Efficient algorithms for ascertaining markers for controlling for population substructure

Oscar Lao
Department of Forensic Molecular Biology
Erasmus Medical Center (Rotterdam)

New Jersey 2009

Workflow

1. Human population substructure

   - How to detect it?

   - How much?

   - Where does it come from?

2. Why does it matter?

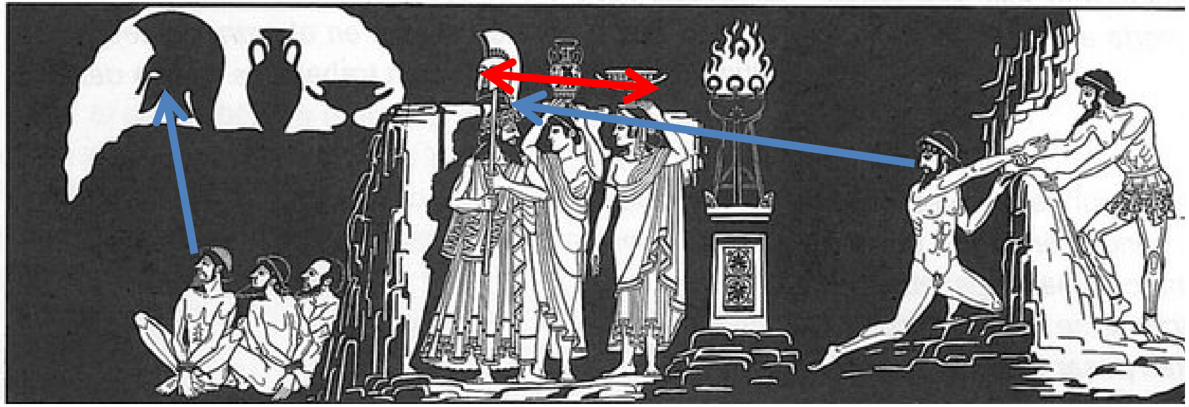
3. Ancestry Sensitive Markers (ASMs) / Ancestry Informative Markers (AIMs)

   - Hypothesis driven. Particular individual clusters are preferred
     - ASMs
     - PhenoASMs

Population substructure

How much there is and how much can be detected. The two sides of the same coin

Plato's cave myth



Population substructure

DETECTION

- – STRUCTURE
- – BAPS
- – FRAPPE
- – GENELAND
- – PCA/MDS + K means
- – Neural Networks
- – …

Sometimes results are NOT reproducible

Population substructure

HOW MUCH?

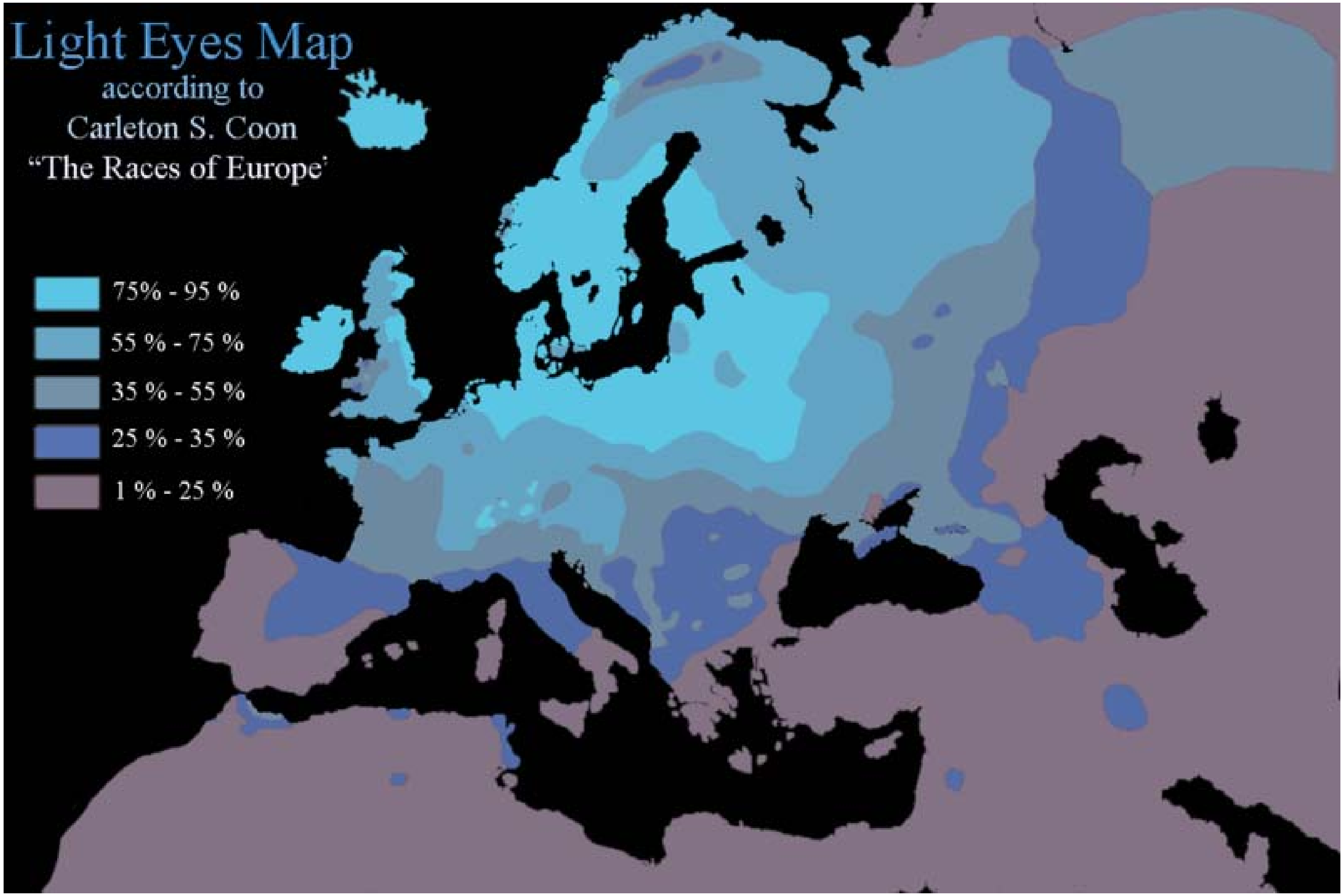
Which type?

- Phenotype
- Genotype
  - Y chromosome
  - mtDNA
  - Autosomal markers

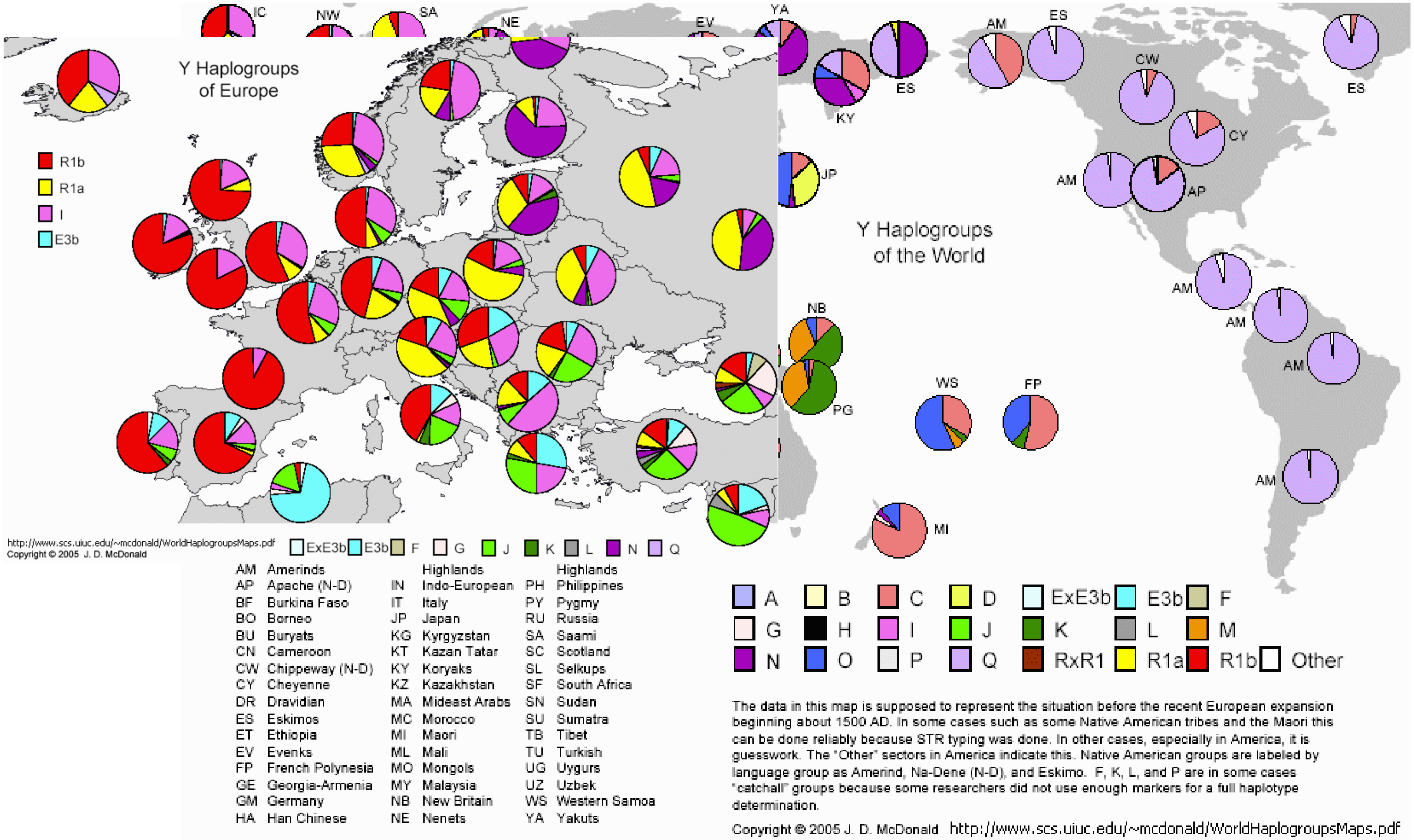
Where?

- Worldwide
- Regional (I will focus on Europe)

Phenotypic substructure



Y chromosome



Y Haplogroups of Europe

Y Haplogroups of the World

🟥 R1b
🟨 R1a
▨ I
▨ E3b
□ ExE3b ▨ E3b ▨ F □ G 🟩 J ▨ K ▨ L ▨ N ▨ Q
IC
NW
SA
NE
EV
YA
ES
KY
AM
CW
CY
AP
JP
NB
PG
WS
FP
MI
□ A □ B ▨ C ▨ D □ ExE3b ▨ E3b ▨ F
□ G ■ H ▨ I 🟩 J ▨ K ▨ L ▨ M
▨ N ▨ O □ P ▨ Q ▨ RxR1 🟨 R1a 🟥 R1b □ Other

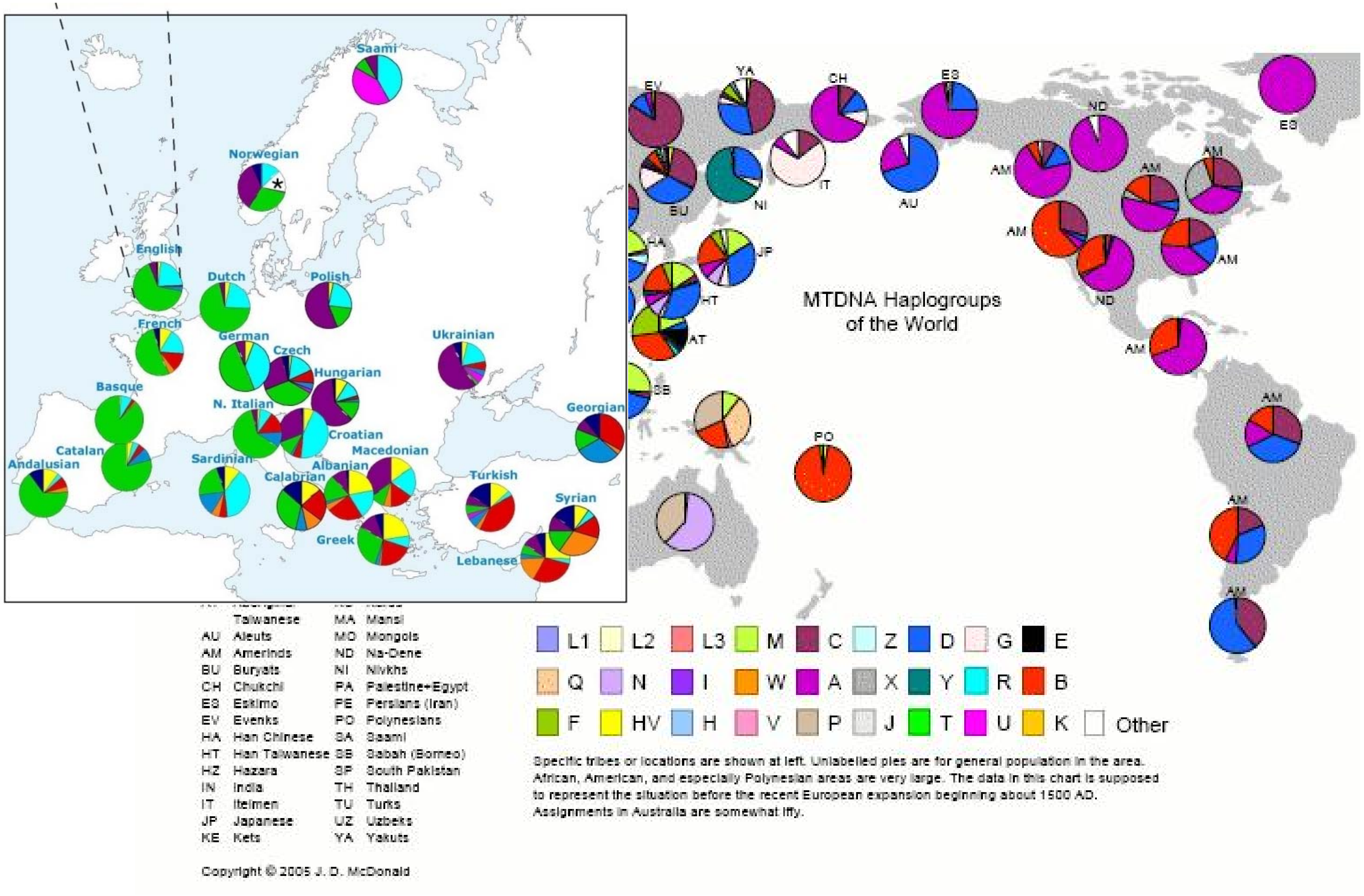
http://www.scs.uiuc.edu/~mcdonald/WorldHaplogroupsMaps.pdf
Copyright © 2005 J. D. McDonald

AM	Amerinds		Highlands		Highlands
AP	Apache (N-D)	IN	Indo-European	PH	Philippines
BF	Burkina Faso	IT	Italy	PY	Pygmy
BO	Borneo	JP	Japan	RU	Russia
BU	Buryats	KG	Kyrgyzstan	SA	Saami
CN	Cameroon	KT	Kazan Tatar	SC	Scotland
CW	Chippeway (N-D)	KY	Koryaks	SL	Selkups
CY	Cheyenne	KZ	Kazakhstan	SF	South Africa
DR	Dravidian	MA	Mideast Arabs	SN	Sudan
ES	Eskimos	MC	Morocco	SU	Sumatra
ET	Ethiopia	MI	Maori	TB	Tibet
EV	Evenks	ML	Mali	TU	Turkish
FP	French Polynesia	MO	Mongols	UG	Uygurs
GE	Georgia-Armenia	MY	Malaysia	UZ	Uzbek
GM	Germany	NB	New Britain	WS	Western Samoa
HA	Han Chinese	NE	Nenets	YA	Yakuts

The data in this map is supposed to represent the situation before the recent European expansion beginning about 1500 AD. In some cases such as some Native American tribes and the Maori this can be done reliably because STR typing was done. In other cases, especially in America, it is guesswork. The "Other" sectors in America indicate this. Native American groups are labeled by language group as Amerind, Na-Dene (N-D), and Eskimo. F, K, L, and P are in some cases "catchall" groups because some researchers did not use enough markers for a full haplotype determination.

