Chapter 18

Detection of Copy Number Variation Using SNP Genotyping

Gregory M. Cooper and Heather C. Mefford

Abstract

Genetic diversity among human genomes comes in many forms, including single nucleotide polymorphisms (SNPs) and small insertions and deletions on the order of one to several basepairs. More recently, large, >1 kb copy number changes have been identified as an important source of normal genomic variation as well as disease-causing variation. The ability to perform genome-wide discovery of large copy number variants (CNVs) has been facilitated by advances in two technologies – array comparative genomic hybridization and SNP genotyping platforms. Here, we discuss the general principles and strategies underlying CNV detection with SNP genotyping platforms, which are widely used and capable of providing both SNP and CNV genotyping information.

Key words: copy number variation, single nucleotide polymorphism genotyping, genomic variation, array comparative genomic hybridization

1. Introduction

Copy number variants (CNVs), defined as insertions, deletions, or duplications of sequence larger than 1 kb, are substantial contributors to human genomic diversity and are important factors in both normal (1) and disease (2, 3) variation. These include environmentally responsive traits like sensory perception (e.g., *opsins* and *olfactory receptors*), immune system function (e.g., Crohn's disease, psoriasis), severe early childhood diseases like developmental delay and autism, and neurological diseases like schizophrenia and epilepsy. Importantly, studies of CNV-trait associations have found evidence for the involvement of both common and rare CNVs in human disease. In the context of pluripotent stem cell development and analysis, the knowledge of CNVs in a given genome can be useful for several reasons. For example, CNVs can

Philip H. Schwartz and Robin L. Wesselschmidt (eds.), *Human Pluripotent Stem Cells: Methods and Protocols*,
Methods in Molecular Biology, vol. 767, DOI 10.1007/978-1-61779-201-4_18,

affect expression of genes within (4) and near (5) the CNV, so expression data for genes affected by CNVs may be interpreted differently.

Owing to their size, heterogeneity, and sequence complexity, the accurate detection of CNVs in human populations is a technically challenging task. There are several methods that may be employed to detect CNVs. For targeted evaluation of one or a few genomic regions of interest, quantitative PCR (qPCR) (6–8) or multiplex ligation probe amplification (MLPA) (9) are commonly used. However, for more extensive, genome-wide analysis, there are two commonly used platforms: array comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) genotyping arrays. Here we focus on the use of SNP arrays to detect copy number variation. The advantages of SNP-based CNV detection include.

1. Simultaneous ascertainment of SNP and CNV data (unavailable from CGH).
2. High-throughput sample processing, treatment, and quality control.
3. High-density of probes, with arrays ranging from hundreds of thousands to multiple millions.
4. Reasonable cost, typically ranging from tens to hundreds of dollars per sample, depending on probe density.

2. Materials

SNP microarray analyses typically require an input of 100 ng to 1 μg of genomic DNA, varying by the specific array/protocol employed and the manufacturer (see Note 1).

3. Methods

3.1. SNP Genotyping

Before discussing how to detect CNVs using SNP genotyping data, an understanding of the basic principles of SNP genotyping is required. SNPs are DNA sequence variants where a single nucleotide can differ among individuals (or chromosomes); most SNPs are di-allelic, meaning that there are two possible alleles (e.g., a "C" or a "T" at a given site). Microarray-based SNP genotyping platforms exploit fluorescence-based visualization of genomic DNA bound in an allele-specific manner to oligonucleotides fixed to a surface. While details vary substantially between different platforms, there are two critical pieces of information gathered for each

