Geographical structure and differential natural selection amongst North European populations

SUPPLEMENTARY FIGURES

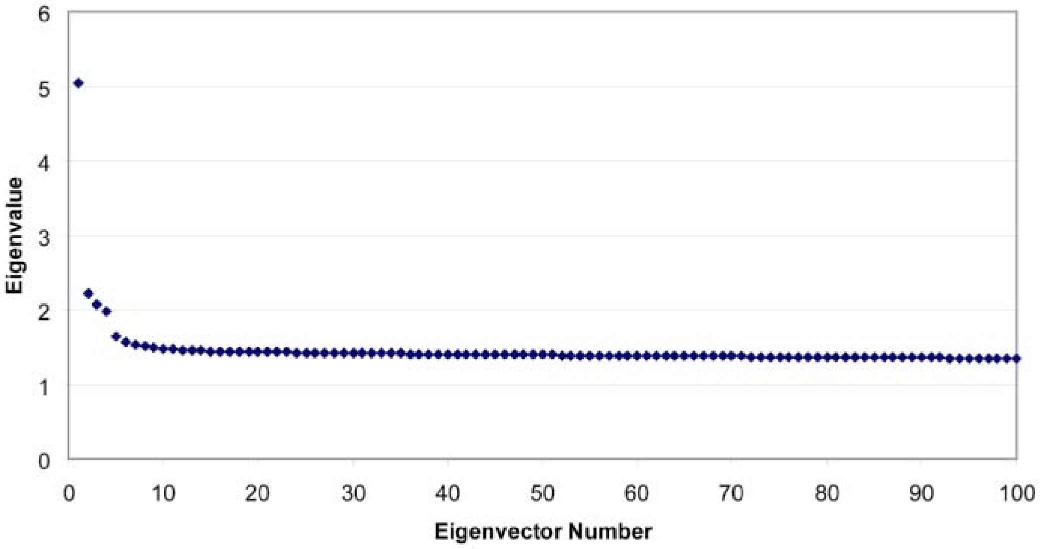

Figure S1- Top 100 Eigenvectors and associated Eigenvalues.

