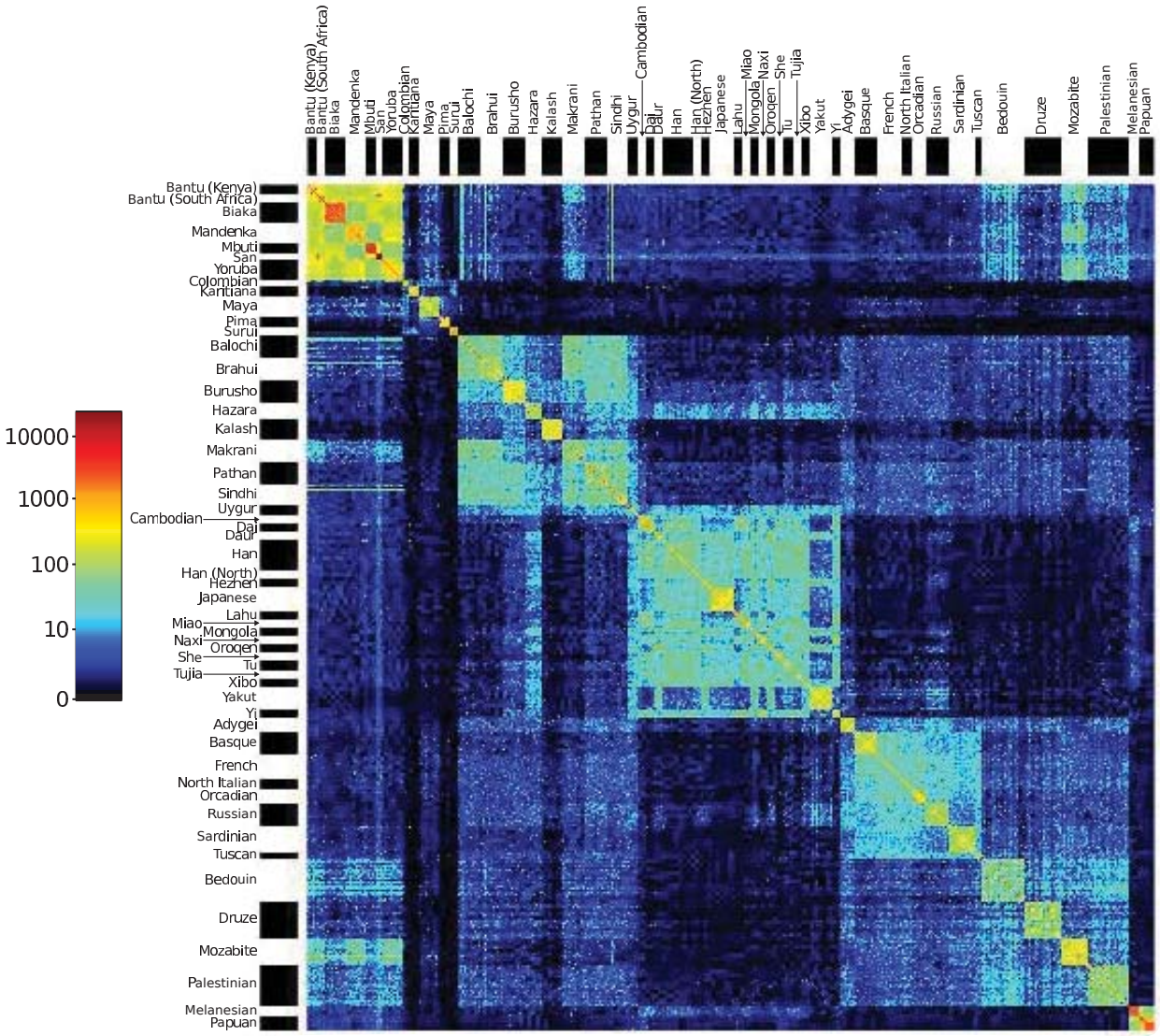

Figure 4.8: Doubleton variant sharing in the HGDP whole-genome sequencing data. *The non-normalized counts of shared variants observed exactly twice across the dataset, for each pair of individuals, displayed on a logarithmic colour scale. The boundaries between populations are indicated with alternating black boxes and empty space on each side of the plot.*

4.7 Conclusions

The high-coverage, whole-genome sequencing data generated from the HGDP-CEPH panel of worldwide populations is likely to be of great use in the study of human population history, and human genetics more generally. Compared to the 1000 Genomes Project (1000 Genomes Project Consortium 2015), which is the most commonly used panel representing worldwide human diversity, this resource will have several advantages. The main one is the more diverse populations represented in the HGDP panel, including many of particular historical interest and utility in analyses. Another one is the high sequencing coverage, enabling unproblematic application of methods that rely on high-quality genotype calls. Lower-coverage sequencing projects have typically relied on genotype imputation to achieve high-quality genotypes; however, there is concern about imputation in population history contexts as the ancestry composition of the panel used for imputation might introduce biases. A disadvantage of the HGDP collection is the relatively small number of

individuals in some populations, e.g. 6-10 for the smallest ones, relative to ∼100 individuals for all populations in 1000 Genomes.

High-coverage sequencing currently provides the highest accuracy genotypes, but this project demonstrates that there are still technical issues that need consideration. The DNA for this sequencing project was extracted from cell lines and in some cases had chromosomal abnormalities, leading to a decision to exclude a small number of individuals. The chromosomal copy number abnormalities have affected array genotypes from the same panel, used for a decade across hundreds of studies; however, it appears that the negative effects are less pronounced for high-coverage sequencing data. Sequencing of individuals across different technologies, i.e. in different centres on different types of Illumina machines, and with PCR-based or PCR-free libraries, results in batch effects in the genotype calls with effects on ancestry analyses. These can be reduced through filtering; however, overly aggressive filtering by necessity means less of the genome is available for use in downstream analyses.

4.8 Materials and methods

Genotype array data from the HGDP-CEPH samples generated on the Illumina 650K array were obtained from (Li, et al. 2008) and chromosomal coordinates were lifted over from GRCh36 to GRCh38 using the NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap), as well as using dbSNP rs IDs. Genotype data on these samples generated on the Human Origins array were obtained from (Patterson, et al. 2012) and lifted over from GRCh37 to GRCh38 using the NCBI Genome Remapping Service. Genotypes for chimpanzee were extracted from the UCSC hg38-panTro4 axtNet alignments. Sequencing coverage along the reference genome was calculated for each sample using the "depth" command from the samtools software (Li, et al. 2009), applying the "-aa" argument. Linkage disequilibrium pruning and principal component analyses were performed using AKT (Arthur, et al. 2017). In order to perform PCA on the patterns of missing genotypes rather than the genotypes themselves, missing genotypes were recoded as 0 and non-missing as 1. The "pilot" and "strict" accessibility masks for GRCh38 were obtained from the 1000 Genomes Project. The rare variant sharing heatmap was made using the heatmap.2 function from the gplots R package.