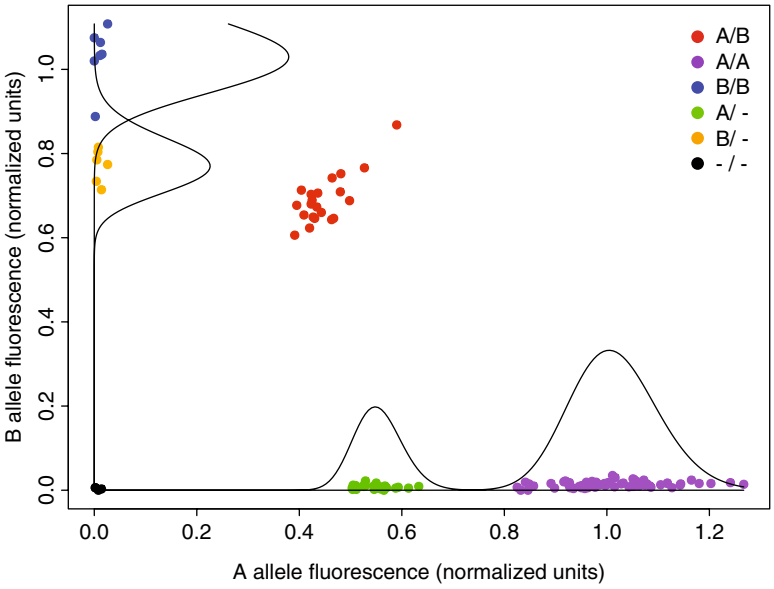

Fig. 1. Scatter plot of intensity information collected for a single SNP (rs10076425) assayed with an Illumina genome-wide SNP array on a collection of 126 (primarily HapMap[21]) samples. Each point corresponds to the intensity for a single sample; the *X*-axis indicates fluorescence intensity for the "A" allele while the *Y*-axis indicates intensity for the "B" allele. The three most populous clusters of samples correspond to the "AA" (*purple*), "AB" (*red*), and "BB" (*blue*) samples who are homozygous for the A-allele, heterozygous, or homozygous for the B-allele. Note the presence of three additional clusters corresponding to the hemizygous A- (*green*) individuals, hemizygous B- (*yellow*) individuals, or homozygous deletion carriers (*black*). The first two are heterozygous for a deletion allele while the latter have a copy number of zero at this location. Superimposed distributions (*black curves*) are estimated from the data and allow statistical separation of diploid from haploid samples. This figure is reproduced from Cooper et al. (12).

targeted genomic site: (1) the total fluorescence intensity for a site which reflects signal combined for both alleles in a given sample and (2) the allelic ratio providing the relative intensity measurements for the two alleles at each site (Fig. 1). For the vast majority of sites, individuals are diploid and will therefore be homozygous for one allele ("AA") or the other ("BB") or heterozygous with one copy of each ("AB"). Note that modern arrays also include many non-SNP (monomorphic sites) probes (e.g., (10)), which only provide total intensity data and are included so as to improve probe density in known or suspected CNV locations.

3.2. CNV Detection from SNP Data: Discovery vs. Genotyping

The task of CNV detection from SNP data can be broken down into two related but distinct challenges: CNV discovery, wherein variants are detected ab initio in a given genome without assumptions about their breakpoints, and CNV genotyping, wherein copy number status is assigned to a set of studied samples for given loci that are known (or suspected) to be copy number variable.

Several important consequences emerge from this distinction. First, CNV discovery is performed sample-by-sample and has the advantage of being able to detect CNVs anywhere in the genome, including for rare and de novo events unique to the given sample. However, owing to the large space of data being examined (if breakpoints are allowed to be anywhere, any pair of analyzed probes within a given chromosome is a candidate set of breakpoints), specificity must be extremely high to avoid an unacceptably large false discovery rate. Such stringent specificity is typically obtained at the expense of sensitivity to small (few probes) or noisy (intensities near detection threshold) sites. On the other hand, CNV genotyping can be applied to many samples simultaneously and can leverage the knowledge that a CNV exists at a given location to achieve both high specificity and sensitivity. However, genotyping is restricted to a priori defined sites, implying that de novo and other rare events outside of the targeted loci will be missed. Below, we outline the basic principles underlying SNP-based CNV detection, contrasting the discovery and genotyping challenges where appropriate.

