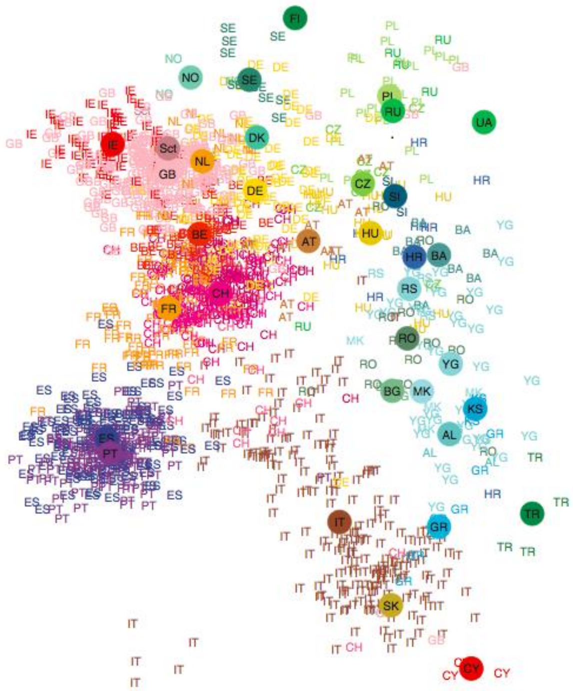
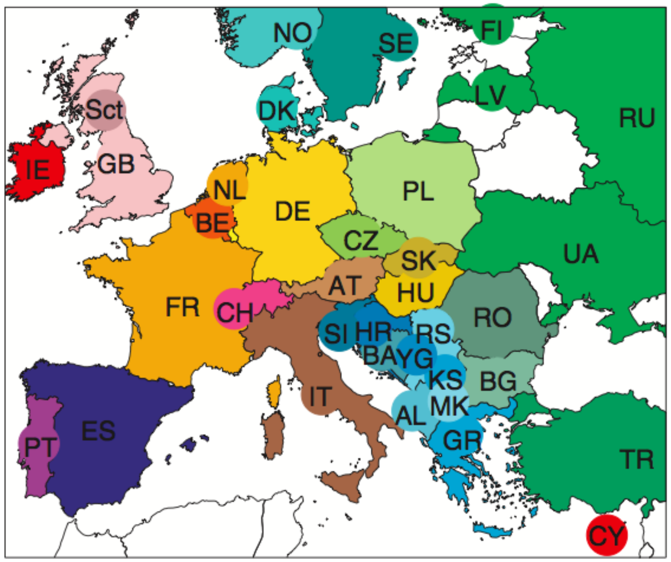




Principal Component Analysis (PCA) on genome-wide SNP data

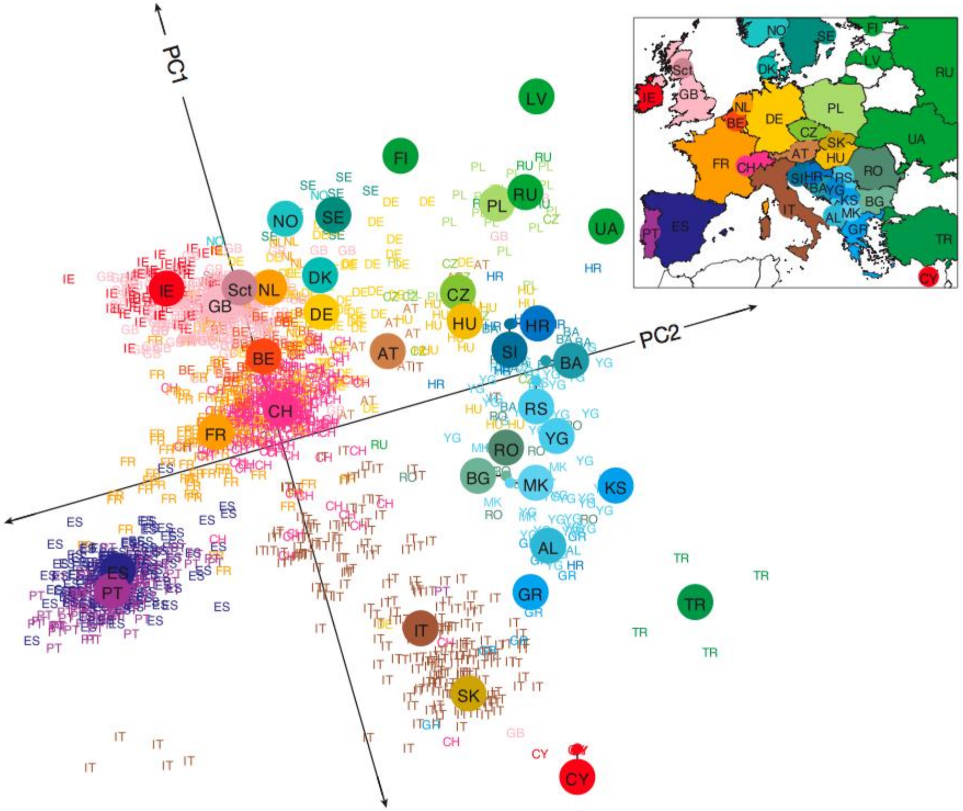
Abdel Abdellaoui

Netherlands Twin Register, VU University Amsterdam

a.abdellaoui@vu.nl

Why PCA? (1)

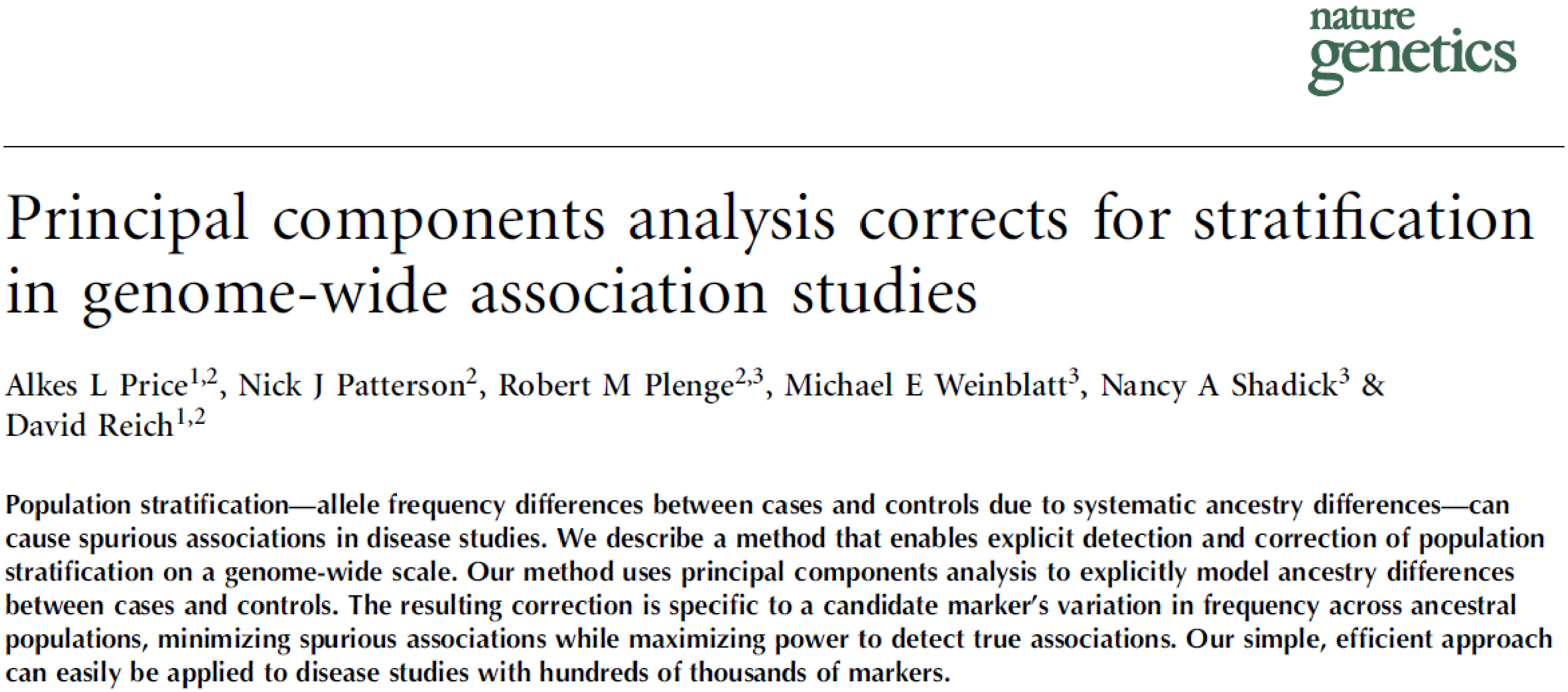
- Before attempting to understand the nitty gritty details of the workings of a complex system of variables (such as SNPs), the main patterns of variation should be understood.
- PCAs show the patterns explaining most variation.
- PCs reflecting ancestry differences usually correlate with geography.