Copyright © 2005 J. D. McDonald http://www.scs.uiuc.edu/~mcdonald/WorldHaplogroupsMaps.pdf

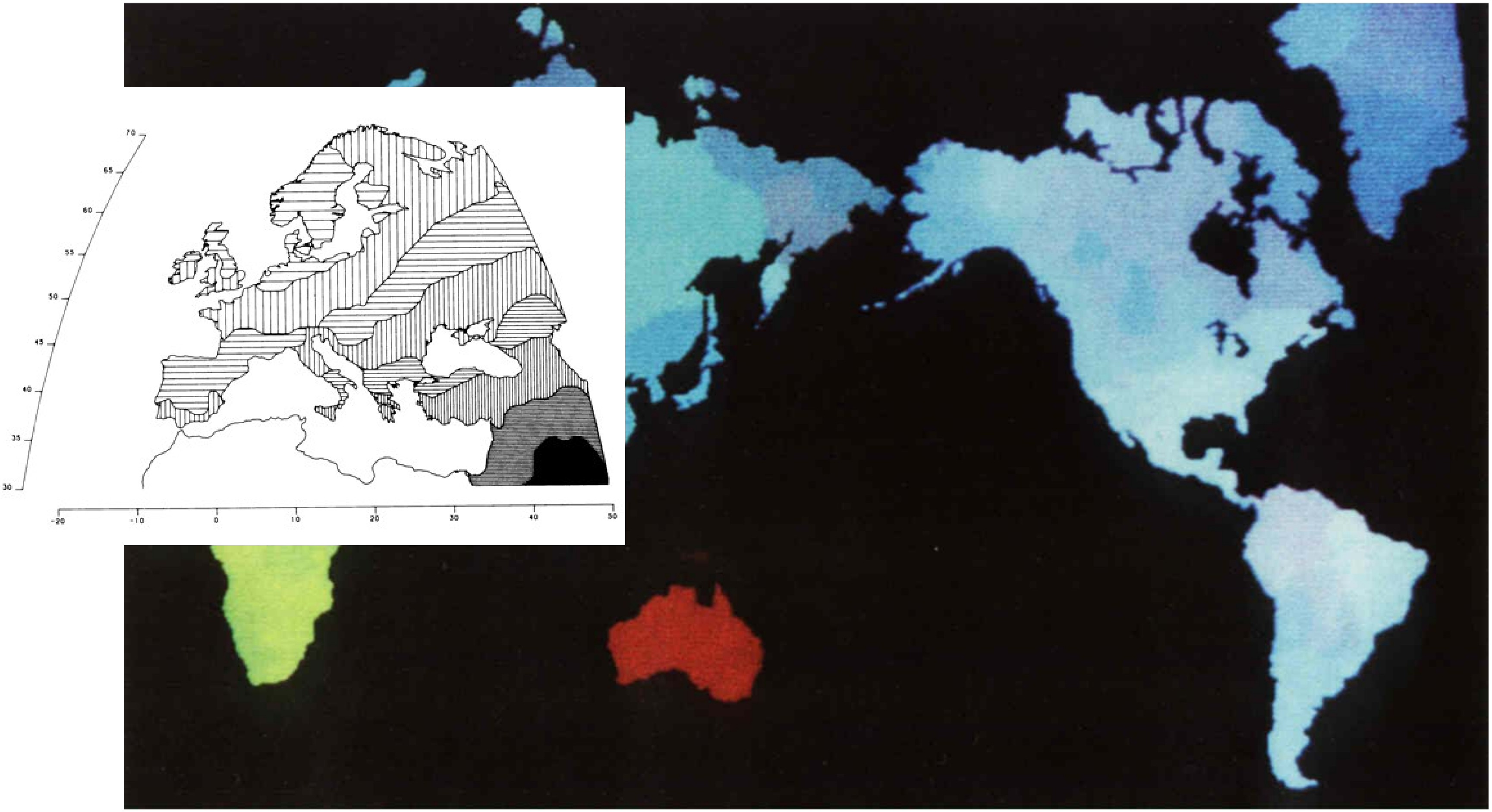
mtDNA



	Taiwanese	MA	Mansi
AU	Aleuts	MO	Mongols
AM	Amerinds	ND	Na-Dene
BU	Buryats	NI	Nivkhs
CH	Chukchi	PA	Palestine+Egypt
ES	Eskimo	PE	Persians (Iran)
EV	Evenks	PO	Polynesians
HA	Han Chinese	SA	Saami
HT	Han Taiwanese	SB	Sabah (Borneo)
HZ	Hazara	SP	South Pakistan
IN	India	TH	Thailand
IT	Itelmen	TU	Turks
JP	Japanese	UZ	Uzbeks
KE	Kets	YA	Yakuts

Specific tribes or locations are shown at left. Unlabelled pies are for general population in the area. African, American, and especially Polynesian areas are very large. The data in this chart is supposed to represent the situation before the recent European expansion beginning about 1500 AD. Assignments in Australia are somewhat iffy.

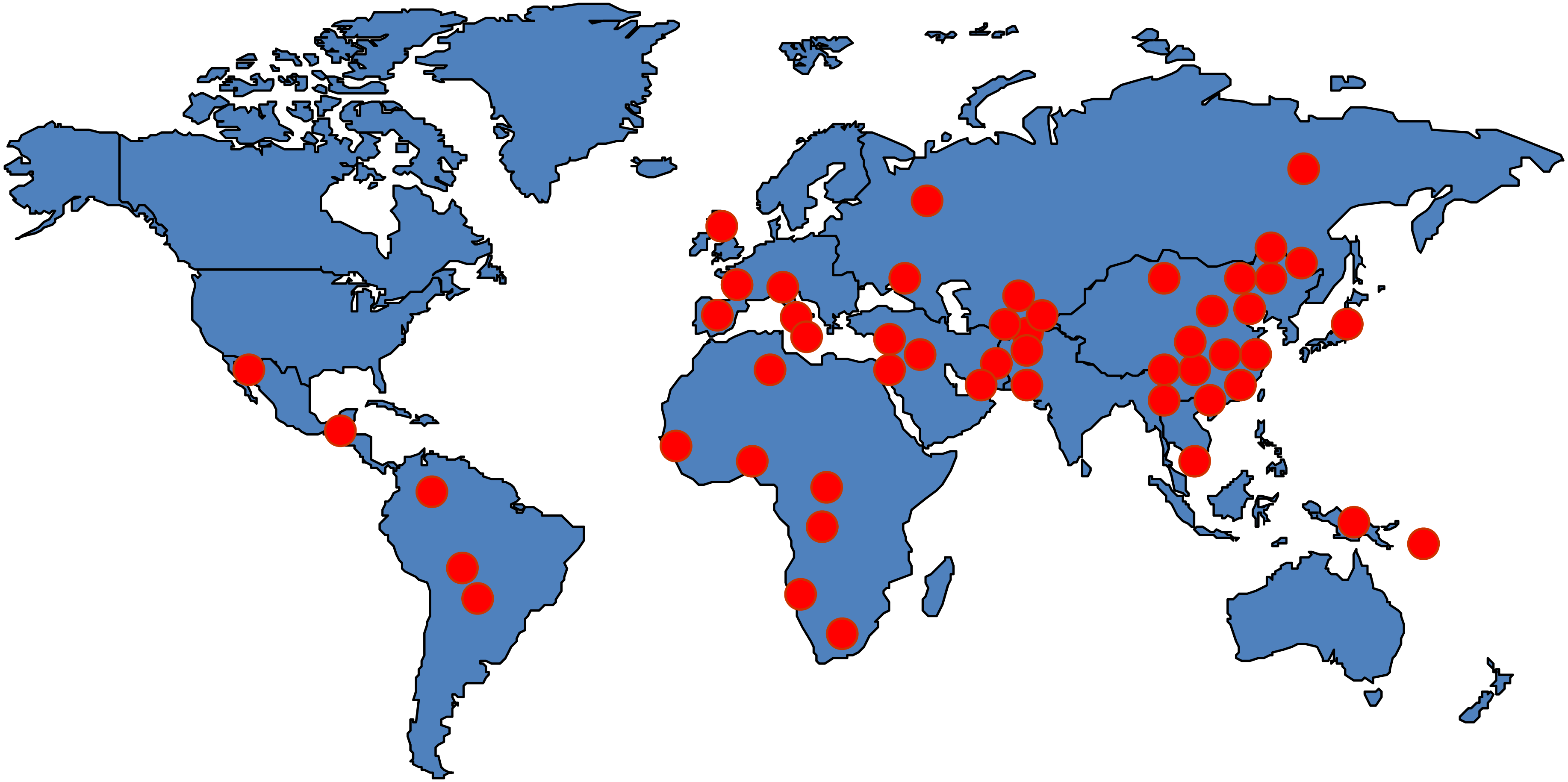
Classical markers



Cavalli-Sforza et al 1994

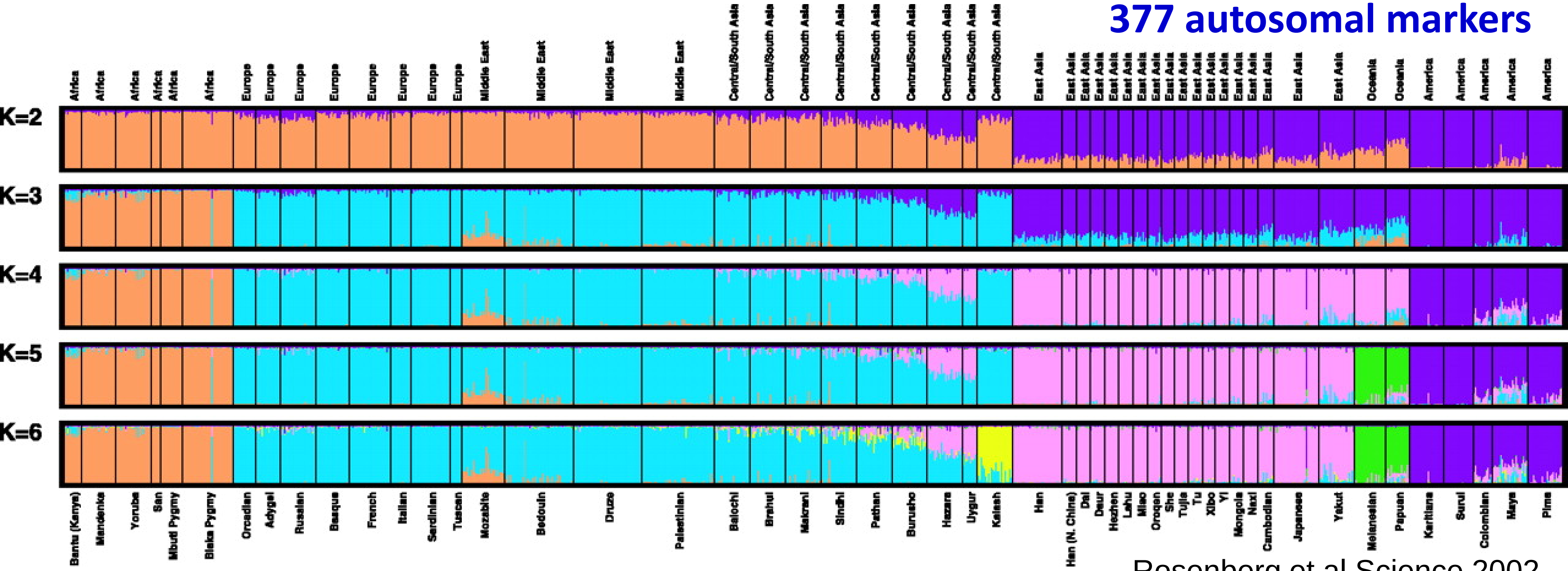
CEPH-HGDP panel

1064 samples
51 human populations of global distribution



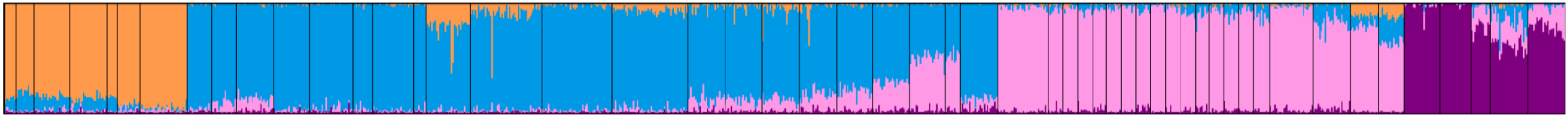
Autosomal STRs

377 autosomal markers



Rosenberg et al Science 2002

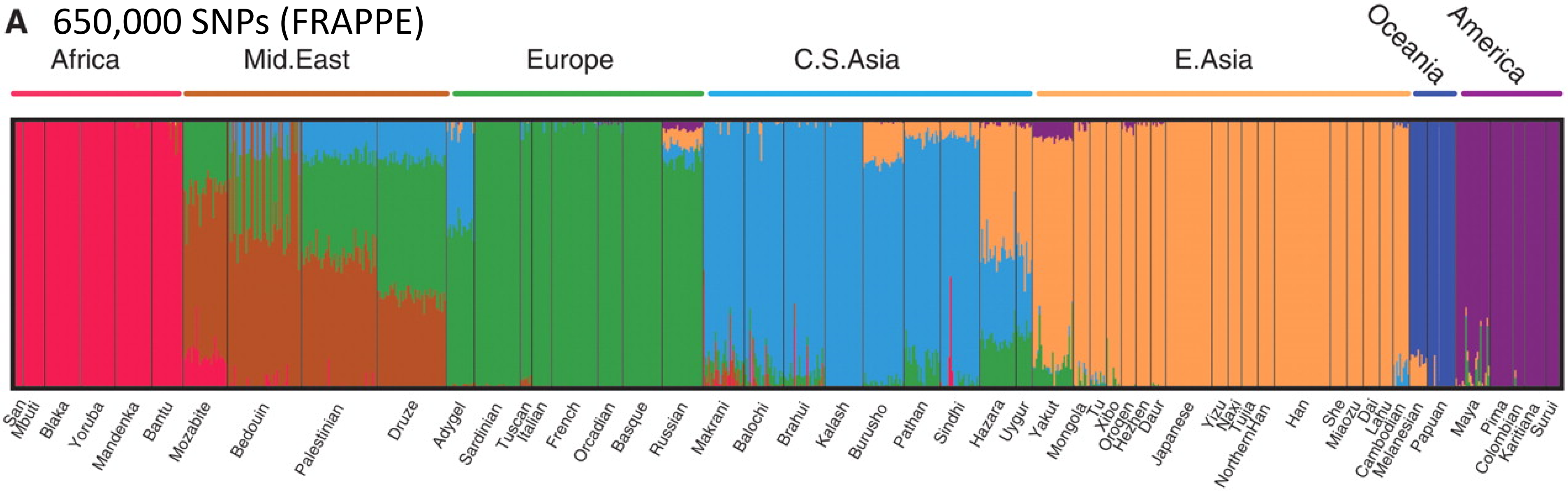
993 autosomal markers



Africa | Europe | M. East | C/S-Asia | E-Asia | O | Ameri

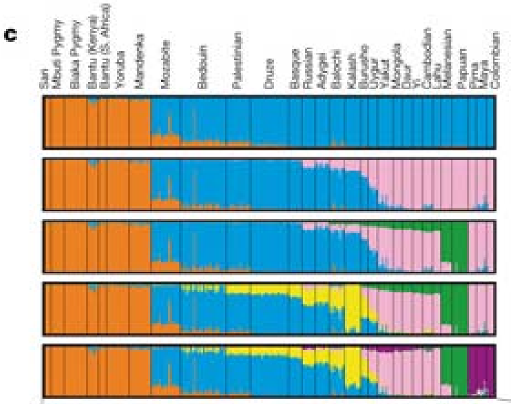
Rosenberg et al Plos Genetics 2005

Autosomal SNPs

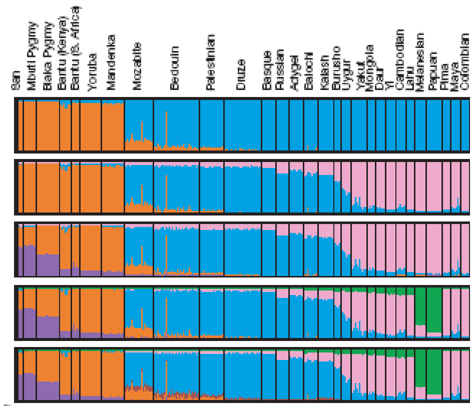

650,000 SNPs (FRAPPE)