3.3. CNV Genotyping

The conversion of intensity information into an estimate of copy number comprises several steps. First, the raw intensity information is normalized to account for systematic effects related to genotyping chemistry (e.g., differences in intensity between fluorophores), microscopy (e.g., location of a probe on a slide), and other factors (e.g., total intensity for a given slide). In addition, measurements are often obtained at multiple physical locations on an array corresponding to the same SNP (or genomic location for nonpolymorphic probes), and this information must be integrated to determine a single measurement for a given site. These steps are heavily sensitive to the specific platform used and typically handled by the manufacturer's software. After these normalization steps, intensity information for a given SNP is comparable across a set of samples, and for polymorphic sites, can be visualized as a two-dimensional scatter plot with intensity information for each of the two alleles (denoted from here forward as "A" and "B" for simplicity) plotted on a separate axis (Fig. 1).

Note that the position of a given sample in this two-dimensional space provides both total intensity (essentially distance to the origin) and allelic ratio (angle of the line joining the datum to the origin). The canonical SNP genotypes (assuming the probe is informative and analyzed in diploid samples) will generally stand out as three distinct clusters, with homozygous (AA or BB) individuals appearing on either of the two axes and heterozygous (AB) individuals appearing as a cluster toward the middle of this space. This clustering information forms the basis for SNP genotyping (10, 11) and can also be used to genotype CNVs via statistical evaluations of the relative locations (separation) and

qualities (variance within a group) of the observed clusters (10, 12). Hemizygotes (A- or B-), for example, will appear as clusters of samples closer to the origin than homozygotes, because relative copy number, and therefore intensity, is reduced, and the allelic ratio indicates that the sample is homozygous (Fig. 1). "Null" individuals (i.e., samples that are homozygous for a deletion event) will typically appear near the origin reflecting the lack of any DNA binding for that sample at that location (Fig. 1). Samples bearing higher copy numbers may yield a diversity of cluster positions depending on total copy number and heterozygosity (Fig. 1). For example, individuals carrying a duplication of a given sequence actually have three copies of that sequence, and at heterozygous locations may be triply homozygous ("AAA" or "BBB") or exhibit a distortion in the allelic ratio ("AAB" or "ABB"). These latter cases can provide a powerful discriminatory signature to define duplication carriers, especially for CNV discovery (see below).

There are several critical caveats to consider. First, CNV genotyping using the above framework is frequency-dependent since it depends on the identification of clusters of individuals with the same copy number. Rare or individual CNV carriers in a sample series appear as outliers rather than in clusters, and alternative approaches are required to identify these events, although it is possible to genotype rare CNVs by exploiting this behavior and specifically looking for outlying samples (caution must be taken to contrast noisy samples from outliers that result from a change in copy number) (13). Second, the CNV genotyping process is sensitive to background intensities and probe-specific noise (e.g., cross-hybridization to other sites in the genome); clustering information from multiple SNPs is typically required to obtain robust copy number genotypes to avoid both probe-specific artifacts and reduce noise in genotype inference. It is also important that genotyping methods either have an automated method to identify "informative" probes or are only applied to predefined probe sets that are known to yield reliable copy number estimates; in most cases it is necessary to combine automated elimination of obviously bad probes with manual curation of potentially good probe sets. Finally, we note that copy number estimation is relative; absolute copy number is typically reliant on the assumption that a diploid copy number is predominant (often but not always true) (see Note 2). Even assignment of zero copies can be confounded by cross-hybridization, for example, and assignment of absolute copy number at high copies (>3) is particularly difficult as the ratios between copy number intensities become smaller (i.e., a change in copy number from 1 to 2 corresponds to a twofold change in intensity, while 5–6 is only a 20% change). In general, SNP-based CNV genotyping has not been shown to be accurate for higher copy ranges.