Genes mirror geography within Europe

NATURE | Vol 456 | 6 November 2008

Why PCA? (2)



nature genetics

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

Overview practical

- Analyses are based on the paper "Population Structure, Migration, and Diversifying Selection in the Netherlands" (Abdellaoui et al, 2013, in press)

Analyses:

- Run PCA on 1000 Genomes, and project PCs on Dutch individuals
  - Goal: identify Dutch individuals with non-European ancestry and exclude
- Run PCA on remaining Dutch individuals
  - Goal: obtain PCs reflecting Dutch ancestry differences

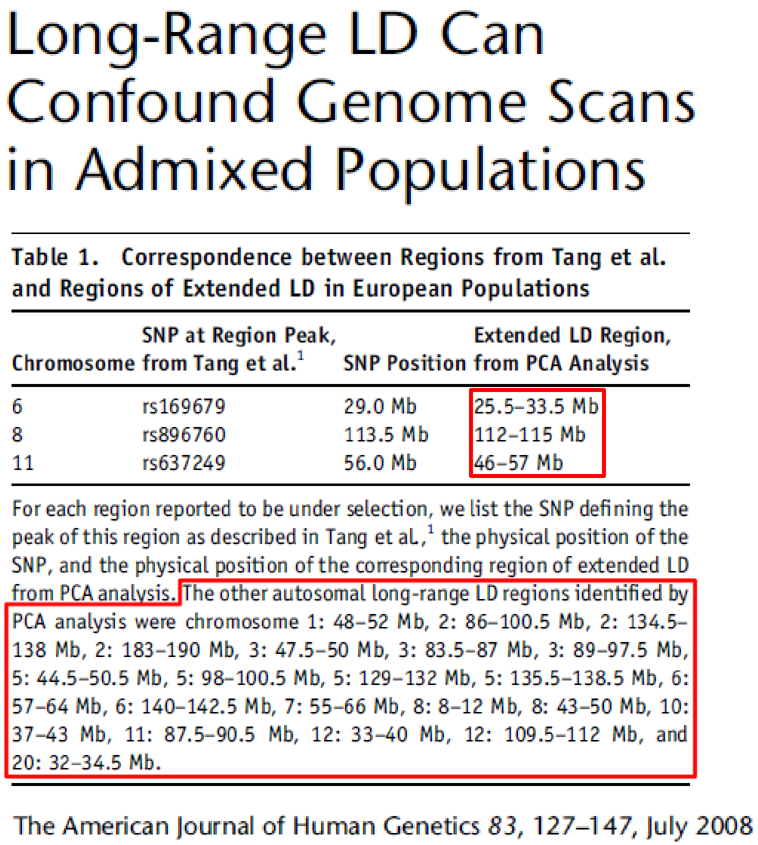
- Software used:
  - Eigenstrat -> http://genepath.med.harvard.edu/~reich/Software.htm
  - Plink -> http://pngu.mgh.harvard.edu/~purcell/plink
  - R -> http://www.r-project.org/

Description of the data

- Individuals:
  - 171 Dutch individuals from the Netherlands Twin Registry (NTR)
  - 221 from 1000 Genomes (Europeans, Africans, and Asians)
    - If you're interested in the 1000 Genomes dataset in plink format (~16 million SNPs): e-mail a.abdellaoui@vu.nl
- SNPs:
  - 113,164 SNPs (from Affy 6.0 chip)
  - Quality Control (done in Plink):
    - MAF > .05
    - HWE *p* > .001
    - SNP missingness < .05 (individual missingness < .02)
    - Excluded long-range LD regions
    - LD Pruned

Why exclude long-range LD regions?

- Elevated levels of LD can be overrepresented in PCs, deluding the genome-wide patterns that reflect the subtle ancestry differences.



Long-Range LD Can Confound Genome Scans in Admixed Populations

Table 1. Correspondence between Regions from Tang et al. and Regions of Extended LD in European Populations

Chromosome	SNP at Region Peak, from Tang et al.[1]	SNP Position	Extended LD Region, from PCA Analysis
6	rs169679	29.0 Mb	25.5–33.5 Mb
8	rs896760	113.5 Mb	112–115 Mb
11	rs637249	56.0 Mb	46–57 Mb

For each region reported to be under selection, we list the SNP defining the peak of this region as described in Tang et al.,[1] the physical position of the SNP, and the physical position of the corresponding region of extended LD from PCA analysis. The other autosomal long-range LD regions identified by PCA analysis were chromosome 1: 48–52 Mb, 2: 86–100.5 Mb, 2: 134.5–138 Mb, 2: 183–190 Mb, 3: 47.5–50 Mb, 3: 83.5–87 Mb, 3: 89–97.5 Mb, 5: 44.5–50.5 Mb, 5: 98–100.5 Mb, 5: 129–132 Mb, 5: 135.5–138.5 Mb, 6: 57–64 Mb, 6: 140–142.5 Mb, 7: 55–66 Mb, 8: 8–12 Mb, 8: 43–50 Mb, 10: 37–43 Mb, 11: 87.5–90.5 Mb, 12: 33–40 Mb, 12: 109.5–112 Mb, and 20: 32–34.5 Mb.

The American Journal of Human Genetics *83*, 127–147, July 2008

Why also prune for LD?

- From EIGENSTRAT paper "*Principal components analysis corrects for stratification in genome-wide association studies*" (Price et al, 2006):

  "Strong LD at a given locus which affects many markers could result in an axis of variation which corresponds to genetic variation specifically at that locus, rather than to genome-wide ancestry. Nonetheless, **we recommend inferring population structure using all markers. This recommendation is based on an analysis of HapMap data which suggests that these potential problems will not affect results in practice.**"

Why also prune for LD?

- PCA was conducted on three sets of SNPs varying in LD on 1000 Genomes populations and Dutch subjects separately
- PCs were identical for 1000 Genomes across the 3 SNP sets. For the Dutch dataset, there were big differences:

SNP set used for PCA	Nr. of SNPs for PCA	Correlations between PCs and North-South gradient (N = 3363)		Correlations between PCs and East-West gradient (N = 3363)		λ for GWASs on height including the North-South PC as a covariate
		Pearson Correlation	Difference test	Pearson Correlation	Difference test	
SNP set 1: All SNPs that passed QC	499,849	$r_{PC2,\updownarrow}$= .428	-	$r_{PC8,\leftrightarrow}$= .205	-	1.03937
SNP set 2: SNP set 1 without the 24 long-range LD regions	487,672	$r_{PC1,\updownarrow}$= .574	$p = 3.9*10^{-46}$ (versus SNP set 1)	$r_{PC3,\leftrightarrow}$= .260	$p = 4.2*10^{-10}$ (versus SNP set 1)	1.03092
SNP set 3: SNP set 2 with genome-wide LD based SNP pruning	130,248	$r_{PC1,\updownarrow}$=.588	$p = 1.9*10^{-4}$ (versus SNP set 2)	$r_{PC2,\leftrightarrow}$=.369	$p = 3.5*10^{-21}$ (versus SNP set 2)	1.02961

- Conclusion: LD pruning is necessary for more homogeneous datasets (i.e., datasets with subjects from a single population)

Copy and unzip files needed for practical

- Open terminal: Applications Menu -> Terminal Emulator
- First run this in your terminal:

```
cp -r /home/abdel/PCA_practical .
cd PCA_practical
unzip dutch_1kG.zip
```

- command.txt contains all the remaining commands we are going to run in the terminal (which are also on the slides PCA_practical.pdf)

Files needed for EIGENSTRAT

- Input files: three files containing information about SNPs and samples (.ped, .map, .fam)
- Parameter file: file containing parameters for the PCA

EIGENSTRAT input files are in plink format

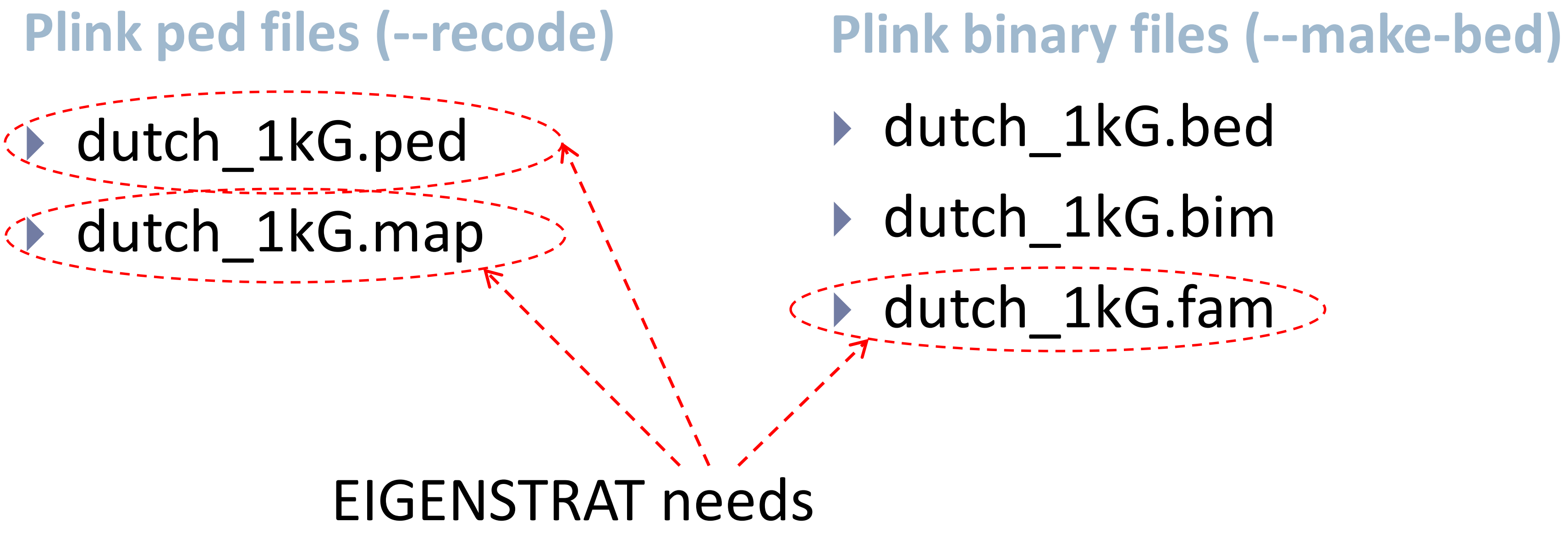
Plink ped files (--recode)

