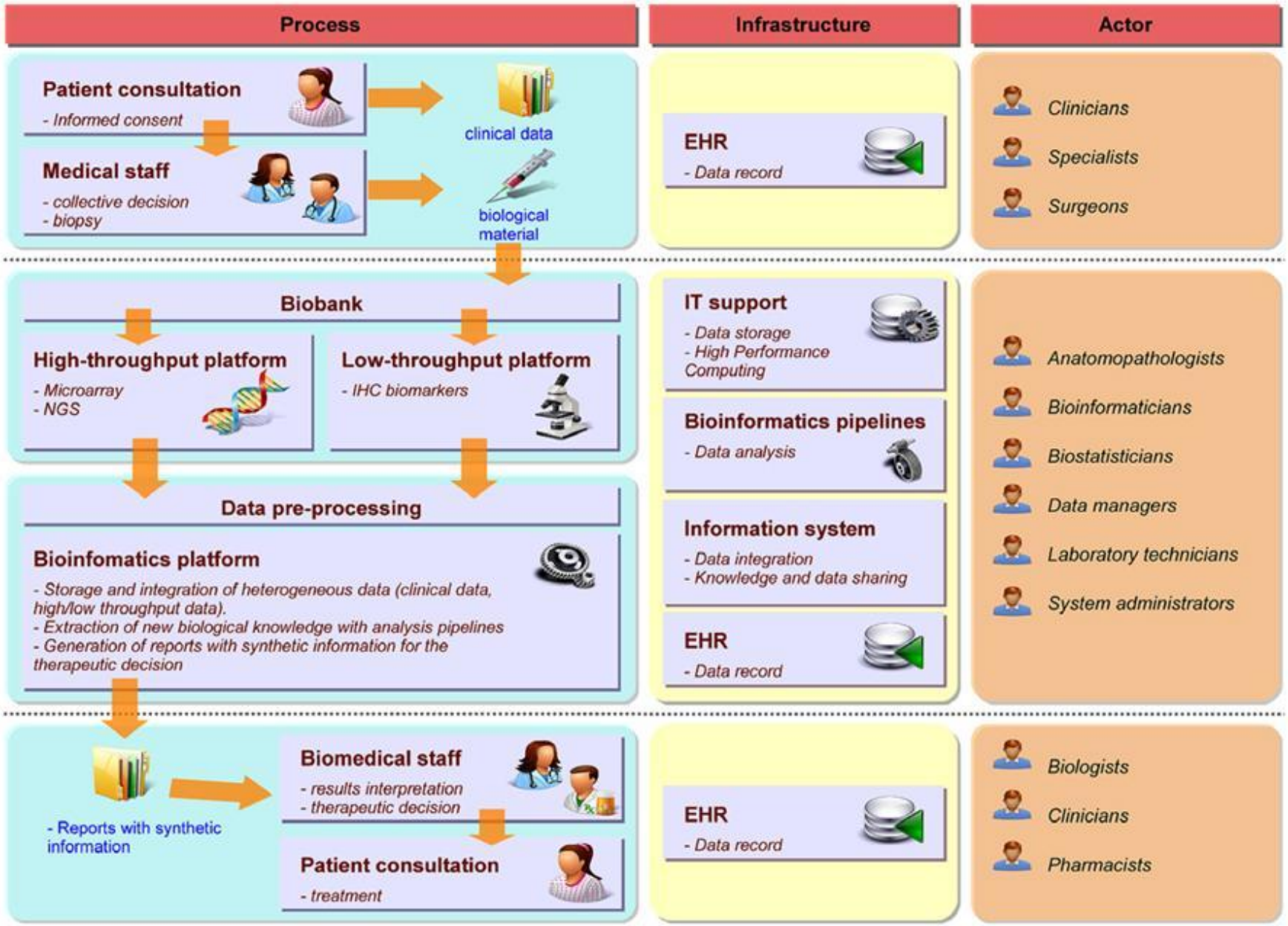
ИСПОЛЬЗОВАНИЕ ЧИПАТОРОВ

Клиническая лаборатория

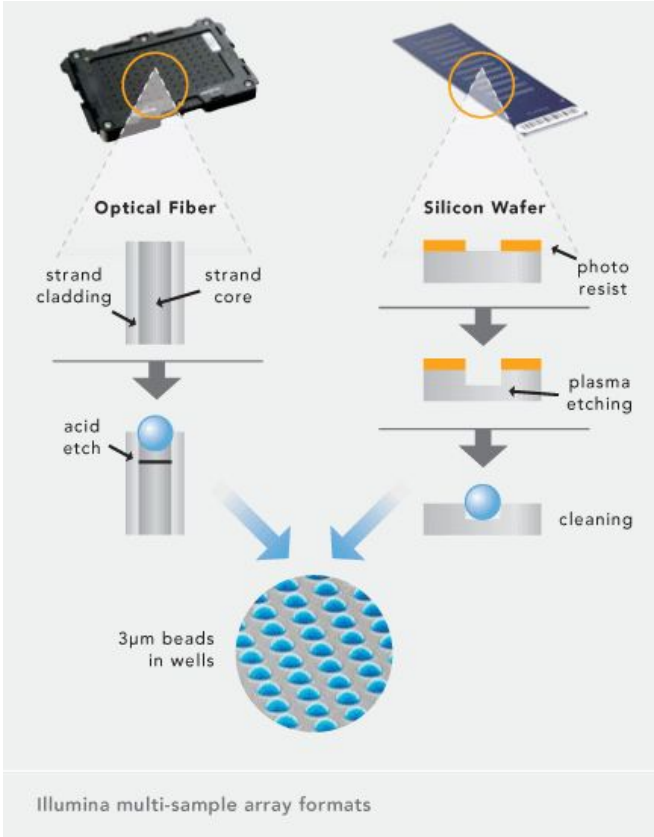

Process
Infrastructure
Actor
Patient consultation
- Informed consent
clinical data
Medical staff
- collective decision
- biopsy
biological material
EHR
- Data record
Clinicians
Specialists
Surgeons
Biobank
High-throughput platform
- Microarray
- NGS
Low-throughput platform
- IHC biomarkers
IT support
- Data storage
- High Performance Computing
Bioinformatics pipelines
- Data analysis
Information system
- Data integration
- Knowledge and data sharing
EHR
- Data record
Anatomopathologists
Bioinformaticians
Biostatisticians
Data managers
Laboratory technicians
System administrators
Data pre-processing
Bioinfomatics platform
- Storage and integration of heterogeneous data (clinical data, high/low throughput data).
- Extraction of new biological knowledge with analysis pipelines