Li et al Science 2008

550,000 SNPs (STRUCTURE)

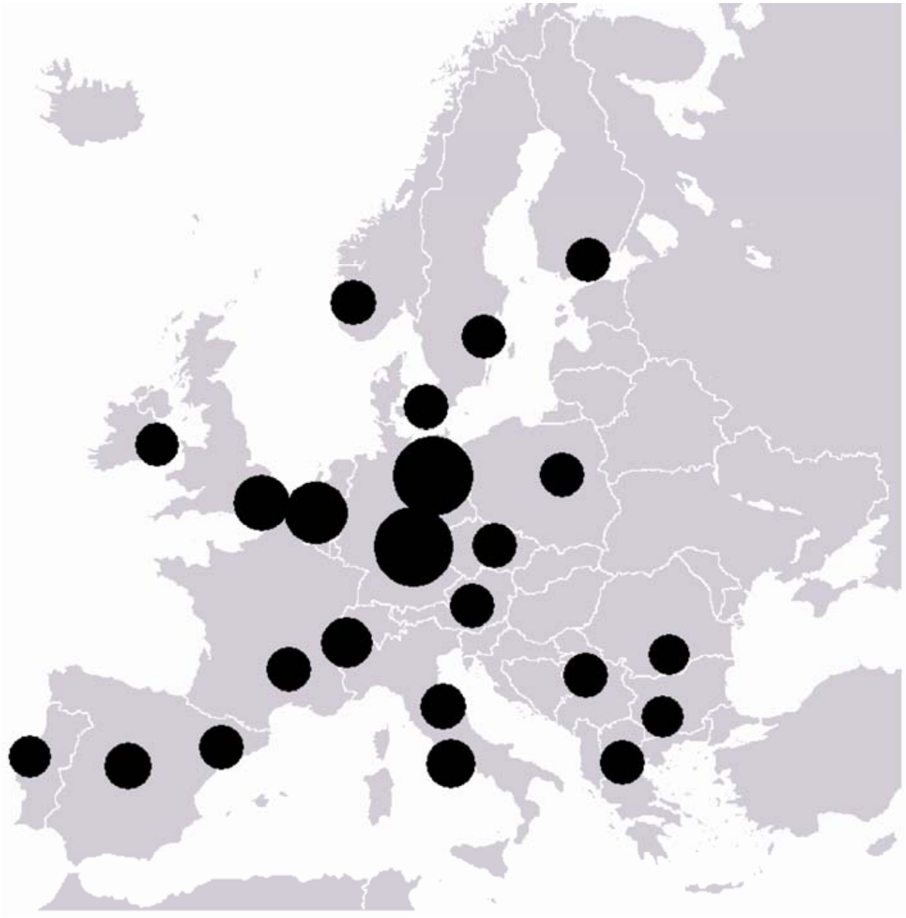


Haplotypes



Jakobsson et al Nature 2008

A set of European populations

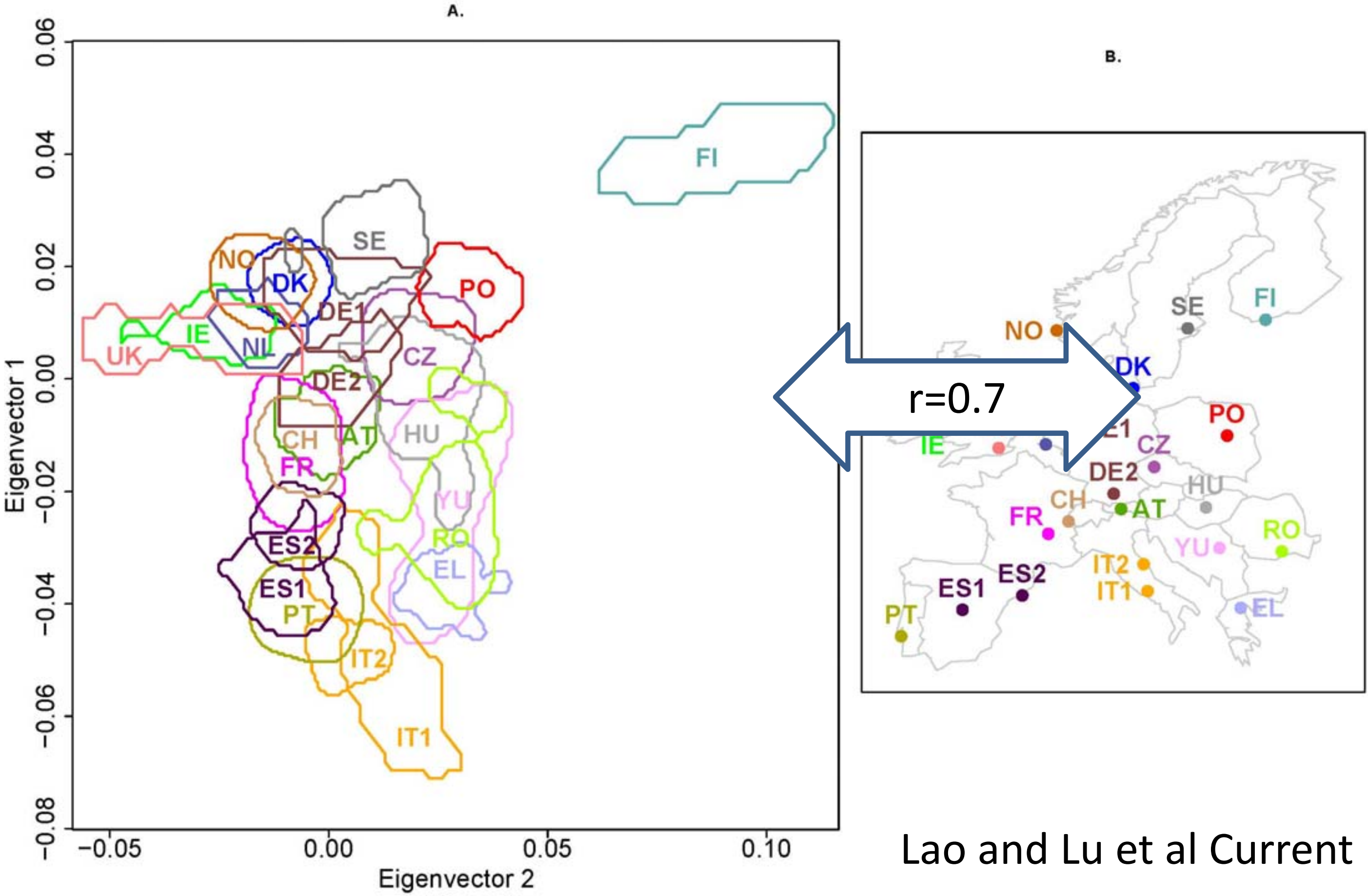


23
populations

500 Affy Array

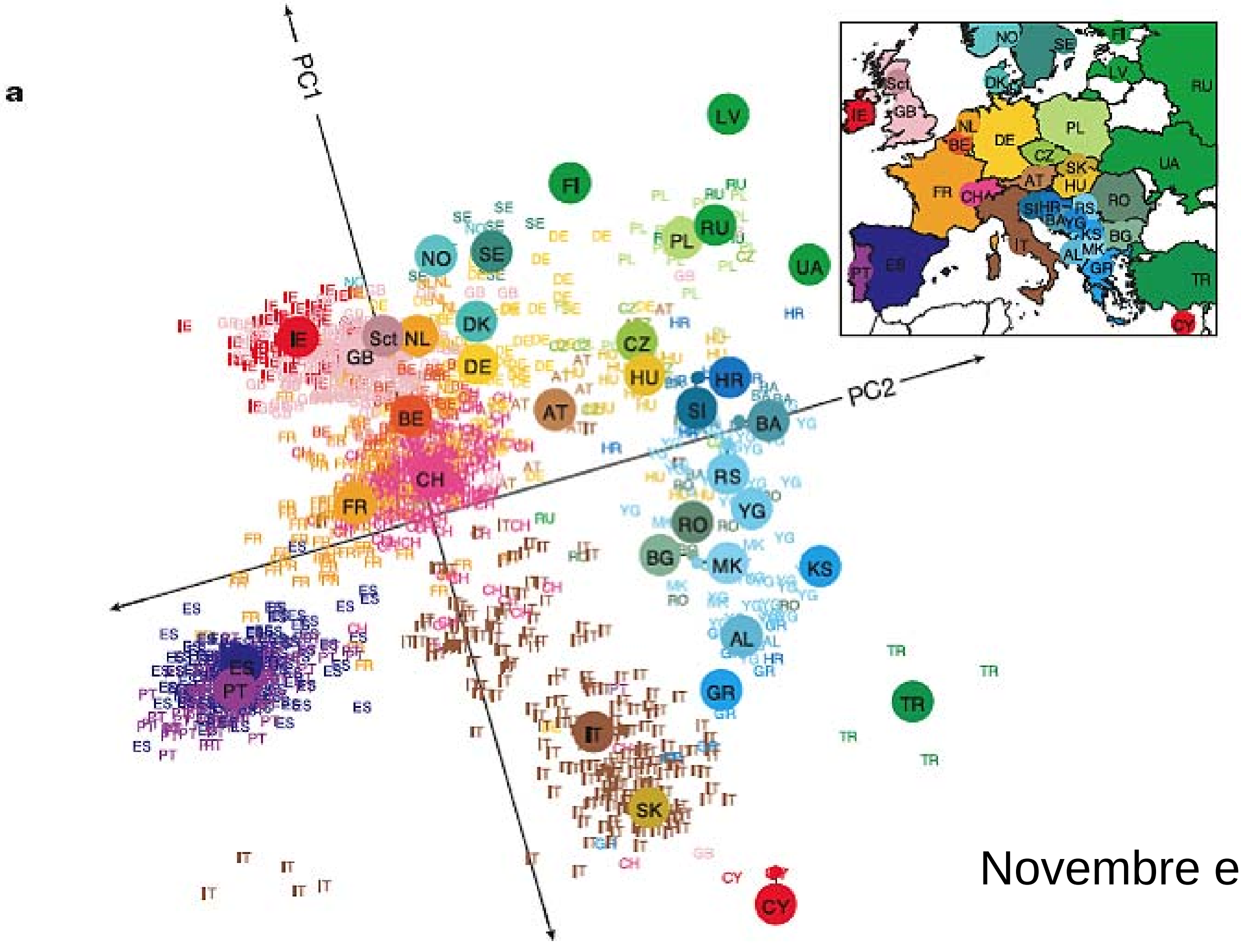
Autosomal SNPs in Europe

300,000 SNPs



Lao and Lu et al Current Biology 2008

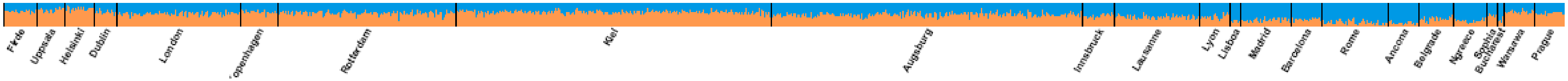
Autosomal SNPs in Europe



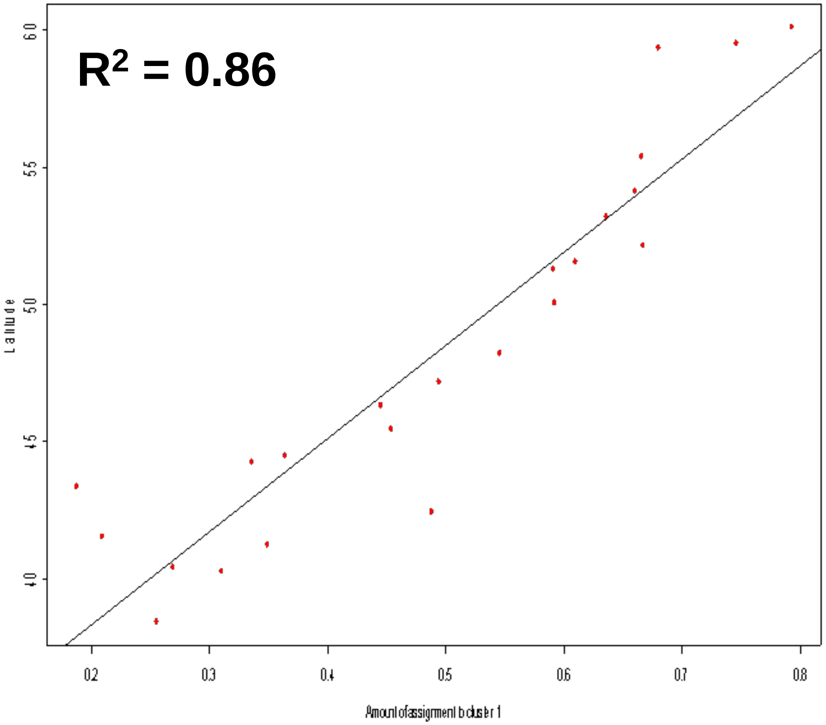
Novembre et al Nature 2008

Autosomal SNPs in Europe

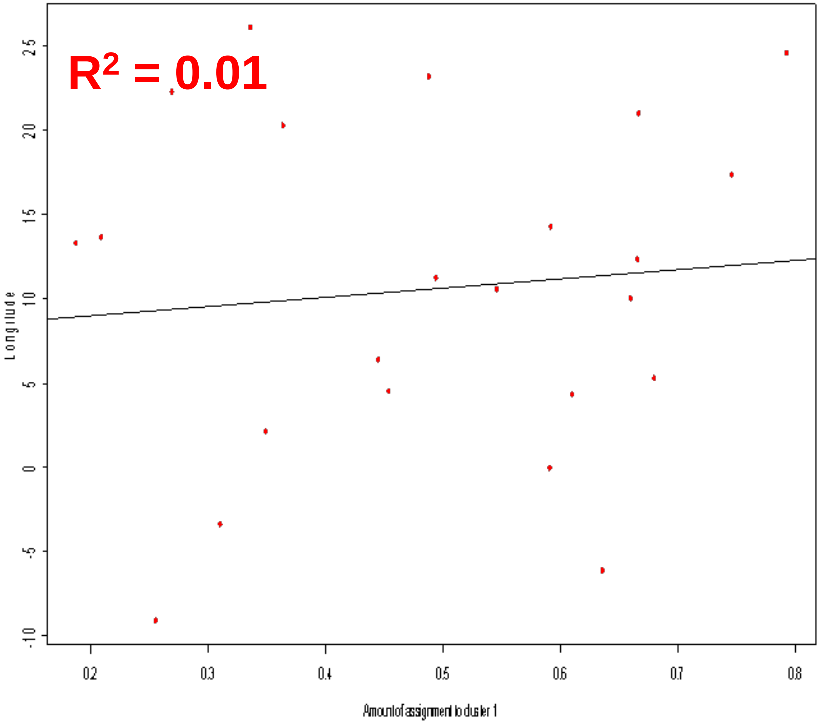
K = 2; Admixture



Correlation with latitude



Correlation with longitude



Autosomal SNPs in Europe

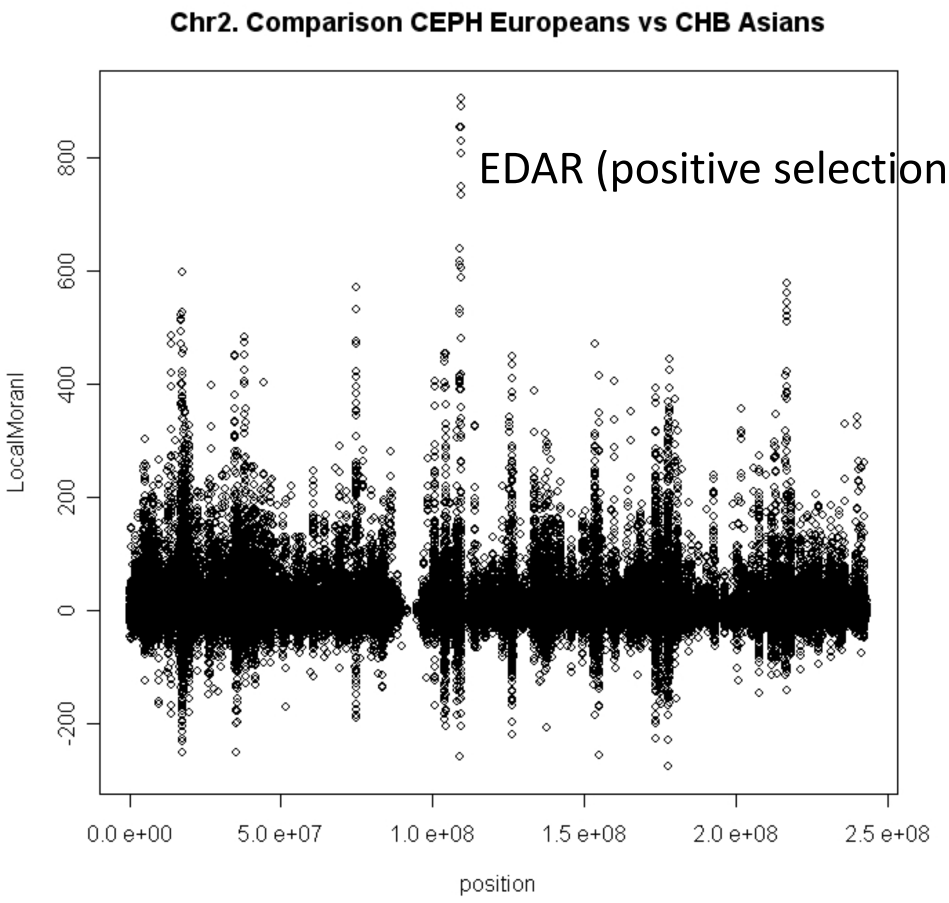
World

Anayet peak (2574 m), Pyrenees

Europe