- dutch_1kG.ped
- dutch_1kG.map

Plink binary files (--make-bed)

- dutch_1kG.bed
- dutch_1kG.bim
- dutch_1kG.fam

EIGENSTRAT input files are in plink format



Plink ped files (--recode)

- dutch_1kG.ped
- dutch_1kG.map

Plink binary files (--make-bed)

- dutch_1kG.bed
- dutch_1kG.bim
- dutch_1kG.fam

EIGENSTRAT needs

Values in the phenotype column (column 6) of .fam file:

3 = Dutch individuals

4 = CEPH individuals

5 = British individuals

6 = Finnish individuals

7 = Iberian (Spain)

8 = Toscan

→ European

9 = Han Chinese in Beijing

10 = Han Chinese South

11 = Japanese individuals

→ Asian

12 = Luhya individuals

13 = Yoruba individuals

→ African

Parameter file (.par)

- The .par file will have the following lines:

genotypename: **dutch_1kG.ped** -> *input genotype file*

snpname: **dutch_1kG.map** -> *input snp file*

indivname: **dutch_1kG.fam** -> *input individual file*

evecoutname: **dutch_1kG.evec** -> *output file of PCs*

evaloutname: **dutch_1kG.eval** -> *output file of all eigenvalues*

numoutevec: **10** -> *number of PCs to output*

numoutlieriter: **0** -> *maximum number of outlier removal iterations (0 turns it off)*

poplistname: **poplist_1kG.txt** -> *file containing population value of individuals (If wishing to infer PCs using only individuals from a subset of populations, and then project to individuals from all other populations; will be used to detect individuals of non-European descent)*

snpweightoutname: **dutch_1kG.snpweight** -> *output file with SNP weightings of each PC*

Parameter file (.par)

- Let’s make the .par file. Run the following commands:

```
echo "genotypename: dutch_1kG.ped" >> dutch_1kG.par
echo "snpname: dutch_1kG.map" >> dutch_1kG.par
echo "indivname: dutch_1kG.fam" >> dutch_1kG.par
echo "evecoutname: dutch_1kG.evec" >> dutch_1kG.par
echo "evaloutname: dutch_1kG.eval" >> dutch_1kG.par
echo "numoutevec: 10" >> dutch_1kG.par
echo "numoutlieriter: 0" >> dutch_1kG.par
echo "poplistname: poplist_1kG.txt" >> dutch_1kG.par
echo "snpweightoutname: dutch_1kG.snpweight" >> dutch_1kG.par
```

- We also need to make the poplistname file (poplist_1kG.txt), containing the population values of the 1000 Genomes populations (4-13). Run the following command:

```
echo "4\n5\n6\n7\n8\n9\n10\n11\n12\n13" > poplist_1kG.txt
```

We’re ready to run EIGENSTRAT

- Run this command:

```
smartpca -p dutch_1kG.par > dutch_1kG.log
```

Let’s look at the PCs in R

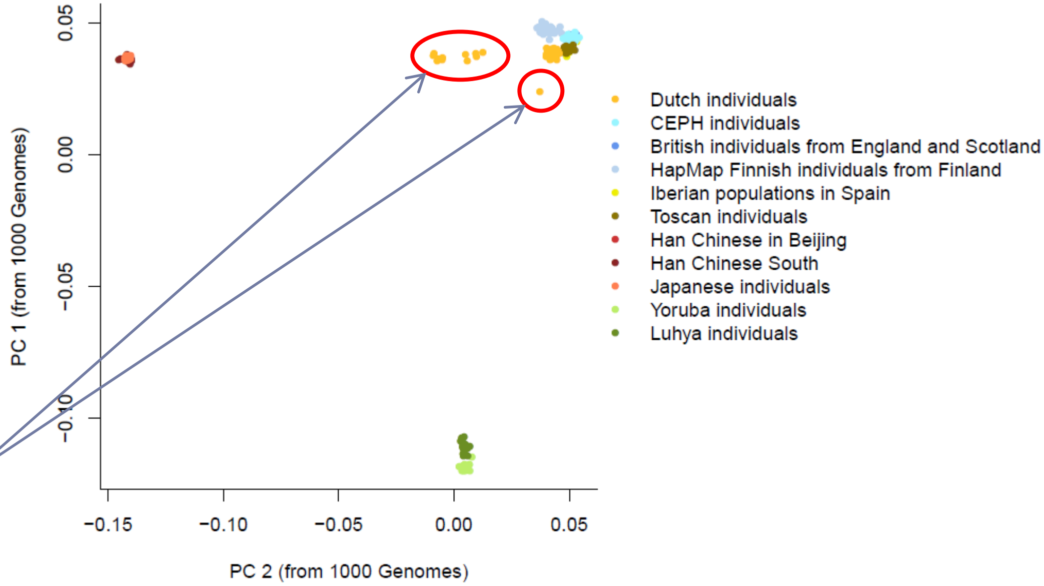
- First, let's make the file readable for R:

```
sed 's/:/ /g' dutch_1kG.evec > dutch_1kG.R.evec
```

- Run R script to make plot and identify outliers:

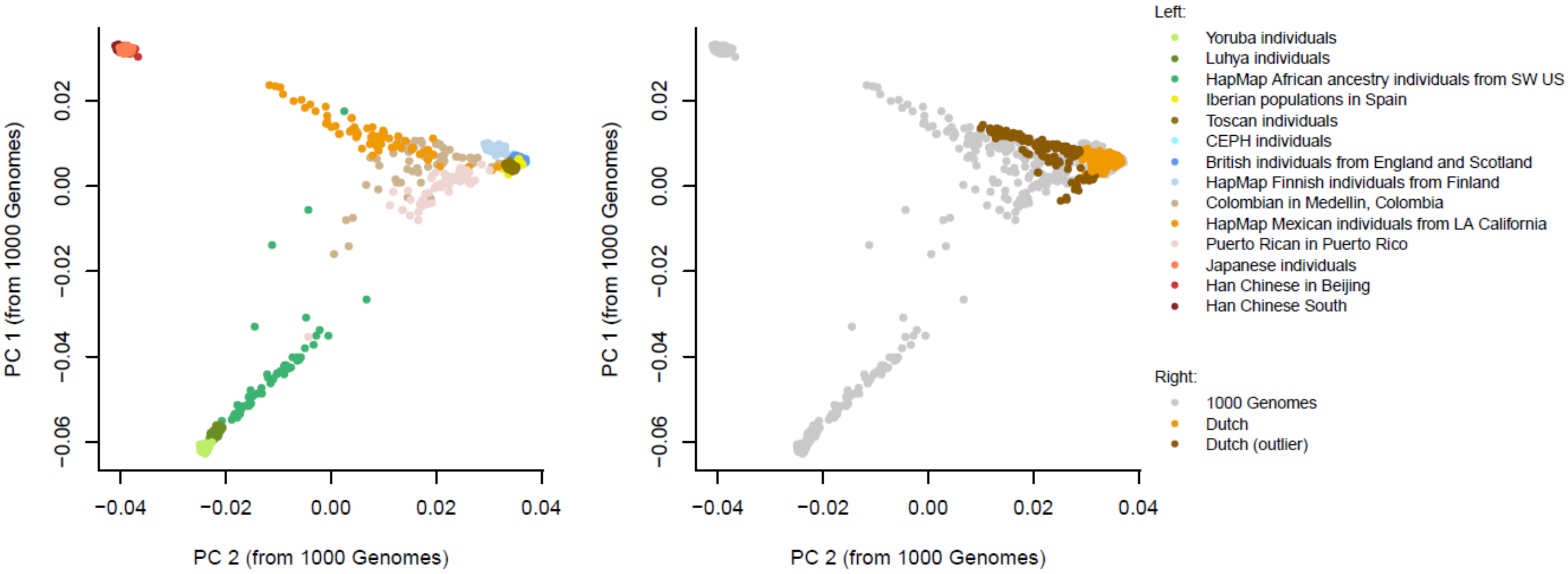
```
R CMD BATCH outliers.R
```

- What does the R script do? (open outliers.R)
  - Read in EIGENSTRAT file
  - Plot PC1 & PC2
  - Write IDs to file of Dutch individuals scoring higher than maximum European or lower than minimum European scores on PC1 or PC2 (to outliers.txt)