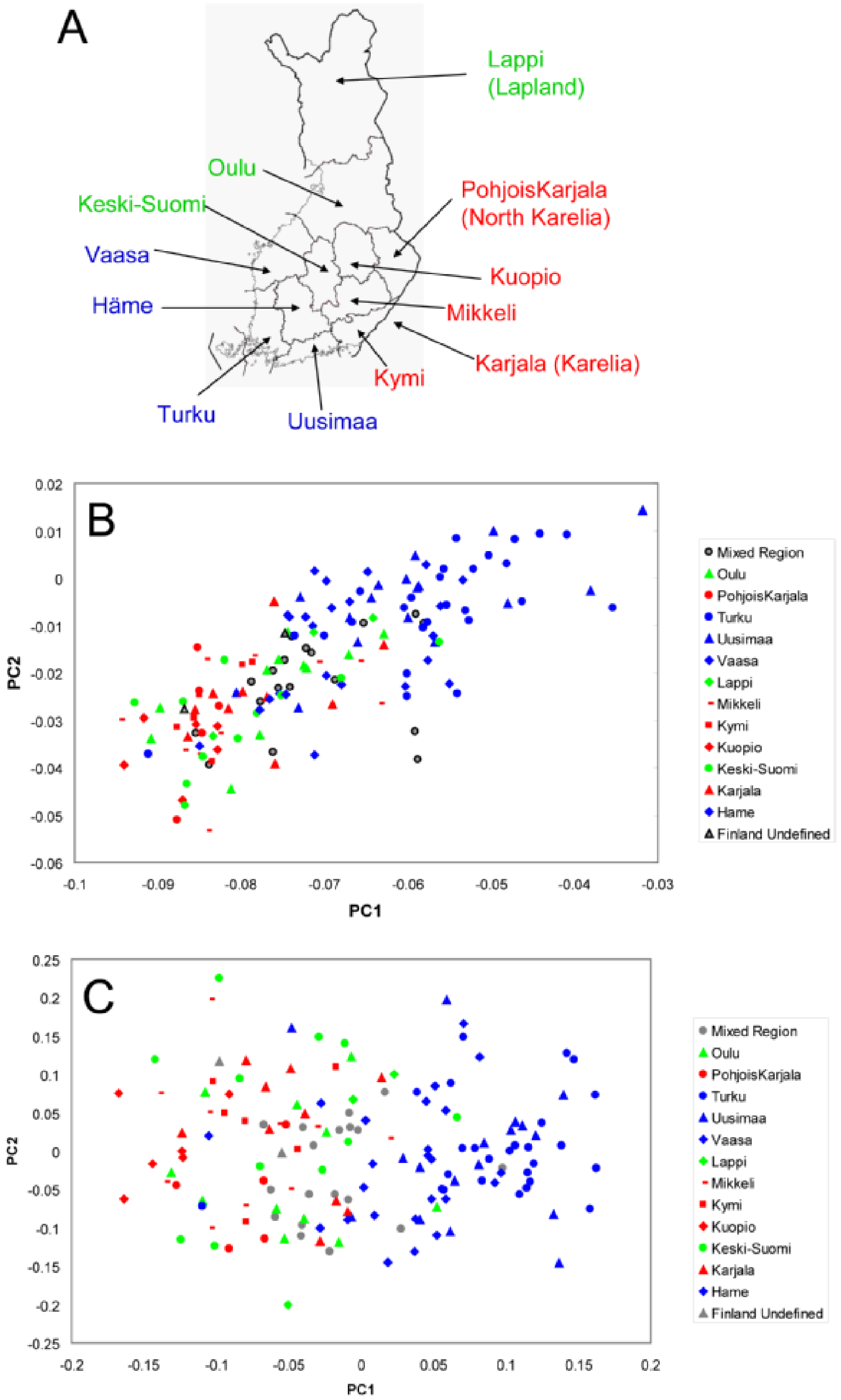

Figure S2 – Finnish Population Sub-structure

(**A**) Historical provinces of Finland. (**B**) PC1 versus PC2 for 149 Finnish individuals. PCA was conducted using the entire Northern European dataset but only the Finnish sample is shown. **(C)** PC1 versus PC2 derived from PCA using only the 149 Finnish individuals. Individuals are divided by province of origin, with further colour coding to indicate region: Blue= West; Red= East; Green=Centre/North (see Map). Grey indicates mixed/unknown provincial ancestry. Province of origin is defined as birth place for an older cohort of individuals (born 1926-1936) and parental province of origin for a younger cohort (born 1975-1979). The latter definition was chosen for younger individuals since it filters out the effects of internal migrations from the late 1950s to 1970s. Younger individuals whose mother and father come from different provinces are designed "Mixed Ancestry". Samples with no location information are labelled "Finland Undefined". Note: Karjala (Karelia) is largely situated in present-day Russia.

