Supplementary Information

Participants

Subjects were registered at the Netherlands Twin Register (NTR, N=5 509; 2 226 males and 3 283 females)[1] or the Netherlands Study of Depression and Anxiety (NESDA, N=2 038; 684 males, and 1 354 females)[2]. The NTR sample consisted of 830 unrelated individuals, 1 431 families with two members, 372 with 3 members, 111 with four, 49 with five, and 2 families with six members (parents, twins, siblings, spouses of twins). The NESDA sample consisted of unrelated individuals only.

Genotyping was performed on the Affymetrix Human Genome-Wide SNP 6.0 Array at two sites (Avera Institute for Human Genetics [AIHG], South Dakota, USA and the Rutgers University Cell and DNA Repository [RUCDR], New Jersey, USA) according to the manufacturer's protocol. Methods for blood and buccal swab collection, genomic DNA extraction, and genotyping have been described previously[3,4].

The birth country of the parents was available for the majority of the subjects (N=4 485) as well as their current living addresses ($N_{relateds}$=7 092, $N_{unrelateds}$=4 103). For the current living addresses, the postal codes were translated into geographic coordinates (longitude and latitude) for each participant using the open source 6PP database,[5] in order to compute correlations between the PCs and North-South/East-West gradients. These coordinates were also used to plot the subjects on the map of the Netherlands in Figure 1. Place of birth (city or municipality of birth) was available for 1 841 subjects who also had current living address available. These were also translated into geographic coordinates using the open source 6PP database (these coordinates are less accurate however than those obtained from postal codes). Adult height (stature; age ≥ 18 years old) was available for the majority of the subjects ($N_{relateds}$=5 914, $N_{unrelateds}$=3 714). Self-reported eye color was available for 3 375 subjects (1 581 unrelated, coded as blue, intermediate or brown). Self-reported hair color was available for 3 380 subjects (1 583 unrelated, coded as blond, red, light brown, dark brown, or black). The numbers reported

here (and in the rest of the manuscript) are excluding 659 individuals that were removed due to a batch effect (see Supplementary Information: *Removing a Batch Effect*).

The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, and an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180, NESDA 03-183). All subjects provided written informed consent.

Quality Control

Autosomal SNPs were analyzed. Quality control (QC) was conducted in Plink[6], by removing all SNPs with a minor allele frequency (MAF) smaller than 5%, missing rate greater than 5%, a Hardy-Weinberg equilibrium (HWE) deviation with a *p*-value smaller than 0.001. SNPs were also removed if there were alleles that were incongruent between datasets, which could also represent an allele flip (this filter was applied twice: when merging the dataset from AIHG with the dataset from RUCDR, and when merging the merged Dutch dataset with a total of 1 094 subjects from the 1000 Genomes dataset). Individuals were removed if they had a missing rate greater than 5%, or excess genome-wide heterozygosity / inbreeding levels (F, as calculated in Plink on an LD-pruned set, must be greater than -0.10 and smaller than 0.10). Only SNPs that passed QC in the genotyped Dutch dataset were analyzed for 1000 Genomes samples (June 2011 release)[7].

Identifying individuals with non-European/non-Dutch ancestry

The 1000 Genomes dataset was used as a reference to aid in identifying and excluding individuals with a non-Dutch ancestry (see Supplementary Figure S2). The 1000 Genomes PCs that were not pruned for LD and did not contain long-range LD regions were used for this goal (SNP Panel 2, in line with the suggestions from Price et al[8,9]). Eight of the top ten 1000 Genomes PCs (all but PC4 and PC7) cluster the European populations together, making them useful for detecting individuals with a

non-European ancestry. A Dutch individual was labeled as a potential outlier with a non-European ancestry if one of the 1000 Genomes PCs of that individual was lower than the minimum or higher than the maximum score of that particular PC of the European 1000 Genomes individuals (CEPH, Finnish, British, Iberian, and Toscan). This yielded 151 outliers from PCs 1, 2, 3, and 5. A good (albeit imperfect) indicator of one's ancestry is the country of birth of the parents, which was available for a subset of the Dutch dataset. When comparing the birthplace of the parents between the outliers and the rest of the Dutch sample, the majority of the outliers (57%) had at least one parent born outside of the Netherlands, as opposed to 4.5% of the rest of the sample. This suggests that (the majority of) these individuals are indeed likely to have a non-European ancestry; hence they were excluded. PCs 4 and 7 are the only two PCs that differentiate between European populations. For PC4, the two populations with the lowest F_{st} when compared to the Dutch, the British (F_{st} = .0005) and CEPH (F_{st} =.0002; see Supplementary Table S1), are the only European populations that cluster with the Dutch. Of the Dutch individuals that fall outside the British and CEPH cluster (N=129), the majority (53.3%) have at least one parent that is born outside of the Netherlands, suggesting these individuals are also likely to have a non-Dutch ancestry component. These individuals were also excluded. The 1000 Genomes PC7 shows three clusters that overlap with each other. PC7 separates British and CEPH individuals from Finnish, Toscan, and Puerto Rican individuals, with the rest of the populations falling in between these two clusters. The Dutch individuals also mainly fell in between the two lateral clusters, but showed a large overlap with all three clusters, making it difficult to interpret who are outliers, therefore no individuals were excluded based on this PC. Eventually a total of 258 of the 7 547 individuals were excluded from the PCA on the Dutch sample. Parental birth place information was available for 132 of these individuals, of which 73 (55.3%) had at least one parent born outside of the Netherlands (as opposed to 4% of the rest of the individuals).

Three SNP sets, varying in LD, for PCAs

Three different SNP sets were created to run the PCAs on, varying in the amount of LD allowed: a SNP set including all SNPs that passed QC (499 849 SNPs; Panel 1); a SNP set excluding 24 long-range LD regions identified by Price et al[9] (487 672 SNPs; Panel 2), and an LD-pruned SNP set without long range LD regions, where SNPs were pruned recursively in a sliding window (window size = 50, number of SNPs to shift after each step = 5) based on a variance inflation factor (VIF) of 2, resulting in a set with 130 248 SNPs (Panel 3). VIF = $1/[1-R^2]$, where R^2 is the multiple correlation coefficient for a SNP regressed on all other SNPs within the window simultaneously[6].

PCA

PCAs were run with the EIGENSOFT package[8] to compute 10 PCs for each of the three LD varying SNP sets using its default parameters. The PCA was run on unrelated individuals only, and projected onto the other subjects. Unrelated individuals were chosen using GCTA[10], by excluding one of each pair of individuals with an estimated genetic relationship of >0.025 (i.e., more related than third or fourth cousin). The genetic relationship matrix was calculated for each population separately. First, PCs extracted from the 1000 Genomes individuals (1014 unrelated individuals, from the SNP set that was not pruned for LD and without long-range LD regions, i.e., in line with the suggestions from Price et al[8,9]) were used to detect individuals with possible non-Dutch or non-European ancestry. After excluding these individuals (N=258), 4 441 unrelated Dutch individuals were extracted with GCTA. PCA was run on these unrelated individuals for each of the three LD varying SNP sets and projected on the rest.