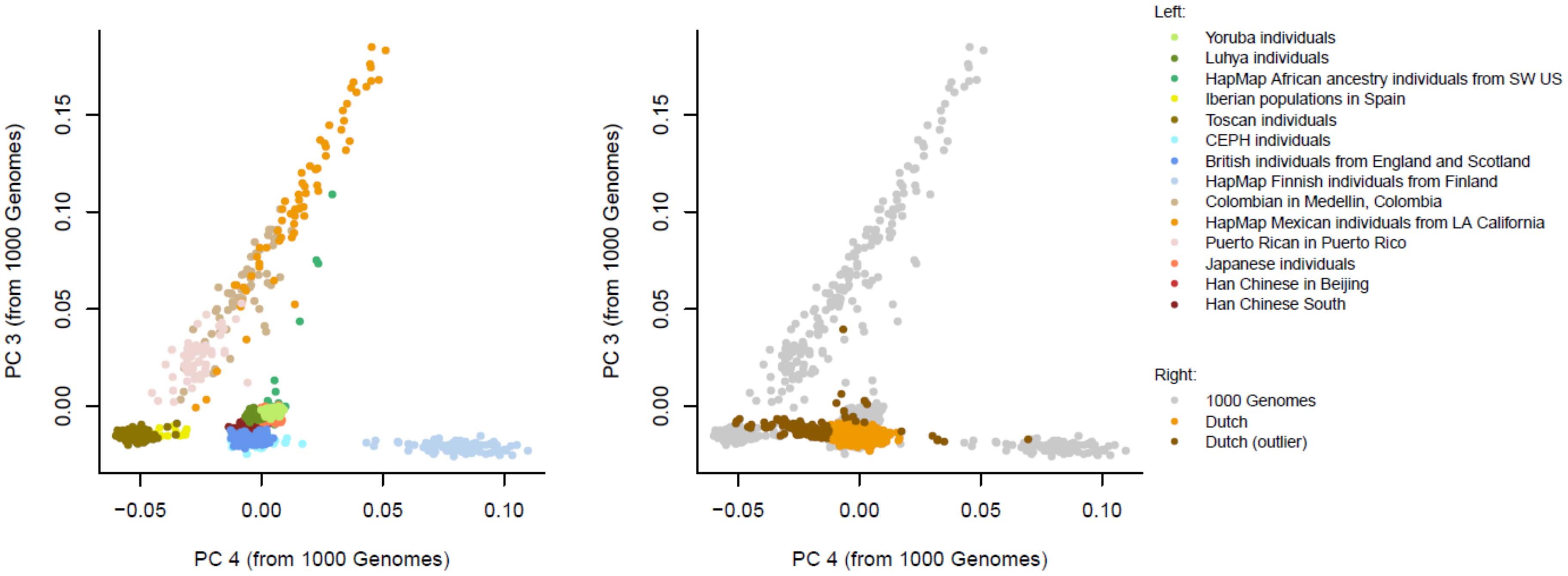
Identifying Dutch with non-European ancestry

- PCs were calculated using a set of 1014 unrelated individuals from 1000 Genomes, and were then projected on ~7500 Dutch individuals.
- 258 individuals were excluded. Parental birth place was available for 132 of these individuals, of which 55.3% had at least one parent born outside of the Netherlands (as opposed to 4% of the rest of the individuals).



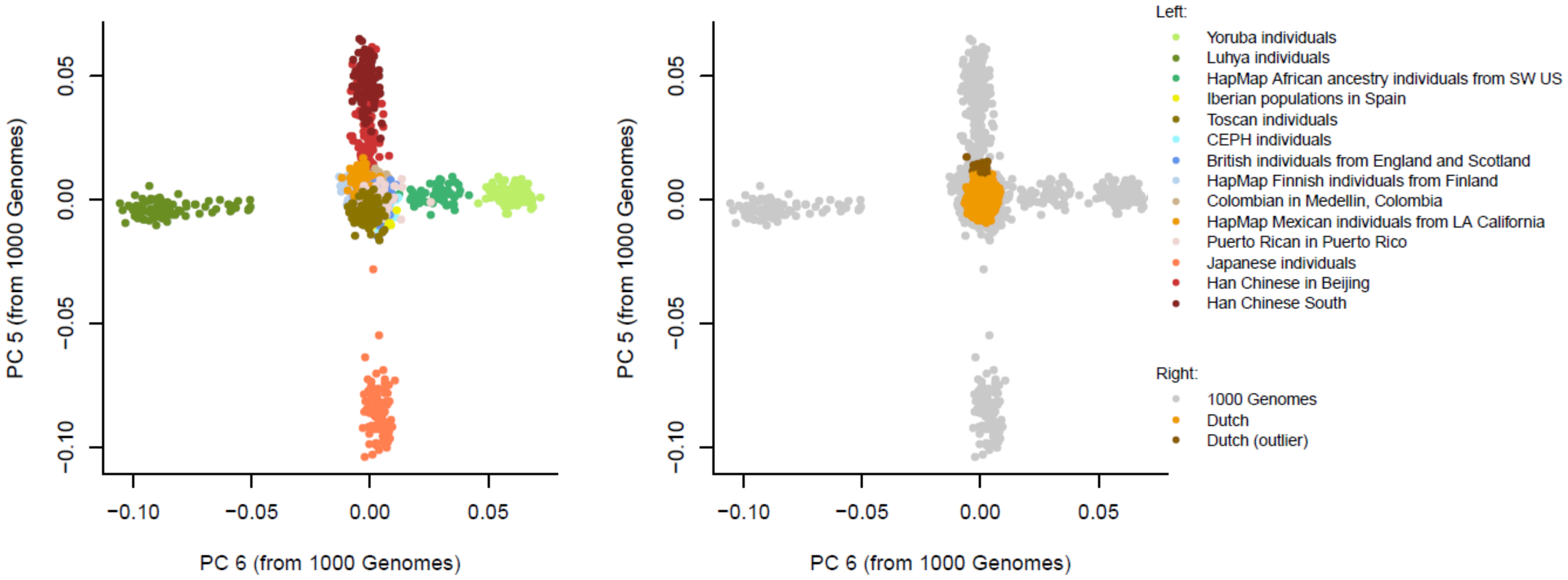
Identifying Dutch with non-European ancestry

- PCs were calculated using a set of 1014 unrelated individuals from 1000 Genomes, and were then projected on ~7500 Dutch individuals.
- 258 individuals were excluded. Parental birth place was available for 132 of these individuals, of which 55.3% had at least one parent born outside of the Netherlands (as opposed to 4% of the rest of the individuals).



Identifying Dutch with non-European ancestry

- PCs were calculated using a set of 1014 unrelated individuals from 1000 Genomes, and were then projected on ~7500 Dutch individuals.
- 258 individuals were excluded. Parental birth place was available for 132 of these individuals, of which 55.3% had at least one parent born outside of the Netherlands (as opposed to 4% of the rest of the individuals).



Exclude Dutch individuals with non-European ancestry and 1000 Genomes

```
awk '$6>3{print $1,$2}' dutch_1kG.fam > 1kG.ids
cat outliers.txt 1kG.ids > remove_outliers.ids

plink --bfile dutch_1kG --remove remove_outliers.ids --make-bed --out dutch
plink --bfile dutch --recode --out dutch
```

Parameter file (.par)

- Let’s make the .par file. Run the following commands:

```
echo "genotypename: dutch.ped" >> dutch.par
echo "snpname: dutch.map" >> dutch.par
echo "indivname: dutch.fam" >> dutch.par
echo "evecoutname: dutch.evec" >> dutch.par
echo "evaloutname: dutch.eval" >> dutch.par
echo "numoutevec: 10" >> dutch.par
echo "numoutlieriter: 0" >> dutch.par
echo "poplistname: poplist_NL.txt" >> dutch.par
echo "snpweightoutname: dutch.snpweight" >> dutch.par
```

- We also need to make the poplistname file (poplist_NL.txt). Run the following command:

```
echo "3" > poplist_NL.txt
```

We’re ready to run the 2nd round of EIGENSTRAT!

- Run this command:

```
smartpca -p dutch.par > dutch.log
```

Let’s plot the first two PCs in R

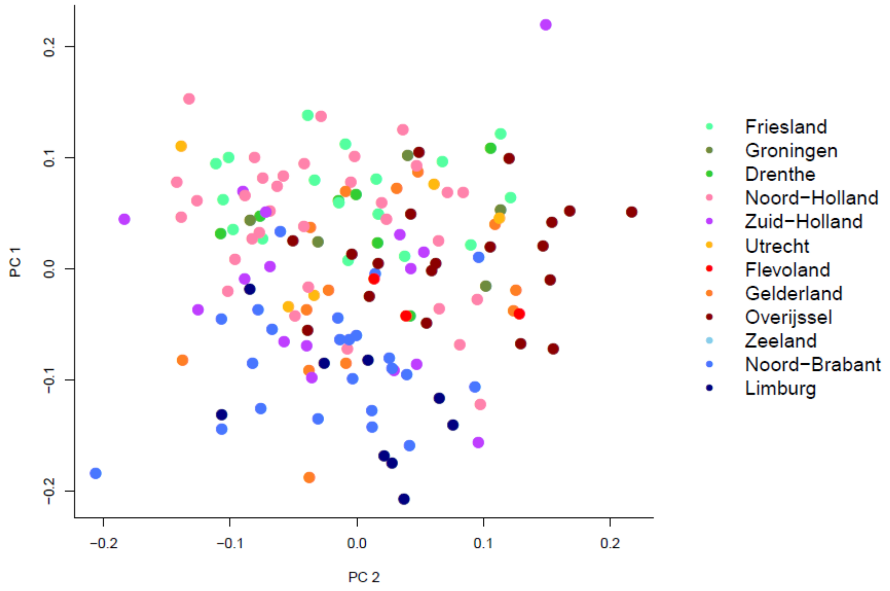
- First, let's make the file readable for R:

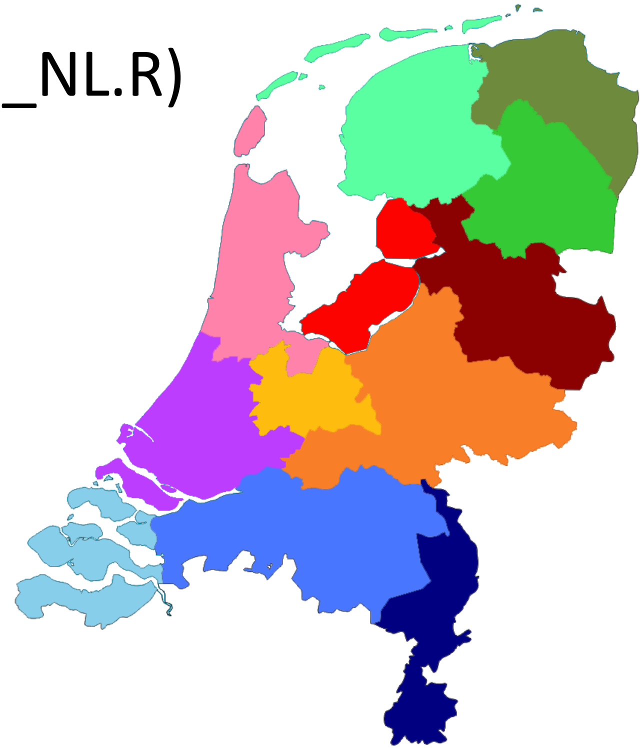
```
sed 's/:/ /g' dutch.evec > dutch.R.evec
```

- Run R script to make plot :

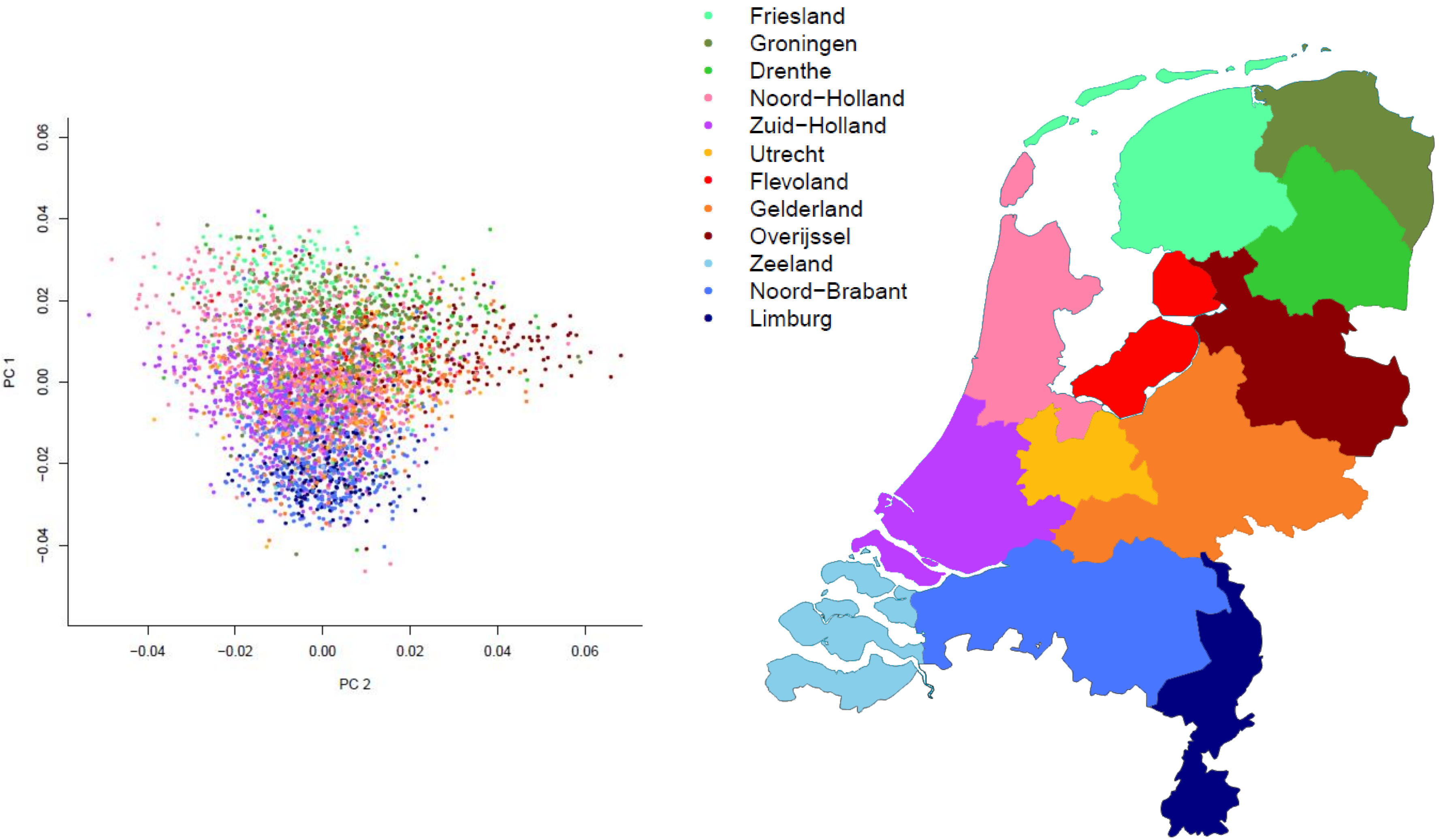
```
R CMD BATCH plot_NL.R
```

- What does the R script do? (open plot_NL.R)

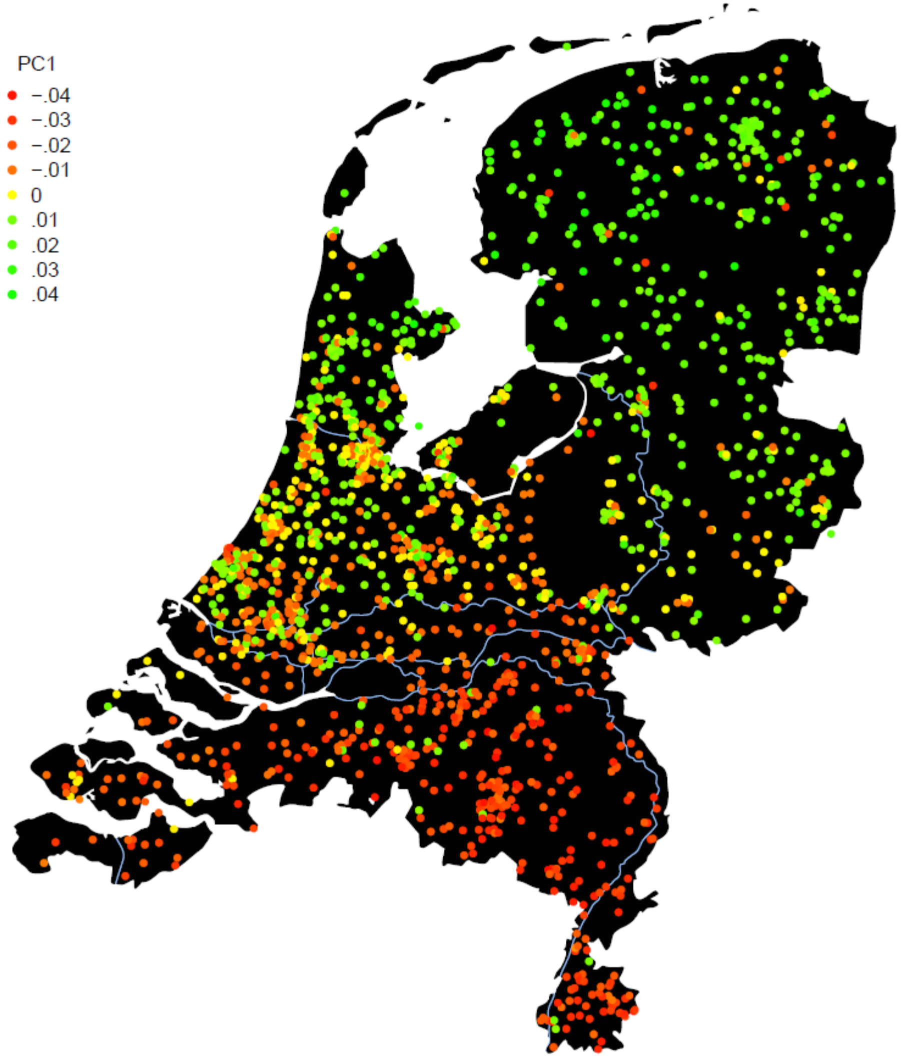




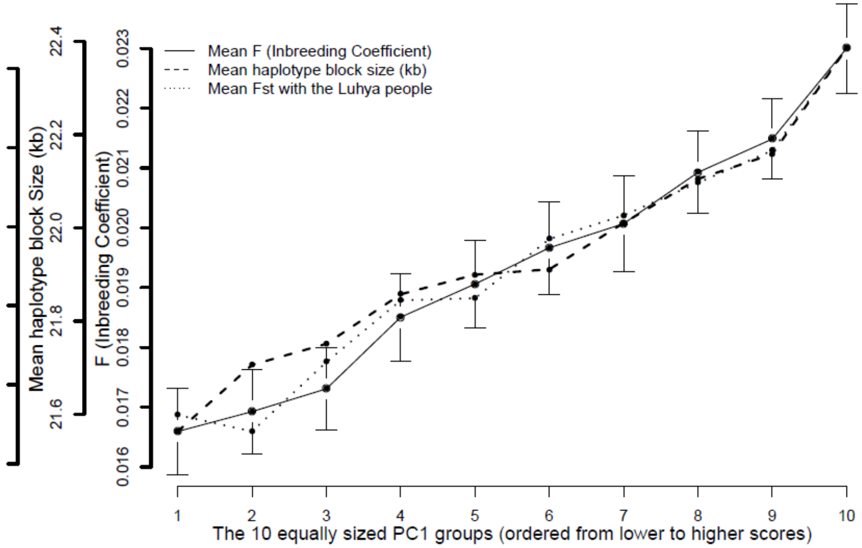
Same plot, with N=4,441



PC1 (N=4,441)