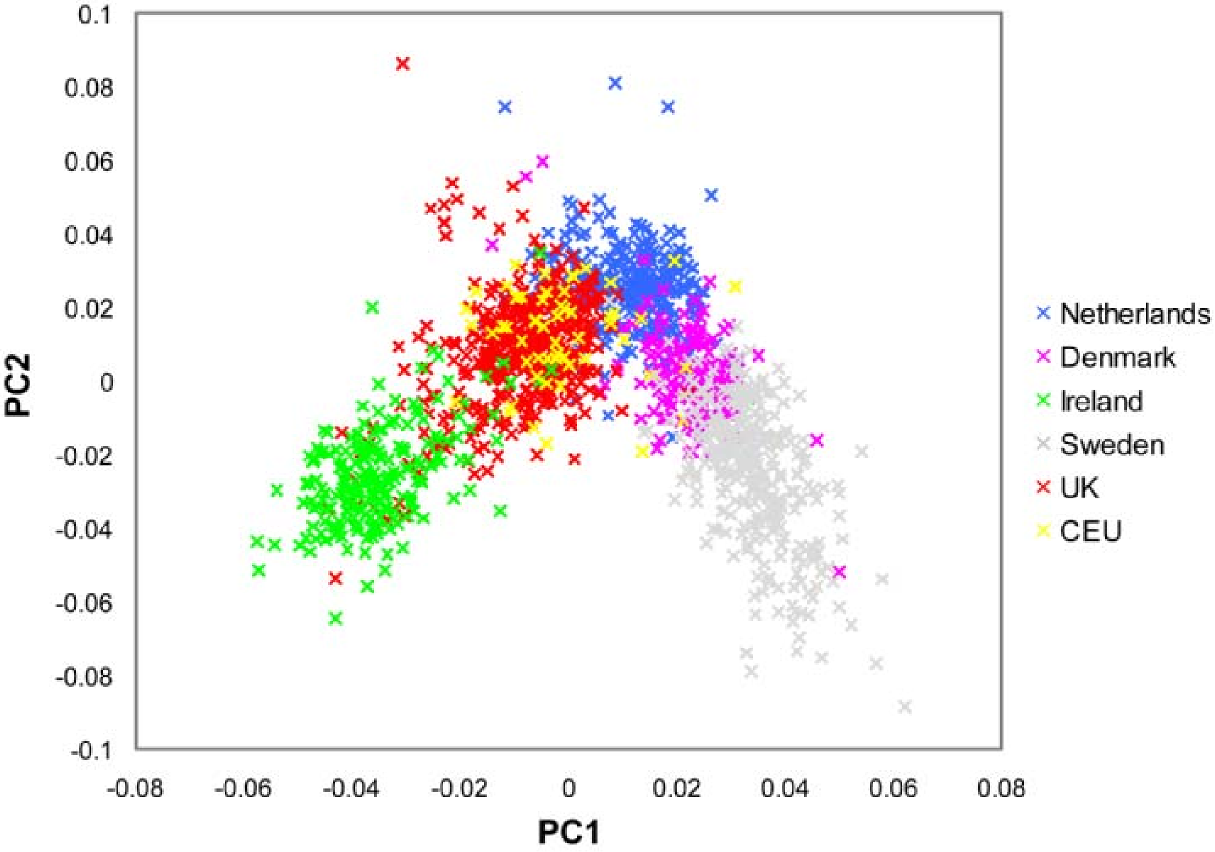

Figure S3 – PCA of Northern European Population Structure excluding Finland. PC1 versus PC2 in 1451 individuals from six Northern European populations. Finnish individuals, who largely create PC1 in the full dataset, were excluded from this PCA. 3036 SNPs from two large genomic regions on chromosome 6 (24MB to 36MB) and chromosome 8 (6MB to 12MB), which form PC3 and PC4 respectively in the original dataset, were also excluded leaving a total of 293517 autosomal SNPs. These genomic regions and the Finnish samples were removed in order to investigate residual population structure in more detail. PC1 here is similar to PC2 derived from the full data while PC2 appears to be cognate with PC5 in the complete dataset.