Delta, F_{st}, mean LD and mean haplotype block size

Delta (δ) and F_{st} were calculated using scripts written in Perl. δ is defined as the absolute allele frequency difference between two groups or populations. In order to determine which SNPs underlie the variation reflected by the PCs, δ was calculated for all SNPs between the individuals with the highest

PC values versus the individuals with the lowest PC values (top and bottom 1000 for the Dutch dataset; top and bottom 250 for the 1000 Genomes dataset).

F_{st} was calculated as a measure for genetic differentiation between populations according to Weir and Cockerham[11] by calculating it for every SNP and then averaging all F_{st} values to obtain a genome-wide point estimate of the genetic distance. F_{st} values normally range between 0 and 1. Note that F_{st} according to Weir and Cockerham (Figure 2, Supplementary Table S1, and analyses on HERC2 in Northern vs. Southern European 1000 Genomes populations) gives slightly different outcomes than the F_{st}'s calculated by Bayescan 2.1[12] and should not be directly compared. For computational reasons, the latter was used only for the analyses on selection pressures, (i.e., in comparing the top 1000 versus bottom 1000 individuals for 3 PCs, described in the paragraph below).

The mean LD and average haplotype block size were calculated in Plink[6] and additional purpose-written perl scripts. To investigate the amount of LD that influenced a PC, Plink was used to calculate an LD matrix of the top 500 SNPs of the PC (determined by δ), after which all LD values (r^2) were averaged (Table 1). To examine the presence of serial founder effects, haplotype blocks were calculated per chromosome in Plink for different groups of individuals. This was done with pair-wise LD calculations for SNPs within 4000 kb (the size of the largest long-range LD region: the chromosome 8p23.1 inversion between 8 and 12 Mb), using the largest SNP set (499 849 SNPs). The sizes of all autosomal haplotype blocks were then averaged.

Identifying variants under selection

Candidate loci that may have been under selection pressures were identified in Bayescan 2.1[12]. A comparison of several algorithms designed to achieve this goal through F_{st} outlier tests concluded that this software package had the lowest false negative and false positive rates[13]. After computing F_{st} values for all 499 849 SNPs between the top 1000 and bottom 1000 individuals for 3 PCs reflecting ancestry, the F_{st} coefficients are decomposed into a population-specific component (β), shared by all loci, and a locus-specific component (α), shared by both populations. If α differs significantly from 0, it is assumed

that the locus was under diversifying ($\alpha > 0$) or balancing/purifying ($\alpha < 0$) selection, although power is usually weak for detecting balancing selection[12-14]. Significance is based on FDR corrected q-values (< .05). Higher false positive rates may be observed when isolated populations are included that underwent a strong bottleneck. Since the subpopulations in our sample are not geographically isolated, we have no strong reasons to assume strong isolation and/or strong bottlenecks within the Netherlands.

Additional Analyses, Software and Bioinformatics

SPSS and additional perl scripts were used for data management. Graphics were created with R. Plink was used for computing F (inbreeding coefficient/genome-wide homozygosity, on an LD-pruned SNP set), and for GWASs (linear regressions on unrelated individuals) on adult height (N=3 714), eye color (N=1 581) and hair color (N=1 583). All reported correlations are Pearson correlations computed with SPSS. All base pair positions are in build 37. SNP annotations and genic information about SNPs were extracted from the Ensembl database (Ensembl Genes 67, GRCh37.7).

Ingenuity Pathway Analysis (Ingenuity Systems, IPA spring release 2012), was used to examine whether particular biological functions were overrepresented among the genes showing significant signals for selection pressures. The Ingenuity database contains a large amount of information about structure, biological function, and subcellular localization of the proteins. Only biological relationships that were experimentally observed were considered in the analysis.

The Randstad

The Randstad is a metropolitan region in the Western part of the Netherlands containing >40% of the Dutch population (~7.1 million out of ~16.8 million). This region includes the four largest cities of the Netherlands (Amsterdam, Rotterdam, Den Haag, Utrecht) and surrounding areas (see Figure 1a). The term Randstad did not exist until the second half of the twentieth century, but migration records since 1800 already show considerable migration flows between rural areas and the urbanized West, as

well as between the major cities in the West[15,16]. This may have lead to more admixture between Dutch subpopulations in this region.

From the 4 155 unrelated Dutch individuals with a known current living address, three selections were made: inhabitants of the four largest municipalities with a population size of >300k (N = 624), inhabitants of the thirteen Randstad municipalities with a population size of >100k (N = 1 086), and of the 26 municipalities from the entire country with a population size >100k (N = 1 630) (see Figure 1a for the locations of the 26 largest municipalities and Supplementary Table S7 for an overview of their population size in April 2012 according to the Central Bureau of Statistics[17]).

Especially for PC1, the Randstad region seems to show most of the intermediate values at face value in Figure 1a. When only including individuals living in the major municipalities in this region, the correlation between PC1 and the North-South axis is not significant (r = -.010, p = .808 for the four major Randstad municipalities with population size >300k; r = .055, p = .074 for the thirteen major Randstad municipalities with population size >100k; see Supplementary Table S8). When excluding these individuals from the entire sample of unrelated Dutch individuals, the correlation with geography increases considerably for PC1 (r = .648, p < .001 excluding the four major Randstad municipalities; r = .669, p < .001 excluding the 13 major Randstad municipalities; see Supplementary Table S8). The correlation between PC1 and genome-wide homozygosity also increases slightly without the major municipalities. As opposed to the correlation with the North-South axis, the correlation with genome-wide homozygosity remains significant for the individuals from the major Randstad municipalities (r = .201, p < .001 excluding the four major Randstad municipalities; .170, p < .001 excluding the 13 major Randstad municipalities). This indicates that the correlation between PC1 and homozygosity observed in the entire sample is not due to local admixture or inbreeding, making the serial-founder effect hypothesis more plausible. When excluding all Dutch municipalities with a population size > 100k, the correlation between PC1 and the North-South axis increases further (r = .678, p < .001), but the correlation for the individuals from these municipalities is also still very significant (r = .451, p < .001).

The correlation of PC2 with the East-West gradient also increases from .378 to .405 as the major municipalities of the Randstad are excluded, and increases further to .439 when all 26 municipalities with a population size > 100k are excluded (*p*'s < .001, see Supplementary Table S8). PC2 still shows a significant correlation with the East-West gradient when only considering the 13 municipalities from the Randstad (r = .145, p < .001), and the 26 municipalities from the entire country (r = .281, p < .001). The correlation of PC3 with the East-West gradient shows little change when excluding these municipalities.

