# 2 MATERIALS AND METHODS

This chapter describes the materials used in this study and the methods of analyses that were applied to the data. The first part presents the source of the DNA samples with information on their geographical origin as well as noting the criteria applied to select the SNPs for the study. The second part describes the laboratory methods applied, primers designed and protocols that were followed. The third and final part lists the programs, databases and scripts used and describes the computational methods that were used in analysing the data – inferring the ancestral state of the alleles, predicting the protein truncation and NMD, looking at the gene ontology and applying summary statistics to search for selection.

## 2.1 THE DATA

### 2.1.1 The Samples

The samples genotyped were derived from 1,191 individuals from 56 geographically diverse populations. 1,064 samples were obtained from the Foundation Jean Dausset, the CEPH Human Genome Diversity Cell Line Panel (HGDP-CEPH) (Cann et al. 2002) and 127 unrelated individuals from the four HapMap populations – CEPH Utah residents with ancestry from northern and western Europe (CEU), Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing (CHB), Japanese in Tokyo (JPT) (The International HapMap Consortium 2005).

The samples used for the re-sequencing analysis were from three HapMap (23 YRI, 23 CHB, 22 CEU) and 23 individuals from one extended HapMap population , the Luhya in Webuye, Kenya (LWK). In addition, one chimpanzee (*Pan troglodytes*) sample was included as an outgroup.

All HapMap samples were purchased from the Coriell Institute for Medical Research (Camden, New Jersey, USA), the HGDP-CEPH collection (Cann et al. 2002) was kindly provided by Howard Cann (CEPH, Paris, France) and the chimpanzee sample was purchased from the ECACC (Salisbury, Wiltshire, UK). The HGDP-CEPH samples were whole-genome amplified before use (see section 2.2.1). The HapMap samples were used as genomic DNA.

In the end, 1,151 of the original 1,191 samples were used in the final genotype analyses. A total of 40 samples were thus excluded. These included 16 samples from HGDP-CEPH which were excluded according to Rosenberg's suggestions for using standardized subsets of the original diversity panel (see Rosenberg 2006). I used the H1048 subset which contains no duplicated samples or individuals that are extremely atypical for their populations. According to this subset 18 individuals should be excluded, but two of these were not found in my dataset. The exclusion of duplicated samples followed the convention of discarding duplicates with higher identification numbers. I followed this rule except when the sample with the lower number yielded more genotype data. A further 24 samples were excluded because their genotyping failed completely.  Of the 91 HapMap samples used in the re-sequencing analysis, 88 were successfully re-sequenced.

The coordinates for the HGDP-CEPH populations were obtained from the CEPH website at http://www.cephb.fr/en/hgdp/diversity.php/table.php and the locations were projected onto a map (see Figure 8A). As exact coordinates were not available for the HapMap samples, their location is not shown on this map. When displaying pie charts with allele frequencies (in chapters 3 and 4) I grouped some closely related populations together to avoid population size bias, resulting in a total of 37 populations instead of 56 (Figure 8B). Additionally, the coordinates of a few HGDP-CEPH populations (in Israel, France, Italy, and Brazil) were changed slightly so that the pie charts would not overlap and the allele frequency proportions could be easily viewed. The HapMap pie charts were inserted separately onto the map. The details of all population names are further displayed in Table 1 and a full list of all samples used is given in Appendix A (on accompanying CD).

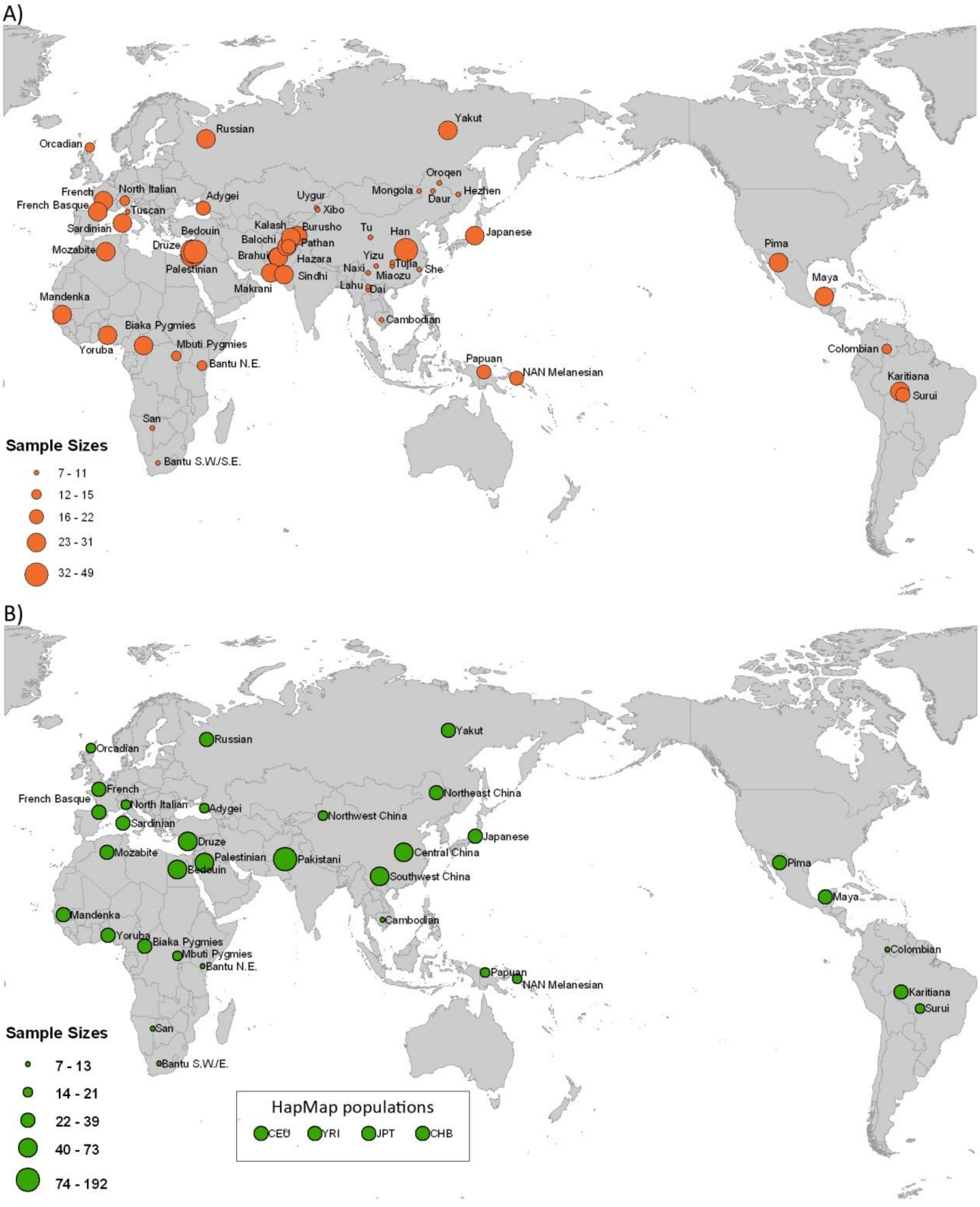

**Figure 8 Population locations of genotyped samples. A)** Geographical locations of the 52 HGDP-CEPH populations genotyped. The diameter of the orange circles is proportional to sample sizes. The HapMap populations are not shown. **B)** Geographical locations of the genotyped populations as they appear in allele frequency pie charts. Some related populations were clustered together to reduce population size bias, resulting in 37 populations displayed on the map. The diameter of the green circles is proportional to sample sizes. Coordinates were slightly shifted for populations too close to each other for the pie charts not to overlap. HapMap populations are inserted at the bottom of the map as they do not have geographical coordinates.