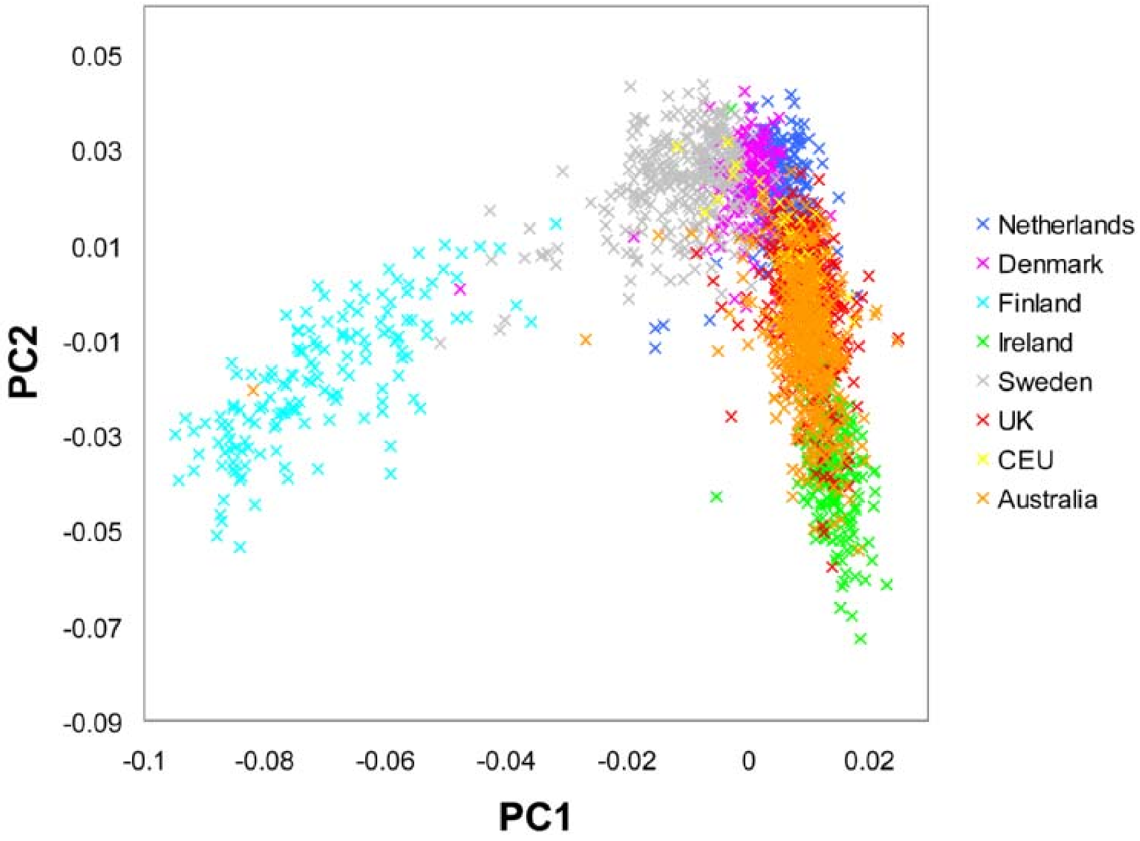

Figure S4 – PCA of Northern European Population Structure including Australia. PC1 versus PC2 from 2051 individuals genotyped for 296553 autosomal SNPs.

