5 Discussion and Future Directions

The main goal of our research was to evaluate the overall evolutionary forces acting on nonsense-SNPs in the human genome and thus provide some insights into the importance of variation in gene number for human evolution. To this end, we embarked on a genome-wide study of loss events and typed a large number of nonsense-SNPs in a set of geographically diverse populations. From this dataset, we hoped to identify candidates for positive selection (to be followed up in more detail by resequencing), and thus provide an evaluation of the relevance of the less-is-more theory for human evolution. We believe we have now accomplished these goals, and in the next few sections will discuss our main conclusions.

5.1 Prevalence and Consequences of Nonsense-SNPs

In chapter 3 we reported the prevalence of nonsense-SNPs in the human genome and their consequences for the protein product. We found that nonsense-SNPs are more prevalent in the human genome than some studies have suggested (Sawyer et al. 2003) and that they are not simply a class of deleterious disease-causing mutations slipping through the system at low frequencies in a heterozygous state. The prevalence of nonsense-SNPs was such that the individuals sampled were found to differ by 24 genes, on average, because of nonsense-SNPs. This will almost certainly be an underestimate and will increase with the findings of large-scale sequencing projects, such as personal genome sequencing projects (Levy et al. 2007; Wheeler et al. 2008) and the more systematic "1,000 genomes" project (http://www.1000genomes.org/). Nevertheless, this is still a higher difference than was reported initially for the more commonly occurring CNVs, where individuals were found to differ by only 11 genes (Sebat et al. 2004).

These nonsense SNPs are made up of a mixture of potentially deleterious variants present only in a heterozygous state (and thus maintained at low frequency

in the population; 70 SNPs, 41%), and near-neutral or advantageous variants that are found in a homozygous state (and can rise to high frequency). For 99 (59%) nonsense-SNPs, at least one stop homozygous sample was reported, showing that both copies of the nonsense-SNP containing genes can be truncated in our sample donors. However, as we have little phenotypic information on the sample donors, we cannot predict the consequences these nonsense-SNPs are having on their health, except to say that they are compatible with survival to adulthood in a state where the individual is competent to provide informed consent for the use of their sample and is sufficiently interested in helping scientists to provide the sample. Direct insights into their phenotypic consequences could potentially be obtained by detailed studies of individuals of known genotype, by the inclusion of these SNPs in association surveys, or from model organisms.

We attempted to predict some consequences of the nonsense-SNPs *in silico* by using bioinformatic information on the SNP position to predict the likely extent of truncation and the triggering, or not, of NMD. These predictions revealed that many of the nonsense-SNPs analysed will cause a large segment of the protein to be truncated (in at least one transcript), and that 55% can trigger NMD. The consequences could thus often be radical: they could lead to the complete loss of the gene product or possibly to an altered function. We therefore attempted to test the consequences by using available gene expression data. This analysis did not leave us with many significant results to make a generalisation about, but most did meet the prediction of reduced expression in cases where NMD was triggered.

With such a large set of genes to consider (167), it was difficult to study each individual gene in full detail. When we came across an interesting outlier in the genome-wide data, we generally did a literature search for information on that particular gene, but many of these genes had not been studied in enough detail to reveal the functional implications of their loss. A detailed experimental approach would be needed to evaluate the true effects of the nonsense-SNPs and then it might be possible to find out the biological effect of these losses. Future work might thus

include some functional studies. Some of the genes were found in the HGM database and have thus already been implicated in disease. This was not unexpected as nonsense-SNPs are known to be the cause of many diseases. However, in the context of human evolution and our interest in the gene loss theory, we were more interested in those that could have been advantageous for our species.

In an attempt to identify the types of genes where nonsense-SNPs can be found, we performed an analysis of the gene ontology. We found that the genes we studied were mainly overrepresented in terms related to olfactory reception and the nervous system, the latter being rather surprising. Annotation of the human genome is incomplete and so not all of our genes were represented in this analysis. We might thus be missing important categories for the less-annotated genes in our dataset. There is a suggestion that the genes containing nonsense-SNPs are more likely to have paralogs that help back up their function should one be lost. However, a study of the representation of genes in segmental duplications (and thus all with paralogs) reported an overrepresentation of genes associated with immunity and defence, membrane surface interactions, drug detoxification and growth/development (Bailey et al. 2002), none of which were found to be overrepresented in our "lost" genes.

5.2 Selective Forces

We wanted to infer the evolutionary forces acting on nonsense-SNPs, i.e. whether they were evolutionarily advantageous, disadvantageous or neutral. Our measures of derived allele frequencies, population differentiation and long-range haplotypes led us to believe that the SNPs are in the main largely neutral or slightly deleterious We did, however, find interesting outliers, some of which we followed up by resequencing. We reported results for *CASP12* and *MAGEE2*. We intended to follow up *SIGLEC12* as well, but the sequence traces were not good enough to use and we may attempt to redo this in the future. Additionally, *SEMA4C* came up as of potential interest to us because of its specificity for the Americas and may also be followed up by genotyping in a larger samples and resequencing. The resequencing

of the two genes enabled us to use neutrality tests and median-joining networks to investigate the region in greater detail in order to infer the selective forces. We found that *CASP12* gave clear evidence for positive selection in most populations by all neutrality tests used, while *MAGEE2* had an interesting phylogenetic structure and may have been subjected to selection more recently (perhaps starting 20,000 – 40,000 years ago) in Asian populations as suggested by its geographical distribution and the value of Fay and Wu's *H* in the CHB sample. While the *CASP12* results are understandable in view of its role in sepsis resistance, no functional information was available for *MAGEE2* and it would thus be interesting to perform extensive functional studies of this gene in the future. *MAGEE2* perhaps illustrates the situation that is most likely to emerge from genome-wide surveys of this kind: despite reasonable evidence for selection, no clues about the nature of the selective force.