| Population (N = 56) | Sample No. | Source | Geographic Origin | Population (N = 37) |
|---|---|---|---|---|
| Mozabite | 30 | HGDP-CEPH | Algeria (Mzab) | Mozabite |
| NAN Melanesian | 19 | HGDP-CEPH | Bougainville | NAN Melanesian |
| Karitiana | 24 | HGDP-CEPH | Brazil | Karitiana |
| Surui | 21 | HGDP-CEPH | Brazil | Surui |
| Cambodian | 11 | HGDP-CEPH | Cambodia | Cambodian |
| Biaka Pygmies | 31 | HGDP-CEPH | Central African Republic | Biaka Pygmy |
| Dai | 10 | HGDP-CEPH | China | Southwest Chinese |
| Daur | 10 | HGDP-CEPH | China | Northeast Chinese |
| Han | 43 | HGDP-CEPH | China | Central Chinese |
| Han Chinese in Beijing | 32 | HapMap | China | CHB |
| Hezhen | 9 | HGDP-CEPH | China | Northeast Chinese |
| Lahu | 10 | HGDP-CEPH | China | Southwest Chinese |
| Miaozu | 10 | HGDP-CEPH | China | Southwest Chinese |
| Mongola | 10 | HGDP-CEPH | China | Northeast Chinese |
| Naxi | 10 | HGDP-CEPH | China | Southwest Chinese |
| Oroqen | 10 | HGDP-CEPH | China | Northeast Chinese |
| She | 10 | HGDP-CEPH | China | Central Chinese |
| Tu | 10 | HGDP-CEPH | China | Central Chinese |
| Tujia | 10 | HGDP-CEPH | China | Central Chinese |
| Uygur | 10 | HGDP-CEPH | China | Northwest Chinese |
| Xibo | 9 | HGDP-CEPH | China | Northwest Chinese |
| Yizu | 10 | HGDP-CEPH | China | Southwest Chinese |
| Colombian | 13 | HGDP-CEPH | Colombia | Colombian |
| Mbuti Pygmies | 15 | HGDP-CEPH | Democratic Republic of Congo | Mbuti Pygmy |
| CEPH Utah residents with ancestry from northern and western Europe | 32 | HapMap | Europe | CEU |
| French | 25 | HGDP-CEPH | France | French |

**Table 1 Genotyped populations.** Shown are the population labels as given by the source (HGDP-CEPH and HapMap) for the 56 populations, the number of samples genotyped in each population, as well as the geographical origin and a broader division of the populations (N=37). The table is sorted by geographical origin.

| Population (N = 56) | Sample No. | Source | Geographic Origin | Population (N = 37) |
|---|---|---|---|---|
| French Basque | 24 | HGDP-CEPH | France | French Basque |
| Druze | 43 | HGDP-CEPH | Israel (Carmel) | Druze |
| Palestinian | 49 | HGDP-CEPH | Israel (Central) | Palestinian |
| Bedouin | 47 | HGDP-CEPH | Israel (Negev) | Bedouin |
| Sardinian | 28 | HGDP-CEPH | Italy | Sardinian |
| Tuscan | 8 | HGDP-CEPH | Italy | Italian (mainland) |
| North Italian | 13 | HGDP-CEPH | Italy (Bergamo) | Italian (mainland) |
| Japanese | 29 | HGDP-CEPH | Japan | Japanese |
| Japanese in Tokyo | 31 | HapMap | Japan | JPT |
| Bantu N.E. | 12 | HGDP-CEPH | Kenya | Bantu N.E. |
| Maya | 24 | HGDP-CEPH | Mexico | Maya |
| Pima | 24 | HGDP-CEPH | Mexico | Pima |
| San | 7 | HGDP-CEPH | Namidia | San |
| Papuan | 17 | HGDP-CEPH | New Guinea | Papuan |
| Yoruba | 25 | HGDP-CEPH | Nigeria | Yoruba |
| Yoruba in Ibadan | 30 | HapMap | Nigeria | YRI |
| Orcadian | 15 | HGDP-CEPH | Orkney Islands | Orcadian |
| Balochi | 25 | HGDP-CEPH | Pakistan | Pakistani |
| Brahui | 25 | HGDP-CEPH | Pakistan | Pakistani |
| Burusho | 24 | HGDP-CEPH | Pakistan | Pakistani |
| Hazara | 24 | HGDP-CEPH | Pakistan | Pakistani |
| Kalash | 23 | HGDP-CEPH | Pakistan | Pakistani |
| Makrani | 25 | HGDP-CEPH | Pakistan | Pakistani |
| Pathan | 22 | HGDP-CEPH | Pakistan | Pakistani |
| Sindhi | 24 | HGDP-CEPH | Pakistan | Pakistani |
| Russian | 25 | HGDP-CEPH | Russia | Russian |
| Adygei | 17 | HGDP-CEPH | Russia Caucasus | Adygei |
| Mandenka | 24 | HGDP-CEPH | Senegal | Mandenka |
| Yakut | 25 | HGDP-CEPH | Siberia | Yakut |
| Bantu S.W./E. | 8 | HGDP-CEPH | South Africa | Bantu S.W./E. |

**Table 1 continued**

### 2.1.2 The SNPs

Nonsense-SNPs were identified from their annotation in dbSNP in early 2005 (build 121), resulting in a list of 1,230. In designing the project, I excluded nonsense-SNPs that were known to be incompatible with the typing method used, but ignored prior information about their frequency if it was available. Synonymous-SNPs were chosen to act as controls in this study; although not perfectly neutral they provide an approximation to neutral variants. They were selected to roughly match the sources (submitter) of the nonsense-SNPs in order to match SNPs that might have been called on the basis of poor sequencing or the use of particular populations.

Most SNP data has been obtained through various different discovery processes that often involve the discovery (ascertainment) of the SNPs in a larger sample (typically non-African) which is then followed by genotyping in a larger sample of different populations. This causes ascertainment bias in the data and often the ascertainment schemes have not been recorded systematically and thus it can be difficult to correct for this bias (discussed in Nielsen et al. 2004). However, since the nonsense-SNPs and synonymous-SNPs were chosen in the same way we expect them to be affected by the same ascertainment bias and the effect of such a bias should therefore be reduced at least when the two types of SNPs are compared.