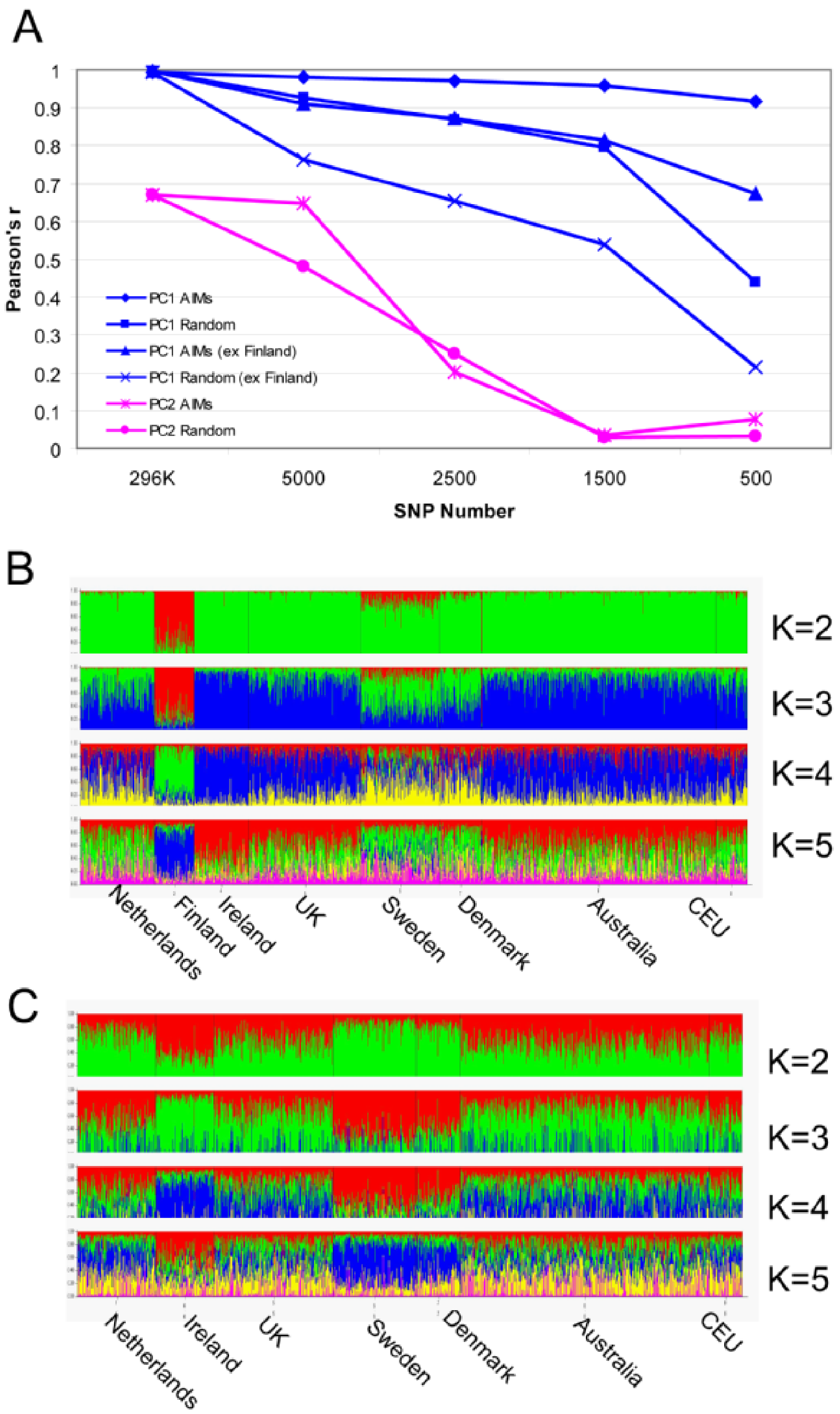

Figure S5- Performance of AIM sets using PCA and STRUCTURE.

(**A**) Correlation (Pearson's r) between individual PC1 or PC2 values obtained in the half-sample test set (n=1281), using 296K SNPs, versus those obtained from smaller marker (5000, 2500, 1500 or 500 SNPs) sets in the same test dataset. The correlation of individual scores in the half-sample test sets versus those from the full-sample sets using the complete SNP set are also shown (296K ‘SNP Number’ categories). Sets were selected by F_{ST} rank (AIMs) or randomly in a second discovery half-sample dataset. The relatively extreme values of Finnish individuals in PC1 may obscure the performance of the AIM sets for the rest of the populations along PC1 and so analysis for PC1 was also repeated excluding Finland. **(B)** STRUCTURE analysis of the test half-sample (n= 1281 which includes all of the Australian and CEU-HapMap individuals) using the set of 5000 top ranked F_{ST} SNPs (AIMs) for K=2 to K=5. Results for 2500, 1500 and 500 AIM SNP sets can be seen in **Figure S6**. **(C)** STRUCTURE analysis for test half sample using 5000 AIMs but excluding Finland. Finnish individuals were also removed from the discovery population set used to select the top 5000 ranked SNPs by F_{ST}. In panels **B** and **C**, each individual is represented by a vertical line divided into K color segments corresponding to the fraction of their genome estimated to be derived from each of the pre-specified K populations or clusters

