Marker selection for genetic case–control association studies

Fredrik H Pettersson, Carl A Anderson, Geraldine M Clarke, Jeffrey C Barrett, Lon R Cardon, Andrew P Morris & Krina T Zondervan

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. Correspondence should be addressed to F.H.P. (fredrikp@well.ox.ac.uk) or K.T.Z. (krina.zondervan@well.ox.ac.uk).

Published online 23 April 2009; doi:10.1038/nprot.2009.38

Association studies can focus on candidate gene(s), a particular genomic region, or adopt a genome-wide association approach, each of which has implications for marker selection. The strategy for marker selection will affect the statistical power of the study to detect a disease association and is a crucial element of study design. The abundant single nucleotide polymorphisms (SNPs) are the markers of choice in genetic case–control association studies. The genotypes of neighboring SNPs are often highly correlated ('in linkage disequilibrium', LD) within a population, which is utilized for selecting specific 'tagSNPs' to serve as proxies for other nearby SNPs in high LD. General guidelines for SNP selection in candidate genes/regions and genome-wide studies are provided in this protocol, along with illustrative examples. Publicly available web-based resources are utilized to browse and retrieve data, and software, such as Haploview and Goldsurfer2, is applied to investigate LD and to select tagSNPs.

INTRODUCTION

The human genome consists of over 3 billion base pairs and was sequenced as part of the Human Genome Project that started in 1990. The 'final' version of the sequence was published in 2003 and estimated at the time to be 92% complete[1]. It showed that the human genome consists of ∼22,000 GENES (see **Box 1** for a glossary of terms) and that human individuals are identical for ∼99.5% of their sequence, with the small remaining part variable to differing extents. As this variation could have an important role in explaining differences in genetic susceptibility to disease, comparing variation between diseased (cases) and healthy (control) individuals from the same population may elucidate which genetic pathways are involved in disease onset and progression[2].

Genetic variation at a specific locus is termed a polymorphism if it occurs with a frequency $>1\%$ in the population and a mutation if it occurs less frequently[3]. The most common class of polymorphisms involves a single base-pair change, also termed single nucleotide polymorphisms (SNPs), which comprise ∼90% of all human variation. Other types of polymorphisms include larger blocks of sequence variation (mini-/micro-satellites) or more complex alterations of the sequence, such as inversions and deletions, or copy number variations (CNVs)[4]. The different variations that a polymorphism can have at a particular locus are termed ALLELES. For SNPs, the single base-pair changes occur predominantly within the two classes of nucleotides, between purines (A ↔ G) and pyrimidines (C ↔ T), which means that most SNPs will only have two alleles in a population. The specific combination of these alleles in an individual counting across the two relevant chromosomes is referred to as a GENOTYPE. In contrast, the combination of consecutive alleles on a single chromosome is termed a HAPLOTYPE.

The potential of using these abundant SNPs in association studies motivated a large-scale international cataloging effort, the SNP Consortium[5]. The database used for submission, dbSNP, currently holds >10 million SNPs. Conducting association studies by genotyping samples for all known polymorphisms is neither financially feasible nor necessary: various studies have shown that alleles of SNPs located close together on a chromosome are not inherited independently within a population but are correlated, such that many individuals share the same haplotype (haplotype blocks)[6]. This correlation between genetic variants within a population is termed 'LINKAGE DISEQUILIBRIUM (LD)', and it is this feature that is utilized in marker selection.

The international HapMap project was initiated to characterize LD across the human genome. In phase I of the study ∼1 million common SNPs were genotyped in four different sample sets of 90 Yoruba in Ibadan from Nigeria (YRI), 45 Japanese in Tokyo from Japan (JPT), 45 Han Chinese in Beijing from China (CHB) and 60 Utah Residents with Northern and Western European Ancestry (CEU)[7]. The study confirmed that alleles of neighboring SNPs are often highly correlated within a population of unrelated individuals and that specific 'tagSNPs'[8] can be selected to serve as proxies for other SNPs in high LD, thus substantially reducing the number of SNPs that need to be genotyped in order to recover most of the information about common variation. The HapMap project provided a map of the LD structure across the genome, together with general statistics for markers such as location and allele frequencies, which allow investigators to select tagSNPs that best capture the common genetic variation in a particular region or across the genome. HapMap phase II increased the number of SNPs genotyped in the four populations to ∼3.2 million[9]; phase III extends the information to an additional seven populations but at a limited number of SNPs. On account of the relatively small number of individuals genotyped in the HapMap project—60 unrelated individuals genotyped in each population sample—only relatively common polymorphisms (with frequency >0.02) can be captured. However, detecting the influence of rare variation underlying common disease through population-based case-control studies—even with 1000s of cases and controls—is difficult, unless effect sizes are very large[2].

When utilizing the HapMap data as a reference for the selection of SNPs for a genetic association study it is important to be aware that the genotypes and associated statistics are based on the

BOX 1 | GLOSSARY

ALLELE – a variant of polymorphism at a locus

ENHANCER – a regulatory region of the DNA to which proteins can bind in order to increase the expression of a gene

EXON – a section of a gene that is represented in the mature form of the mRNA transcript

GENE – a functional unit of the DNA that contains the necessary information for the cell machinery to create a RNA template that is either functional by itself or can be translated to a protein

GENOTYPE – the combination of two alleles across both chromosomes at a particular locus in an individual

HAPLOTYPE – a consecutive sequence of alleles at different loci on a single chromosome in an individual

HAPMAP – An international project to develop a haplotype map of the human genome. The publicly available data consist of ∼3.2 million common SNPs genotyped in four different sample sets of 60–90 individuals with African, Asian and European ancestry. Frequently used as reference data for designing and following up genetic case control studies

INTRAGENIC REGION – a section of the genome that lies between genes

INTRON – a noncoding section of a gene that is removed from the mature mRNA sequence by a process called splicing

LINKAGE DISEQUILIBRIUM (LD) – The population correlation between two (usually nearby) allelic variants on the same chromosome; they are in LD if they are inherited together more often than expected by chance