3.4. CNV Discovery

As noted above, clustering-based approaches to CNV detection can generally only apply to curated sets of probes at previously defined CNVs that are known or suspected to be common. Rare variant discovery is a distinct challenge that is usually accomplished by considering data from each individual separately, scanning across the genome to identify regions (sets of contiguous probes) that exhibit evidence for gain or loss of segmental DNA copy number. However, before such an analysis can be done, it is important that each probe be normalized so that intensity data are comparable between probes within a given sample (the normalization steps described above are taken to make intensities comparable across samples but within a given probe). Clustering of individuals is again applied here, except in this circumstance all individuals of a given SNP genotype are assumed to be the same copy number (ideally, common CNVs would be known and eliminated prior to such an analysis and for X-chromosome SNPs males and females would be treated separately). Subsequently, for each probe within each individual, a normalized intensity measure can be derived that describes the location of a given individual relative to the other samples. In Illumina genotyping, for example, the total intensity for a given probe in a given individual is reported as the "LogR Ratio," where a value of 0 indicates that a sample has a total intensity at the center of the cluster of individuals with the same SNP genotype (i.e., "AA" individuals), while positive and negative values indicate that a sample is above or below the mean intensity (14). Related to this is a normalized measure of the allelic ratio; in Illumina genotyping, this value is the "B-allele frequency" (BAF), so-called because it is inferred to be the fraction of the total intensity at a given probe that is derived from the "B" allele. For example, the BAF for "AA" individuals should be 0, for "AB" individuals should be 0.5, and for "BB" individuals should be 1, because in these samples 0, 50, or 100% of the total intensity comes from the B-allele, respectively. In practice, the center of the "AB" heterozygote cluster is used to define a BAF of 0.5, again because the absolute position of this cluster varies from probe to probe while it is assumed that probes are all capturing diploid locations.

After this probe-by-probe normalization is applied, the challenge lies in the identification of segments within a given genome that exhibit intensity and BAF information that is consistent with the presence of a CNV. Deletions, for example, should result in both negative LogR values and complete loss of heterozygosity (BAF of only 0 or 1; Fig. 2a). Duplications should manifest as positive LogR values and a skew in BAF at heterozygous sites to either 1/3 ("AAB") or 2/3 ("ABB") since the alleles are no longer in one-to-one proportion (Fig. 2b). Note that homozygous sites in a duplicated segment ("AAA" or "BBB") would still have BAF values of 0 or 1, and because of this polymorphic sites