Removing a Batch Effect

Genotyping was done in 8 207 individuals, of which 320 individuals were excluded from the initial PCA because of a non-European/non-Dutch ancestry (identified as described under *Identifying individuals with non-European/non-Dutch ancestry*). PCA was run on 125,303 LD based pruned SNPs (parameters as described in Methods) in 4 666 unrelated individuals and projected onto the remainder of the sample. PC1 from this analysis showed a strong correlation with F (r = .755) and the Contrast QC (CQC, a quality metric from Affymetrix representing how well allele intensities separate into clusters; r = .596). The strong correlation was caused by a subset of individuals. We calculated the distance between the mean value of PC1 (-.0022) and the highest observed PC1 value (.0116), and subtracted this value from the mean. All individuals who scored below this value were considered outliers and excluded from subsequent analyses (N=659; see Supplementary Figure S6). All numbers reported in the main manuscript and Supplementary Information are excluding these 659 individuals.

References

1. Boomsma DI, De Geus EJC, Vink JM *et al*: Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* 2006; **9:** 849-857.

2. Penninx BWJH, Beekman ATF, Smit JH *et al*: The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int J Method Psych* 2008; **17:** 121-140.

3. Boomsma DI, Willemsen G, Sullivan PF *et al*: Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur J Hum Genet* 2008; **16:** 335-342.

4. Willemsen G, de Geus EJC, Bartels M *et al*: The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* 2010; **13:** 231-245.

5. Broek Kvd: 6PP database (http://www.d-centralize.nl/projects/6pp/downloads/), 2012.

6. Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-575.

7. Oxford: Mathematical Genetics and Bioinformatics Groups; 1000 Genomes June 2011 Readme file: http://mathgen.stats.ox.ac.uk/impute/README_1000G_phase1interim_jun2011.txt, 2011.

8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38:** 904-909.

9. Price AL, Weale ME, Patterson N *et al*: Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 2008; **83:** 132.

10. Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2010.

11. Weir BS: *Genetic data analysis II*. Sunderland, MA: Sinauer, 1996.

12. Foll M, Gaggiotti O: A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 2008; **180:** 977-993.

13. Narum SR, Hess JE: Comparison of FST outlier tests for SNP loci under selection. *Mol Ecol Resour* 2011; **11:** 184-194.

14. Beaumont MA, Balding DJ: Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 2004; **13:** 969-980.

15. Lesger C: *Noord-Hollanders in beweging: economische ontwikkeling en binnenlandse migratie, ca. 1800-1930*. CGM, 2003.

16. Suurenbroek F: *Binnenlandse migratie naar en uit Amsterdam (1870-1890)*. Centrum voor de Geschiedenis van Migranten, 2001.

17. CBS: Centraal Bureau voor de Statistiek, Bevolkingsontwikkeling; regio per maand, April 2012, 2012.

Supplementary Tables

Supplementary Table S1: Population pairwise F_{st} values for the Dutch and 1000 Genomes populations. NLD = Dutch individuals from the NTR and NESDA; ASW = HapMap African ancestry individuals from SW US; CEU = CEPH individuals; CHB = Han Chinese in Beijing; CHS = Han Chinese South; CLM = Colombian in Medellin, Colombia; FIN = HapMap Finnish individuals from Finland; GBR = British individuals from England and Scotland; IBS = Iberian populations in Spain; JPT = Japanese individuals; LWK = Luhya individuals; MXL = HapMap Mexican individuals from LA California; PEL = Peruvian in Lima, Peru; TSI = Toscan individuals; YRI = Yoruba individuals.

	NLD	ASW	CEU	CHB	CHS	CLM	FIN	GBR	IBS	JPT	LWK	MXL	PEL	TSI	YRI
NLD	0														
ASW	.08363	0													
CEU	.00020	.07524	0												
CHB	.08551	.11905	.08450	0											
CHS	.08650	.11916	.08547	.00094	0										
CLM	.01604	.06074	.01370	.06904	.07007	0									
FIN	.00615	.07925	.00589	.07820	.07937	.01691	0								
GBR	.00050	.07552	.00019	.08455	.08554	.01393	.00639	0							
IBS	.00122	.06898	.00199	.10833	.10999	.01062	.01007	.00215	0						
JPT	.08659	.11974	.08574	.00625	.00796	.06978	.07931	.08574	.11083	0					
LWK	.11488	.00951	.10414	.13597	.13658	.09144	.10707	.10443	.11163	.13715	0				
MXL	.03089	.07247	.02820	.06013	.06158	.00705	.02834	.02829	.02618	.06085	.10242	0			
PEL	.01176	.05149	.00985	.07656	.07739	.00548	.01498	.01006	.00618	.07741	.08168	.01531	0		
TSI	.00373	.07336	.00326	.08428	.08527	.01403	.01109	.00350	.00154	.08553	.10106	.02956	.00902	0	
YRI	.12253	.01051	.11282	.14222	.14321	.10133	.11524	.11311	.12653	.14375	.00678	.11224	.09134	.10960	0

Supplementary Table S1: The correlations between the SNP set including all SNPs that passed QC (Panel 1), and the SNP set excluding the 24 long-range LD regions (Panel 2), for the 1000 Genomes dataset.

Pearson Correlations		1000 Genomes PCs: Panel 1 (all SNPs)									
		PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
1000 Genomes PCs: Panel 2 (all SNPs excluding 24 long-range LD regions)	**PC1**	**1.000**	.020	.031	.006	.017	.059	.001	.001	-.006	.000
	PC2	.018	**1.000**	.003	-.006	-.014	.026	-.002	.002	.002	-.003
	PC3	.032	.002	**1.000**	.059	.014	.001	-.025	.023	-.049	.001
	PC4	.007	-.006	.056	**1.000**	.001	.003	-.009	.002	-.017	.006
	PC5	.018	-.014	.007	.003	**.999**	.012	.001	.003	.000	.004
	PC6	.059	.026	.001	.001	-.005	**1.000**	.003	-.002	.000	-.006
	PC7	.001	-.002	-.021	-.010	-.005	.001	**.997**	-.011	.010	.024
	PC8	.002	.002	.021	.003	.000	-.001	.009	**.997**	.039	.013
	PC9	-.006	.002	-.048	-.015	-.003	.002	.003	-.045	**.991**	-.070
	PC10	-.001	.002	.002	.000	.000	.001	-.001	.003	.011	.071

Supplementary Table S2: The correlations between the SNP set excluding the 24 long-range LD regions (Panel 2), and the LD pruned SNP set excluding the 24 long-range LD regions (Panel 3) for the 1000 Genomes dataset.