In the end, assays were designed for 805 nonsense-SNPs and 732 synonymous-SNPs, a total of 1,536 SNPs which is the number required for one bundle of an Illumina BeadArray™. All SNPs were genotyped in the HGDP-CEPH and HapMap samples using a multiplexed genotyping assay, the GoldenGate™ assay (Fan et al. 2003). The genotyping is further described in section 2.2.3.

The genotyping results were subjected to sequential quality control filters by the Sanger Genotyping Platform Group (Team 67). Each plate contained three duplicates, and SNPs with more than 33% discrepancies between duplicates were excluded. The Gene Call (GC) score which gives the confidence of the genotype read (intensity) was then estimated. A very low value is not to be trusted. Genotypes without call, individual genotypes with a GC score less than 0.25, assays with a median GC score lower than 0.3 and assays with less than 80% data were also discarded. 406 SNPs (181 nonsense and 225 synonymous) were excluded because they failed these quality control filters. A further 494 SNPs (387 nonsense and 107 synonymous) were excluded by me as they were monomorphic in the combined samples. The SNP had to show variation in at least one individual to be kept. Lastly, I excluded 183 SNPs (68 nonsense and 115 synonymous) that did not pass my manual reassessment of gene annotation incorporating information that became available after the assays were designed. For manual assessment I looked to see whether the nonsense-SNP genes overlapped with Vega pseudogenes (manually annotated and curated by the international vertebrate genome annotation (VEGA) project)(Ashurst et al. 2005) and excluded them if they were found to do so. I used the Tblastx tool to search for the ORF of the sequence surrounding SNPs that had "Stop lost" listed as a consequence and removed those where the ancestral state (chimpanzee) was found to be the PTC and the derived state (human) was found to be a read through of the protein. One SNP, in the *PCDH11XY*, was excluded as the variation observed was found be due to variation between the X and Y chromosomes and not because of polymorphism of the SNP. In addition I excluded those synonymous-SNPs that were found to be intronic. As derived allele information is essential to the analysis, I also excluded SNPs were the ancestral state could not be inferred (1 nonsense- and 8 synonymous SNPs). My final dataset consisted of 452 polymorphic SNPs, 169 nonsense SNPs in 167 genes and 283 synonymous SNPs, and this was used in subsequent analyses. Table 2 lists the number of SNPs kept after each of the above filtering steps.

| SNP Status | Nonsense | Synonymous | Total |
|---|---|---|---|
| Original number of SNPs | 805 | 731 | 1536 |
| Successfully genotyped | 624 | 506 | 1130 |
| Polymorphic in our dataset | 237 | 399 | 636 |
| Passed manual assessment* | 169 | 283 | 452 |

**Table 2 The number of SNPs kept in the dataset after the various filtering stages.** *See description in text.

## 2.2 Laboratory Methods and Protocols

### 2.2.1 Whole Genome Amplification

The samples from the HGDP-CEPH panel had low amounts of DNA and were thus subjected to whole-genome-amplification (WGA) on 11/10/2004 by Yali Xue at the WTSI using the GenomiPhi DNA Amplification Kit by GE Healthcare (formerly Amersham Bioscience) and the protocol was performed according to the manufacturer's guidelines. The resulting stock was then stored at -20°C as there is some indication that WGA DNA degradation in time is temperature dependent.

### 2.2.2 DNA Quantitation

The Illumina GoldenGate™ assay for genotyping required 22μl of ≥50 ng/μl DNA. I performed quantiation on the DNA samples with the Quant-iT™ PicoGreen® dsDNA Assay Kit from Molecular Probes (Invitrogen) and diluted the samples accordingly to ~50ng/μl. The assay was performed according to the manufacturer's guidelines, with the following modification. For the DNA standard curve the Lambda DNA standard was diluted to 5 μg/ml, instead of to 2 μg/ml.

The assay plates were read and fluorescence was measured using a Cytofluor 4000 Fluorescence Plate Reader (MTX Lab Systems, Inc.) with excitation light and filter settings set for excitation at 480 nm and emission at 520 nm. Using the DNA standards, the amount of DNA versus fluorescence intensity was plotted and a line was fitted to the points. This standard curve was then used to determine the amount of DNA from the fluorescence intensity for each sample.

### 2.2.3 Genotyping

Once the DNA samples had been diluted to a concentration of ≥50 ng/μl I submitted them to the Sanger Genotyping Platform Group (Team 67). 1,536 SNPs were genotyped in 1,191 samples (but see later sample and SNP exclusions in sections 2.1.1 and 2.1.2) with the GoldenGate™ assay protocol (Illumina) (Fan et al. 2003) according to the manufacturer's instructions. The GoldenGate™ assay workflow is displayed and described in Figure 9.

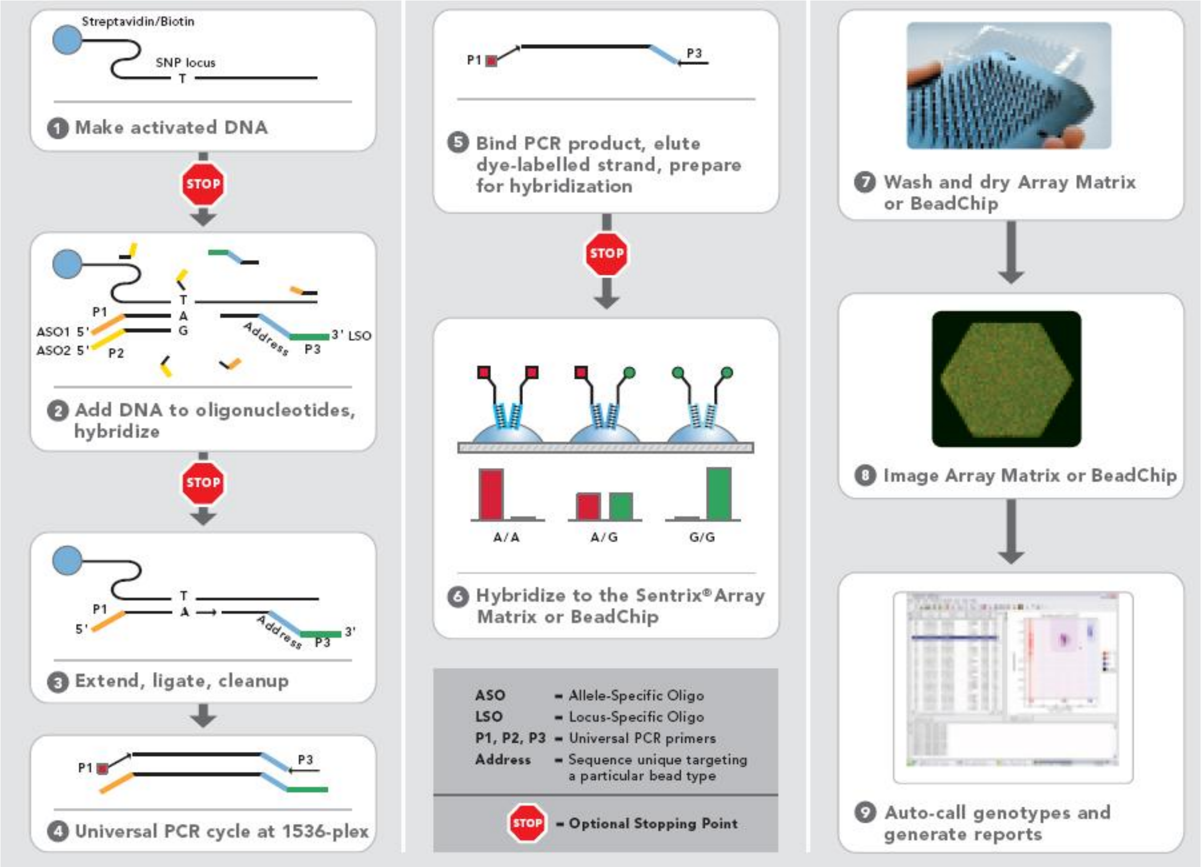
Streptavidin/Biotin
SNP locus
T
1 Make activated DNA
STOP
P1
ASO1 5'
ASO2 5'
P2
T
A
G
Address
P3
3' LSO
2 Add DNA to oligonucleotides, hybridize
STOP
P1
5'
T
A
Address
P3
3'
3 Extend, ligate, cleanup
P1
P3
4 Universal PCR cycle at 1536-plex
P1
P3
5 Bind PCR product, elute dye-labelled strand, prepare for hybridization
STOP
A/A
A/G
G/G
6 Hybridize to the Sentrix® Array Matrix or BeadChip
ASO - Allele-Specific Oligo
LSO - Locus-Specific Oligo
P1, P2, P3 - Universal PCR primers
Address - Sequence unique targeting a particular bead type
STOP - Optional Stopping Point
7 Wash and dry Array Matrix or BeadChip
8 Image Array Matrix or BeadChip
9 Auto-call genotypes and generate reports