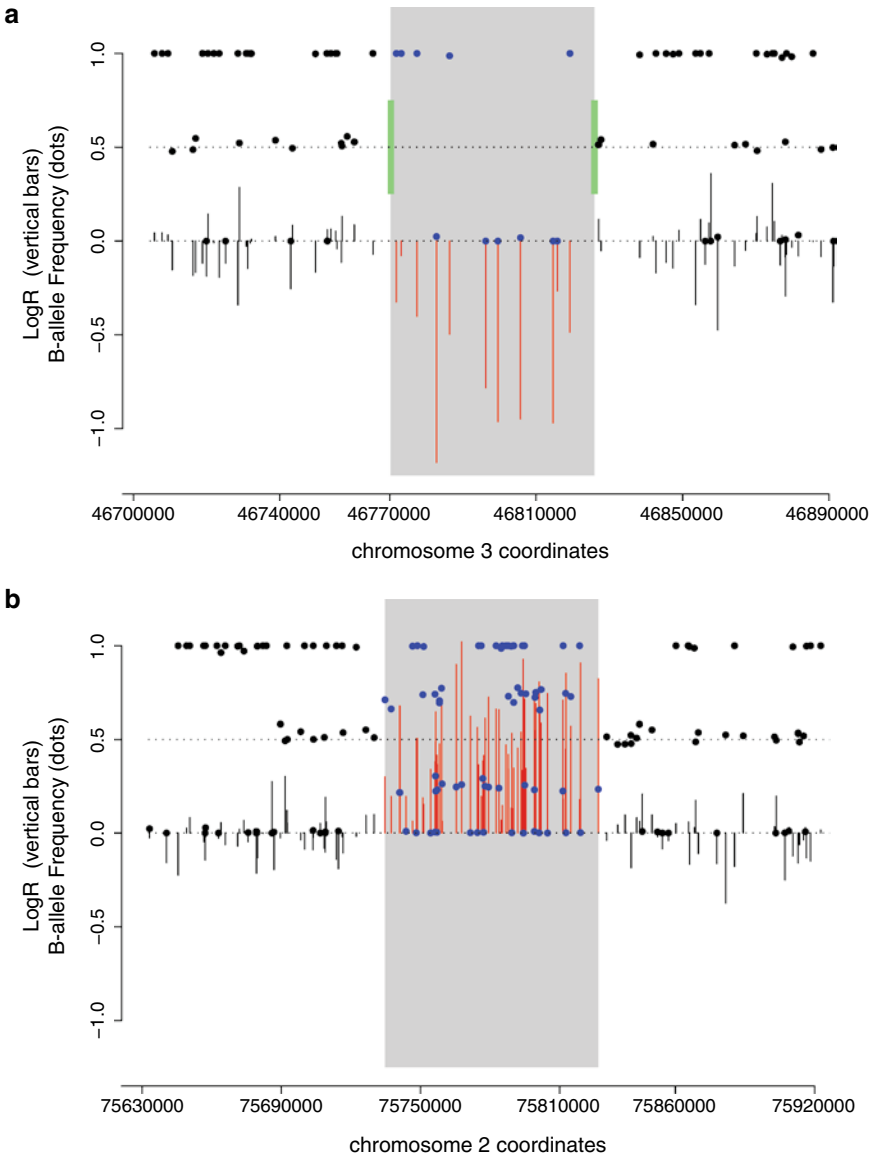

Fig. 2. CNVs "discovered" within Illumina genome-wide SNP array intensity data, adapted from Cooper et al. (12). (**a**) Example of a deletion event. Intensity data for all probes within the indicated genomic interval (*X*-axis) for a single sample are plotted. "LogR Ratio" and "B-allele Frequency" (14) are plotted as *vertical bars* and *filled dots*, respectively. The *gray box* indicates the deletion span inferred by computational segmentation of the SNP array data; probes internal to this box are colored *red* ("LogR Ratio") or *blue* ("B-allele Frequency"). *Green vertical bars* indicate the deletion borders defined by resequencing. Note that the LogR drops within the deletion and the B-allele Frequency values indicate a loss of heterozygosity. (**b**) Similar to (**a**), except an example of a duplication is highlighted. Note that "heterozygous" SNPs within the duplicated segment have a B-allele frequency of either ~1/3 or ~2/3, indicating that this individual carries three copies of this segment of the genome ("AAB" or "ABB").

are in general more informative than non-polymorphic sites; equivalently for deletions, loss of heterozygosity information is only useful at polymorphic sites. There are a variety of methods available to perform segmentation, with the most commonly

used methods leveraging well-established statistical methods like Hidden Markov Models (for several examples of methods to detect CNV breakpoints see (12, 15–17)). There are numerous details that are important to consider in all these methods (for example, typically even normalized intensity data, e.g., "LogR" values, are subjected to further rounds of normalization or manipulation to account for effects related to allele frequency, probe specificity, etc., prior to segmentation). However, a general rule is that, because the breakpoints are unknown, the search space is very large (many potential pairs of breakpoints) and extremely high specificity is required. This typically results in reduced sensitivity, especially to small CNVs (i.e., variants spanning ten probes can be more robustly inferred than those spanning only two probes) and those (as for genotyping) embedded in more complex sequence (e.g., CNVs that change copy number state from 5 to 6).

3.5. Additional Considerations

There are a variety of critical contextual factors that influence the accuracy and reliability of CNV information inferred using SNP array data, including DNA quality, quantity and concentration; normalization methods; quality-control measures; and CNV calling algorithms used (see Notes 1–3). Perhaps most importantly, all the critical quality-control measures that are intrinsic to well-designed array-based experiments, such as uniform treatment of samples, randomization of cases and controls, controlling for batch artifacts, etc are also important to studies of CNVs using SNP array data; in fact, assignment of SNP genotypes is generally more robust than assignment of CNV genotypes and interpretation of data should be treated accordingly.