- Generation of reports with synthetic information for the therapeutic decision
- Reports with synthetic information
Biomedical staff
- results interpretation
- therapeutic decision
Patient consultation
- treatment
EHR
- Data record
Biologists
Clinicians
Pharmacists

DISTRIBUTED VS CONSOLIDATED

One machine per lab
Optimized usage time and PI control
Sample prep and data analysis is done inside the lab

All equipment is in one center
Data analysis is conducted by the same team of bioinformaticians
Sample prep by the same technicians

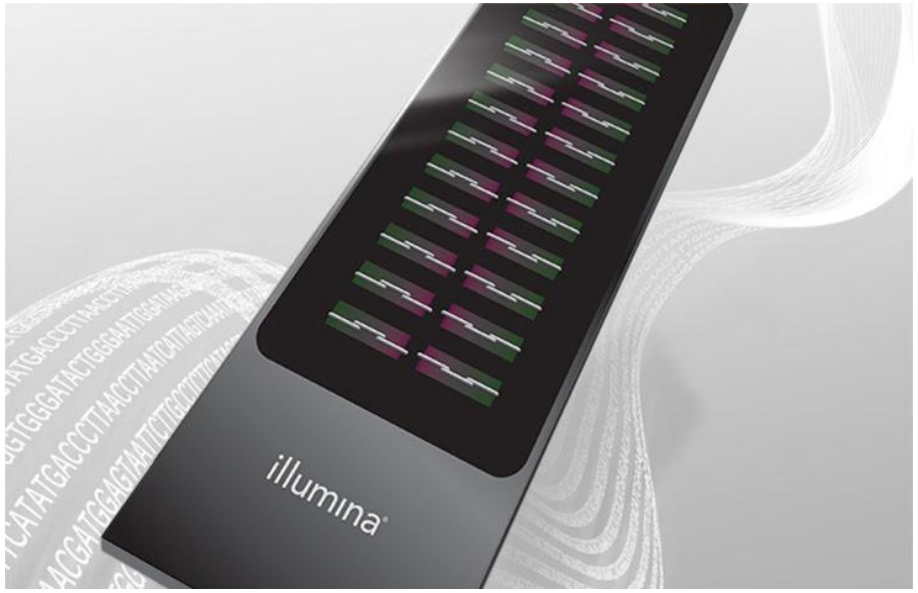
ТЕХНОЛОГИЯ ISCAN



Illumina multi-sample array formats

СКОЛЬКО НАДО ИЗМЕРЯТЬ СНИПОВ

- High Density - До 1.2 миллиона маркеров
  - Можно измерять много снипов
  - Структурные изменения на геноме
  - Мало непокрытых участков
- Low density – 15 тысяч снипов
  - Дешево
  - Точно
  - Зачем?