Pearson Correlations		1000 Genomes PCs: Panel 2 (all SNPs, no long-range LD)									
		PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
1000 Genomes PCs: Panel 3 (LD pruned SNPs, no long-range LD)	**PC1**	**.998**	-.043	.034	.007	.019	.057	.000	.001	-.005	-.001
	PC2	.084	**.998**	.009	-.003	-.011	.030	-.005	.001	.002	.002
	PC3	.029	-.001	**.999**	.050	.013	-.001	-.023	.021	-.047	.002
	PC4	-.008	.008	-.066	**-.997**	-.002	-.010	.004	-.004	.023	.005
	PC5	-.020	.013	-.009	-.002	**-.993**	-.060	.008	.004	.002	-.010
	PC6	.061	.026	.003	-.005	-.049	**.993**	-.007	-.006	.003	-.001
	PC7	.002	.002	-.022	-.013	.004	.012	**.983**	.017	.011	.006
	PC8	-.003	-.003	-.022	-.002	-.007	-.005	.017	**-.970**	-.019	-.011
	PC9	-.006	.003	-.046	-.009	-.003	-.001	.004	-.023	**.957**	.021
	PC10	.001	.001	.003	.005	-.011	.001	-.005	.005	-.039	**.704**

Supplementary Table S3: The correlations between the SNP set including all SNPs that passed QC (Panel 1), and the SNP set excluding the 24 long-range LD regions (Panel 2), for the Dutch dataset.

Pearson Correlations		Dutch PCs: Panel 1 (all SNPs)									
		PC1	**PC2(↕)**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8(↔)**	**PC9**	**PC10**
Dutch PCs: Panel 2 (all SNPs, no long-range LD)	**PC1(↕)**	-.111	**-.750**	**.634**	-.069	.054	-.043	.009	.002	.034	.019
	PC2	-.031	.001	.002	-.091	**-.726**	**-.638**	-.187	.052	.024	-.023
	PC3(↔)	-.013	-.019	-.030	.093	.081	-.102	-.262	**-.865**	.157	.114
	PC4	-.051	-.022	-.044	-.061	-.049	.066	.073	-.066	.241	.026
	PC5	-.001	-.020	-.027	-.086	.036	.000	-.089	-.024	.102	**-.758**
	PC6	-.005	-.005	.000	.031	.018	-.026	-.048	.053	-.166	.037
	PC7	.020	-.008	-.004	-.050	.005	-.011	-.041	-.037	-.260	-.010
	PC8	-.016	.013	.009	.077	.035	-.052	.076	.072	.034	.089
	PC9	.005	.035	.043	.024	.015	-.006	-.007	-.001	.248	-.061
	PC10	.004	.030	.042	.025	.010	-.019	-.045	-.001	-.060	.060

↕: The PC with the highest correlation with the North-South gradient.

↔: The PC with the highest correlation with the East-West gradient.

Supplementary Table S5: The correlations between the SNP set excluding the 24 long-range LD regions (Panel 2), and the LD pruned SNP set excluding the 24 long-range LD regions (Panel 3), for the Dutch dataset.

Pearson Correlations		Dutch PCs: Panel 2 (no long-range LD)									
		PC1(↕)	**PC2**	**PC3(↔)**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Dutch PCs: Panel 3 (LD pruned, no long-range LD)	**PC1(↕)**	**-.964**	-.024	-.026	-.024	.014	.003	-.002	.002	.002	-.004
	PC2(↔)	.025	.010	**-.752**	.011	-.073	-.098	.183	-.063	.009	-.062
	PC3(↗)	.009	-.035	-.112	.007	.061	.193	-.216	-.106	.139	-.118
	PC4	.017	-.016	.050	-.097	.031	-.112	.061	.006	.027	-.013
	PC5	.012	.018	.003	-.080	-.005	-.031	-.026	-.063	.008	-.163
	PC6	-.007	.013	.006	.007	.001	-.064	.016	-.025	.074	-.008
	PC7	-.005	.001	.032	.020	-.063	-.064	.080	-.151	.026	.091
	PC8	.002	-.019	-.011	-.046	-.083	.001	-.006	-.083	-.007	.050
	PC9	.000	.051	-.012	-.030	.004	.019	.031	.121	.027	.052
	PC10	.000	.009	.021	-.011	.033	.046	.034	-.106	.078	.034

↕: The PC with the highest correlation with the North-South gradient.

↔: The PC with the highest correlation with the East-West gradient.

↗: The PC that also showed a correlation with the East-West gradient, and separates individuals from the middle of the Netherlands from individuals from the rest of the country (illustrated in Figure 1d).

Supplementary Table S6: Genotype frequencies (%) of rs8039195 and rs12913832 (from the HERC2 gene) for the Dutch population and the 1000 Genomes populations.

Population	rs8039195 (HERC2)			rs12913832 (HERC2)		
	CC	CT	TT	AA	AG	GG
HapMap Finnish individuals from Finland	.0	6.5	93.5	.0	19.4	80.6
Northern Dutch individuals (top 1000 PC1)	.4	13.1	86.5	-	-	-
British individuals from England and Scotland	1.2	21.4	77.4	3.4	28.1	68.5
Southern Dutch individuals (bottom 1000 PC1)	2.3	23.9	73.7	-	-	-
CEPH individuals	1.2	29.4	69.4	3.4	39.1	57.5
Iberian populations in Spain	.0	50.0	50.0	42.9	50.0	7.1
Colombian in Medellin, Colombia	9.4	43.4	47.2	51.7	41.7	6.7
Toscan individuals	16.8	42.1	41.1	30.6	53.1	16.3
HapMap Mexican individuals from LA California	16.3	46.9	36.7	75.8	15.2	9.1
Puerto Rican in Puerto Rico	9.1	56.4	34.5	65.5	29.1	5.5
Han Chinese South	39.3	48.3	12.4	99.0	1.0	.0
HapMap African ancestry individuals from SW US	43.8	45.8	10.4	70.5	27.9	1.6
Japanese individuals	68.5	23.6	7.9	100.0	.0	.0
Yoruba individuals	77.3	18.2	4.5	100.0	.0	.0
Han Chinese in Beijing	49.5	46.4	4.1	100.0	.0	.0
Luhya individuals	69.3	26.7	4.0	100.0	.0	.0

Supplementary Table S7: The 26 Dutch municipalities with a population size > 100k in April 2012 according to the Central Bureau of Statistics[17].

Municipality	Population Size	Randstad
Amsterdam	790 654	✓
Rotterdam	616 525	✓
Den Haag	502 683	✓
Utrecht	317 540	✓
Eindhoven	217 235	
Tilburg	207 398	
Almere	193 615	✓
Groningen	192 871	
Breda	176 835	
Nijmegen	165 262	
Enschede	158 020	
Apeldoorn	157 132	
Haarlem	152 260	✓
Arnhem	149 361	
Amersfoort	148 595	✓
Zaanstad	148 542	✓
Haarlemmermeer	143 885	✓
‘s-Hertogenbosch	141 981	
Zoetermeer	122 334	✓
Zwolle	121 733	
Maastricht	121 008	
Leiden	119 028	✓
Dordrecht	118 723	✓
Ede	108 802	
Emmen	108 779	
Westland	101 670	✓

Supplementary Table S8: The correlations of the first three Dutch PCs with the North-South gradient (PC1), the East-West gradient (PC2 & PC3), and F (genome-wide homozygosity) with and without individuals from the major municipalities in the Netherlands.