In summary, SNP arrays are widely available, relatively affordable and can provide up to millions of SNPs in a single experiment. The data collected can be used to discover and genotype CNVs ranging from several kilobases in size to whole chromosome abnormalities, in addition to the value of the SNP data. Furthermore, it is expected that this information will improve as maps of known CNVs become more comprehensive (10, 20), as density of SNP arrays increases, and as genotyping algorithms and data normalization approaches become more accurate (21).

4. Notes

1. DNA quality, quantity and concentration can affect fluorescence intensity levels and therefore inference of CNVs. Furthermore, whole-genome amplification steps can introduce systematic noise that may overwhelm legitimate CNV signal; this may be true even when SNP genotypes can be reliably inferred.

2. Data normalization must always be considered when interpreting CNV information, especially in the context of common CNVs. For example, if a deletion event is at high frequency in the population, then the assumption that most samples have a copy number of 2 does not hold and inappropriate assignment of absolute copy number may result.
3. Many algorithms that have been and continue to be developed for the discovery and/or genotyping of CNVs from SNP microarray data. It is important to note that some studies explicitly differentiate these tasks while others attempt to perform both simultaneously. Examples of algorithms include (but are not limited to) QuantiSNP (16), PennCNV (15), SCIMM (12), SCOUT (13), BirdSuite (10, 18), and others (19). Choice of any given algorithm and sets of parameters to apply is complex, but should take into consideration the platform used (e.g., Illumina vs Affymetrix), the goals of the study (e.g., common CNV genotyping vs. rare variant discovery), the density and spacing of probes, and the respective costs of false positive and false negative CNV assignments.

References

1. Perry GH, Dominy NJ, Claw KG, et al. (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**:1256–60.
2. Mefford HC, Eichler EE. (2009) Duplication hotspots, rare genomic disorders, and common disease. *Curr Opin Genet Dev* **19**:196–204.
3. Fanciulli M, Petretto E, Aitman TJ. (2009) Gene copy number variation and common human disease. *Clin Genet* 77:201–13.
4. Henrichsen CN, Chaignat E, Reymond A. (2009) Copy number variants, diseases and gene expression. *Hum Mol Genet* **18**:R1-8.
5. Merla G, Howald C, Henrichsen CN, et al. (2006) Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am J Hum Genet* **79**:332–41.
6. Porcher C, Malinge MC, Picat C, Grandchamp B. (1992) A simplified method for determination of specific DNA or RNA copy number using quantitative PCR and an automatic DNA sequencer. *Biotechniques* **13**:106–14.
7. Livak KJ, Schmittgen TD. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) Method. *Methods* **25**:402–8.
8. Wong A, Cortopassi G. Reproducible quantitative PCR of mitochondrial and nuclear DNA copy number using the LightCycler. (2002) *Methods Mol Biol* **197**:129–37.
9. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* **30**:e57.
10. McCarroll SA, Kuruvilla FG, Korn JM, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**:1166–74.
11. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. (2006) Whole-genome genotyping with the single-base extension assay. *Nat Methods* **3**:31–3.
12. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* **40**:1199–203.
13. Mefford HC, Cooper GM, Zerr T, et al. (2009) A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. *Genome Res* **19**: 1579–85.
14. Peiffer DA, Le JM, Steemers FJ, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* **16**:1136–48.
15. Wang K, Li M, Hadley D, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**:1665–74.

16. Colella S, Yau C, Taylor JM, et al. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35**:2013–25.
17. Shaikh TH, Gai X, Perin JC, et al. (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* **19**:1682–90.
18. Korn JM, Kuruvilla FG, McCarroll SA, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**:1253–60.
19. Gamazon ER, Zhang W, Konkashbaev A, et al. (2009) SCAN: SNP and Copy number Annotation. *Bioinformatics* **26**:259–62.
20. Kidd JM, Cooper GM, Donahue WF, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**:56–64.
21. International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437**:1299–320.