Marker Category	Category Description	Number of Markers
ADME	Pharmacogenomics, from PharmADME.org	1009
AIM	Ancestry Informative markers from exome array (http://genome.sph.umich.edu/wiki/Exome_Chip_Design#Ancestry_Informative_Markers)	2910
Blood group	From NCBI's dbRBC database covering 51 blood group defining genes http://www.ncbi.nlm.nih.gov/projects/gv/rbc/xslcgi.fcgi?cmd=bgmut/systems	1659
Fingerprint	High MAF SNPs unlikely to be in LD with each other, from http://www.cstl.nist.gov/strbase/SNP.htm and http://alfred.med.yale.edu/alfred/index.asp	477
Linkage	Linkage Panel by Illumina, contains heterozygous SNPs to test for Mendelian disorders, from Linkage 12 array	5486
Extended MHC	Variants from extended major histocompatibility complex MHC covering 8 Mb region containing immune markers	930
Mitochondrial	Determination of mtDNA haplogroups	141
Sex chromosomes	X-chromosome specific Y-chromosome specific Pseudoautosomal Regions	1840 1401 535

Pharmacogenomics biomarkers (N = 1,009) were selected from the PharmADME.org database according to the list of most common requests from Illumina's collaborators.

Ancestry Informative Markers (AIMs) (N = 2,910) were comprised from two sources. The first source, "African American vs. European Ancestry", is a grid of 3,388 markers with more or less even distribution on chromosomes, with an approximate density of one per Mb, and the strong ability to differentiate samples of African and European ancestry deposited in the 1000 Genomes Project [1]. Among these, the markers previously included in the Illumina Omni 2.5M array were favored, and the markers represented by A/T or G/C alleles avoided. The second set, capable to sort out Native American versus European Ancestry, contains 1,000 markers selected to be in low linkage disequilibrium to one another (defined as $R^2 \leq 0.1$ in Native American populations) and at least 250 kb apart from each other. SNPs with a significant heterogeneity of the frequencies in same-continent populations were excluded. In this subset, all markers were previously genotyped in three samples of European ancestry and six samples of Native Americans. Among 2,910 AIMs, there was a bias for autosomal locations within coding regions.

Blood Group Markers (N = 1,659) were retrieved from the Blood Group Antigen Gene Mutation Database (dbRBC) [2] maintained by NCBI. This set of markers covers 51 genes and is capable to differentiate 34 blood groups including less common ones, such as the Chido/Rodgers Blood Group System and C4 complement.

Sex chromosomes. This group includes 1,840 variants located on the X chromosome, 1,401 on Y chromosome and 535 from the pseudoautosomal regions PAR1, PAR2, and PAR3 present in both sex chromosomes.

Fingerprinting markers (N = 477). These “high MAF no LD” variants were submitted by the Population Architecture using Genomics Epidemiology (PAGE) Consortium http://www.cstl.nist.gov/strbase/SNP.htm and http://alfred.med.yale.edu/alfred/index.asp.

Linkage markers (N = 5,486) were taken from a previous Illumina product HumanLinkage and represent common variants most likely to be correctly imputed. http://support.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_imputation.pdf

Extended set of MHC markers (N = 930). These markers reside within an extended Major Histocompatibility Complex (MHC) region (8Mb) and are indispensable to determine histocompatibility and predisposition to a variety of chronic diseases.