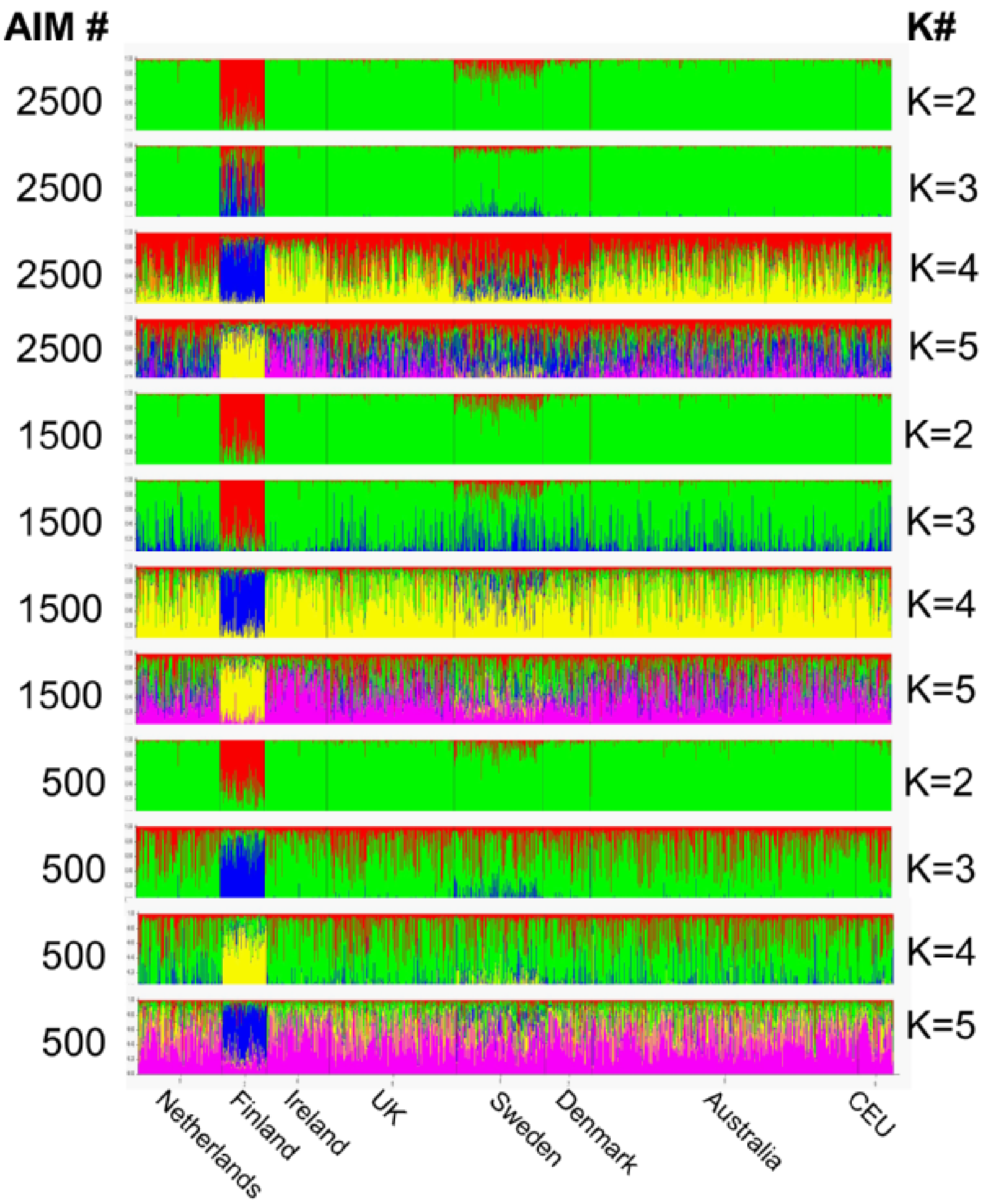

Figure S6 - STRUCTURE Analysis of Northern Europeans.

Results of STRUCTURE runs based on 2500, 1500 or 500 top ranked F_{ST} SNPs (AIMs) applied to the test half-sample set (n= 1281 which includes all of the Australian and CEU-HapMap samples) for *K*=2 to *K*=5. Each individual is represented by a vertical line divided into *K* colour segments corresponding to the fraction of their genome estimated to be derived from each of the pre-specified *K* populations or clusters.

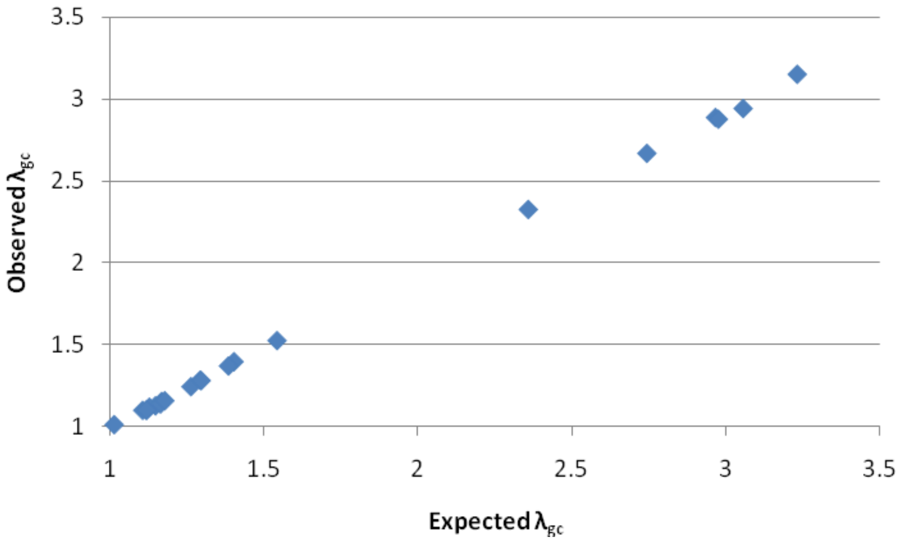

Figure S7 - Correlation of Observed and Expected λ_{gc} Values. Observed values were calculated for each pairwise population comparison (excluding the HapMap CEU) based on the median genotypic χ^2 association statistic across the full set of SNPs. Calculations were conducted using a truncated sample size of n=149 for all populations, selected at random from larger samples, to match the smallest actual (Finland) population size. Expected λ_{gc} were determined according to $E(\lambda_{gc}) \sim 1 + (n^*F_{ST})$ using the pairwise F_{ST} genetic distances in **Table 1** and an average sample size of n=149. There is a strong relationship between observed and expected values both when Finland is included (r≈1) and excluded (r≈1).