**Figure 9 GoldenGate™ Assay Overview. Step 1:** The DNA sample is activated for binding to paramagnetic particles. **Step 2:** Assay oligonucleotides (oligos), hybridization buffer, and paramagnetic particles are then combined with the activated DNA. Three oligos are designed for each SNP locus. Two oligos are specific to each allele of the SNP site (Allele-Specific Oligos, ASOs). A third oligo that hybridizes several bases downstream from the SNP site is the Locus-Specific Oligo (LSO). All three oligo sequences contain regions of genomic complementarity and universal PCR primer sites; the LSO also contains a unique address sequence that targets a particular bead type. The hybridization is followed by several wash steps. **Step 3:** Extension of the appropriate ASO and ligation of the extended product to the LSO joins information about the genotype present at the SNP site to address the sequence on the LSO. **Step 4:** These joined, full-length products thus provide a template for PCR using universal PCR primers P1, P2 and P3. **Step 5:** Universal PCR primers P1 and P2 are Cy3- and Cy5-labeled. **Step 6:** After downstream processing, the single-stranded, dye-labeled DNAs are hybridized to their compliment bead type through their unique address sequences. **Step 7:** Hybridization of the GoldenGate assay products onto the BeadChip allows for the separation of the assay products in solution, onto a solid surface for individual SNP genotype readout. **Step 8:** After hybridization, the BeadArray reader is used to analyze the florescent signal on the BeadChip. **Step 9:** GeneCall software is then used for automated genotype clustering and calling. Figure and assay description were obtained from http://www.illumina.com/

The Illumina primer sequences are given in Appendix B.1. (on accompanying CD). The genotypes were inferred from genotype clusters in GeneCall (from Illumina) and the quality control filters have already been described in section 2.1.2.

#### *2.2.3.1 Problems with Genotype Clusters*

At the WTSI it is customary for the genotype clustering and quality control for large scale surveys to be handled by Team 67, as was the case here. However, at a later stage in the analyses I detected a number of cases causing a deviation from Hardy-Weinberg Equilibrium (HWE) and when I looked back at the raw data I noticed some odd genotype calls. This problem was subsequently resolved and will now be explained with the example of one SNP (rs2233919). The alleles observed at this SNP are A/G, and our samples showed the following numbers of genotypes – 29 AA, 5 AG and 1105 GG – which deviates significantly from HWE (chi-square, P<0.0001). I then looked at the genotype clusters as they appear in GeneCall (Figure 10). At this point it should be noted that plates containing the samples were submitted in batches to Team 67 for genotyping, starting with plate 1 (containing only HapMap samples), then plates 2-5 (containing only HGDP-CEPH samples) were submitted and finally plates 6-13 (containing both HapMap and HGDP-CEPH samples). In Figure 10 each dot represents a sample and the genotype clusters are revealed with different colours, where the pink area designates AA homozygotes, the purple area the heterozygotes (AG) and the blue represents clusters of GG homozygotes. I will not go into the details of the clustering method performed, but note that the clusters observed in Figure 10A and B returned the expected genotypes in our dataset, e.g. in Figure 10A you see 4 dots in the purple area and these corresponded to the 4 (out of 5) heterozygotes observed for this SNP.