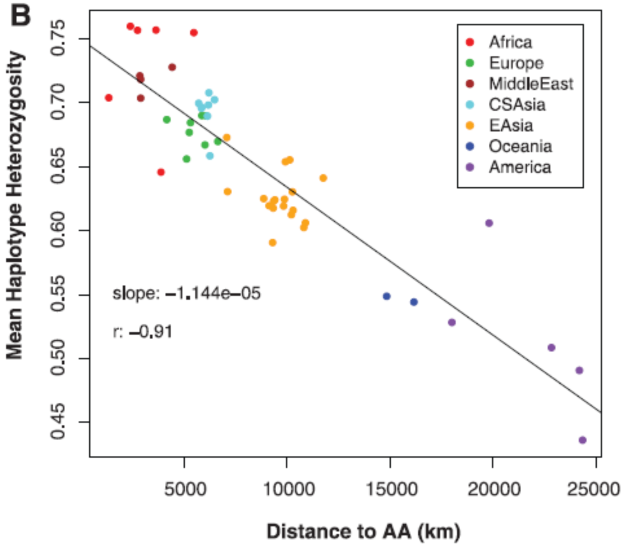
- $r_{\updownarrow}$ = .588
- Correlates .656 with European North-South PC.
- Spouse correlation = .555
- Serial founder effect? (correlation with F: .245)



Serial founder effect: heterozygosity decreases (F increases) as you move away from Addis Ababa, Ethopia

Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation

Jun Z. Li,[1,2]*† Devin M. Absher,[1,2]* Hua Tang,[1] Audrey M. Southwick,[1,2] Amanda M. Casto,[1]
Sohini Ramachandran,[4] Howard M. Cann,[5] Gregory S. Barsh,[1,3] Marcus Feldman,[4]‡
Luigi L. Cavalli-Sforza,[1]‡ Richard M. Myers[1,2]‡



We compared SNP haplotype heterozygosity across populations and found, consistent with earlier reports (*22*), that it is highest in sub-Saharan Africa and decreases steadily with distance from this region (Fig. 3B). The mean heterozygosity across autosomal haplotypes (using 295 haplotype blocks in Chr16) (*14*) is negatively correlated with distance from Addis Ababa, Ethiopia (*5*, *23*), with a correlation coefficient r of –0.91 and a slope of -1.1×10^{-5} per km (Fig. 3B). This trend is consistent with a serial founder effect, a scenario in which population expansion involves successive migration of a small fraction of individuals out of the previous location, starting from a single origin in sub-Saharan Africa.

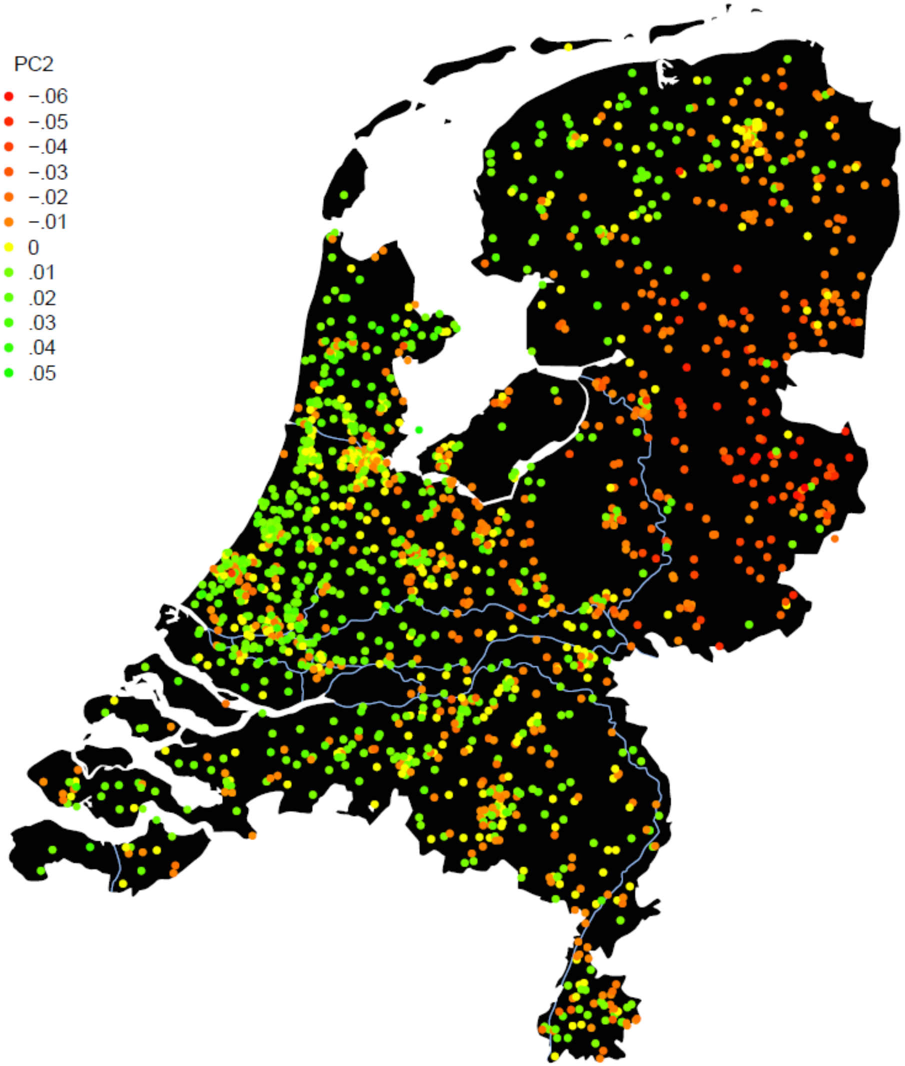
22 FEBRUARY 2008 VOL 319 **SCIENCE** www.sciencemag.org

- Genome-wide homozygosity (F) can be computed in plink with *--het*

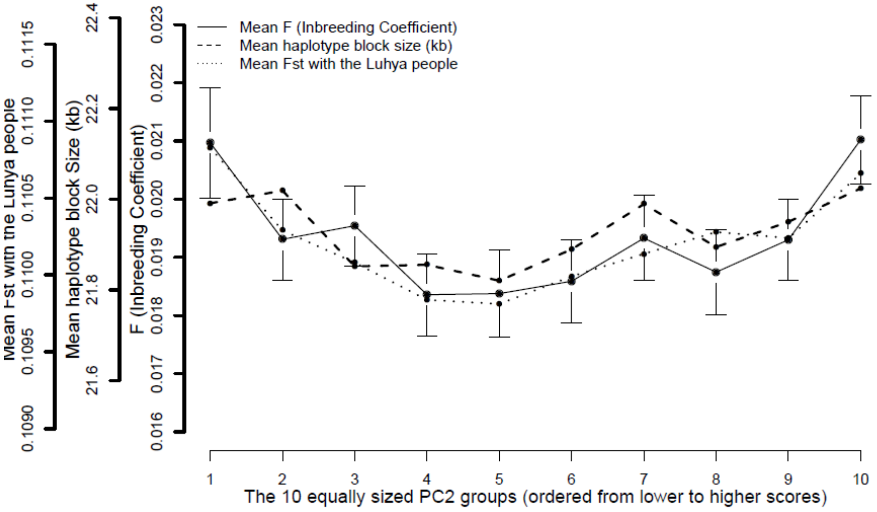
Height

- Northern Dutch are known to be taller on average than the Dutch from the Southern parts of the Netherlands. Also within Europe, Northern Europeans are taller on average than Southern Europeans.
- In our sample, height does not correlate very high with the North-South gradient of the current living address:
  - males: *r* = .036, *p* = .232; females: *r* = .050, *p* = .020
- Height however correlates higher and more significantly with the North-South PC in both genders:
  - males: *r* = .142, *p* < .001; females: *r* = .153, *p* < .001
- This confirms that the PC is a better measure for ancestral origin than the geographical location, and that the height differences are indeed genetic.