REGULATORY REGION – a section of the DNA sequence that directly or indirectly affects the expression of a gene

PROMOTER – a regulatory element often located immediately upstream of the gene whose expression it regulates

SINGLE NUCLEOTIDE POLYMORPHISM (SNP) – a genetic variant that consists of a single DNA base pair change, resulting in two possible allelic identities at that position

TAGSNP – a SNP in a region of the genome featuring high LD, which is a proxy for others in close proximity and which can be used to genotype individuals at a reduced cost, while maintaining power

UNTRANSLATED REGION (UTR) – A section of the mRNA transcript that is not translated to amino-acid sequence. Two examples of such regions are the 5′UTR and the 3′UTR, which are located before and after the coding parts of the transcript, respectively

characterized sample populations and may not be directly transferable to the individuals in a given study. Thus, investigators should view summary statistics for the HapMap population, which they believe is most similar in terms of ancestry to the study population. Moreover, certain SNPs will either not have been genotyped in HapMap or may be monomorphic in the HapMap population of 60 individuals. Better characterization of such SNPs may be based on a more complete resource such as dbSNP, although this database has the limitation that SNPs may not have been validated and that no genotype data is publicly available. For studying specific candidate regions, other publicly available SNP databases such as SeattleSNPs and GeneSNPs (http://www.genome.utah.edu/genesnps/) sometimes offer better coverage, as they contain genotype information from resequencing of genes or candidate regions. An international collaborative effort launched in 2008, The 1,000 Genomes Project aims to resequence 1,000 individuals using third-generation high-throughput resequencing technologies. The outcome of this project will greatly facilitate the follow-up of genetic studies and will be a main resource for the development of new analytical and experimental methods.

The types of genetic association studies conducted are commonly divided into candidate gene/region and genome-wide association studies. Both these approaches involve genotyping SNPs in large collections of cases and controls (the focus of this protocol) or in large collections of individuals characterized by continuous quantitative traits. Although the general principles by which markers are selected in candidate gene versus genome-wide studies are the same, i.e., based on optimal genomic coverage, the investigator-driven methodology that is implemented in a candidate gene/region approach implies that in practice they are very different. A candidate gene study is based on an earlier hypothesis suggesting a potential role of the gene(s) in a particular phenotype or disease. The support for the selection of a candidate gene or region is typically based on its biological function or on its location in a region implicated in a previous linkage or association study. In a candidate gene study the aim is to obtain the highest possible coverage of genetic variation within specified genomic boundaries, taking account of any known functional genomic characteristics. In contrast, genome-wide association studies utilize predesigned genotyping panels containing selected sets of SNPs distributed throughout the whole genome. Genome-wide association studies are useful for hypothesis-generating purposes, but rarely provide the same coverage for a candidate region as a well-designed candidate gene study.

Candidate gene/region marker selection

The advantage of performing a study focused on one or few candidate genes or regions is cost efficiency. A relatively small number of markers will need to be genotyped to capture most of the common variation in the candidate region. However, *de novo* candidate gene studies, in which candidates are selected entirely on the basis of their supposed biological significance without any previous statistical evidence that the region they are located in is implicated, have had very low success rates even when well designed[10]. Candidate gene/region studies should therefore, follow genomic linkage or association studies that have already implicated the region. Linkage studies are designed to indicate evidence of relatively rare variants with large effects, whereas association studies are powered to pick up common variants with modest effects. Following up linkage signals with population-based association studies might not be the optimal strategy, and this needs to be taken into account when selecting both the individuals and the markers to genotype in the follow-up study. Alternatively, candidate genes can be selected from biological pathways that harbor other previously associated risk loci. Various types of other data can also be useful to provide further insight on functionality, e.g., whether a gene is expressed in tissues relevant to the biological mechanism of the disease.

A genetic polymorphism may act by affecting the regulation of expression of the candidate gene or by leading to a changed composition and function of the protein that it encodes. The selection of SNPs in or around the candidate gene or region should take account of in-depth characterization of the functional elements identified therein, and how their localization in relation to the gene might affect its function. Nevertheless, it is important to note that functional elements both within and outside the boundaries of a gene may well be unknown at the time of study, and our knowledge of functionality is expanding continuously. REGULATORY regions, such as PROMOTERS and ENHANCERS are essential for controlling the extent to which the gene is transcribed. Polymorphisms in coding EXONS and regulatory regions are less frequent than variations in a part of a sequence with no clear functional role, such as INTERGENIC REGIONS and INTRONS. This is explained by the fact that altering the function of the fine-tuned biological system will likely have a detrimental effect on the functioning and survival of the individual, the result of which is a smaller likelihood that the variation is passed on and established in the population (although a mutation can also have a positive effect on survival and, therefore, increase in frequency).

Through customized genotyping, the coverage can be specifically increased for the candidate genes/regions. In addition, the investigator has the opportunity to customize SNP selection, ensuring the inclusion of known functional or rare variants that might have been removed from other commercial panels. Therefore, the first step in SNP selection for a candidate gene or region is to find out relevant genomic information: identify all known SNPs, identify those known to be located in functional regions and characterize the LD structure, using publicly available online databases such as Ensembl, UCSC Genome Browser[11] and HapMap.

Depending on the size of the candidate region and the cost implications, the investigator can choose to include all potentially functional SNPs (localized in predicted regulatory regions, splice sites, intergenic sequence, introns or coding exons) in marker selection, supplemented by a selection of tagSNPs or rely on the use of tagSNPs alone to cover the gene/region. Selecting markers on the basis of functional annotation alone is not recommended, since the causative polymorphism may not have been identified yet or may be in a region that has to date been deemed 'non-functional'[12]. Of course, if a specific study is aimed at replicating and/or fine mapping a candidate region suggested by the results of a previous study, it is imperative that the SNP associated in the original study is included in the design.