	All unrelated individuals (N = 4 441)	Excluding inhabitants of 4 Randstad municipalities with population size >300k (N = 3 531)	Inhabitants of 4 Randstad municipalities with population size >300k (N = 624)	Excluding inhabitants of 13 Randstad municipalities with population size >100k (N = 3 069)	Inhabitants of 13 Randstad municipalities with population size >100k (N = 1 086)	Excluding inhabitants of 26 municipalities with population size >100k (N = 2 525)	Inhabitants of 26 municipalities with population size >100k (N = 1 630)
$r_{PC1,\updownarrow}$	.603	.648	-.010	.669	.055	.678	.451
$r_{PC1,F}$	.245	.247	.201	.259	.170	.246	.233
$r_{PC2,\leftrightarrow}$	.378	.404	.067	.405	.145	.439	.281
$r_{PC2,F}$	.017	.033	.066	.031	.033	.019	.028
$r_{PC3,\leftrightarrow}$	.162	.171	.092	.166	.085	.164	.162
$r_{PC3,F}$	.009	.008	.063	.011	.018	.004	.036

Supplementary Figure Legends

Supplementary Figure S1: The distance between birthplace and current living address for 1,841 Dutch individuals. The mean distance is 33.33 km (SD=43.60) or 20.71 mi (SD=27.09).

Supplementary Figure S2: Identifying individuals with a non-European/non-Dutch ancestry with the projection of the 1000 Genomes PCs on the Dutch PCs. See *Supplementary Information: Identifying individuals with non-European/non-Dutch ancestry* for a more detailed description.

Supplementary Figure S3: The eigenvalues of the Dutch PCs from the LD pruned dataset without long-range LD regions or ethnic outliers.

Supplementary Figure S4: The X-Y plots for the first three PCs, with each individual colored by province of the current living address (note that PC2 has been inversed [PC2*-1], to illustrate the geographic correlation with East-West).

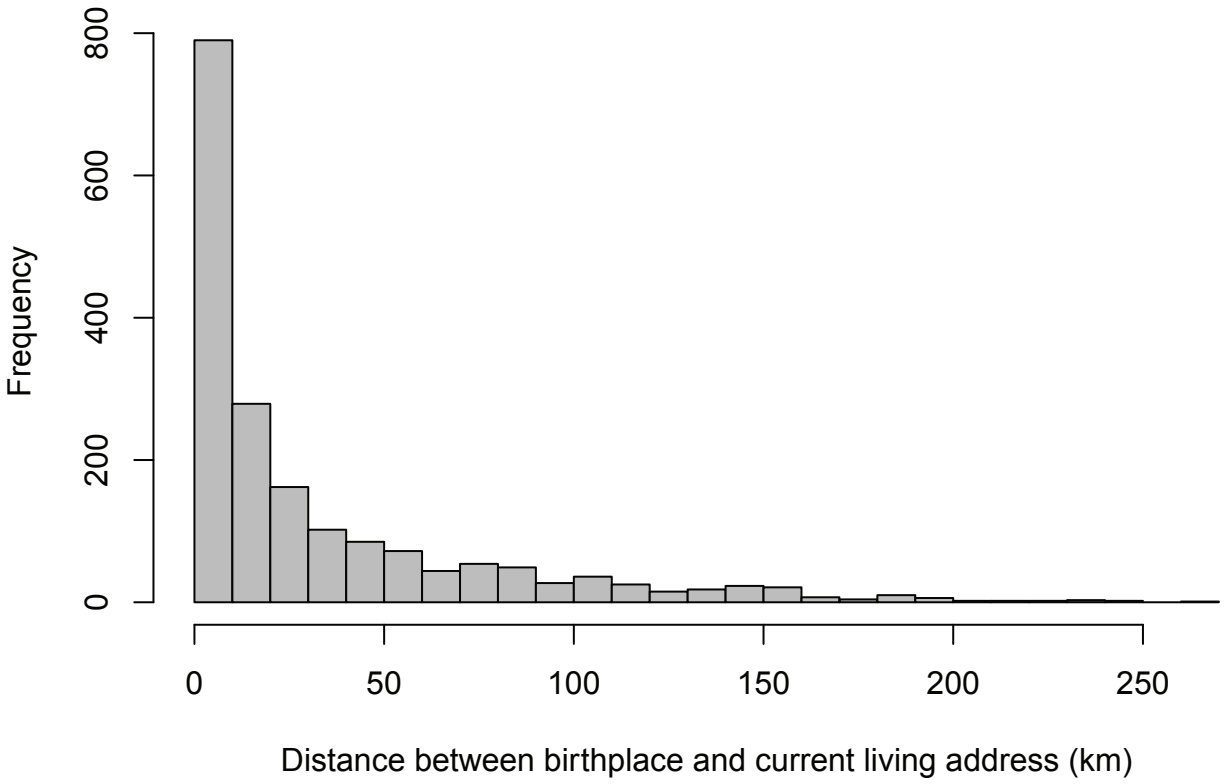
Supplementary Figure S5: PC1 and PC2 plotted against Latitude and Longitude respectively, and PC3 plotted against both Latitude and Longitude, with each individual colored by province of the current living address (note that PC2 has been inversed here as well [PC2*-1]).

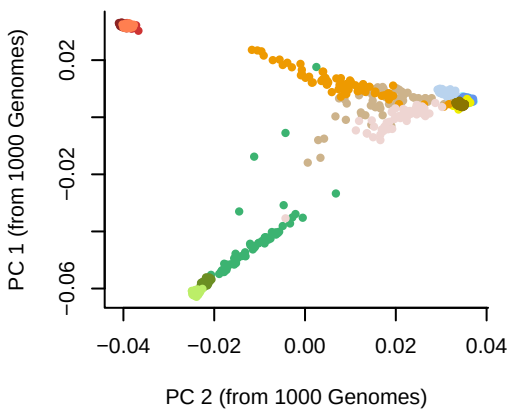
Supplementary Figure S6: Scatterplot of PC1 from the PCA on 8 207 individuals and CQC.