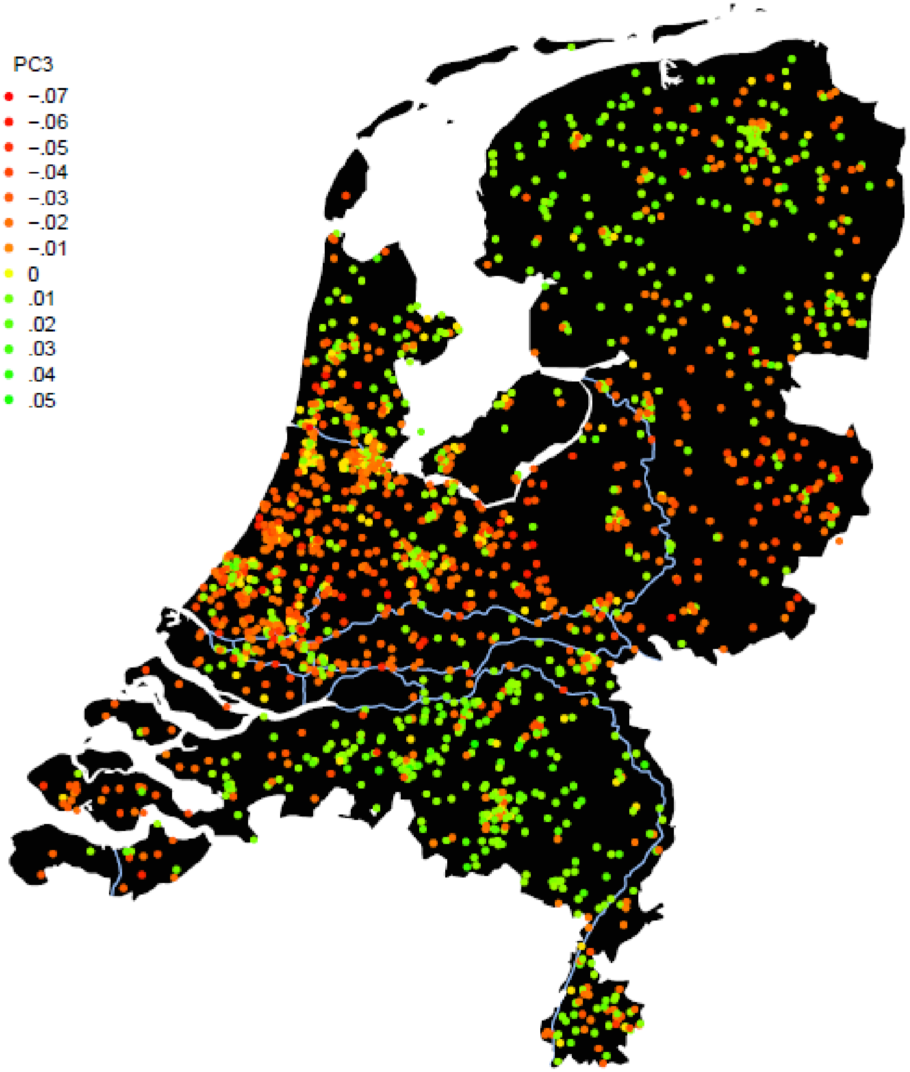
PC2 (N=4,441)



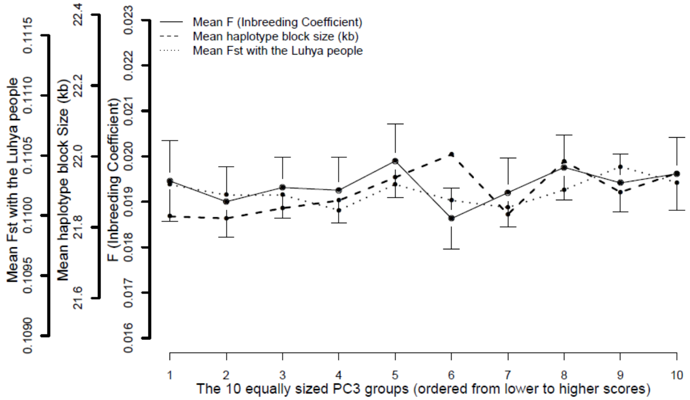
- $r_{\leftrightarrow}$ = .369
- Spouse correlation = .164



PC3 (N=4,441)

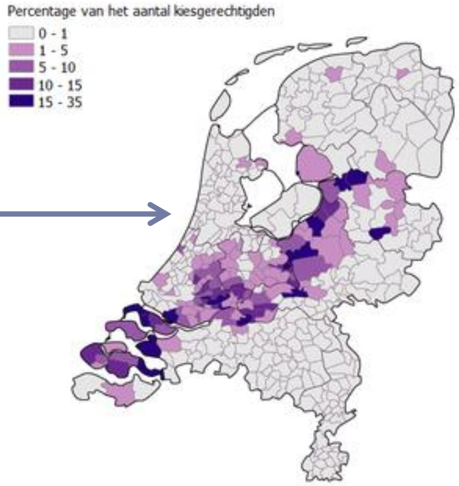


- Was only observed with minimized LD
- Spouse correlation = .174



- Bible belt?

Votes for conservative Christian party (SGP)



Using PCs to identify loci under selection

- Bayescan 2.1 was used to calculate F_{st} values for all SNPs and identify outliers with a Bayesian approach
- F_{st}'s were computed between top 1000 and bottom 1000 individuals for each ancestry-informative PC
- F_{st} is then decomposed into 2 components:
  - population-specific component (β), shared by all loci
  - locus-specific component (α), shared by both populations
- If α is significantly different from 0, the locus may have been under selection:
  - $\alpha > 0$ = diversifying selection
  - $\alpha < 0$ = balancing selection (power to detect this is weak)
- Significance is based on FDR corrected q-value (< .05)

Using PCs to identify loci under selection: results

- 499,849 SNPs in total (51.4% within genes):
  - PC1 (North-South): 273 significant SNPs (59% within 88 genes)
  - PC2 (East-West): 172 significant SNPs (58.1% within 55 genes)
  - PC3 (Middle-Band): 100 significant SNPs (75% within 41 genes)

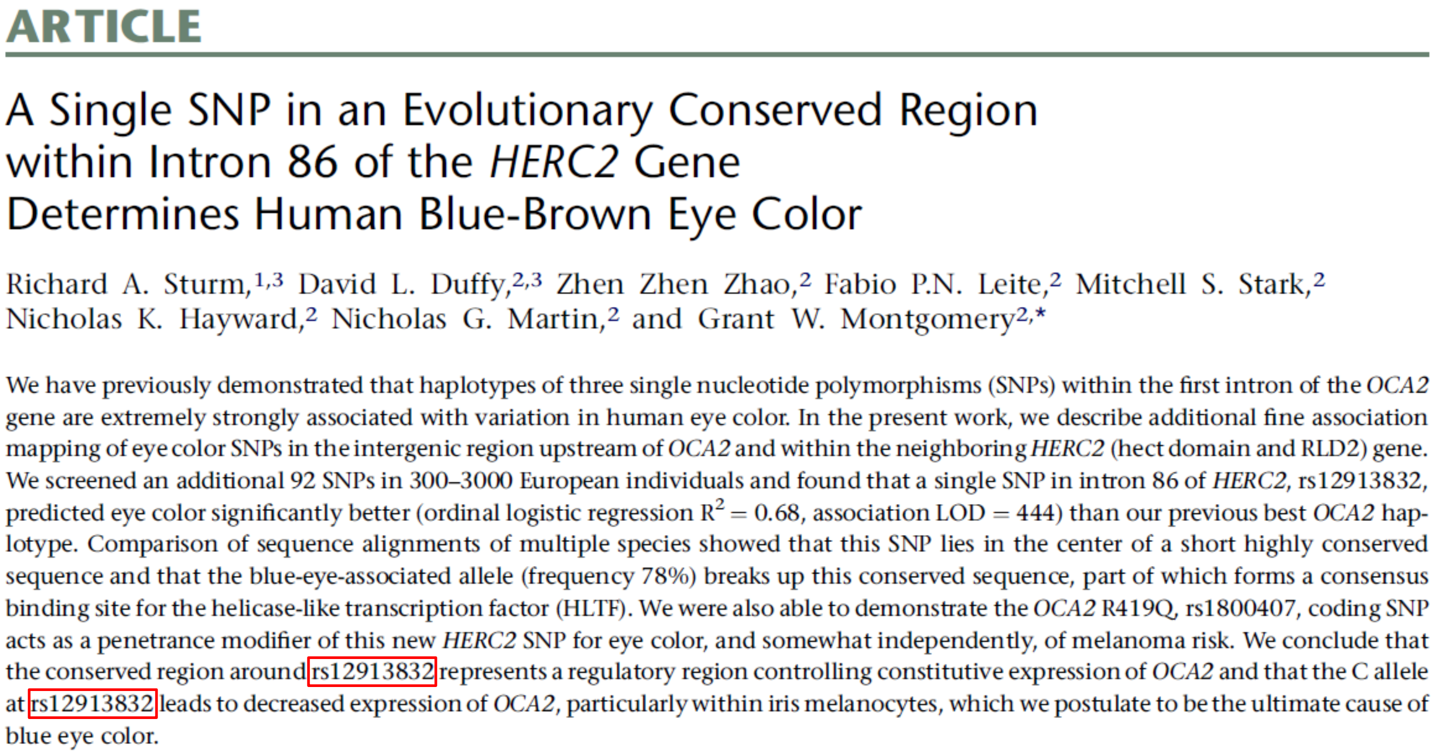
- Several of the genes with significant SNPs have been observed to be strongly differentiated within Europe in previous studies:
  - *LCT* (PC1), *HERC2* (PC1), *CADPS* (PC1), *IRF1* (PC1), *SLC44A5* (PC1), *R3HDM1* (PC1), *ACOXL* (PC3), and *BTBD9* (PC3)

HERC2 & eye color

- Highest F_{st} observed in PC1 for SNP in HERC2 gene (rs8039195). Strongly associated with eye color in several GWASs ($p = 7.8 \times 10^{-112}$ in current dataset).
- F_{st}'s were calculated for 3495 SNPs in and around HERC2 between Northern European populations (British and Finnish) and Southern European populations (Iberian and Toscan) from 1000 Genomes.