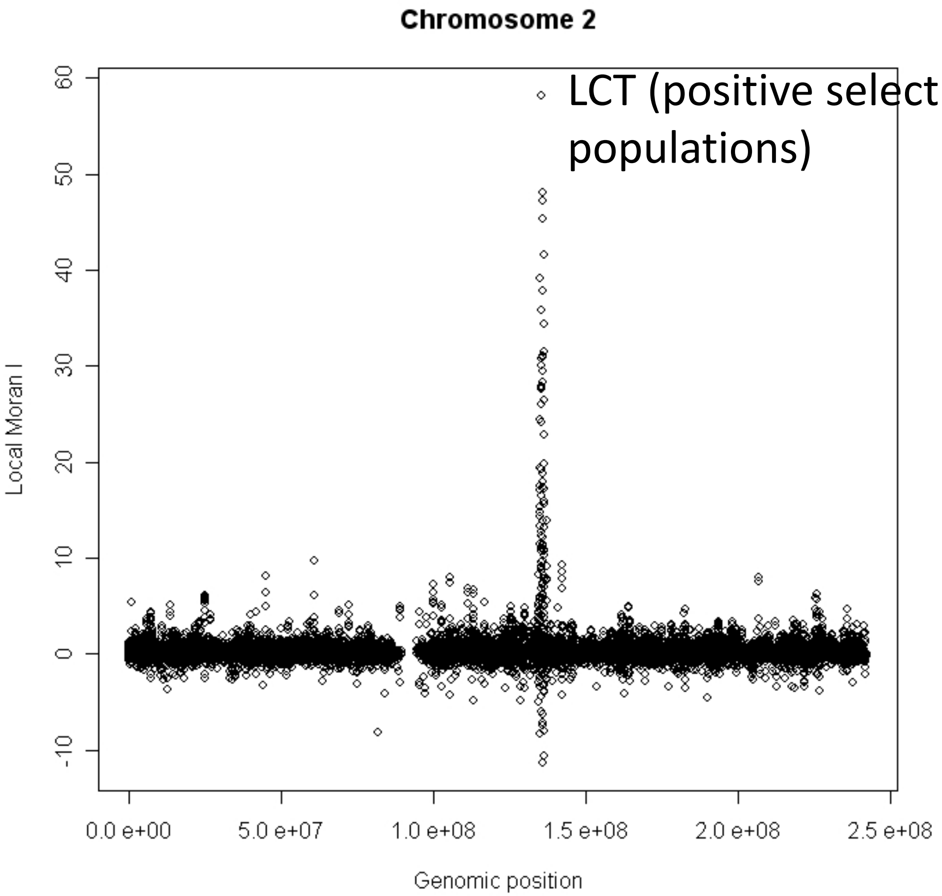
Keukenhoof garden(-2 m), Netherlands

Non random distribution of population substructure

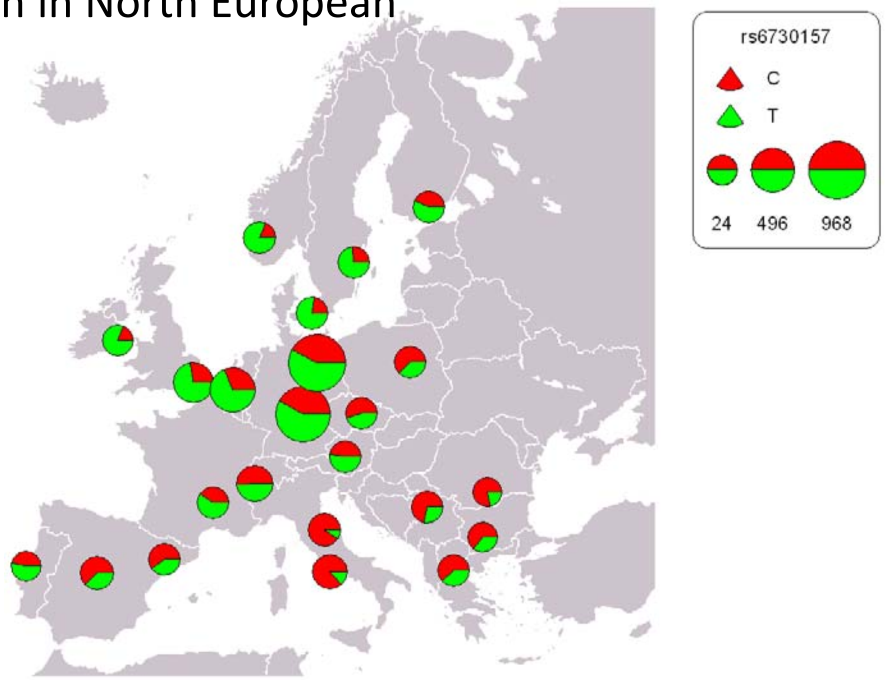


EDAR (positive selection in Asians)

Non random distribution of population substructure

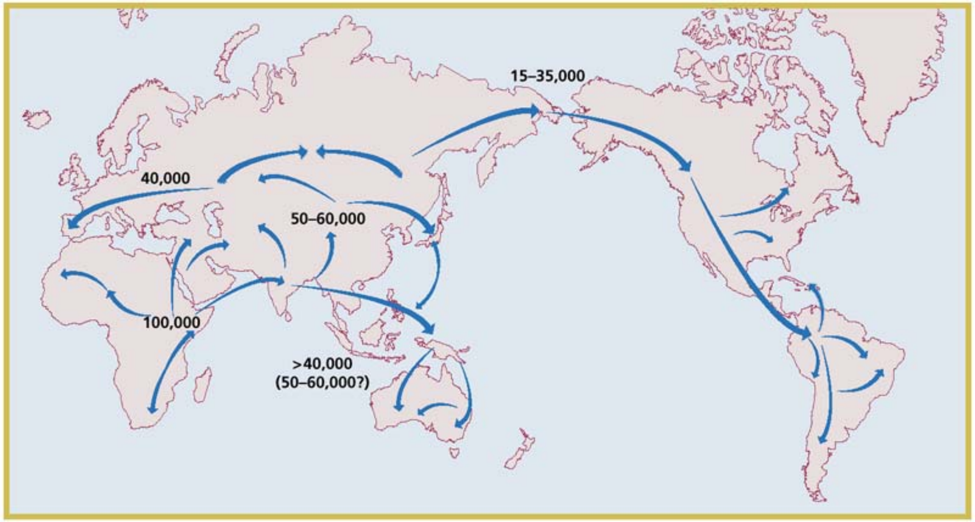


LCT (positive selection in North European populations)

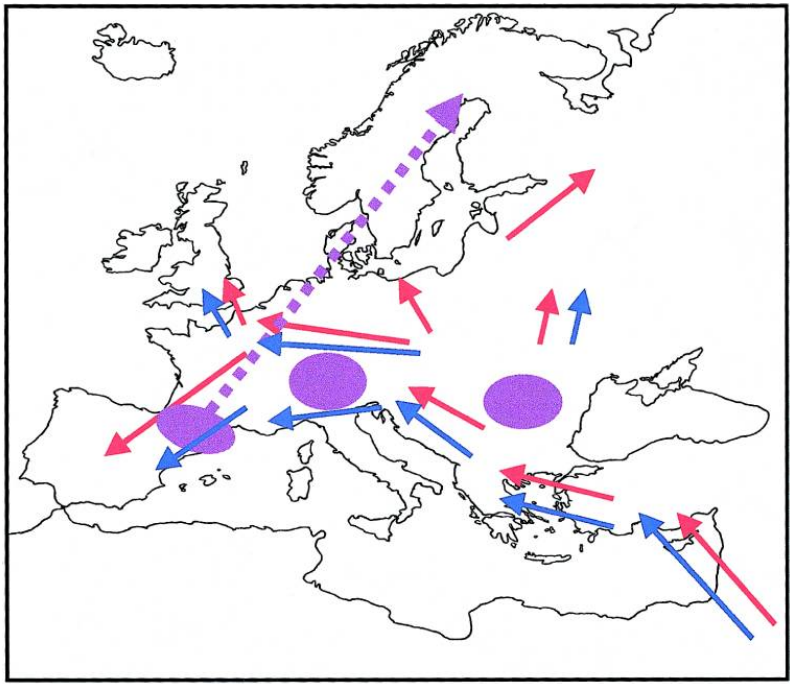


Lao and Lu et al Current Biology 2008

Demography shapes the population substructure



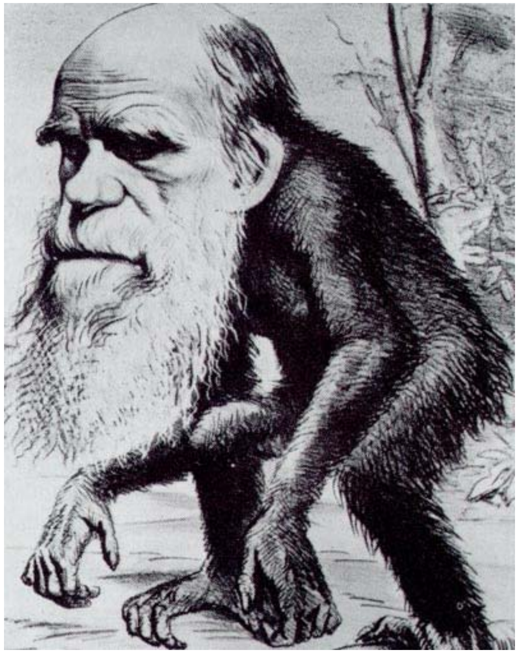
Cavalli-Sforza & Feldman Nature Genetics 2003



Simoni et al AJHG 2000

Selection shapes the population substructure

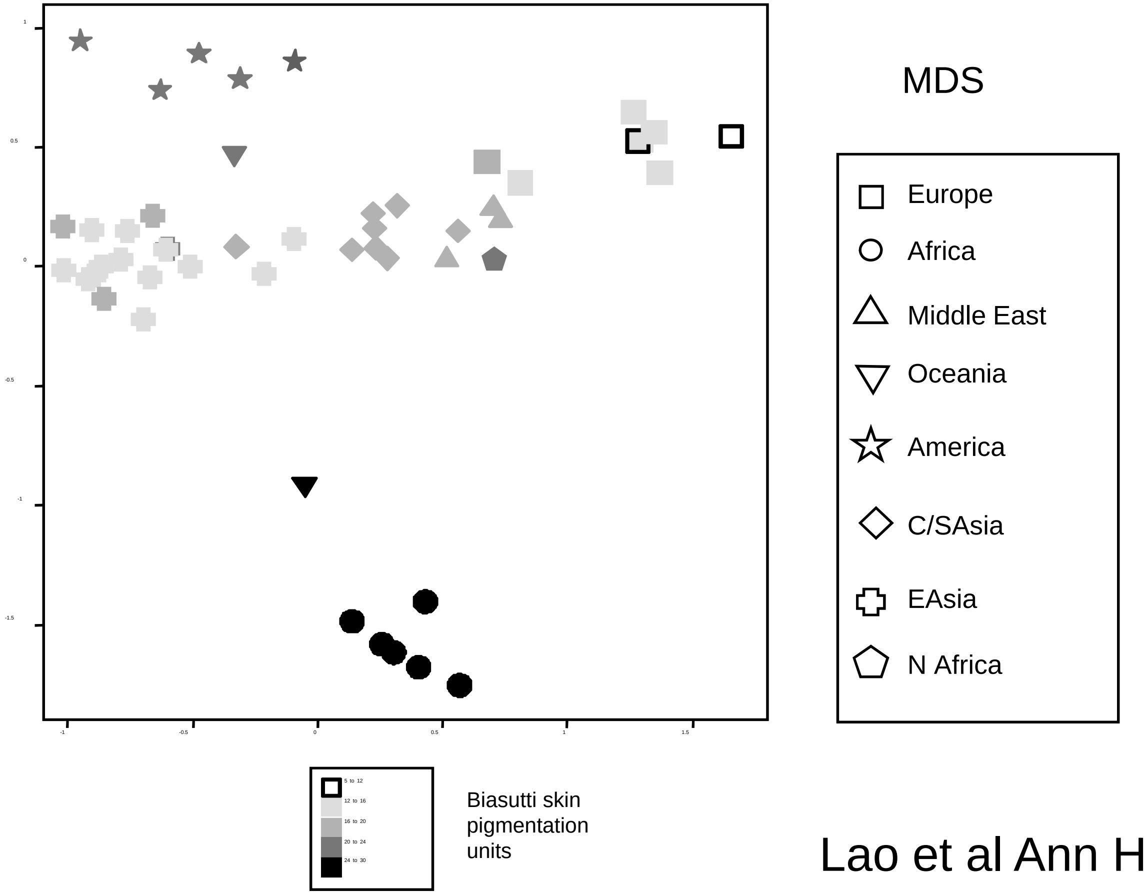
- Selective pressures within the species (locus specific)