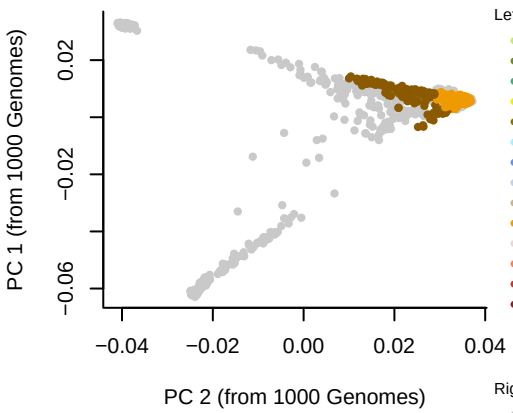
Supplementary file

The supplementary file (supplementary_file.xls) contains all SNPs that showed significant signals for diversifying selection ($q < .05$) for all three PCs (note: each PC is on a separate sheet of the file). The following columns are included:

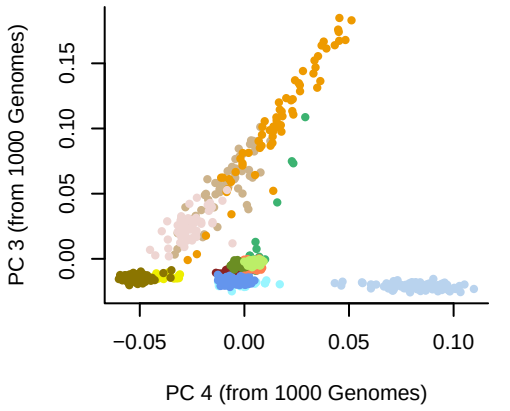
- **chr**: chromosome.
- **bp**: base pair position (according to build 37).
- **SNP_ID**: rs ID of the SNP.
- **fst**: the F_{st} coefficient averaged over populations. In each population F_{st} is calculated as the posterior mean using model averaging.
- **alpha**: the estimated alpha coefficient indicating the strength and direction of selection. A positive value of alpha suggests diversifying selection, whereas negative values suggest balancing or purifying selection.
- **prob**: the posterior probability for the model including selection.
- **log10PO**: the logarithm of Posterior Odds to base 10 for the model including selection. Note that this value is arbitrarily fixed to 1000 when the posterior probability is 1 (should be infinity).
- **qval**: the q-value for the model including selection.
- **Gene_ID**: The name of the gene the SNP falls in.

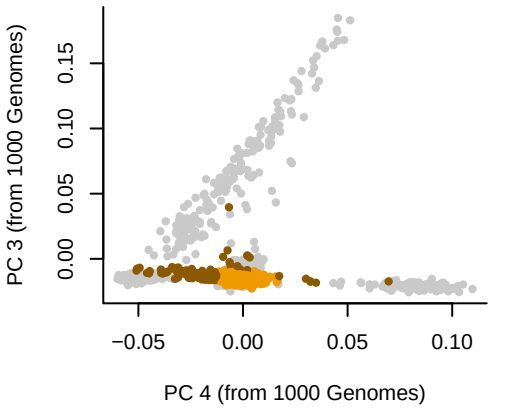

Frequency
0
200
400
600
800
0
50
100
150
200
250
Distance between birthplace and current living address (km)

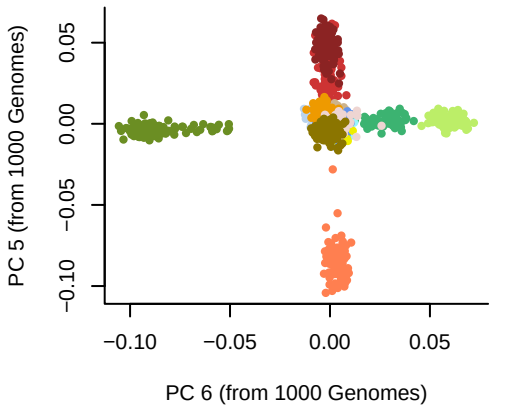

PC 1 (from 1000 Genomes)
0.02
−0.02
−0.06
−0.04
−0.02
0.00
0.02
0.04
PC 2 (from 1000 Genomes)

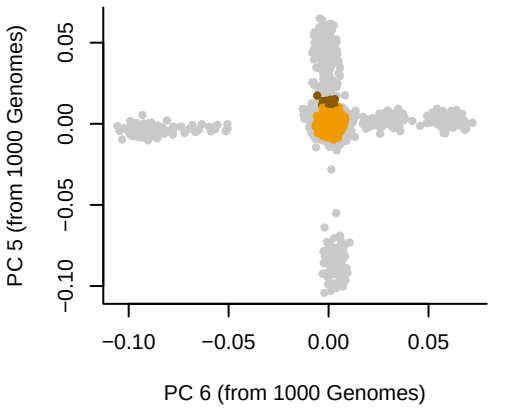

PC 1 (from 1000 Genomes)
0.02
−0.02
−0.06
−0.04
−0.02
0.00
0.02
0.04
PC 2 (from 1000 Genomes)

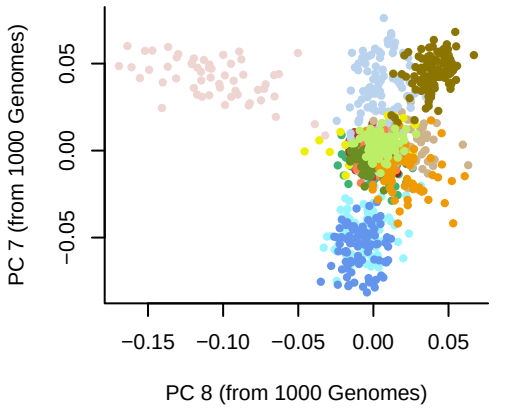

Left:
Yoruba individuals
Luhya individuals
HapMap African ancestry individuals from SW US
Iberian populations in Spain
Toscan individuals
CEPH individuals
British individuals from England and Scotland
HapMap Finnish individuals from Finland
Colombian in Medellin, Colombia
HapMap Mexican individuals from LA California
Puerto Rican in Puerto Rico
Japanese individuals
Han Chinese in Beijing
Han Chinese South
Right:
1000 Genomes
Dutch
Dutch (outlier)

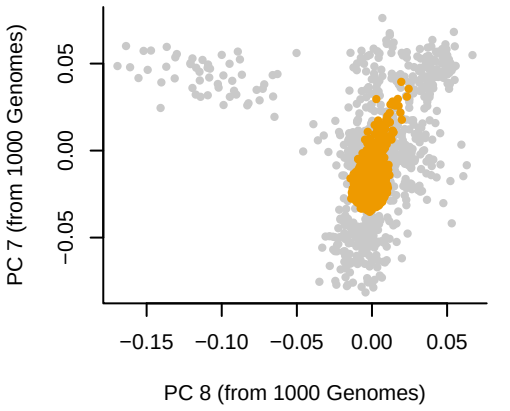

PC 3 (from 1000 Genomes)
0.15
0.10
0.05
0.00
−0.05
0.00
0.05
0.10
PC 4 (from 1000 Genomes)