A final step in selecting SNPs is, if possible, to check whether there is any evidence that genotyping is at risk of failing. For instance, it is normally considered difficult to genotype or sequence in genomic regions of low complexity, containing numerous repetitive elements. As the genotyping could fail for some markers in these regions, it is useful to incorporate additional SNPs in order to capture local variation. This is particularly important in a replication study, where the ability to replicate hinges on the genotyping accuracy of the SNP in question. In this situation, it is advisable to include the next most highly associated SNP in high LD ($r^2 > 0.7$) and/or a SNP in complete LD with the initial hit using data from the original study or HapMap information if the samples are of similar ancestry.

When genotyping SNPs in a candidate gene, the number of SNPs to be genotyped is unlikely to exceed the hundreds. Such quantities are usually easily genotyped using low-throughput, PCR-based methods[13]. However, when genotyping a candidate region the number of SNPs is likely be in the thousands. With increasingly large numbers of SNPs, it will be more cost-efficient to conduct genotyping using high-throughput array-based products, similar to those used in Genome-wide association (GWA) studies. Such custom-made arrays are normally designed as multiplexes with a fixed number of markers that can be genotyped per chip (e.g., Illumina's custom arrays are currently designed in panels of up to 1,536 SNPs or > 7,800 SNPs and Affymetrix at present offers solutions for customized genotyping in configurations of 3K, 5K, 10K and 20K markers). The manufacturers of these panels also publish scores indicating the likelihood of genotype failure of individual SNPs on the relevant panel, thus helping the investigator choose a set of SNPs that will suffer minimal failure.

Genome-wide association panels

With recent advances in the efficiency of array-based high-throughput SNP genotyping technology[14], hundreds of thousands of SNPs are now routinely genotyped on sample sizes necessary to detect the modest genetic effects we expect for complex diseases[2]. GWA studies have been undertaken to study a wide variety of diseases and phenotypes, often involving big consortia to obtain the necessary sample sizes. The success of recently published GWAs has shown, that this approach can be powerful in helping to identify disease related loci[15]. One of the largest published examples highlighting the success of GWA studies was conducted by the Wellcome Trust Case Control Consortium (WTCCC) where 17,000 individuals were genotyped with the GeneChip 500K Mapping Array Set (Affymetrix chip) panel to study seven different diseases[16]. Subsequent studies showed that most of the significant associations in the WTCCC, as well as those found in other GWA studies, could be replicated and are thus very likely to be real. A summary of recent GWA publications attempting to assay at least 100,000 SNPs is available online and lists the results from hundreds of publications with thousands of SNPs significantly associated to a wide variety of diseases and phenotypes (http://www.genome.gov/gwastudies/).

The main difference between currently available GWA panels involves the number of SNPs, for which probes are included and the resulting genomic coverage obtained by this selection. Some of the latest panels also contain probes for analyzing copy number variations (CNVs), although there are also methods for identifying these *in silico* from genotyping intensities. Current commercially available genotyping panels typically range in capacity from 300K to 1M SNPs, but there are also focused panels that only capture a few hundred SNPs. The design of the panels mainly falls into three categories containing SNPs (1) more or less randomly selected, (2) chosen because they are tagSNPs and (3) chosen as part of a focused selection based on previously implicated functional importance or for lying within or nearby genes that could have a role for different diseases such as cancer. TagSNP panels have been shown to provide better coverage of common variation[17]. An important note is that the calculated coverage by the selection of tagSNPs is currently on the basis of coverage of all common SNPs genotyped in HapMap, not of the remaining unknown genetic variation. Therefore, coverage achieved is likely to be overestimated and may not necessarily be sufficient for fine mapping and studying specific regions of interest. Data from the ENCODE project, aimed at identifying all genetic variation in ten selected genomic regions for 48 HapMap individuals, however, showed that HapMap phase II data generally provide a high coverage of all common genetic variation[18]. The 1,000 Genomes Project, aimed at uncovering genomic variation in 1,000 individuals around the world through resequencing, should enable a more accurate evaluation of coverage.

Among the currently available genotyping panels, aiming for very high coverage at the expense of sample size will result in a decrease rather than an increase of power[19]. Choosing a panel with fewer markers will reduce genomic coverage, but the funds saved by choosing the cheaper panel can, instead, be spent on genotyping more individuals. This will increase the statistical power to a greater extent than adding more markers would do. In addition, recently published methods enable imputation (prediction) of ungenotyped SNPs on the basis of those that are genotyped, utilizing LD structures calculated from HapMap[20]. This offers a potentially attractive approach for increasing coverage, while spending more funds on increasing sample size[19]. However, imputation is not necessarily a successful solution to increase coverage in all situations. Accurate prediction of an ungenotyped marker depends on the availability of an already genotyped marker to act as a proxy. If a SNP does not have any markers in high LD with it, it cannot be predicted accurately . This implies that, in practice, well-covered regions can easily be filled in, but poorly covered regions cannot be improved to any major extent. Given that imputation methods utilize the haplotype structure of reference datasets for the inference of genotypes, The 1,000 Genomes Project will also contribute highly to our understanding of genotype imputation methods and their accuracy for imputation of polymorphisms of different frequencies.

Another strategy to save funds, which can be spent instead on choosing a greater-coverage genotyping panel or preferably genotyping more cases, is to use publicly available genotyping data for common controls from another study such as the WTCCC project (provided that they have been sampled from the same underlying population as cases and that population structure is investigated and adjusted for at the analysis stage[2]). Publicly available controls, however, are likely to have been genotyped on different panels, and if so, imputation would need to be performed to allow combined analyses.

In this protocol, we present two scenarios for SNP selection, involving a candidate-gene and a GWA study, following the example of type 2 diabetes[2]. Marker selection for the candidate gene approach will center around the investigation of the influence of Peroxisome Proliferator-Activated Receptor Gamma (PPARG) on type 2 diabetes in a population of European ancestry. As PPARG (rs1801282) was previously implicated in this disease[21], we ensure that the associated variant is included in the final marker selection. Although the example shown relates to a candidate gene, the same methods are applicable when selecting markers for a larger candidate region.