Lactose tolerance
Malaria resistence
Human pigmentation
…

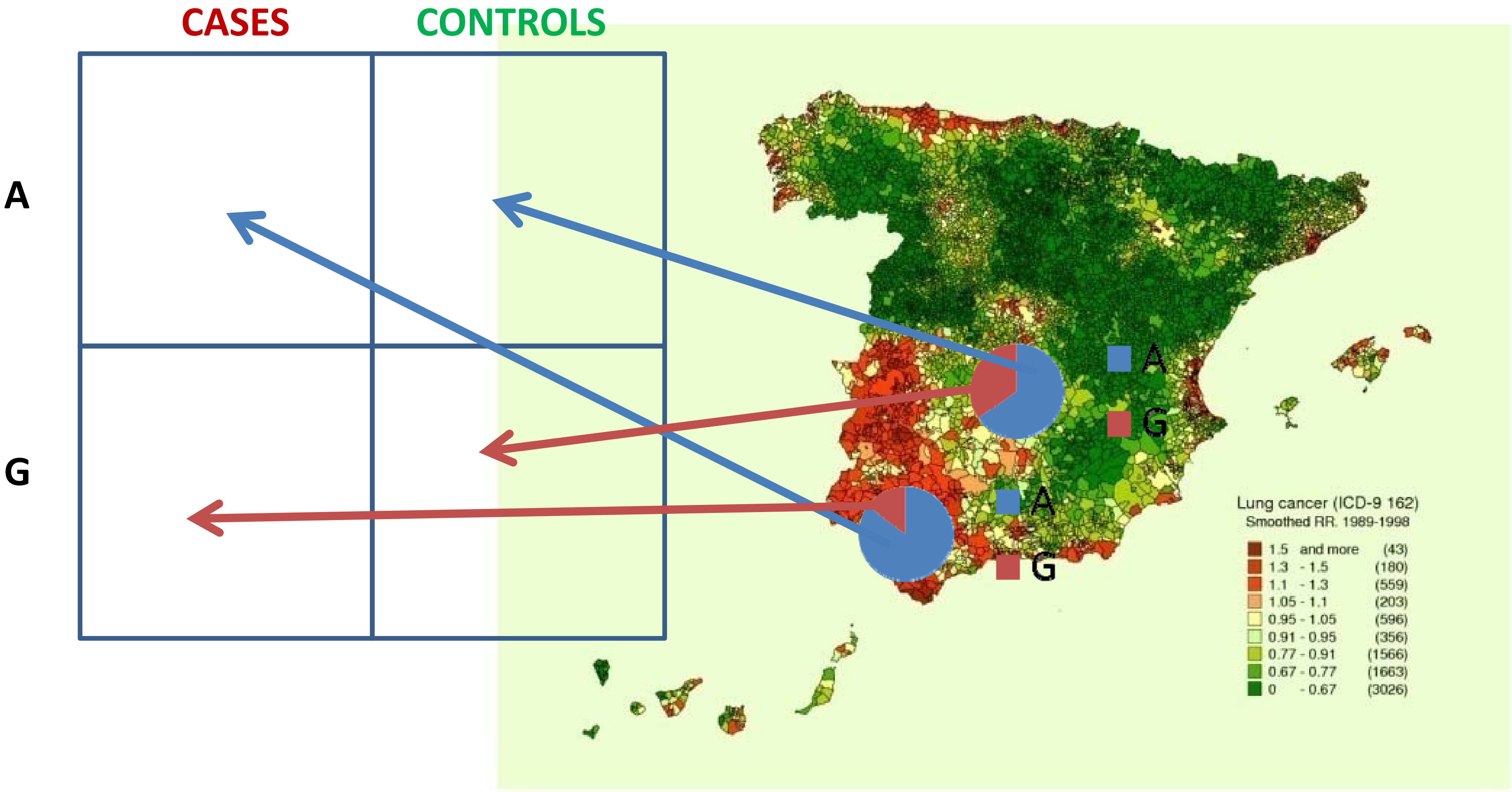
Selection shapes the population substructure

- Population substructure & pigmentation (5 SNPs)



Lao et al Ann Hum Genet 2007

Why population substructure: Confounding factor



CASES

CONTROLS

A

G

Population substructure: improving the detection

Plato’s cave myth

CHANGE THE ALGORITHM
FOR DETECTING
POPULATION
SUBSTRUCTURE

Population substructure: improving the detection

Plato’s cave myth

INCREASE THE
RESOLUTION TO
SEE THE OBJECTS

AIMs/ASMs

- Markers that capture most of the genetic ancestry
  - – Estimate ancestry
  - – Reduce the number of markers to test for genetic homogeneity
    - Time cost (clustering algorithms can be extremely computational intensive)
    - Economical cost (i.e exclude individuals BEFORE doing the GWA)

Strategies to ascertain ASMs

- Based on the existing diversity between individuals (i.e Paschou et al 2008)
- Based on predefined groups of individuals
  - No phenotype linked
    - Large Genetic distances
    - Signals of positive selection
  - Phenotype linked
    - Covariates with the phenotype of interest

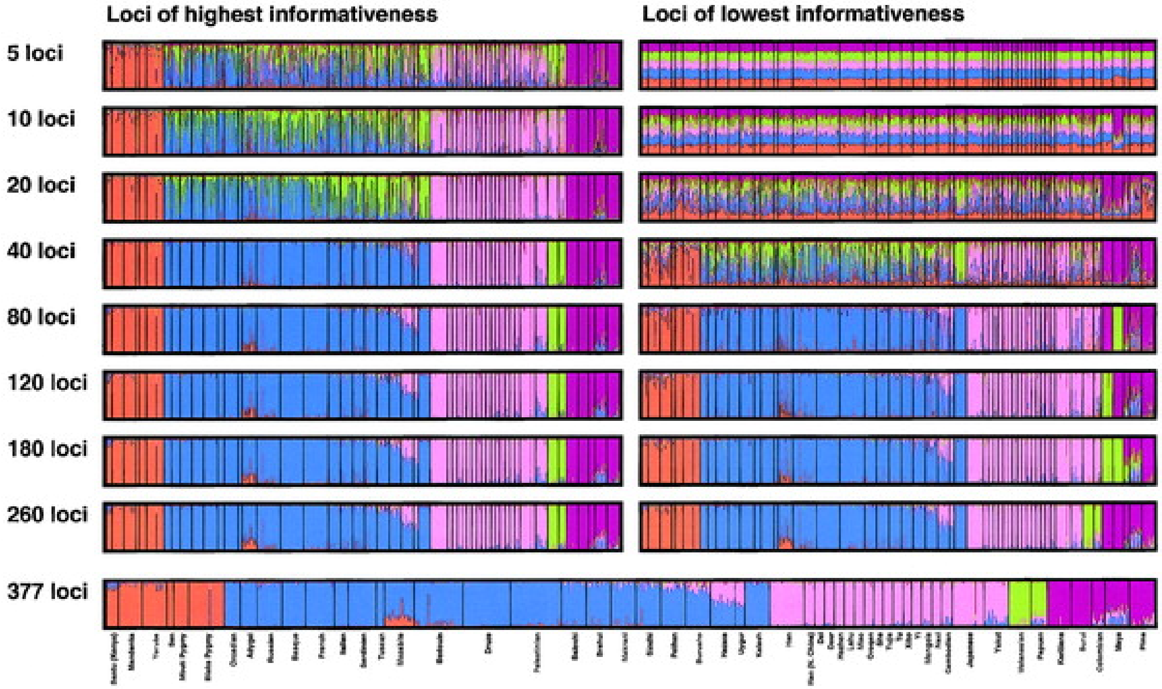
A basic algorithm to ascertain ASMs

- Use a statistic to quantify the amount of differentiation between populations
- Compute the OVERAL non-redundant amount of In between set of SNPs
- Take the best combination of markers from all the possible combinations
- Repeat the process until the information of the set of markers is maximum

A statistic to ascertain ASMs

informativeness for assignment

$$I_n(Q;J)=\sum_{j=1}^{N}\left(-p_j\log p_j+\sum_{i=1}^{K}\frac{p_{ij}}{K}\log p_{ij}\right)$$



Am J Hum Genet. 2003 Dec;73(6):1402-22

A statistic to ascertain ASMs

- How much information a marker contains about the ancestry of one individual (measured in *nats*)
- Ranges from 0 to the natural logarithm of the number of clusters and it is proportional to the number of differentiated clusters

A statistic to ascertain ASMs

- Computes the non-redundant amount of information when considering more than one marker
- Requires computing the frequency of **ALL** the allelic combinations when considering more than 1 locus

A way to compute In

- Problem: The number of combinations increases exponentially with the number of markers.
  - Number of allelic combinations considering 50 SNPs:

$$2^{50} = 1{,}125{,}899{,}906{,}842{,}624$$

A way to compute In

$$I_n(Q;J)=\sum_{j=1}^{N}\left(-p_j\log p_j+\sum_{i=1}^{K}\frac{p_{ij}}{K}\log p_{ij}\right)$$

$$I_n(Q;J)=\sum_{j=1}^{N}\left(\overline{H_j}-\sum_{i=1}^{K}\frac{H_{ij}}{K}\right)$$

$$H\approx\frac{1}{N}\sum_{i=1}^{N}\ln(p)$$

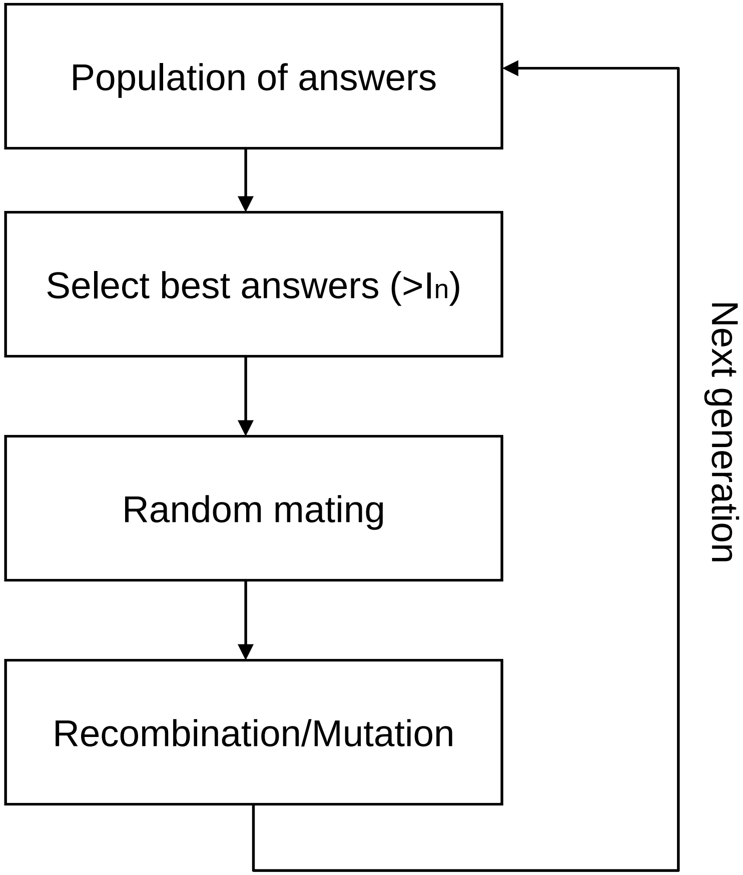
By applying the Asymptotic Equipartition Property of Entropy

A method to ascertain ASMs

- Problem: Considering 8,000 markers, ascertaining the best set of 50 markers requires computing :

$$N_{combinations} = \frac{8,000!}{50!\left(8,000 - 50\right)!} \approx 4 \times 10^{130}$$

A method to ascertain ASMs



Population of answers

Select best answers (>I_n)

Random mating

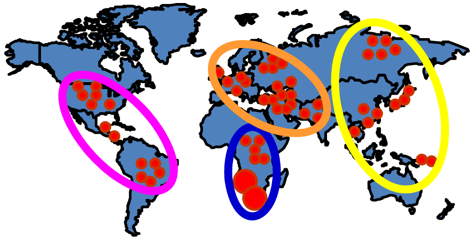
Recombination/Mutation

Next generation

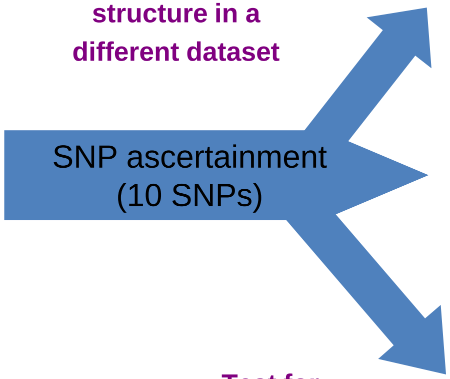
ASMs for continental differentiation using Affy 10k

YCC-panel

76 human individuals
21 sampling localities



10k Affymetrix Array
(~9000 SNPs after excluding X-SNPs & missing SNPs)



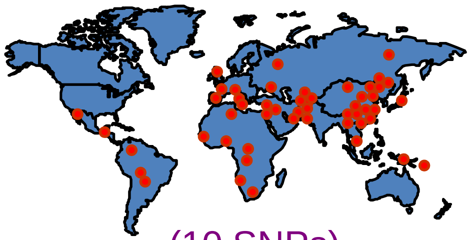
SNP ascertainment
(10 SNPs)

Reproducibility of geographic structure in a different dataset

Test for signatures of positive selection (EHH test)

CEPH-HGDP panel

1064 samples
51 human populations of global distribution



(10 SNPs)