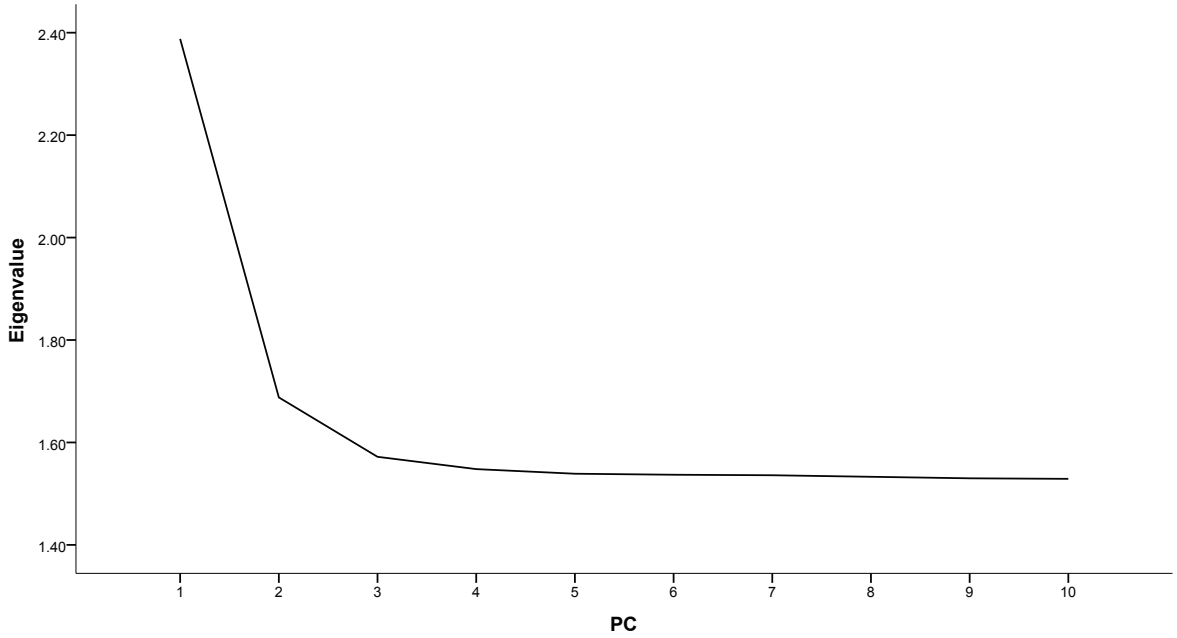

PC 3 (from 1000 Genomes)
0.15
0.10
0.05
0.00
−0.05
0.00
0.05
0.10
PC 4 (from 1000 Genomes)

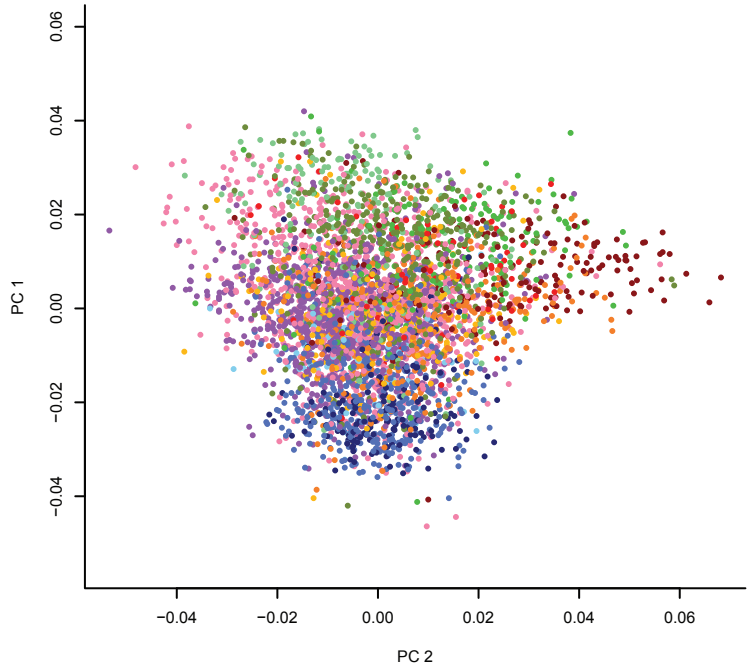

PC 5 (from 1000 Genomes)
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
PC 6 (from 1000 Genomes)

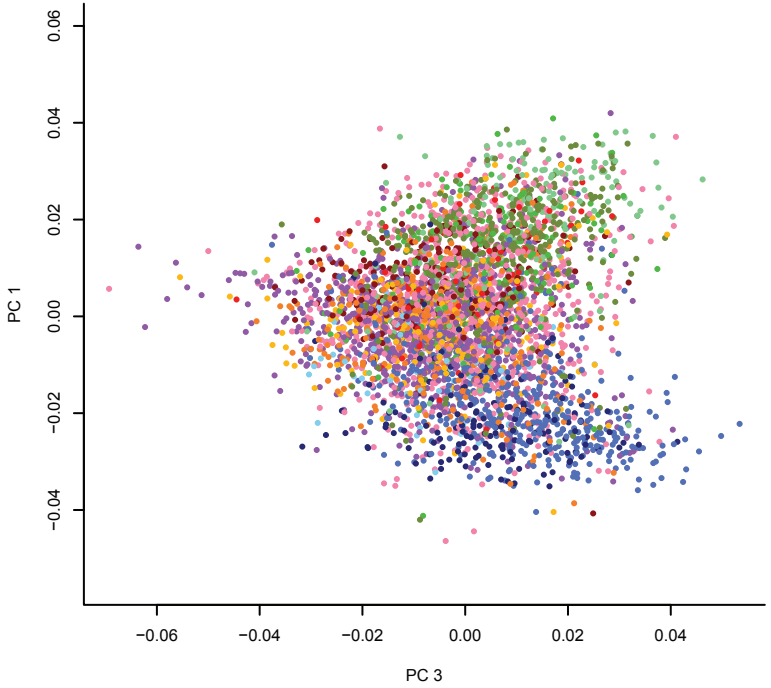

PC 5 (from 1000 Genomes)
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
PC 6 (from 1000 Genomes)


PC 7 (from 1000 Genomes)
0.05
0.00
−0.05
−0.15
−0.10
−0.05
0.00
0.05
PC 8 (from 1000 Genomes)