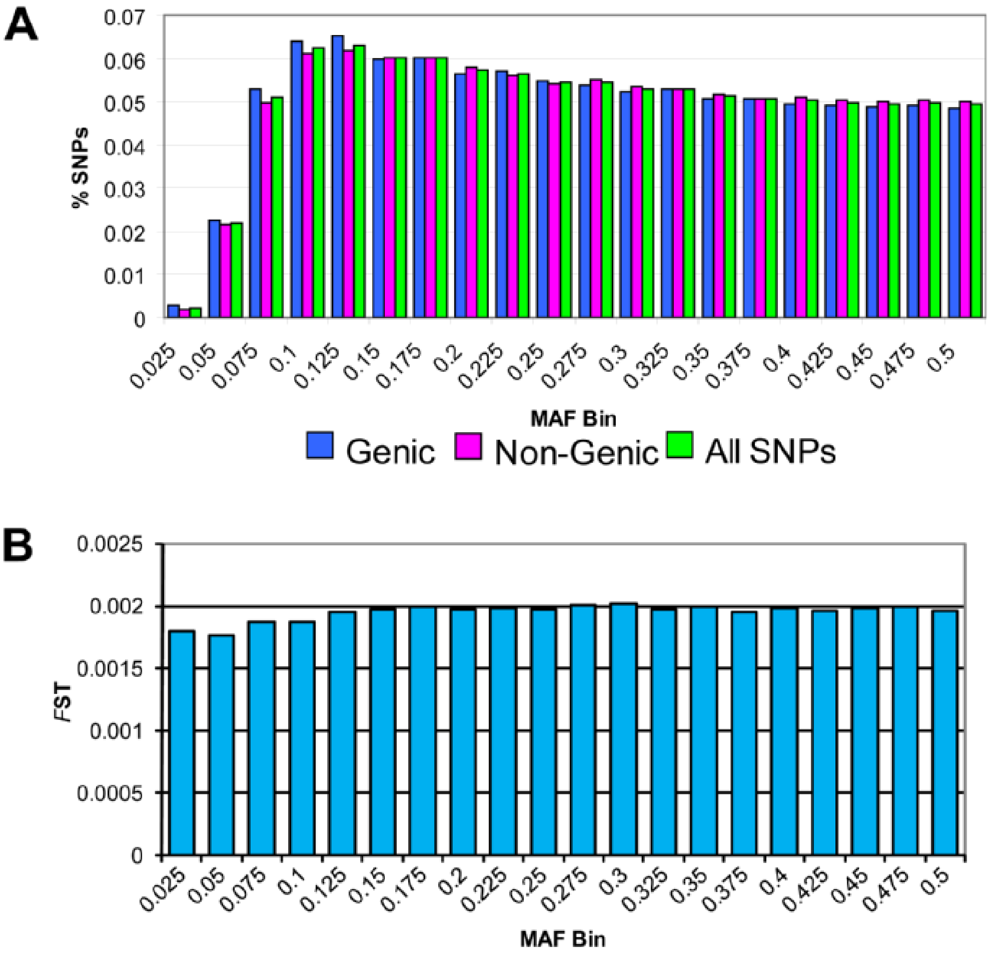

Figure S8 – Minor Allele Frequency (MAF) Spectra.

(**A**) Frequency distribution of genic (n≈120K), non-genic (n≈176K) and all autosomal SNPs (n≈296K) by MAF bin. MAF calculated from the populations used in this study. (**B**) Average F_{ST} values by MAF bin.