Commercially available products mainly differ in the coverage to price ratio, which rapidly changes with the development and release of new panels. In the GWA procedure (see **Box 2**) we explain how to calculate the coverage of available genotyping panels for the markers genotyped in the HapMap project. The script that will be used for calculating the coverage implements the algorithm published by Barrett and Cardon[17]. A new release of the HapMap project with a more complete selection of SNPs genotyped for a larger number of individuals is likely to be released in the near future as well as new commercially available genotyping panels. The coverage calculations can easily be repeated for additional panels by running the script provided, using updated reference data and new SNP lists. Guidelines on how to quality-check and analyze generated GWA and candidate gene data will be presented in subsequent *Nature Protocols*.

MATERIALS

EQUIPMENT

- Computer (PC or workstation) with web browser
- Computer with Java Runtime Environment (JRE) 1.5 or later installed
- Computer with Perl script interpreter installed
- Unzipping tool such as WinZip or gunzip
- Program to select tagSNPs (http://www.broad.mit.edu/mpg/haploview/)
- Program to visualize genes, SNPs and LD structure (http://www.well.ox.ac.uk/gs2)
- Scripts for calculating coverage for current genotyping panels (http://www.well.ox.ac.uk/~carl/gwa/cost-efficiency/)

BOX 2 | GENOME-WIDE ASSOCIATION STUDY

Calculating genotyping coverage

1. To download a script for calculating coverage (estimate-coverage.pl) and necessary auxillary files, go to http://www.well.ox.ac.uk/~carl/gwa/cost-efficiency/. Scroll down to the 'coverage of genome-wide SNP platforms' section and download files by clicking on the first four links. If downloading does not start automatically, right-click on each of the links and choose to download linked files. Save the files in a suitable directory. Unzip the 'whole-genome.rsq.gz' file. The script accepts several arguments as inputs together with the list of SNP rs identifiers, for which coverage is to be estimated. Precompiled lists for currently available genotyping panels can be downloaded by clicking on any of the links in the next section on the webpage.

2. In this protocol, we will calculate the autosomal coverage (at an r2 ≥ 0.8) in HapMap release 35 for the Affymetrix SNP array 5.0 panel. Download the list by clicking on the link named accordingly and save it in the same folder as the files downloaded in step 1. Open a terminal in which to type the command to start the calculation. In the terminal, go to the directory where you downloaded the files. Start the calculations by typing '`perl [path_to/]estimate-coverage.pl -file=affymetrix-SNP-array-5.0.snps -r2=0.8 -skipX`'

▲ CRITICAL STEP If a .txt extension has been added to filenames during downloading, remove it before running the script

▲ CRITICAL STEP Make sure that the scripting language perl is installed on the computer. To list the options for the script, type 'perl estimate-coverage.pl' without arguments in the terminal.

PROCEDURE

Finding and visualizing genomic information for a candidate gene (e.g., PPARG)

1| The main procedure focuses on the selection of tagSNPs for candidate gene studies. In **Box 2**, directly after this procedure, a separate example illustrates how to calculate the genomic coverage for a typical GWA study obtained by using any of the commercially available genotyping panels. Start the candidate gene procedure by going to the UCSC Genome Browser homepage[12] to get an overview of the available genes with a specific search term. Click on the 'Genome Browse' link on top of the menu on the left side. By default, the genome loaded is that for 'clade = mammal, 'genome = human' and 'release = March 2006' (the latest genome release).
▲ CRITICAL STEP There are two main ways of accessing the UCSC database, either through the Genome Browser, which gives an interactive graphical representation of the information or the Table Browser showing the data in table mode. It should be noted that we use UCSC Genome Browser in our example; however, there are many websites with similar genomic information. Another useful website is Ensembl.

2| In the box 'position or search term', enter the name of the candidate gene, in this procedure: 'PPARG' (case insensitive) to search for all entries in the UCSC database with the search pattern in their name, description or connected information. Scroll down to view the 'RefSeq' and 'Known Genes' tracks. Under the 'RefSeq Genes' section four hits for the PPARG gene are listed. They represent different splice variants of the gene, indicating that the mRNA transcript is composed of varying combinations of exons or reflects different gene predictions with potentially different regulatory regions.
▲ CRITICAL STEP Make sure to use the consensus name of the gene of interest, as alternative names are not necessarily included in every browser. The RefSeq id is commonly used and can be queried and cross-referenced in most databases. When searching by position, make sure that the coordinates entered refer to the same genome release as selected in the browser. The search result is presented in four sections, one for each scanned sequence database. The mRNA search results are obtained by scanning data associated with the GenBank record for mRNAs. These are often redundant but occasionally contain interesting hits not yet deposited to RefSeq. The mRNA information can also be used to link to the people who deposited the mRNA into GenBank.

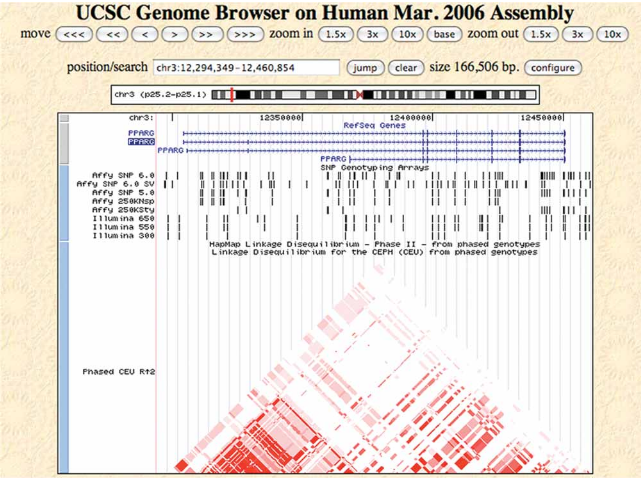
3| Choose the RefSeq gene that covers the largest area of the gene of interest. In our case, we choose the first hit of the RefSeq genes, 'NM_138712'. Under 'RefSeq Genes' click on the first link: 'PPARG at chr3:12304349-12450855' to view the genomic region. The name of the entry that was clicked is highlighted and surrounded by the other variants of the gene listed in the previous results page. The view shows that PPARG is located on chromosome 3 between p25.2-p25.1. The tracks under the RefSeq Genes show the SNPs that are captured by each panel selection of commercially available genotyping panels.
▲ CRITICAL STEP Web-based resources such as bioinformatics databases are dynamic and are likely to change. Keep track of the version used and when it was accessed. Results published earlier gives support for 'NM_015869' rather than 'NM_138712' being the biologically functional variant in humans[21]. For other less well-studied genes the choice may not be as obvious, which is why opting for the longest version is generally recommended.