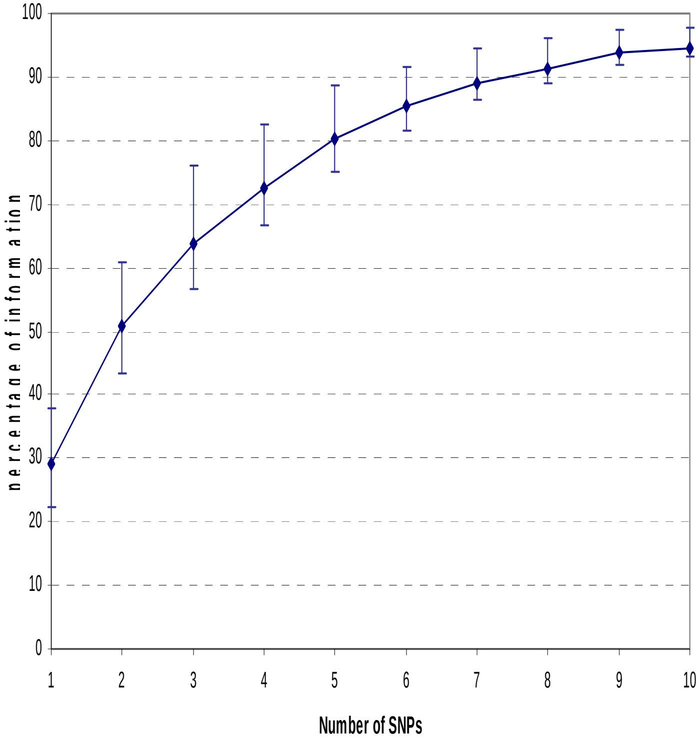
Perlegen Database

3 Human populations
~1,500,000 SNPs
(most informative 5 SNPs)

Lao et al. Am J Hum Genet. 2006 Apr;78(4):680-90

ASMs for continental differentiation using Affy 10k

The genetic algorithm was applied increasing every time the number of selected SNPs



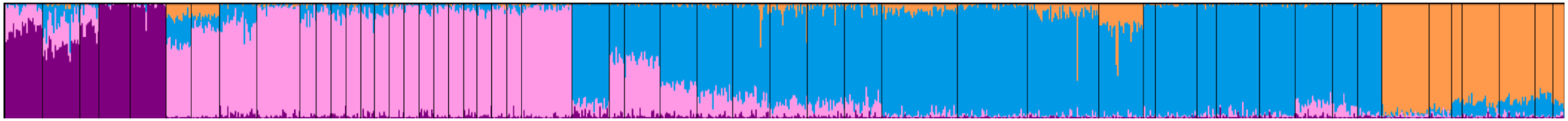
Selected SNPs in the final 10 SNPs run

Marker name	*Chromosome*	*Gene name*	*I_N (%) from 4 groups YCC panel*	*I_N (%) from 7 groups CEPH-HGDP*
rs722869	14	VRK1	29.066	7.960
rs1858465	17		25.637	9.228
rs1876482	2	LOC442008	24.589	10.290
rs1344870	3		22.810	11.074
rs1363448	5	PCDHGB1	19.418	4.552
rs952718	2	ABCA12	18.739	9.472
rs2352476	7		18.317	5.603
rs714857	11		18.083	6.157
rs1823718	15		17.845	5.451
rs735612	15	RYR3	14.315	5.530

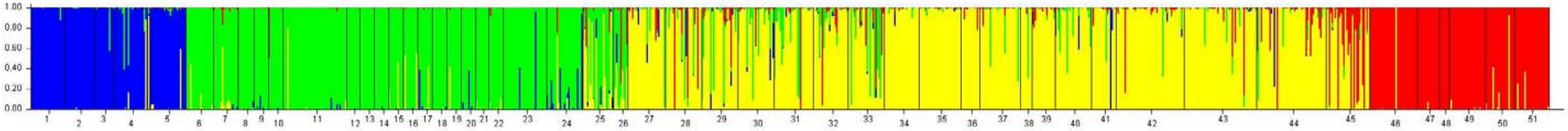
Lao et al. Am J Hum Genet. 2006 Apr;78(4):680-90

ASMs for continental differentiation using Affy 10k

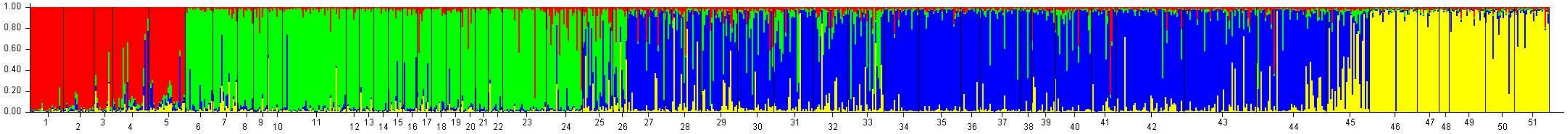
993 autosomal markers



10 SNPs **No admixture**



Admixture

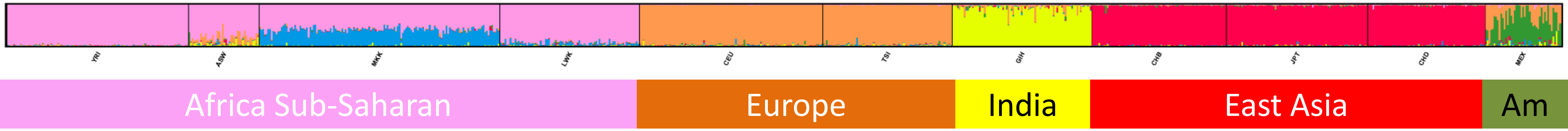


Ameri | O | E-Asia | C/S-Asia Europe M. East | Africa

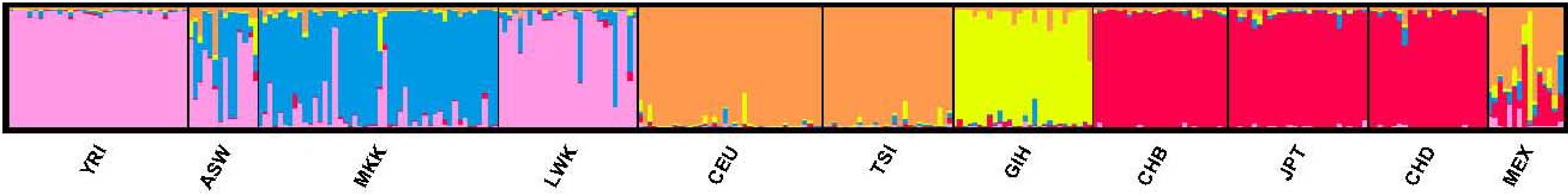
Lao et al. Am J Hum Genet. 2006 Apr;78(4):680-90

ASMs for continental differentiation using HapMap III

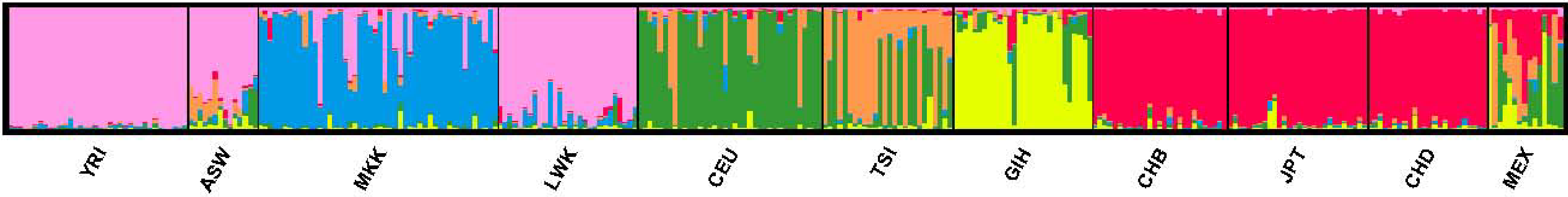
K = 6 (1000 (randomly ascertained) markers, Admixture, 10,000 burning, 10,000 retained simulations)



K = 5 (50 markers, Admixture, 500,000 burning, 500,000 retained simulations)

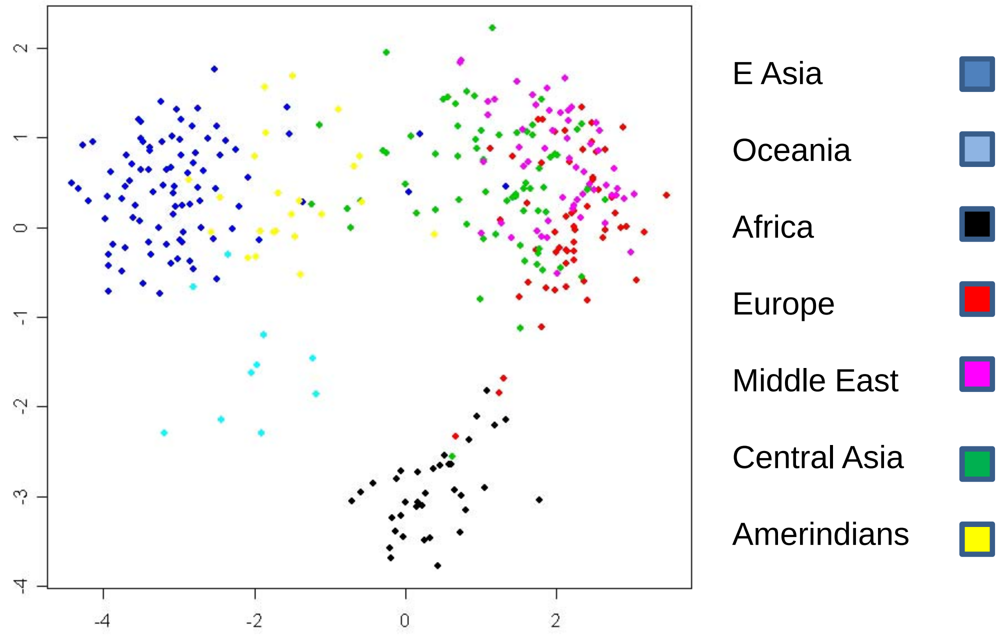


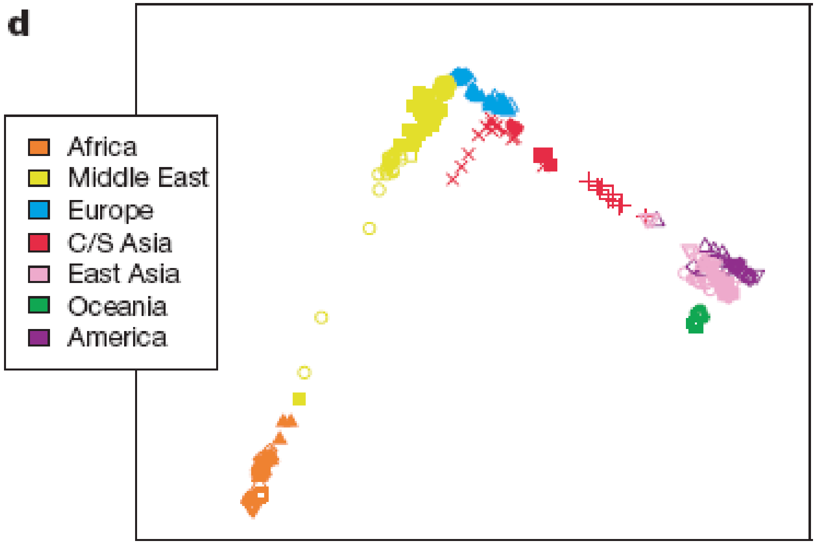
K = 6 (100 markers, Admixture, 100,000 burning, 100,000 retained simulations)



ASMs for continental differentiation using Illumina 650k

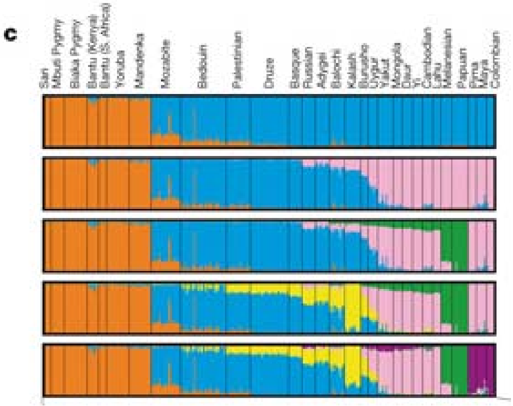
25 ascertained markers. PCA





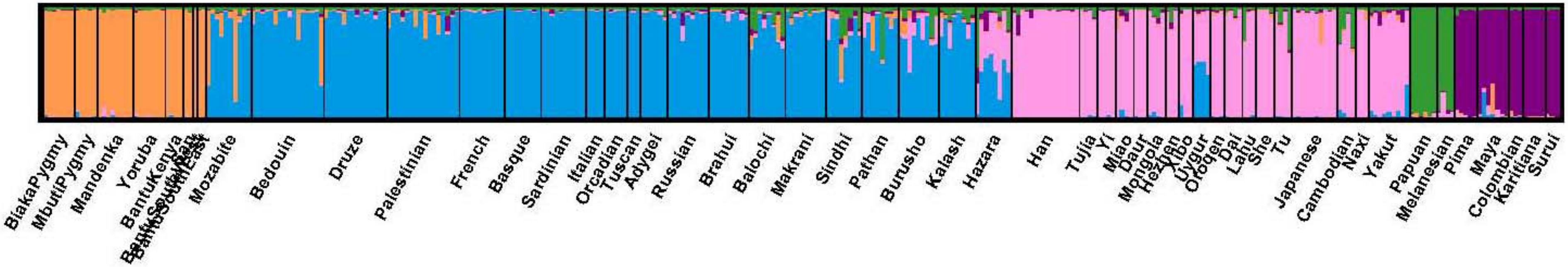
ASMs for continental differentiation using Illumina 650k

- CEPH

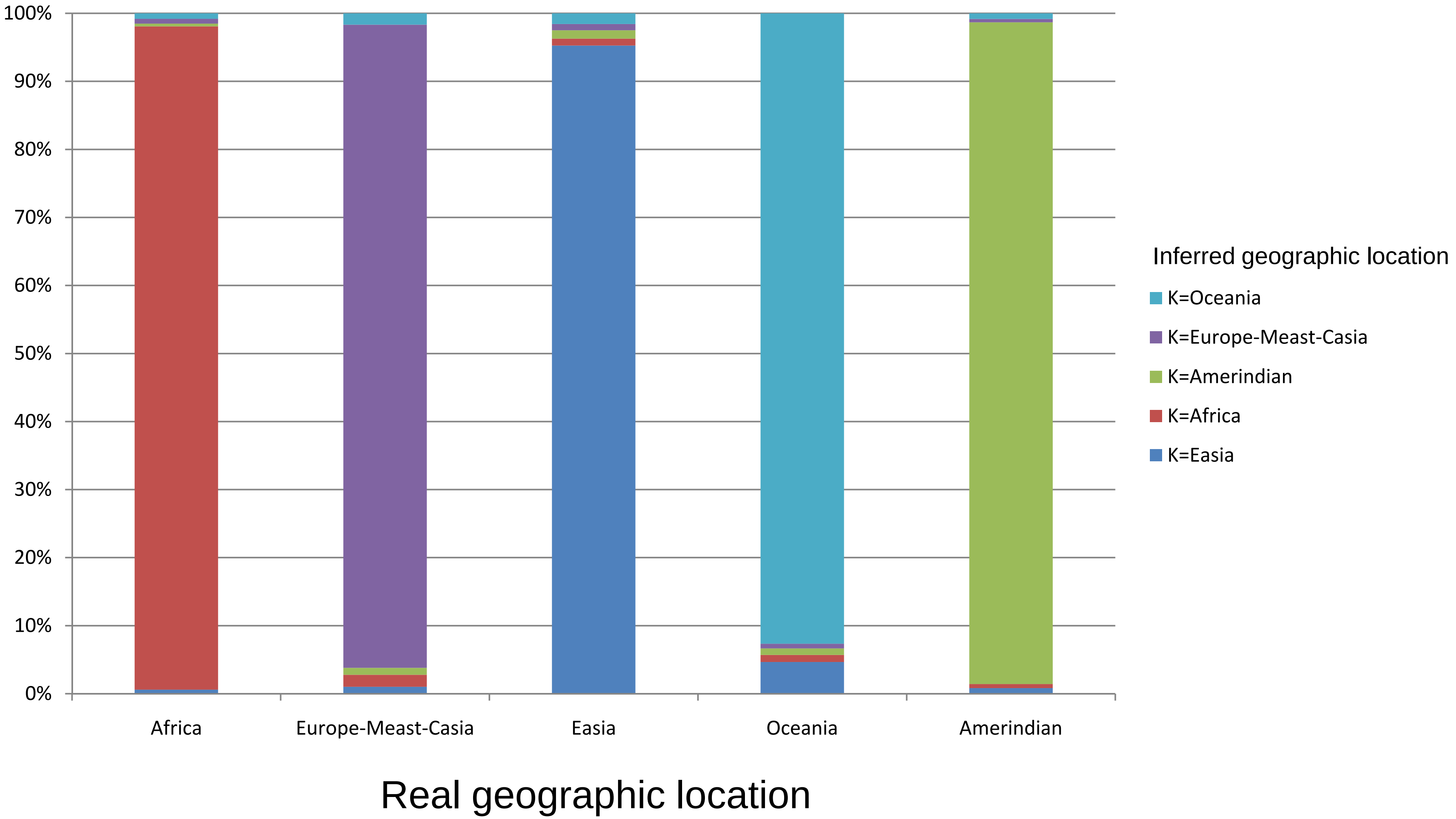


550,000 SNPs

K = 5 (50 ascertained markers, Admixture, 500,000 burning, 500,000 retained simulations)