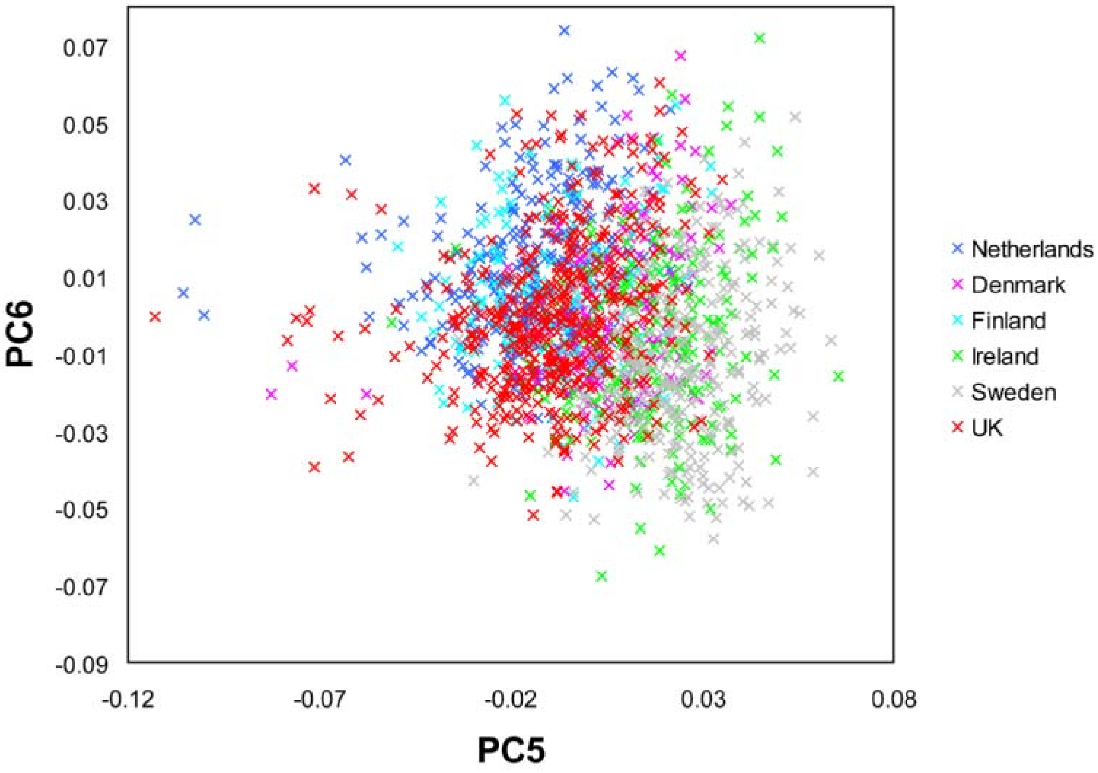

Figure S9– PC5 and PC6 in Northern European populations.

PC5 verus PC6 derived from 2051 individuals genotyped for 296553 autosomal SNPs. For increased clarity only the individuals from the *in situ* European populations are included (therefore Australians and CEU Europeans are not shown).

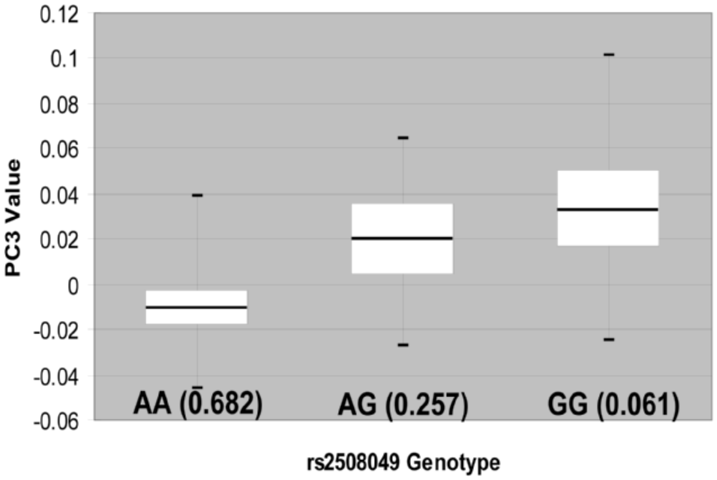

Figure S10 – PC3 values and rs2508049 Genotype.

Median, inter-quartile and full range of PC3 values by rs2508049 genotype over 2051 individuals. The frequency of each genotype is shown in parenthese