4| There is a large selection of controls for customizing the appearance of the browser. One interesting feature is to view the LD structure around a candidate gene. Scroll down to the blue bar entitled 'Variation and repeats'. Click on '+' to display its options. Click on the 'Hapmap LD Phased' link above the combo box. This brings up a page with different LD display options. Change 'display mode' to 'Full' and select as LD values 'r^2'. Under 'populations', deselect 'Linkage Disequilibrium for the Yoruba from phased genotypes' and 'LD for the Han Chinese + Japanese from Tokyo (JPT+CHB) from phased genotypes', leaving only 'LD for CEPH (CEU)' ticked. Leave all other options as they are, and click on the 'submit' button. The genome browser page now shows two extensive blocks of LD structure across the PPARG region, with red representing high LD values in terms of r^2, and paler pink representing lower LD and white representing no LD.
▲ CRITICAL STEP The default settings for web-based tools such as the UCSC genome browser, and the information in the databases themselves are likely to change over time. Therefore, the layout presented in this protocol might not be exactly the same as that viewed by a reader.

5| It is often useful to look beyond the boundaries of a gene to get an overview of its proximal potential regulatory regions and to see how far a surrounding LD block extends. For PPARG, increase the visualized region by 40kb (an arbitrary value) both upstream and downstream, by entering 'chr3:12,264,349-12,490,855'. The LD structure for the CEU shows that the PPARG gene includes two blocks of high LD, that expands approximately to the boundaries of the largest variant of the gene (**Fig. 1**). Given that LD does not extend outside the gene implies that there is no need to increase the main interval within which we select our markers. For other candidate regions it might be necessary to extend the interval to get an overview of functional elements in the region.

6| Another useful tool for visualizing the candidate region and to gain access to underlying genotype data is the HapMap genome browser. Go to the International HapMap project homepage. Click on the 'HapMap Genome Browser (Phase I & II—latest genotypes & frequencies)' in the left-hand menu under the 'project data' section. This opens a browser for viewing the data in the context of the latest genome release. In 'Landmark or Region', enter PPARG (case insensitive), and in the results click on 'NM_138712 to zoom in on the largest variant of the gene.

▲ CRITICAL STEP In this protocol, we have chosen to use phase II HapMap data, even though phase III has recently been released. Although phase III includes a more comprehensive set of individuals (seven additional populations), the number of genotypes is reduced compared with phase II. The optimal release to be used will depend on the sample population being studied and the candidate region of interest.


UCSC Genome Browser on Human Mar. 2006 Assembly

position/search: chr3:12,294,349-12,460,854 — size 166,506 bp.

Figure 1 | View of the candidate genomic region around PPARG using the UCSC genome browser. The plot shows the different versions of the candidate gene and the LD structure in the region as measured in r^2 for the CEU HapMap sample.

Retrieving HapMap SNP genotypes for the region

7| To capture any upstream regulatory elements we increase the size of the region by adding 10k (arbitrary and depends on the candidate region) at both the start and the end of the region. In the 'Landmark or Region' textfield enter 'chr3:12,294,349-12,460,854' and press 'search' to refresh the view (**Fig. 2**). Under 'reports and analysis' choose 'Download SNP genotype data' and click on 'Configure'. In the window that opens up choose 'CEU', 'rs' and 'Save to Disk' and click on 'go'. Save the genotype file under a suitable name (here: pparg_dump.txt). This file now contains the genotypes of all SNPs in PPARG ± 10kb that were genotyped in HapMap phase II (**Supplementary Table 1** online).

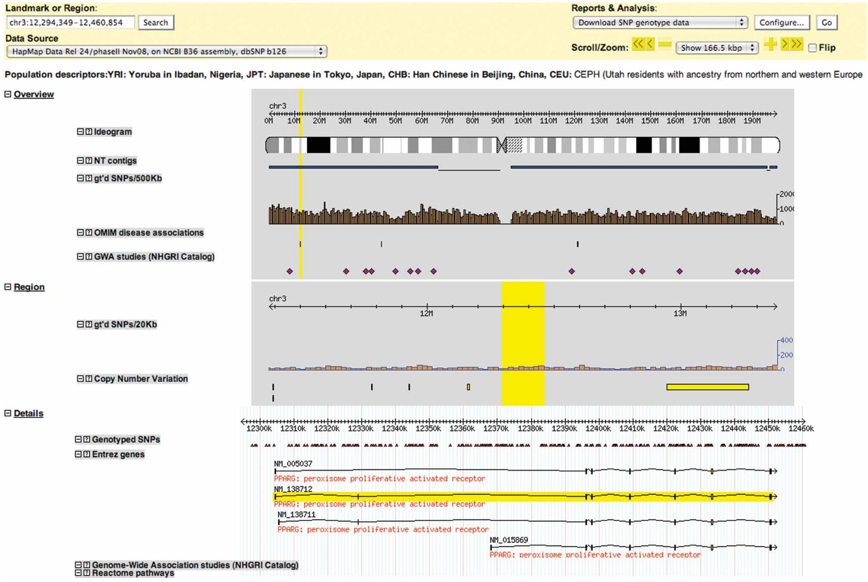
▲ CRITICAL STEP It has been suggested that the most distal enhancer element is found at a distance of approximately 1 Mb from the gene it is regulating. Ensuring to catch such effects within the scope of a standard candidate gene study is likely to involve genotyping a very high number of SNPs. It should be noted that genotypes can also be downloaded using the HapMart facility or can be bulk downloaded from the FTP server, both accessible from the HapMap webpage.