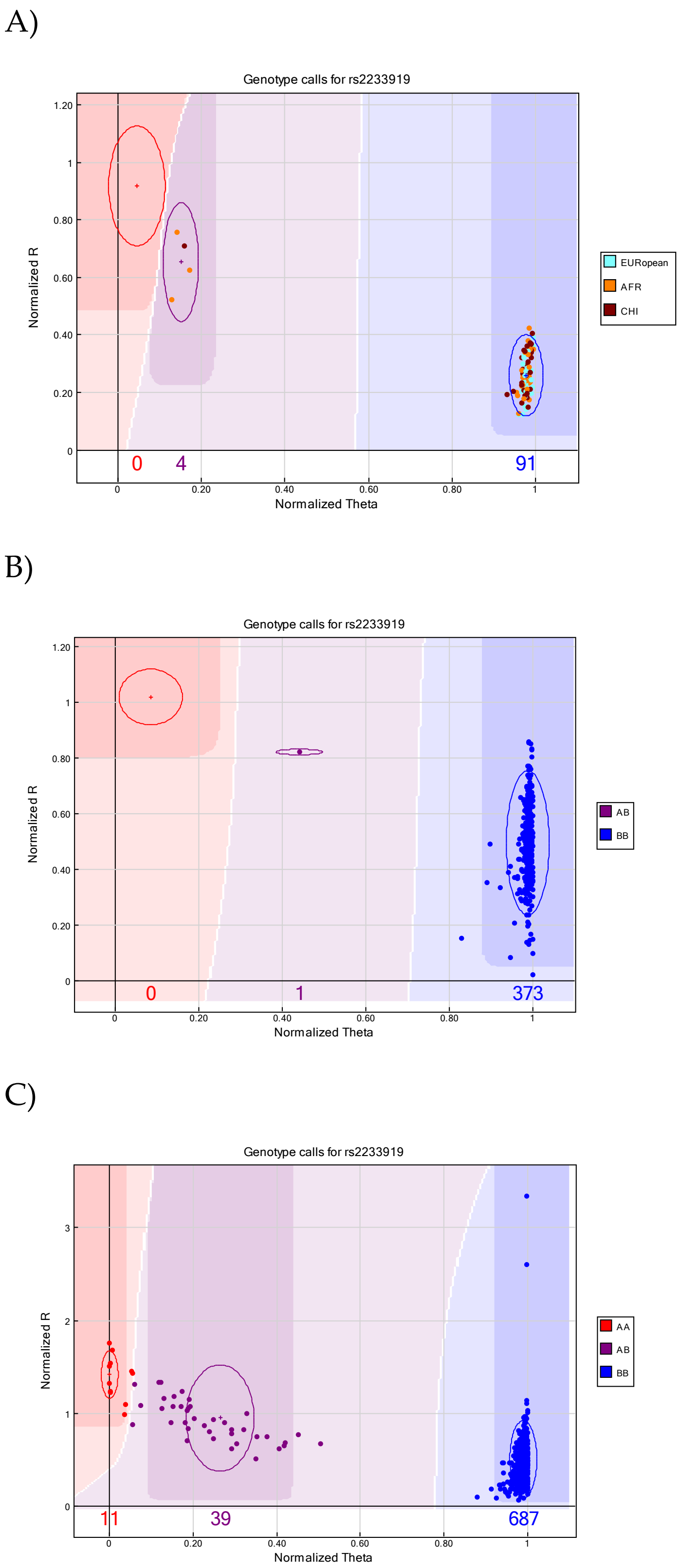

**Figure 10 Genotype clusters for SNP rs2233919 as displayed in GeneCall. A)** Plate 1 (containing only HapMap samples) **B)** Plates 2-5 (containing only HGDP-CEPH samples) **C)** Plates 6-13 (which contained samples from both HGDP-CEPH and HapMap). Each cluster has a plus sign to indicate the mean of the data.

However, the clustering in Figure 10C looked odd, as I only had one more heterozygote reported for this SNP although the purple cluster has a large number of dots filling the purple area. When I extracted the sample names for these dots, one was indeed the expected heterozygote, but the rest were reported as AA homozygotes which should then be represented in the pink area. After various discussions with several members of Team 67 we came to the conclusion that the problem came from analysing the clusters of both HapMap (genomic) and HGDP-CEPH (WGA) samples together and that genomic DNA and WGA DNA should not be analysed together because of different properties. As a consequence, the clustering and subsequent quality control was redone for the whole data set, by analysing the HapMap and HGDP-CEPH samples separately. Unfortunately, the SNP given as an example above was consequently excluded by the quality control filters and so I am unable to represent its new genotype calls here.

#### *2.2.3.2 Additional Quality Control*

When we got the new genotype results back I performed additional quality controls of my own to investigate the genotype calls further. I checked for deviation from HWE for each SNP in the individual populations and found none that deviated from HWE. I also decided to compare my genotyping results to the publicly available genotypes of the HapMap. I used the SNP IDs (rs numbers) of my 452 SNPs and extracted their genotypes for the four HapMap populations (CEU, YRI, CHB and JPT) using the HapMart tool from the HapMap website at http://hapmart.hapmap.org/BioMart/martview and then compared those genotype results to the genotypes of my typed HapMap samples. 77% of my SNPs were included in HapMap Phase II, and I only found inconsistencies for 0.692% of the genotype comparisons (i.e. ~seven inconsistencies per 1000 genotypes SNPs), which is similar to the reproducibility of the HapMap results and other comparisons with HapMap data carried out in our team, and therefore acceptable. Thus, I conclude that the quality control filters were satisfactory and that the genotype calls are to be trusted.

In the end, 452 SNPs were successfully genotyped in 1,151 samples and the whole dataset is available as a tab delimited text file on the accompanying CD (Appendix D).

### 2.2.4 Resequencing

In addition to genotyping 1,536 SNPs, we decided to follow up on two nonsense-SNPs, rs1343879 in *MAGEE2* and rs16982743 in *SIGLEC12*, which were observed as outliers in the nonsense-SNP data set, by resequencing the genes. We also followed up on rs497116 in *CASP12*, but as the re-sequencing of *CASP12* was not performed for this project, the methods used are described elsewhere (Xue et al. 2006).

All primers were ordered from Sigma-Genosys and their sequence is given in Appendix B. The machine used for all PCRs was Alpha™ Unit Bloc Assembly for DNA Engine System, ALS1296, BIO-RAD.

#### *2.2.4.1 Long-Range Polymerase Chain Reaction*

The regions we chose to analyse were around 13 kb in length for each gene with the nonsense-SNP in the middle. Primers were designed for human and chimpanzee with Primer3 (Rozen and Skaletsky 2000) and a custom Perl script, pcr_overlap.pl (see description in section 2.3.3.1), and were selected to amplify two long polymerase chain reaction (PCR) fragments, ~6.5 kb, for each gene. The sequences of the long-range PCR primers for *MAGEE2* and *SIGLEC12* are given in Appendix B.2.

The Platinum® Taq DNA polymerase High Fidelity (Invitrogen) was used for all long PCRs. A long PCR reaction mastermix sufficient for the number of reactions to be carried out was prepared and the recipe for one reaction is given in Table 3.

| Reagent | Volume (µl) x1 |
|---|---|
| $ddH_2O$ | 8.96 |
| 10X High Fidelity PCR Buffer | 1.50 |
| $MgSO_4$ (50 mM) | 0.60 |
| dNTPs (25mM each) | 0.12 |
| Forward Primers (10 µM) | 0.60 |
| Reverse Primers (10 µM) | 0.60 |
| Platinum Taq High Fidelity (5 U) | 0.12 |
| **Total volume added to plate** | **12.50** |
| DNA template (50 ng/µl) | 2.50 |
| **Total volume** | **15.00** |

**Table 3 Recipe for amplification of long PCR products.**

The following PCR cycle conditions were used for the long PCR reactions:

94°C for 2min

94°C for 30sec
68°C for 30sec (decrease 0.5C/cycle) } 15 cycles
68°C for 6min

94°C for 30sec
58°C for 30sec } 20cycles
68°C for 6min

68°C 7min
4°C forever

### *2.2.4.2 Nested PCR*

In order to get good quality sequence traces, it is better to re-amplify segments of the long PCR product with overlaps rather than sequence the long PCR product directly. Therefore, a set of nested primers was designed using a perl script, pcroverlap.pl (see description in section 2.3.3.1). The primers were conditioned to amplify nested PCR products of 500x(1±15%) bp length overlapping by 240x(1±30%) bp. The sequences of the nested primers for *MAGEE2* and *SIGLEC12* are listed in Appendix B.3.