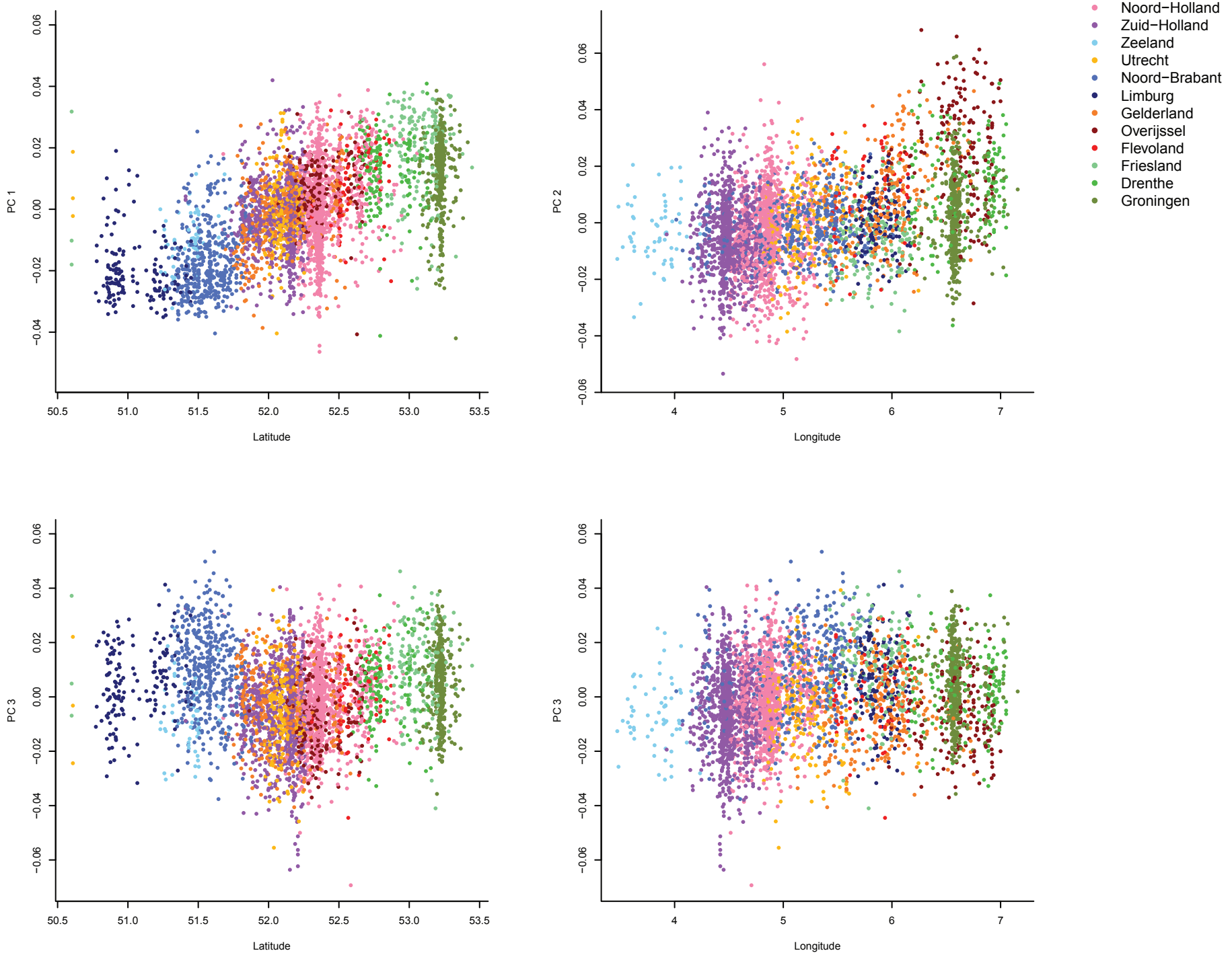

PC 7 (from 1000 Genomes)
0.05
0.00
−0.05
−0.15
−0.10
−0.05
0.00
0.05
PC 8 (from 1000 Genomes)

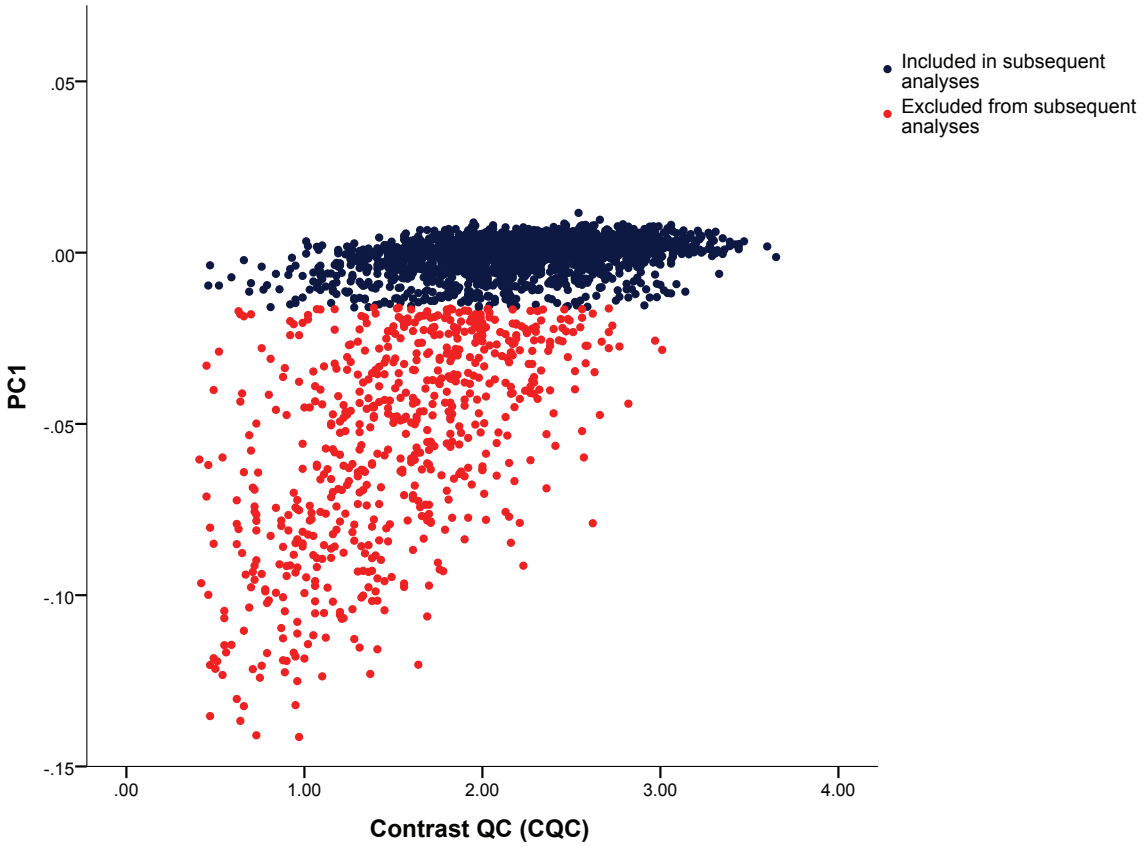

Eigenvalue
2.40
2.20
2.00
1.80
1.60
1.40
1
2
3
4
5
6
7
8
9
10
PC


PC 1
PC 2
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.04
−0.02
0.00
0.02
0.04
0.06


PC 1
PC 3
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.04
−0.02
0.00
0.02
0.04


Noord−Holland
Zuid−Holland
Zeeland
Utrecht
Noord−Brabant
Limburg
Gelderland
Overijssel
Flevoland
Friesland
Drenthe
Groningen


PC 1
Latitude
PC 2
Longitude
PC 3
Latitude
PC 3
Longitude
Noord−Holland
Zuid−Holland
Zeeland
Utrecht
Noord−Brabant
Limburg
Gelderland
Overijssel
Flevoland
Friesland
Drenthe
Groningen


Included in subsequent analyses
Excluded from subsequent analyses
PC1
.05
.00
-.05
-.10
-.15
.00
1.00
2.00
3.00
4.00
Contrast QC (CQC)