Creating a list of functional SNPs to be forced into a tagSNP selection

8| The HapMap SNPs are all a subset of the dbSNP database, which can be queried to retrieve information about their functional annotation. The UCSC table browser is a useful tool for retrieving functional annotation from this and other databases. Go to the UCSC genome browser webpage and click on 'Table browser'. Start with the preset default settings and select 'group: Variation and Repeats', 'tracks: SNPs (129)' and 'table: snp129'. In the textfield to the right of 'region:' enter the interval used in the last step and click on 'position'. Enter 'snpAnnotation.txt' in the 'output file:' text field and click on 'get output' to save the results (**Supplementary Table 2**).

▲ CRITICAL STEP There are many publicly available tools for in-depth evaluation of the functional annotation of genetic variants. One useful example is HuGE navigator (http://hugenavigator.net/HuGENavigator/startPageMapper.do), which facilitates interactive searches across databases and interpretation of results by cross-referencing.

▲ CRITICAL STEP In the table browser entry page click on 'describe table schema' to view a description of the information available in the selected table.


Landmark or Region: chr3:12,294,349-12,460,854

Data Source: HapMap Data Rel 24/phaseII Nov08, on NCBI B36 assembly, dbSNP b126

Reports & Analysis: Download SNP genotype data

Scroll/Zoom: Show 166.5 kbp

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

Search
Configure...
Go
☐ Flip
Overview
chr3
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M 160M 170M 180M 190M
Ideogram
NT contigs
gt'd SNPs/500Kb
OMIM disease associations
GWA studies (NHGRI Catalog)
Region
chr3
12M
13M
gt'd SNPs/20Kb
Copy Number Variation
Details
12300k 12310k 12320k 12330k 12340k 12350k 12360k 12370k 12380k 12390k 12400k 12410k 12420k 12430k 12440k 12450k 12460k
Genotyped SNPs
Entrez genes
NM_005037
PPARG: peroxisome proliferative activated receptor
NM_138712
PPARG: peroxisome proliferative activated receptor
NM_138711
PPARG: peroxisome proliferative activated receptor
NM_015869
PPARG: peroxisome proliferative activated receptor
Genome-Wide Association studies (NHGRI Catalog)
Reactome pathways

Figure 2 | View of the candidate genomic region around PPARG using the HapMap genome browser. The plot shows the genotyped SNPs in the region and the different versions of the candidate gene together with other information. The 'GWA studies (NHGRI Catalog)' track shows SNPs that have previously been found to be associated with a studied trait in published GWA literature.

9| Open up the file in Excel, sort it by the 'func' column, which contains information about the functional annotation, and keep entries which for this column start with 'coding' or 'cds'. Copy the

column with the remaining 'rs' names to a text editor. The marker 'rs1801282', which has earlier been implicated to be associated with type 2 diabetes is added to the list and the filed is saved as 'forceList.txt' (**Supplementary Table 3**).
▲ CRITICAL STEP The final force list is compiled from the entries in the dbSNP database, which contains many SNPs that might not have been validated or that have not been genotyped in the HapMap project. When later importing the forceList for the selection of tagSNPs, the markers that are not available in the dataset are automatically excluded. The choice of criteria to use for selecting markers differs between studies. Our selection is somewhat arbitrary and was chosen for illustration purposes. More advanced selection, forcing in markers in, e.g., promoters and UTRs, involves more in-depth data analysis of the candidate regions and the elements therein[18].

Selecting tagSNPs

10| Haploview[22] is a popular software for selecting tagSNPs and visualizing LD. TagSNPs can be selected by using the Haploview through the graphical user interface (GUI, option A) or through a terminal running it in command line mode (option B). Running the Haploview with a GUI has many advantages in terms of ease-of-use and display functions. However, when tagging a larger genomic region, it is advisable to use a server where the application can run in the background for computational reasons.

(A) Haploview with GUI

(i) Start up Haploview by clicking on the haploview.jar file. In the 'Welcome to Haploview window', select the 'HapMap Format' tab on the left side. Click the gray 'browse' button, and select the SNP genotype file downloaded at step 7: pparg_dump.txt. Leave the options as they are (ignore pairwise comparison of markers > 500kb apart and exclude individuals with > 50% missing genotypes). Click 'OK'. After import, the data can be accessed through 4 tabs. In the 'LD Plot' tab, the LD structure in the region is interactively visualized, the 'Haplotypes' tab shows the predicted haplotypes, 'Check Markers' summarizes in a table the statistics for the 296 imported SNPs and the 'Tagger' tab contains the controls for selecting the tagSNPs.
▲ CRITICAL STEP Depending on which format the genotype data is saved in, it might be necessary to select a different tab when importing the data into Haploview.

(ii) In the 'Tagger' tab, load the list of markers to force include that which was generated in step 9 by clicking on 'Load Includes'. Click on 'Run Tagger' to start the calculation. The results from the tagger are summarized under the 'Results' tab. The selected tagSNPs are shown in the lower left list that can be saved by clicking on 'Dump Tags File'. The resulting list of tagSNPs will be utilized as an input for the quality control and association testing in the following Nature Protocols (**Supplementary Table 4**).
▲ CRITICAL STEP It is recommended to select tagSNPs using the 'pairwise tagging only' approach. The alternative, multimarker tagging approach, utilizes haplotype structure for more efficient selection of tagSNPs. However, this algorithm requires greater genotyping quality and completeness and may result in loss of statistical power; also, data analysis will need to be haplotype-based.
▲ CRITICAL STEP If only common SNPs are of interest, it is advisable to set the minor allele frequency of tagSNPs to 0.05 in the 'Check markers' window, and click on 'Rescore markers'.
▲ CRITICAL STEP The LD cut-off specifies that all SNPs with a maximum pair-wise r^2 below the limit are automatically included in the selection of tagSNPs; the tagger algorithm picks a minimum set of tagSNPs from the remaining SNPs for maximum coverage. A cut-off of $r^2 = 0.8$ gives a similar coverage to that for a typical genome-wide association study. For better coverage increase the cut-off, though the extent of improvement is region-specific. In our example, applying a cut-off of $r^2 \geq 0.9$ results in a selection of 50 tagSNPs as compared with 42 tagSNPs selected on applying a cutoff of 0.8.