The Platinum® Taq DNA Polymerase (Invitrogen) was used for the nested PCRs. A nested PCR reaction mastermix sufficient for the number of reactions to be carried out was prepared and the recipe for one reaction is given in Table 4.

| **Reagent** | **Volume (µl) x1** |
|---|---|
| $ddH_2O$ | 9.65 |
| Platinum Buffer 10x | 1.50 |
| $MgCl_2$ (50 mM) | 0.48 |
| dNTPs (25mM each) | 0.12 |
| Forward primers (100uM) | 0.10 |
| F&R primers (100uM) | 0.10 |
| Platinum Taq (5 U) | 0.05 |
| **Total volume added to plate** | **12.00** |
| 400x diluted long PCR products | 3.00 |
| **Total volume** | **15.0** |

**Table 4 Recipe for amplification of nested PCR products.**

The following PCR cycle conditions were used for the nested PCR reactions:

94°C for 15min

94°C for 45sec
61°C for 45sec } 15 cycles
72°C for 45sec

72°C for 7min
4°C forever

### *2.2.4.3 Electrophoresis*

Products were analysed by electrophoresis on a 1.5% agarose gel containing ethidium bromide to check that a band of the expected size was present at an adequate concentration. ~20% of each plate was checked.

### *2.2.4.4 PCR-Product Purification*

The PCR-products were purified before they were sent off for re-sequencing. A mastermix of Shrimp Alkaline Phosphatase (USB) and Exonuclease I (USB) sufficient for the number of reactions to be cleaned was prepared and the recipe for one reaction is given in Table 5.

| Reagent | Volume (µl) x1 |
|---|---|
| $ddH_2O$ | 1.380 |
| ExoSAP buffer* | 0.670 |
| Exonuclase I (20U/ul) | 0.033 |
| Shrimp Alkaline Phosphatase (1U/ul) | 0.670 |
| Total volume added to plate | 2.000 |
| PCR product | 8.000 |
| Total volume | 10.00 |

**Table 5 Recipe for one reaction of mastermix required for PCR-product clean-up.** *ExoSAP buffer: 1M Tris (PH8.0) 20ml, 1M $MgCl_2$, $ddH_2O$ 70ml

The following PCR conditions were used for the clean-up of PCR products:

Step 1. Incubate at 37°C for 1 hour

Step 2. 80°C for 20 min

Step 3. 4°C forever.

Products were sequenced on both strands by the Sanger Large Scale Sequencing Pipeline using BigDye Sanger sequencing technology with an 3730 *xl* DNA Analyzer (Applied Biosystems).

## 2.3 Computational Methods

### 2.3.1 Programs and Databases

The complete data set was stored in a Microsoft Access database and was handled and queried using SQL query language implemented therein. Many online databases enabled us to browse, extract data and use various tools supplied. The most commonly used were NCBI, Ensembl, HapMap, UCSC Genome Browser (Kent et al. 2002), The Human Gene Mutation Database, SNP2NMD (Han et al. 2007) and DAVID (Dennis et al. 2003); the usage of some of these is described in other sections.

In order to visualise the geographical distribution of alleles, the geographical coordinates of the sampled individuals were imported into the ESRI ArcGIS 8.2 software (projected with the Gall Stereographic coordinate system with the central meridian set at 145) and pie charts were then produced from allele frequencies.

Basic statistical analyses were performed in Microsoft Excel, Minitab® (release 14) and in R. To test for the significance of the differences in the distribution of values observed for the nonsense-SNPs versus the synonymous-SNPs we applied the Kolmogorov-Smirnov test with an online calculator, http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html. $F_{ST}$ was calculated using the R package HIERFSTAT (Goudet 2005) for autosomal SNPs and in Arlequin (Schneider et al. 2000) for X-chromosomal SNPs. Pairwise difference was calculated using Arlequin (Schneider et al. 2000). Calculations of summary statistics were performed in DnaSP (Rozas et al. 2003). The LRH-test (Sabeti et al. 2002) was performed for the whole SNP dataset (with extra controls) with a java version of Sweep™ and individual SNPs were visualised in Haplotter (Voight et al. 2006). Haplotypes were inferred using PHASE 2.1 (Stephens and Donnelly 2003; Stephens et al. 2001), and median-joining networks (Bandelt et al. 1999) were constructed with Network (http://www.fluxus-engineering.com/sharenet.htm). The use of these programs is further described in the appropriate sections in 2.3.8.

### 2.3.2 Detection of Variants

Potential variable positions in sequence traces were flagged by Mutation Surveyor® v. 2.0. (SoftGenetics, LLC., PA, USA) and checked manually. A Perl script, merge_sts.pl, was then used to check the SNP calling consistency between the overlapping sequence tag sites (STSs) as well as the four duplicates (see description in section 2.3.3.1). Unfortunately, at this stage it was apparent that we could not use the resequenced data from the *SIGLEC12* gene as the sequence traces were unreadable and full of complications. This gene was thus not analysed in the end.

### 2.3.3 Programming Scripts

Several custom computer scripts written in the Perl and Java programming languages were used. All input files were tab delimited. The scripts are found on the

CD accompanying this thesis (Appendix C), with a detailed description of the input files and command lines required.

### *2.3.3.1 Perl Scripts*

**pcroverlap.pl:** This program takes large tracts of sequence data in FASTA file format, and produces PCR products in overlapping segments to span the entire region. It divides up the given sequence, based on the user's criteria for PCR product size (e.g. 500-700 bp) and overlap between adjacent segments (e.g. 200-400 bp), and passes these choices to the PCR primer-selecting program Primer3 (Rozen and Skaletsky 2000). Primer3 then chooses a set of nested primers based on specific selection criteria. The output file is a list of nested primers consisting of the primer sequence, melting temperature (Primer3 calculated), "quality" of nested primers (lower is better; Primer3 calculated), primer positions, primer lengths, PCR product length, and amount of overlap between adjacent fragments. The script was originally obtained from the SeattleSNPs website (http://droog.gs.washington.edu/PCR-Overlap.html) and was modified slightly by Yuan Chen & Cara Woodwark.

**hgdp2sweep.pl:** This program takes a genotype file as input and gives you as output the .snp and .many input files needed to run Sweep™. The PHASE program (Stephens and Donnelly 2003; Stephens et al. 2001) needs to be installed as this script will take the genotype input file, run PHASE to infer the haplotypes, and then use the phased data to create the Sweep input file. The input file should be space delimited and contain the following information: SNP id, chromosome, position and genotypes for all samples. This script was originally created by Yuan Chen and modified by myself.

**create_fstat_input.pl:** This program takes a tab delimited text file with the following information: SNP name, SNP number, sample name, population number, Genotype Code (i.e. 11 = homozygote for first allele, 22 = homozygote for second allele and 12 = heterozygote) and converts it into the file input required by HIERFSTAT (Goudet 2005). This script was created by Jim Stalker.