Population	rs8039195 (HERC2)		
	CC	CT	TT
Finnish	.0	6.5	93.5
Northern Dutch	.4	13.1	86.5
British	1.2	21.4	77.4
Southern Dutch	2.3	23.9	73.7
Iberian	.0	50.0	50.0
Toscan	16.8	42.1	41.1

- Of the SNPs genotyped in the Dutch, rs8039195 had the highest F_{st}.
- Of all 3495 SNPs, highest F_{st} was observed for rs12913832 (LD with rs8039195: $r^2 = .394$, $D' = .993$), the SNP with the largest effect on human blue/brown eye color.

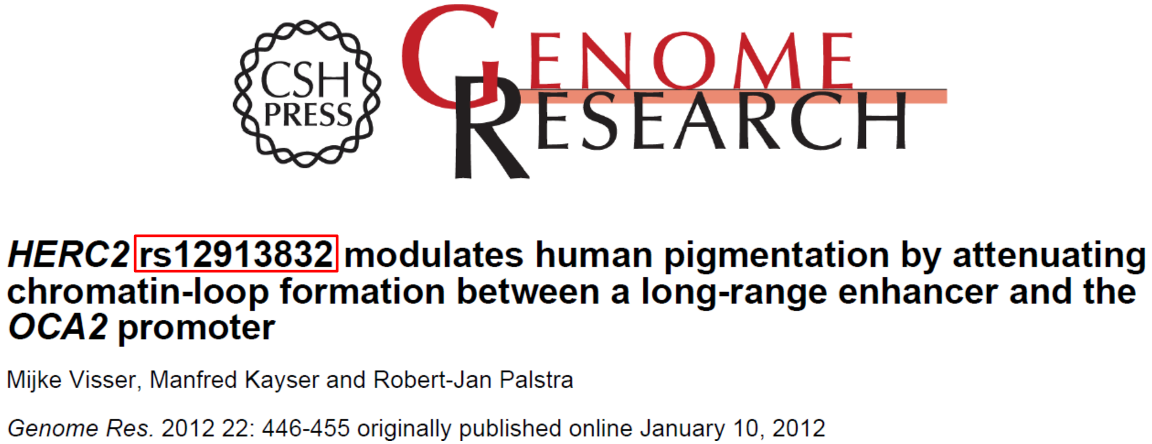


ARTICLE

A Single SNP in an Evolutionary Conserved Region within Intron 86 of the *HERC2* Gene Determines Human Blue-Brown Eye Color

Richard A. Sturm,[1,3] David L. Duffy,[2,3] Zhen Zhen Zhao,[2] Fabio P.N. Leite,[2] Mitchell S. Stark,[2] Nicholas K. Hayward,[2] Nicholas G. Martin,[2] and Grant W. Montgomery[2,*]

We have previously demonstrated that haplotypes of three single nucleotide polymorphisms (SNPs) within the first intron of the *OCA2* gene are extremely strongly associated with variation in human eye color. In the present work, we describe additional fine association mapping of eye color SNPs in the intergenic region upstream of *OCA2* and within the neighboring *HERC2* (hect domain and RLD2) gene. We screened an additional 92 SNPs in 300–3000 European individuals and found that a single SNP in intron 86 of *HERC2*, rs12913832, predicted eye color significantly better (ordinal logistic regression $R^2 = 0.68$, association LOD = 444) than our previous best *OCA2* haplotype. Comparison of sequence alignments of multiple species showed that this SNP lies in the center of a short highly conserved sequence and that the blue-eye-associated allele (frequency 78%) breaks up this conserved sequence, part of which forms a consensus binding site for the helicase-like transcription factor (HLTF). We were also able to demonstrate the *OCA2* R419Q, rs1800407, coding SNP acts as a penetrance modifier of this new *HERC2* SNP for eye color, and somewhat independently, of melanoma risk. We conclude that the conserved region around rs12913832 represents a regulatory region controlling constitutive expression of *OCA2* and that the C allele at rs12913832 leads to decreased expression of *OCA2*, particularly within iris melanocytes, which we postulate to be the ultimate cause of blue eye color.



CSH PRESS GENOME RESEARCH

HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the *OCA2* promoter

Mijke Visser, Manfred Kayser and Robert-Jan Palstra

Genome Res. 2012 22: 446-455 originally published online January 10, 2012

Using PCs to identify loci under selection: results

- Significant signals were observed within genes that play major roles in brain function, such as:
  - GRM7 (PC1; encodes a metabotropic glutamate receptor)
  - GRIN2A (PC1; encodes a subunit for the NMDA receptor)
  - BDNF (PC2; encodes the brain-derived neurotrophic factor)
  - SLC6A4 (PC3; a.k.a. SERT, encodes the serotonin transporter)
  - AUTS2 (PC3; autism susceptibility candidate 2)

- When including genes from all 3 PCs in Ingenuity Pathway Analysis, the top 11 biological functions are brain related ($p \leq 1.26 \times 10^{-4}$) with the most significant being *neurotransmission of nervous tissue* with 11 molecules and $p = 2.2 \times 10^{-6}$.

Using PCs to identify loci under selection: results

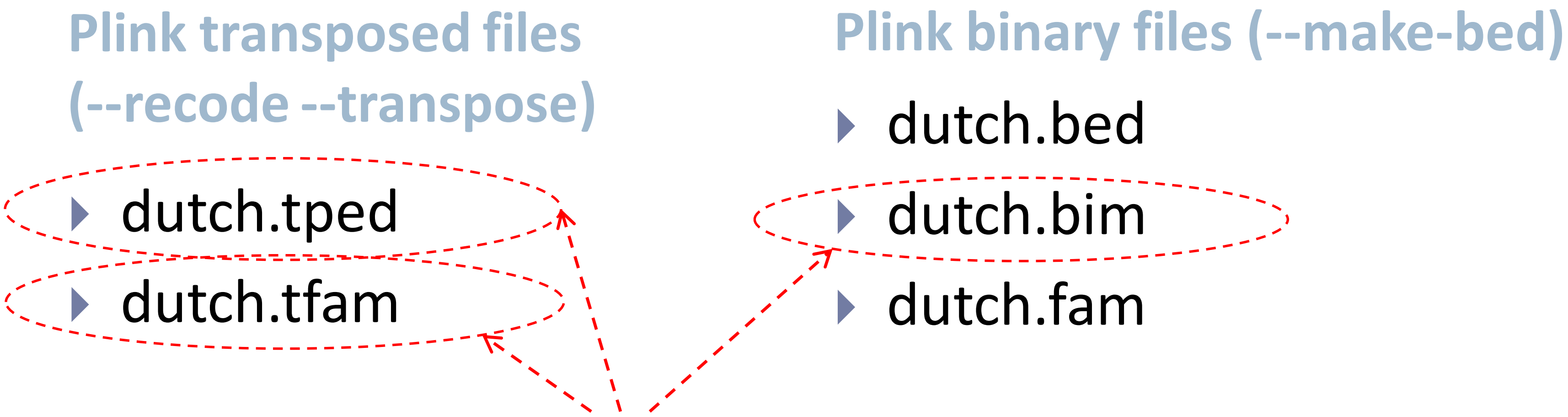
- Other notable genes include:
  - FTO (PC1): has been associated with obesity many times.
  - LCT (PC1): influences the ability to digest lactose into adulthood
  - HCP5 (HLA Complex P5 gene) from the MHC region. One of two genes that appear in multiple PCs (PC1 & PC2), and plays a role in the immune system.

    Strong divergence of genes from the HLA complex has been observed in many human populations.

    Other immunity-related genes that showed significant signals of selection in this study as well as previous studies are: *IRF1* (PC1), *ACE* (PC1), *LRRC4C* (PC2), *PLCL1* (PC3), and *HSPD1* (PC3).

- Bayescan can be found here: http://cmpg.unibe.ch/software/bayescan/

Converting plink files to Bayescan format with the script convert_to_bayescan.pl



Plink transposed files (--recode --transpose)

- dutch.tped
- dutch.tfam

Plink binary files (--make-bed)

- dutch.bed
- dutch.bim
- dutch.fam

convert_to_bayescan.pl needs

- The populations you want to compare have to be coded as 1 and 2 in the phenotype column (6th column) of the .tfam file.
- Use --pheno to update phenotypes : http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#pheno

Usage:

```
perl convert_to_bayescan.pl dutch dutch_outputfile
```