(B) Haploview from command line

(i) Start by opening a terminal in which the commands are typed. On Windows open the command prompt, which is most easily accessed through the start menu by clicking on run and typing cmd. If using Mac OSX, go through Applications-> Utilities and click on Terminal.app. For Unix/Linux check the documentation for the distribution used.

(ii) To run the Tagger algorithm type:

```
java -jar [path_to/]Haploview.jar -nogui -memory 1800 -hapmap [path_to/]pparg_-
dump.txt -pairwisetagging -tagrsqcutoff 0.8 -minmaf 0.05 -includetagsfile [path_to/
]forceList.txt
```

(where `[path_to/]` is the path from the directory where the command is typed to the directory where the haploview.jar file is stored). The arguments after java –jar specify, in order: the location of the haploview program: that no graphical interface should be created: the amount of memory allocated by the system: path to the HapMap genotypes: apply pairwise tagging: use 0.8 as cut-off for LD (r^2): exclude markers with minor allele frequency lower than 0.05: and finally path and filename containing a list with markers to force in. In this example specify the list that was compiled in step 9.

(iii) After running Tagger, two files are created, both with the prefix of the input file and ending with .TAGS or .TEST. The .TEST file contains the identifiers of the markers selected by the tagging algorithm (**Supplementary Table 4**).
▲ CRITICAL STEP To list the options for Haploview, type:
`'java -jar [path_to/]Haploview.jar –h'` in the terminal.

Visualizing the tagSNPs in the context of PPARG and regional LD structure

11| Start Goldsurfer2 (ref. 23) by clicking on the gs2.jar file or the gs2.exe (only for windows) in the unzipped GsWinExe folder. Gs2 can also be opened from the command line by typing:
`'java –jar –Xms64m –Xmx1024m gs2.jar'`. The –Xmx argument specifies the maximum amount of memory to be allocated by the system. If bigger datasets are to be analyzed this setting can be increased accordingly. Open the dataset by clicking on 'File'>'Load markers' and choosing 'HapMap(FTP)' as input format and specifying file to import by clicking on 'Browse & import'. The imported markers are summarized in the central table and in the summary plots in the right hand side of the window. To get an overview of the data in a genome browser view, click on the plot tab. Plot summary statistics by clicking on 'Plot>show/hide summary plots'. Select the rare SNPs (maf <0.05) by clicking on the bar showing markers with maf between 0 and 0.05 in the top right histogram. Right-click in the table and choose 'exclude' to remove the markers.
▲ CRITICAL STEP If importing genotypes are saved from other sources than those specified in the procedure, then choose import format accordingly.

12| Import information about genes in the region by clicking on 'Bio' in the main menu and choosing 'Parse genes (local)' > 'Release 36'. Only one variant of the gene is shown since the reference file is based on the 'knownCanonical' table in the UCSC, where redundancies and uncertain gene predictions have been removed. Calculate and show LD by going through 'Plots' in the main menu and choosing 'LD' > '2D' > 'Recalculate' or 'LD' > '3D' > 'Show plot'. To show the 3D view of LD, the SNPs to be visualized need to be selected before creating the plot. Adjust the tracks that are to be displayed in the genome view by right-clicking on the plot and selecting 'Plot settings'. Hide all the tracks except for ucscGenes36.txt, snp marks and LD. The vertical and horizontal bars can be used to navigate the view. To illustrate the selection of tagSNPs and the markers kept from forceList.txt, repeat the import of the dump file and again exclude the markers with maf < 0.05. Select the first dataset again and click on 'Stats>Link to other node' and choose the last imported dataset by double-clicking on it. Select the first dataset and paste rs1801282 in the 'Name (java regexp)' text field in the Selection dialog, accessed through the 'Selection' menu. For the second dataset, repeat the procedure but paste instead the identifiers for the tagSNPs. Click on the first dataset to view the resulting plot (**Fig. 3**).
? TROUBLESHOOTING

● TIMING

Candidate gene study

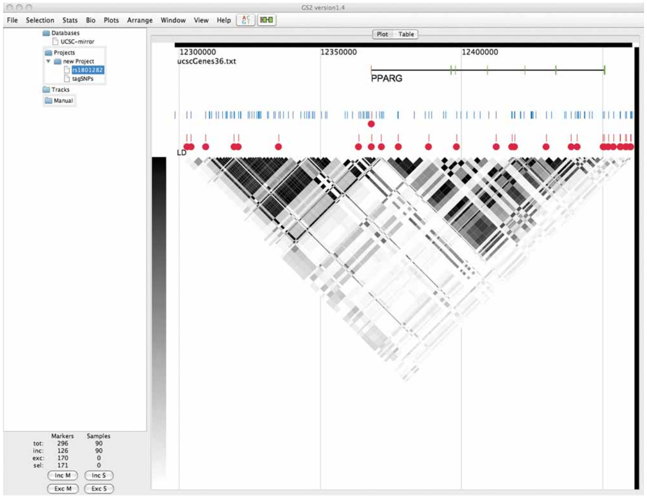
None of the programs used take longer than a few minutes to run. Exploring the different websites and displaying and interpreting the relevant information are the rate-limiting steps.

Genome-wide association study

The reference file with the LD measurements is relatively large (650Mb), with the downloading time depending on the speed of the Internet connection. With a standard broadband connection, downloading is likely to be completed in a few minutes. To run the estimate-coverage.pl script for a panel with ~500k SNPs takes around 15 min on a standard desktop computer.