**merge_sts.pl:** This script was used to check the SNP calling consistency between the overlapping STSs as well as the four duplicates. When the callings were consistent, the script joined the different segments together to reconstruct the whole resequenced region. It then created a table with the variable positions listed for each sample (a SNP table) This script was created by Ni Huang.

**snptab2phase.pl**: This script converts the SNP table produced by merge_sts.pl into the PHASE input file format population by population. Additionally, it requires a file with sample id for each population. This script was created by Ni Huang.

**phase2fasta.pl**: This script converts the PHASE output files from different populations into FASTA format and converts them into a format that can be read into the DNaSP program for the neutralisty tests. A file containing all the PHASE output file names is needed. This script was created by Cara Woodwark.

**phase2network.pl**: This script converts the PHASE output files from different populations into .rdf format and converts them into a format required for the Network program in order to create median-joining networks. A file with the all PHASE output file name list is needed. This script was created by Ni Huang.

#### *2.3.3.2 Java Scripts*

**InputFileTransformer.java**: This program will convert a crosstab table created in Access with homozygote and heterozygote codes (00, 11 and 01) into the format required in Arlequin (Schneider et al. 2000) to calculate the number of pairwise differences. This script was created by Bjarki Holm.

**DelimitedFileTransformer.java**: This program was designed to convert the HapMart output from HapMap so that it would correspond to the format of our genotyping results in order to make the comparison between the two easier. This script was created by Bjarki Holm.

**SweepFileConversion.java:** This program collects the HapMap phased data from a URL for a region of choice and outputs the .snp and .many files required by Sweep™ for each SNP and each HapMap population. This script was created by Bjarki Holm.

### 2.3.4 Inferring the Ancestral State

In order to calculate the derived allele frequency (DAF), we needed to know the ancestral state of each allele. The chimpanzee (*Pan troglodytes*) base was primarily used as the ancestral state, but when the chimp sequence was not available or differed from both the observed human alleles, we accepted sequence from other primates (*Macaca mulatta* or *Lagothrix lagotricha*). The derived allele was then defined as the other observed human allele.

We used the Table Browser on the UCSC Genome Browser website (http://genome.ucsc.edu/cgi-bin/hgTables) and retrieved the ancestral allele for ~98% (445 SNPs) from the "snp126OrthoPanTro2RheMac2" table. We then looked manually for the ancestral state of the missing 2% (8 SNPs). We obtained FASTA sequences surrounding the SNPs and used the NCBI Blastn algorithm to find the best hit with a primate reference sequence and thereby identified the ancestral allele for 6 of these at the appropriate position.

The derived allele frequency was obtained by direct allele counting and a Kolmogorov-Smirnov test was used to evaluate the difference between the distributions of nonsense- and synonymous-SNPs.

### 2.3.5 Predicted Truncations and Calculations of NMD

In order to visualize the predicted effect of these nonsense-SNPs on the gene product, we first estimated the proportion of protein truncation each SNP would cause. 112 genes bearing nonsense-SNPs were found to code for a single transcript. The remaining 57 nonsense-SNPs were found in genes undergoing alternative splicing and were reported in more than one transcript. For such SNPs we used the transcript showing the largest truncation. The truncation was calculated as a percentage of the ancestral sequence ORF length (100-(SNP protein position/protein length*100)).

The nonsense SNP could lead to a truncated protein with an altered function but if it is located more than 50-55 nucleotides upstream of the 3′-most exon-exon junction the transcript will be eliminated by NMD (Maquat 2004). In order to assess whether our nonsense-SNPs were likely to trigger NMD we used the SNP2NMD database (Han et al. 2007) available from http://bioportal.kobic.re.kr/SNP2NMD. This database contains human nonsense-SNPs with an estimate of whether or not NMD is expected to be triggered according to the 50-55 nucleotide rule. 107 (~63%) of our nonsense-SNP were in SNP2NMD and we used the default setting of the "NMD distance" (distance between a SNP and the 3′-most exon-exon junction) to be >50 nucleotides for the NMD pathway to be triggered. As the transcripts used in SNP2NMD were obtained from different sources from our data, we applied the same rule as mentioned above and selected the transcript with the maximum truncation when having to choose from multiple transcripts. For the remaining 62 (~37%) SNPs missing from SNP2NMD we extracted information on the location of the nonsense- SNP with respect to exon-intron boundaries from Ensembl (release 37 and 43) and calculated the prediction for NMD manually.

### 2.3.6 Gene Expression

In collaboration with Barbara Stranger and Manolis Dermitzakis, of Team 16 (Population and Comparative Genomics) at the WTSI, we used their available expression data to test the association between nonsense-SNP genotypes and expression levels. Gene expression quantification and normalization had already been performed by Barbara Stranger *et al* (Stranger and Dermitzakis 2006; Stranger et al. 2007b)

Gene expression data were obtained for approximately 48,000 transcripts, including a subset of 14,456 probes (13,643 unique autosomal genes) that were highly variable among lymphoblastoid cell lines of the 210 unrelated HapMap individuals (Stranger et al. 2007b). Hybridization intensity values were normalized on a log2 scale using a quantile normalization method (Kuhn et al. 2004) across all replicates of a single individual followed by a median normalization method across all 210 individuals. A subset of 14,456 probes (13,643 unique autosomal genes) that were highly variable within and between populations was selected from the 47,294 probes on the array, and were used for the analysis. A detailed description can be found in Stranger *et al* (2007b).

We first attempted to test our set of 169 nonsense-SNPs for association with expression of these variable genes, but found that only 57 of the SNPs mapped within the genes corresponding to the 14,456 probes, and of these, only 19 were polymorphic and genotyped in the HapMap (The International HapMap Consortium 2005). This gave us little power to draw any conclusions and we thus resorted to using all available nonsense-SNPs (dbSNP126) which gave us a starting dataset of 1,624 SNPs instead of our original 169. In the end, 588 of these had been typed in HapMap and 105 of those could be mapped within genes corresponding to the expression probes exhibiting variable gene expression.

We tested the nonsense-SNP genotype for association with expression levels of the gene by using an additive linear regression model (Stranger et al. 2005; Stranger et al. 2007a; Stranger et al. 2007b) applied to each population separately. Our association analysis employed: 1) nonsense-SNP genotypes for the unrelated individuals of each HapMap population (MAF<0.05) from the HapMap phase II map for each population (version 21, NCBI Build 35) and 2) normalized log2 quantitative gene expression measurements for the 210 unrelated individuals from the original four HapMap populations (60 CEU 45 CHB, 45 JPT, 60 YRI).

To assess the significance of association between nonsense-SNP genotypes and expression variation of the gene harbouring the nonsense-SNP, we performed 10,000 permutations of each expression phenotype relative to the genotypes (Stranger et al. 2007b). An association to gene expression was considered significant if the nominal p-value from the linear regression test was lower than the 0.01 tail of the distribution of the minimal p-values (among all comparisons for a given gene) from each of the 10,000 permutations of the expression phenotypes. For genes containing more than one nonsense-SNP, the most stringent permuted p-value was retained.