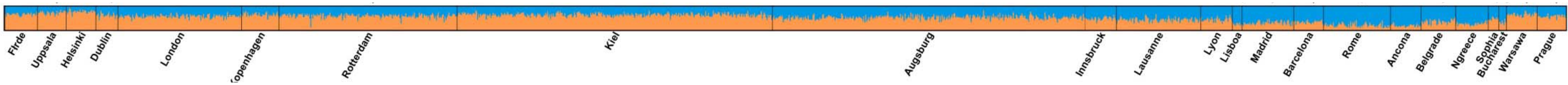


Specificity of the ascertained markers

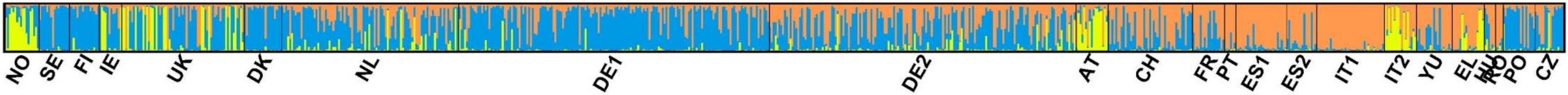


ASMs for population differentiation in the European continent using Affymetrix 500k

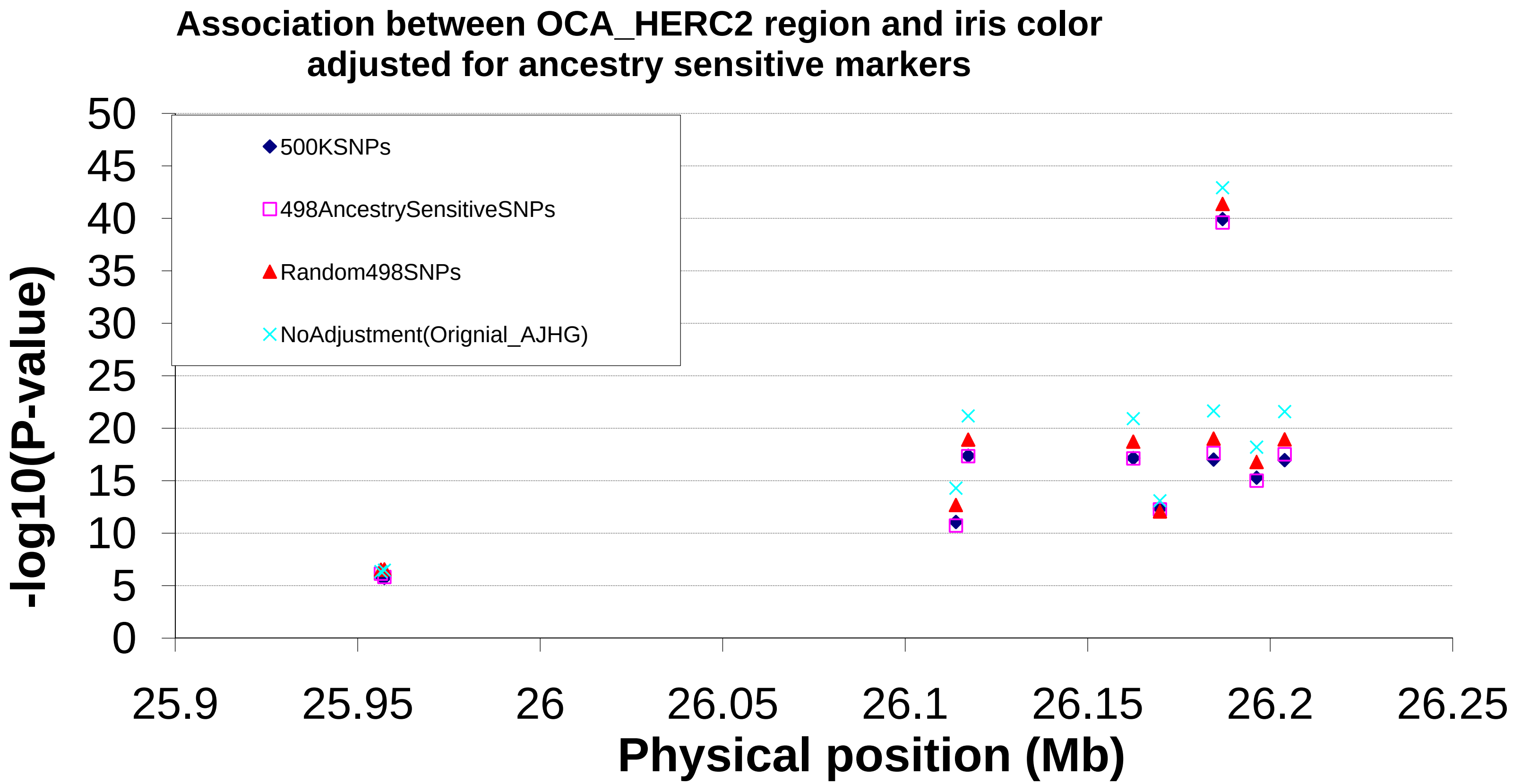
K = 2 (5000 random markers, Admixture, 10,000 burning, 10,000 retained simulations)



K = 3 (500 ascertained markers, Admixture, 10,000 burning, 10,000 retained simulations)



Use of the 500 ASMs for correcting the effect of population substructure



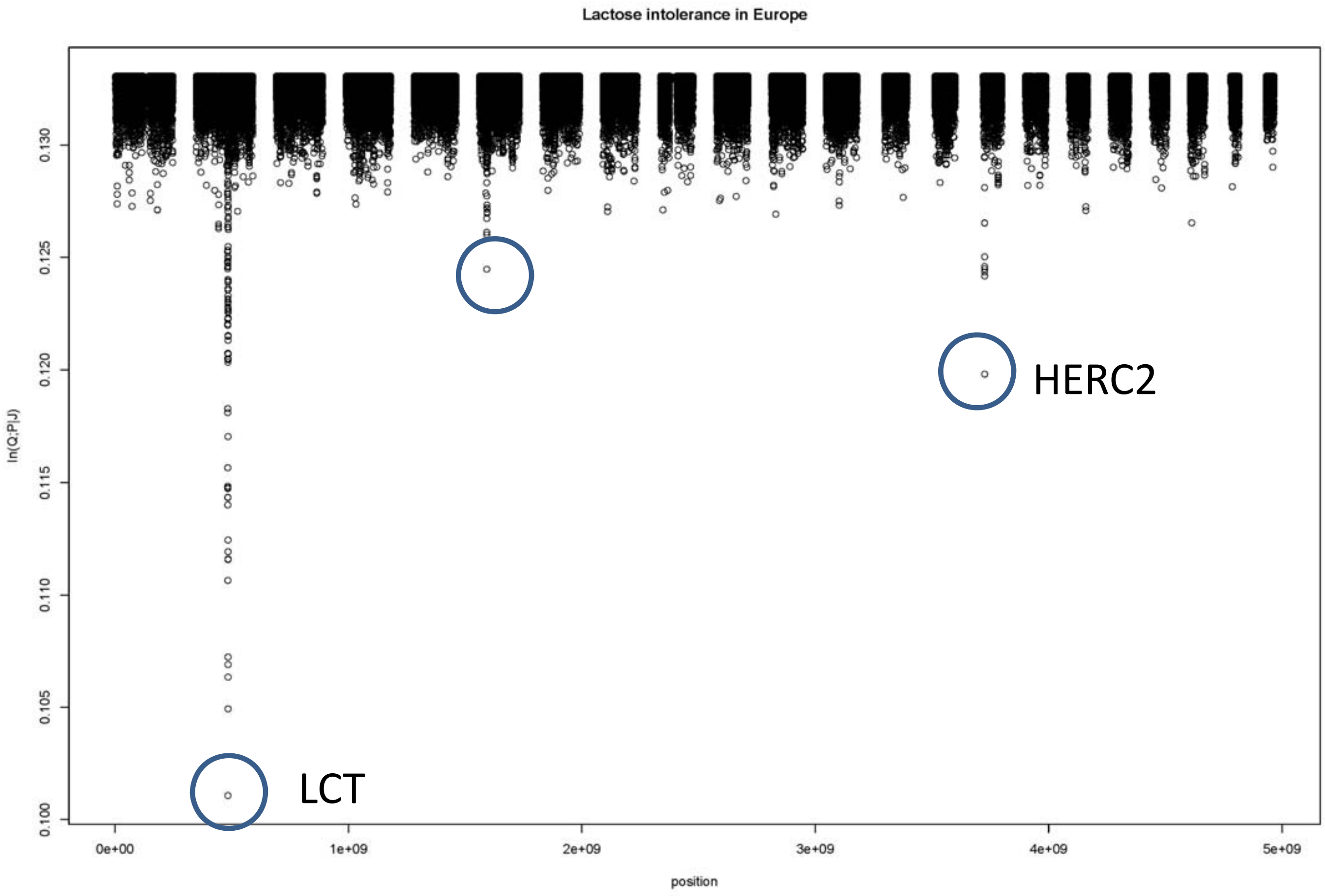
PhenoASMs

- Recall
  - Population substructure is only a problem when PHENOTIPIC and GENOTYPIC variation covariates
  - Why not ascertaining markers that are associated to the particular spatial pattern of the phenotype?

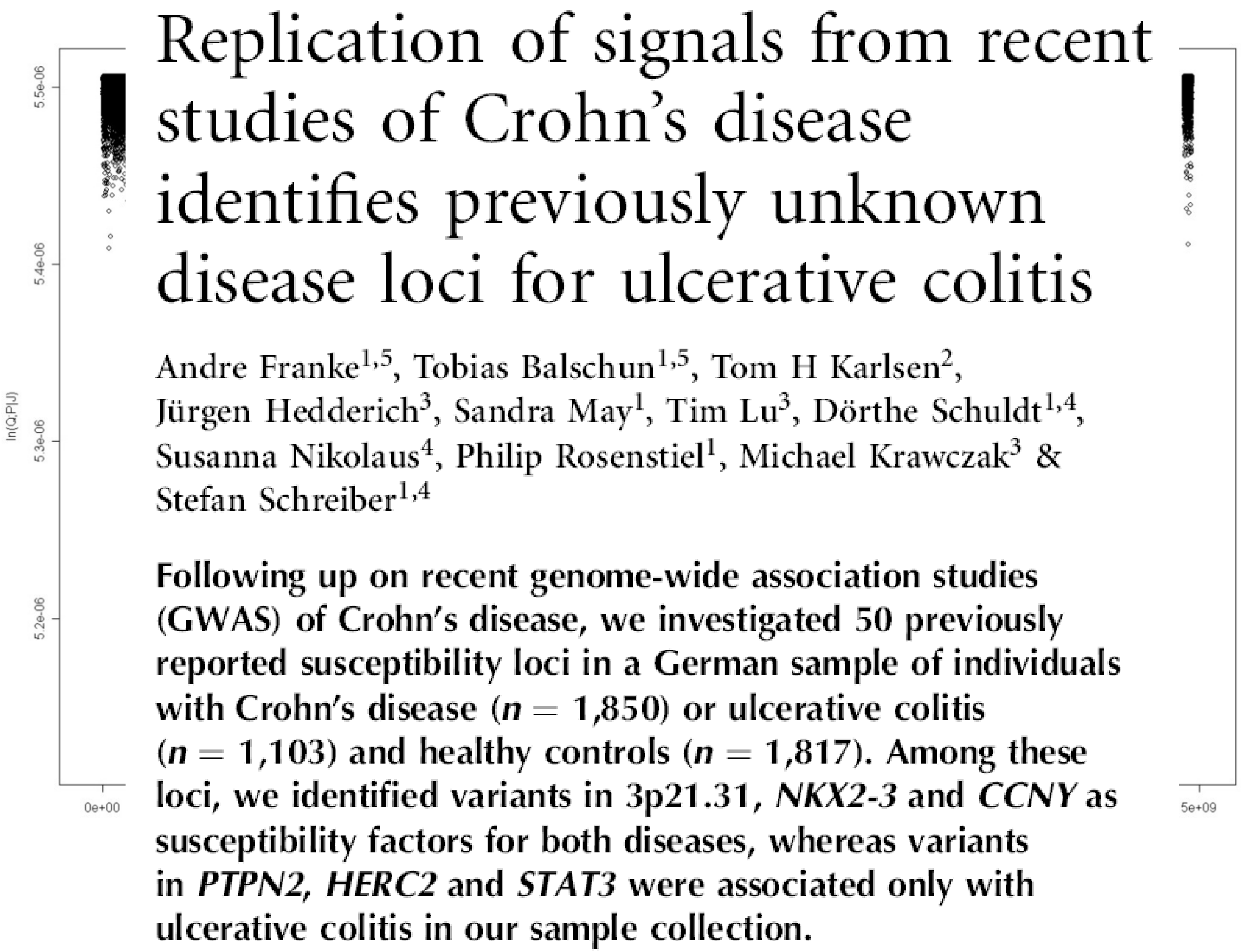
$$I_n(Q;P\,|\,J) = I_n(Q;P;J) - I_n(Q;J)$$

“Amount of information of the phenotype (P) conditional on the genotype (J): How well could we correctly classify one individual given that we know his phenotype if we already know his genotype in a particular locus”

PhenoASMs for lactose tolerance



PhenoASMs for Crohn disease



Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis

Andre Franke[1,5], Tobias Balschun[1,5], Tom H Karlsen[2], Jürgen Hedderich[3], Sandra May[1], Tim Lu[3], Dörthe Schuldt[1,4], Susanna Nikolaus[4], Philip Rosenstiel[1], Michael Krawczak[3] & Stefan Schreiber[1,4]

Following up on recent genome-wide association studies (GWAS) of Crohn's disease, we investigated 50 previously reported susceptibility loci in a German sample of individuals with Crohn's disease (*n* = 1,850) or ulcerative colitis (*n* = 1,103) and healthy controls (*n* = 1,817). Among these loci, we identified variants in 3p21.31, *NKX2-3* and *CCNY* as susceptibility factors for both diseases, whereas variants in *PTPN2, HERC2* and *STAT3* were associated only with ulcerative colitis in our sample collection.