? TROUBLESHOOTING

For help on the programs and websites used in this protocol, refer to the relevant websites:
GoldSurfer2, http://www.well.ox.ac.uk/gs2/tutorial.html
HaploView, http://www.broad.mit.edu/haploview/user-manual
UCSC, http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html
HapMap, browser http://www.hapmap.org/gbrowse_help.html
Estimate-coverage.pl, http://www.well.ox.ac.uk/~carl/gwa/cost-efficiency/



	Markers	Samples
tot:	296	90
inc:	126	90
exc:	170	0
sel:	171	0

Figure 3 | Genomeview from GoldSurfer2 showing the selection of tagSNPs (track 2) in relation to PPARG and all HapMap SNPs with minor allele frequency >0.05 (track 1). The SNP, rs1801282, previously implicated in type 2 diabetes, is highlighted near the start of the PPARG gene. LD is measured in r^2.

It should be noted that bioinformatics websites are continually subject to change, and therefore, their precise functionality and resulting displays may differ from the exact wording in this Protocol.

ANTICIPATED RESULTS

Visualizing genomic information in the candidate gene region (PPARG)

Figure 1 shows the PPARG genomic information displayed using the UCSC database by focusing on the region 40 kb up and downstream of the RefSeq track NM_138712 (Step 5). The LD structure for the CEU shows that the PPARG gene includes two blocks of high LD that expand approximately to the boundaries of the largest variant of the gene. Given that LD does not extend outside the gene means that there is no need to increase the main interval within which we select our markers. Furthermore, substantial cost reductions can be achieved by selecting the tagSNPs from within the two large LD blocks. The SNPs captured by different commercially available panels are highlighted, and with some of these designed using a tagging approach, it gives an idea about the result obtained by this approach and shows the applicability of the genotype data obtained from a GWA panel.

Figure 2 shows the PPARG genomic information displayed using the HapMap Genome Browser (step 6). Multiple types of information can be visualized and downloaded by using the combobox under 'Reports & Analysis' and the controls under the plot window. Useful features to look at are copy number variations (CNVs) and results from previously published GWA studies by default presented as separate tracks in the plot. By clicking on the question mark beside the 'GWA studies (NHGRI Catalog)' label, it is possible to get directly linked to published GWA studies.

Retrieving HapMap SNP genotypes for the PPARG region

The saved HapMap genotype dump file (step 7) contains genotype data for 296 SNPs in 90 individuals (**Supplementary Table 1**).

Creating a list of functional SNPs to be forced into a tagSNP selection

The full list downloaded in step 8 contains 832 SNPs (**Supplementary Table 2**). After filtering, the forceList.txt file (step 9) contains 12 SNPs, including the SNP, rs1801282, implicated earlier (**Supplementary Table 3**).

Selecting tagSNPs

After tagging using HaploView, a set of 42 tagSNPs (**Supplementary Table 4**) was selected including two SNPs that were force-included (steps 10A(ii) and 10B(iii)). The selection of SNPs in the force list was based on the dbSNP database in which a high proportion of the entries are neither validated nor have been genotyped in the HapMap project. This explains why only two of the 12 SNPs in the list enter into the final selection of tagSNPs.

Visualizing the tagSNPs in the context of PPARG and regional LD structure

Figure 3 shows the tagSNP visualization using GoldSurfer2 (step 12). The upper track with SNP marks shows all common (maf $>$ 5%) HapMap SNPs in the region with rs1801282 highlighted in red. The second SNP track shows the coverage of the selected tagSNPs. The knownCanonical gene table from UCSC is cleaned for false hits and redundancies with the effect that only the suggested functional version of PPARG is visible in the plot. LD is measured in r^2.

Calculating coverage

The calculated coverage is 0.648, as output by the coverage program.

Note: Supplementary information is available via the HTML version of this article.

ACKNOWLEDGMENTS F.H.P., C.A.A., G.M.C., A.P.M., and K.T.Z. are funded by the Wellcome Trust, UK.

Published online at http://www.natureprotocols.com/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Zondervan, K.T. & Cardon, L.R. Designing candidate gene and genome-wide case-control association studies. *Nat. Protoc.* **2**, 2492–2501 (2007).
3. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
4. Scherer, S.W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* **39**, 7–15 (2007).
5. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
6. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
7. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
8. Johnson, G.C.L. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237 (2001).
9. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
10. Hirschhorn, J.N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
11. Kuhn, R.M. *et al.* The UCSC Genome Browser Database: Update 2007. *Nucleic Acids Res.* **35**, 668–673 (2007).
12. Evans, D.E., Barrett, J.C. & Cardon, L.R. To what extent do scans of non-synonymous SNPs complement denser genome-wide association studies? *Eur. J. Hum. Genet.* **16**, 718–23 (2008).

13. Perkel, J. SNP genotyping: six technologies that keyed a revolution. *Nat. Methods* **5**, 447–453 (2008).
14. Syvanen, A.C. Toward genome-wide SNP genotyping. *Nat. Genet.* **37**, 5–10 (2005).
15. Kingsmore, S.F., Lindquist, I.E., Mudge, J., Gessler, D.D. & Beavis, W.D. Genome-wide association studies: progress and potential for drug discovery and development. *Nat. Rev. Drug Discov.* **7**, 221–30 (2008).
16. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
17. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659–662 (2006).
18. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
19. Anderson, C.A. *et al.* Evaluating the effects of imputation on the power, coverage and cost-efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.* **83**, 112–119 (2008).
20. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
21. Altshuler, D. *et al.* The common PPAR gamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **26**, 76–80 (2000).
22. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
23. Pettersson, F., Morris, A.P., Barnes, M.R. & Cardon, L.R. Goldsurfer2 (Gs2): A comprehensive tool for the analysis and visualization of genome wide association studies. *BMC Bioinformatics* **9** (2008).