### 2.3.7 Gene Ontology Term Enrichment Analysis

To find out if the set of genes containing nonsense-SNPs have an overrepresentation of a particular molecular function (MF) or biological process (BP), their relevant gene ontology (GO) (Ashburner et al. 2000) terms were identified. We performed the GO term enrichment analysis with the DAVID chart analysis tool in DAVID (Dennis et al. 2003) (http://david.abcc.ncifcrf.gov/summary.jsp, 26/05/08). All available GO terms were used and all human genes (implemented in DAVID) were defined as the background. Ensembl gene IDs were collected for each of the 169 nonsense-SNPs (167 genes) with the BioMart query system (http://www.ensembl.org/biomart/index.html, 26/05/2008) and these were used as input for the enrichment analysis. P-values were calculated by the EASE score which is a modified conservative adjustment of the one-tailed Fisher Exact test (Hosack et al. 2003) and is implemented in DAVID. Terms with values below 0.05 were considered to be enriched. While a multiple correction is often applied for these tests, the authors of DAVID attest that it will be too conservative on the cost of the biological importance (revealed in a personal communication through their website). Thus, while the Bonferroni correction is given with our results, it should not be taken too seriously. Of the total 167 genes analysed, 71 were not included in the output for BP and 88 for MF. For the 71 (BP) and 88 (MF) missing, 26 (BP) and 59 (MF) had GO terms associated with the genes but the terms did not pass the filter of the EASE score (enrichment analysis), while 45 (BP) and 29 (MF) did not have any GO annotation because the functional annotation of the human genome is incomplete.

### 2.3.8 Population Genetic Calculations

#### *2.3.8.1 Population Differentiation Calculations* ($F_{ST}$)

$F_{ST}$ was used as a measurement of population differentiation. $F_{ST}$ values were calculated by conventional F-statistic methods with the HIERFSTAT (Goudet 2005) package for R using the *varcomp* function to calculate the $F_{ST}$ (theta) from Weir and Cockerham (1984). This F-statistic uses the allele frequencies to quantify the proportion of the total variance among the human populations. $F_{ST}$ values were calculated for each SNP across the 37 populations (see division in Table 1). The Kolmogorov-Smirnov test was used to assess whether there was a significant difference between the distributions of nonsense- and synonymous-SNPs. For comparison with empirical data we downloaded the genotypes for the HapMap phase II SNPs and for a set of 650K publicly available SNPs genotyped in the HGDP-CEPH populations and calculated their $F_{ST}$ values to find out if our SNPs were significant outliers (i.e. lying above the 95th or 99th percentiles). The values calculated for the HGDP-CEPH were calculated from the 32 HGDP-CEPH populations as well as for the combination of those 32 populations into five major groups to match the K=5 division in Rosenberg (2002).

Traditionally, the range of $F_{ST}$ is between 0 and 1, where 0 would imply no differentiation between populations and 1 complete differentiation. However, it is possible for the unbiased estimate of $F_{ST}$ to give negative values. When this occurred, we assigned negative values of $F_{ST}$ to zero as suggested by Nei (1987).

#### *2.3.8.2 Heterozygosity*

Nei's measure of heterozygosity (Nei 1987), the probability that any two randomly chosen samples from a population are the same, was calculated for each SNP by:

$$H = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right)$$

Where $n$ is the number of alleles, $k$ is the number of haplotypes and $p_i$ is the frequency of the $i$th haplotype.

#### *2.3.8.3 Pairwise Differences*

To estimate how much human individuals differ with respect to the nonsense-SNPs we calculated the mean number of pairwise differences as implemented in Arlequin (Schneider et al. 2000).

#### *2.3.8.4 Long-Range Haplotype Test*

To gain a better insight into the possible action of natural selection, we applied the REHH test (Sabeti et al. 2002). This test has been implemented in the Sweep™ program which requires phased haplotype data as input and analyses haplotype structure in the genome by determining the frequency and long-range LD for each allele. The method uses LD to measure the association between a single allele at one locus with multiple loci at various distances (see 1.3.1.6). We identified our nonsense-SNPs as the so-called "core" haplotype (SNP) and then increasingly distant SNPs were added to quantify the decay of LD from the core. The assumption is that a positively selected SNP will be found at a high frequency on an unusually long haplotype.

We used the SweepFileConversion.java programme to collect the phased haplotypes from the HapMap populations (CEU, YRI, CHB+JPT) and to convert them into the format required to run Sweep (see section 2.3.3.2). We caution that the CHB+JPT phased haplotypes were later withdrawn from the HapMap as they were under scrutiny and were not available again to use in time for this thesis.

We used the Phase II data (Build 36) which contained 131 out of the 169 nonsense-SNPs. We chose to use Build 36 as it contained a higher number of our SNPs than did Build 35, 131 compared to 106. As the current version of Sweep will only accept coordinates from a Build as high as 35, we used Build 35 coordinates for the Build 36 SNPs when available, and collected the coordinates for the 25 SNPs present in Build 36 but not in 35 manually using the Ensembl Genome Browser archive (Ensembl release 42).

For each of the 131 nonsense-SNPs genotyped in HapMap we chose to use a 100 kb region on each side of the SNP to infer the haplotypes. In addition, we chose 30 ENCODE random regions, which are assumed to be neutral, to act as controls. The coordinates of these were obtained from the UCSC Genome Browser. Each ENCODE region was roughly ~500 kb in length. REHH was calculated with the default setting of a 0.04 marker breakdown from the core SNP.

To evaluate whether or not our nonsense-SNPs found at high frequencies with unusually extended haplotypes were significant, we plotted the SNP frequency against its REHH value for both the nonsense-SNPs and the ENCODE SNPs (used as empirical controls), calculated the 95th and 99th percentiles, and considered a nonsense-SNP significant if it was above those.

### 2.3.9 Neutrality Tests

Two genes (*MAGEE2* and *SIGLEC12*) were re-sequenced, but only the *MAGEE2* sequence was of good enough quality to be further analysed (see explanation in section 2.3.2). We used DnaSP (Rozas et al. 2003) to calculate traditional neutrality tests (discussed in section 1.3.1.2). These included Tajima's *D* (Tajima 1989c), Fu and Li's *D*, *D**, *F* and *F** (Fu and Li 1993), Fu's $F_S$ (Fu 1997) and Fay and Wu's *H* (Fay and Wu 2000). Null distributions were obtained by the custom modified ms program (Hudson 2002) incorporating the best-fit demographic model (Schaffner et al. 2005).

### 2.3.10 Median-Joining Network

Haplotypes for the resequenced data were inferred using PHASE 2.1 (Stephens and Donnelly 2003; Stephens et al. 2001). Median-joining networks (Bandelt et al. 1999) were constructed from the inferred haplotypes with Network (http://www.fluxus-engineering.com/sharenet.htm).