To conclude, we do find some nonsense-SNPs that may be taken to support Olson's less-is-more hypothesis, and thus that gene loss has contributed to human evolution, but do not find evidence that such loss has been a major evolutionary force in human history.

5.3 The Effectiveness of Our Methods

When SNP data are used, one always has to be aware of ascertainment bias in their discovery as allele frequencies and distributions will depend greatly on this. Indeed, many global SNP projects, such as the HapMap, have displayed a deficit of rare and an excess of intermediate frequency SNPs (The International HapMap Consortium 2005). Furthermore, as many of the SNPs used were initially discovered in non-African populations, the HapMap data may be missing out variation within Africa. As we looked at the geographical distribution of our rare stop alleles, we did not find much difference between African and non-African populations. This might be an indication that the expected excess of variation in African populations is not found in our dataset, and thus that African-specific variants are under-represented.

An additional concern related to ascertainment is that tests that depend on allele frequency data, such as population differentiation measures (F_{ST}), should be interpreted with caution. Our genome-wide survey of nonsense-SNPs using genotype data enabled us to pick up signals (in population differentiation and otherwise) that, when followed up by resequencing, were revealed to be of evolutionary interest. So while F_{ST} should not be taken alone as evidence for selection (Xue et al. submitted), it may provide us with useful clues which can then be followed up by more trustworthy methods. Indeed, resequencing data will not be affected by ascertainment bias. Furthermore, as our two SNP classes, nonsense- and synonymous-SNPs, were chosen in the same way, they should also be subject to the same ascertainment bias. Therefore, comparison of the two classes is justifiable as a way to identify nonsense-SNP outliers compared to the assumed neutral synonymous-SNPs, as well as comparing nonsense-SNPs to other nonsense-SNPs to identify those of special interest.

However, while resequencing data are without ascertainmet bias, the neutrality tests are still potentially subject to erroneous conclusions as they rely on population genetic models that make specific (and undoubtedly too simplistic) assumptions about the demography of the populations. In particular, these models often make the assumption that population size is constant and that there is no population structure. Neutrality tests have even been shown to reject neutrality in the absence of selection (reviewed in Nielsen 2005). Indeed most interpretation of genetic diversity is highly sensitive to demographic assumptions. For example, it has been shown that Tajima's *D* (Tajima 1989c) will reject the neutral model in the presence of population growth (Simonsen et al. 1995). Population growth may give a similar effect to a selective sweep. Tests based on patterns of LD may be particularly sensitive, because they rely on assumptions about demography as well as the underlying recombination rates and these can vary greatly between regions (McVean et al. 2004). Thus we are also concerned with distinguishing between the signal given by demographic and selective processes. However, demography will have a similar

affect on the whole genome, whereas selection will have locus-specific effects. Therefore, this problem can be overcome in a number of ways: by modelling demography more realistically (Schaffner et al. 2005) or by the use of empirical comparisons and data from multiple loci as was done with our survey of nonsense-SNPs.

In the end we find that the combination of tests based on genotype (multiple loci) and resequencing (free of ascertainment bias) data currently provides the best way to distinguish a real selective signal from an apparent one based on ascertainment or demography. If accompanied by biological insights into the nature of the phenotype that might be under selection, a convincing case for selection can be made.

5.4 The importance of Knowing one's Nonsense-SNPs

The extent of gene content variation in the healthy population is starting to be appreciated, and it can be seen to be made up of copy number variation (Jakobsson et al. 2008; Redon et al. 2006; Sebat et al. 2004), nonsense-SNPs, other truncating variants such as indels and splice site alterations, and polymorphisms in regulatory elements that ablate expression (Stranger et al. 2007b). Together, these lead to substantial differences in the number of active genes carried by different healthy humans. The number of genes affected in this way is still largely unknown, but this study provides a minimum estimate of the variation due to nonsense-SNPs, and suggests that the total must be a substantial proportion of the entire gene content.

We see that the set of nonsense-SNPs documented in this thesis can be particularly significant for three areas of genetics and medicine. First, sequencing is starting to be used to survey genes or genomes for disease-associated variants, and to inform genetic risk counselling, including whole-genome resequencing for personalized medicine. Nonsense-SNPs discovered in such studies would merit particular attention, but at least the 99 found here in the homozygous state are not associated with mendelian disorders, have no overt influence on the phenotype and are compatible with healthy life. Second, there are nevertheless some situations

where generally-neutral differences in gene content have medical consequences: allogeneic hematopoietic stem cell transplantation, where a donor lacking a gene may mount an immune reaction against the tissues of a recipient with that gene, leading to graft-versus-host disease (Murata et al. 2003). Donors and recipients should be screened for potential gene differences, including those resulting from these nonsense-SNPs. Third, a general treatment for a wide variety of genetic disorders caused by nonsense-SNPs has been proposed: administration of the drug PTC124 which promotes read-through of premature but not normal termination codons (Welch et al. 2007). Such treatment would, if effective, also promote the expression of endogenous non-target genes carrying nonsense-SNPs, and the consequences of this should be evaluated. We need to understand the full extent of human genetic variation in order to reap the full benefits of present and future medicine.