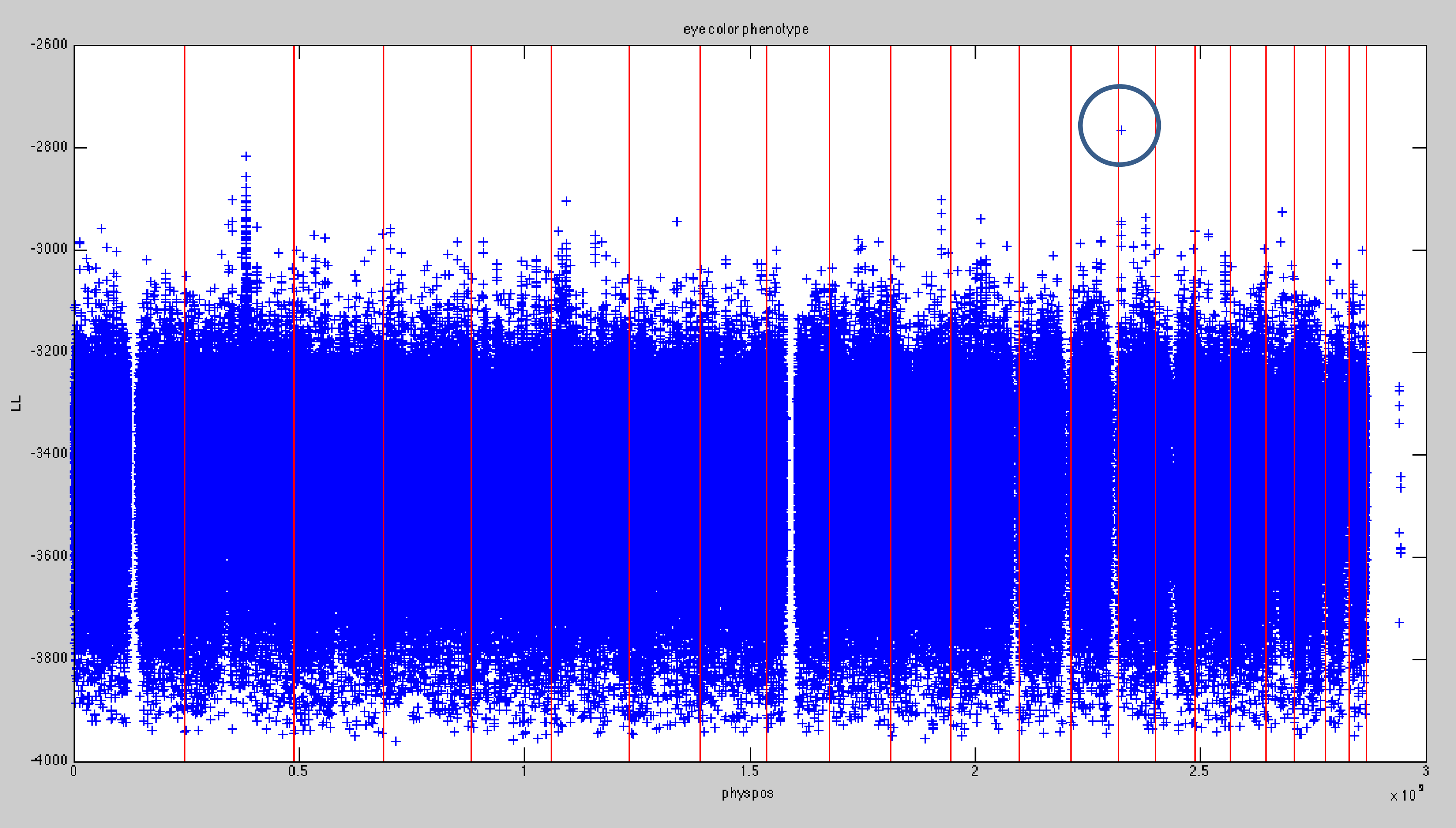
PhenoASMs: a little bit further

	AA	AB	BB	Marginal phenotype
C	$P(AA)P(C \mid AA)$	$P(AB)P(C \mid AB)$	$P(BB)P(C \mid BB)$	$\sum P(g)P(C \mid g)$
D	$P(AA)P(D \mid AA)$	$P(AB)P(D \mid AB)$	$P(BB)P(D \mid BB)$	$\sum P(g)P(D \mid g)$

PhenoASMs: a Bayesian approach

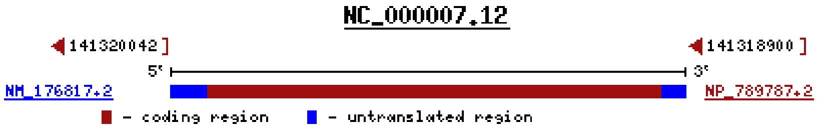
- Update θ with a Metropolis algorithm
- Update the covariance matrix of the proposal distribution by means of a “*quasi-perfect adaptive MCMC*” (Andrieu and Atchade)
- Compute the harmonic mean of the likelihood in order to obtain a rough estimate of P(M|D)

Phenotype-genotype association for eye color

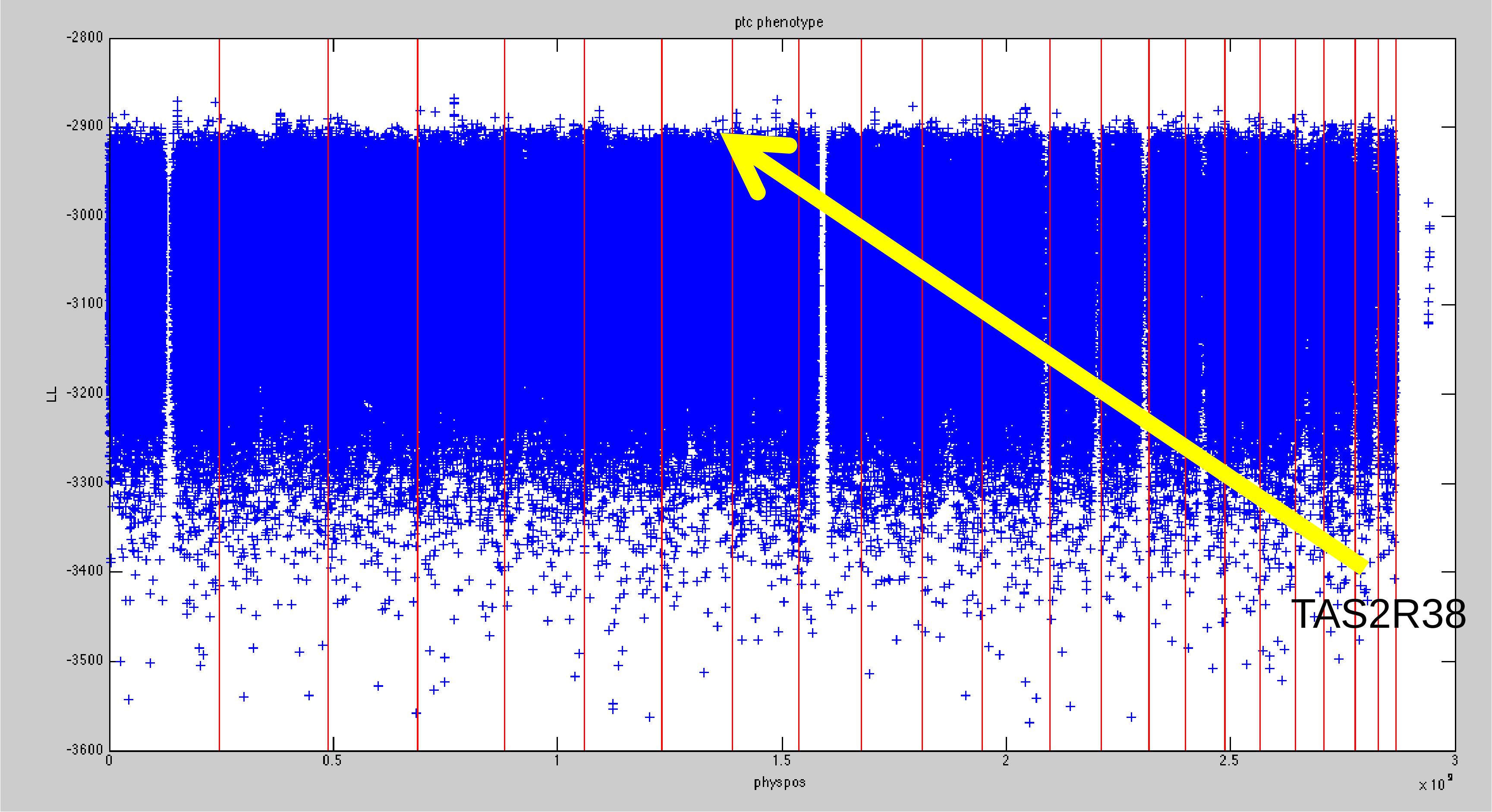


Phenotype-genotype association for bitter taste

TAS2R38

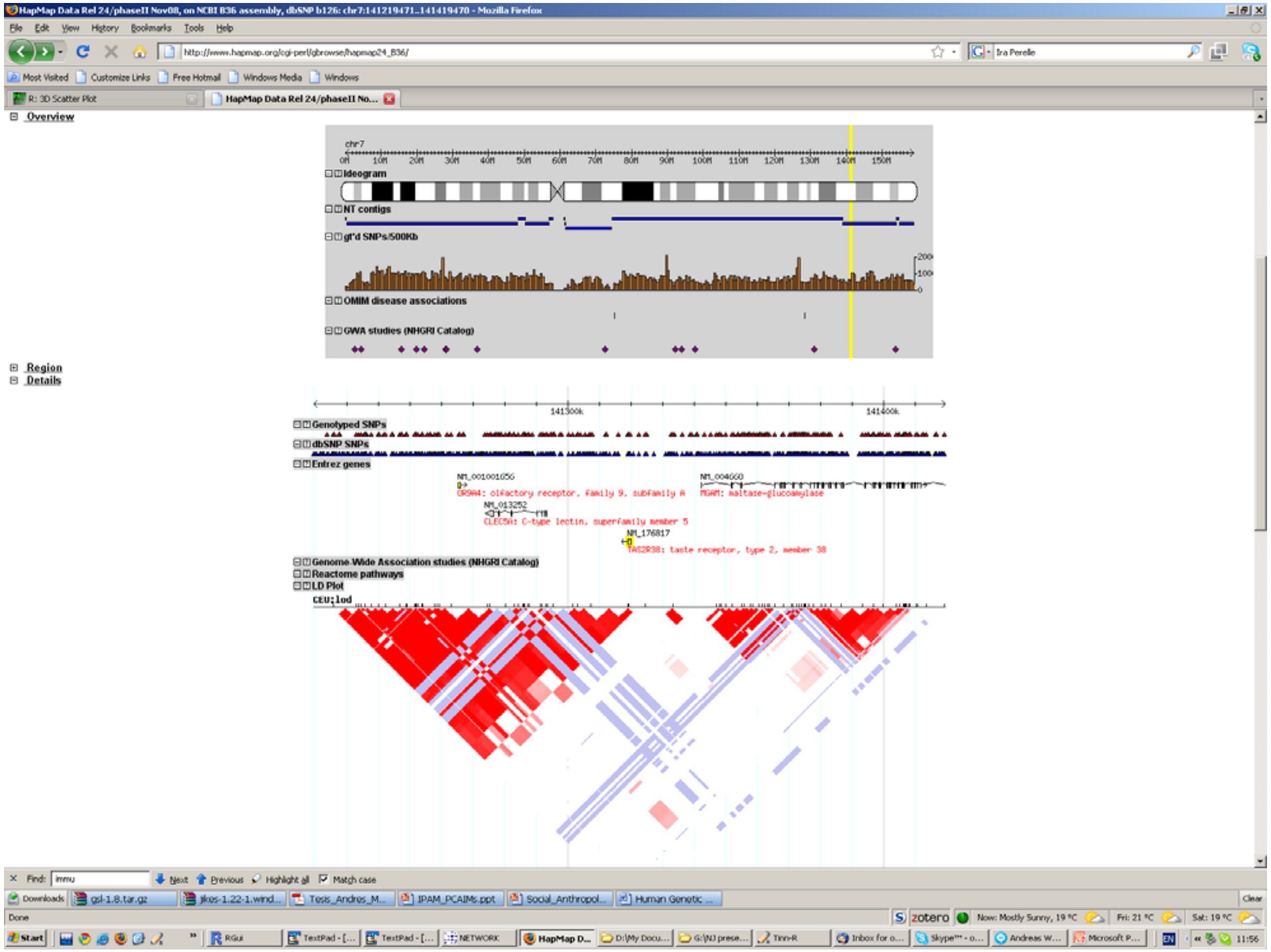


Phenotype-genotype association for bitter taste



TAS2R38

Phenotype-genotype association for bitter taste



HapMap Data Rel 24/phaseII Nov08, on NCBI B36 assembly, dbSNP b126: chr7:141219471..141419470 - Mozilla Firefox

http://www.hapmap.org/cgi-perl/gbrowse/hapmap24_B36/

File Edit View History Bookmarks Tools Help
Most Visited
Customize Links
Free Hotmail
Windows Media
Windows
R: 3D Scatter Plot
HapMap Data Rel 24/phaseII No...
Overview
chr7
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M
Ideogram
NT contigs
gt'd SNPs/500Kb
200
100
0
OMIM disease associations
GWA studies (NHGRI Catalog)
Region
Details
141300k
141400k
Genotyped SNPs
dbSNP SNPs
Entrez genes
NM_001001656
OR9A4: olfactory receptor, family 9, subfamily A
NM_004668
MGAM: maltase-glucoamylase
NM_013252
CLEC5A: C-type lectin, superfamily member 5
NM_176817
TAS2R38: taste receptor, type 2, member 38
Genome Wide Association studies (NHGRI Catalog)
Reactome pathways
LD Plot
CEU:lod
Find: immu
Next
Previous
Highlight all
Match case
Done

Conclusions

- Low to moderate human population differentiation
- Mainly associated to geography
- No sharp discontinuities, except in particular genomic regions (selection?)
- Results depend on the clustering algorithm
- ASMs can improve the detection of population substructure

Conclusions

- I_n is a good statistic for ascertaining markers to differentiate predefined populations
- If a prior definition of a population is used, ASMs will tend to differentiate such population, independently of the biological meaning
- PhenoASMs as the next level of ASMs?

In collaboration with

M. Balascakova, C. Becker, J. Bertranpetit, L.A. Bindoff, D. Comas, U. Gether, C. Gieger, G. Holmlund, A. Kouvatski, M. Macek, I. Mollet, M. Nelson, P. Nuernberg, W. Parson, R. Ploski, A. Ruether, A. Sajantila, S. Schreiber, A. Tagliabracci, A. Uiterlinden, T. Werge, and E. Wichmann.

Acknowledgements

Tim Lu

Manfred Kayser

Michael Krawczak

Andreas Wollstein

Petros Drineas

Peristeia Paschou

Thank you very much!