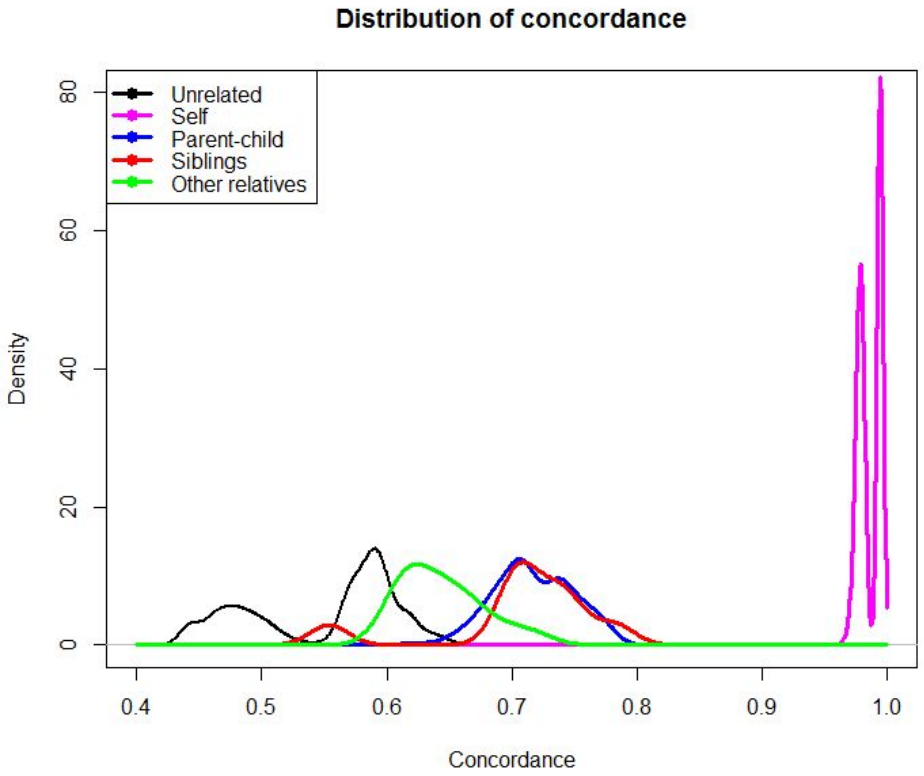
Mitochondrial markers (N = 141). Maternally inherited variations of a mitochondrial genome constitute a set of distinct signatures known as mitochondrial haplogroups proven extremely valuable in discerning human evolution and migration patterns, and thus used extensively in forensics.

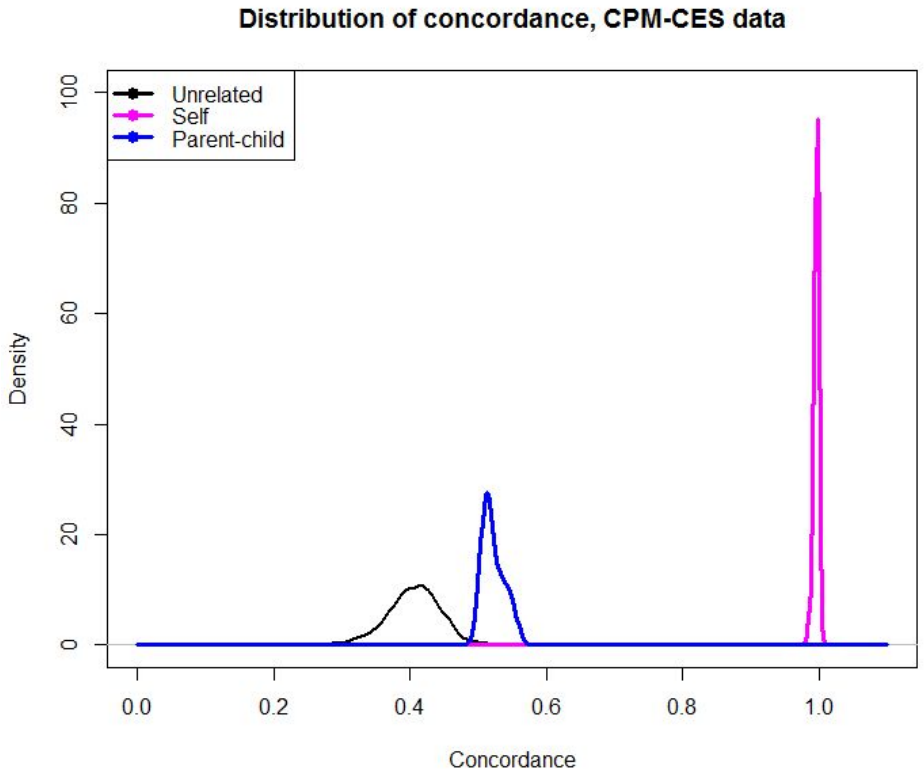
CONCORDANCE OF VARIANT CALLING BETWEEN PLATFORMS

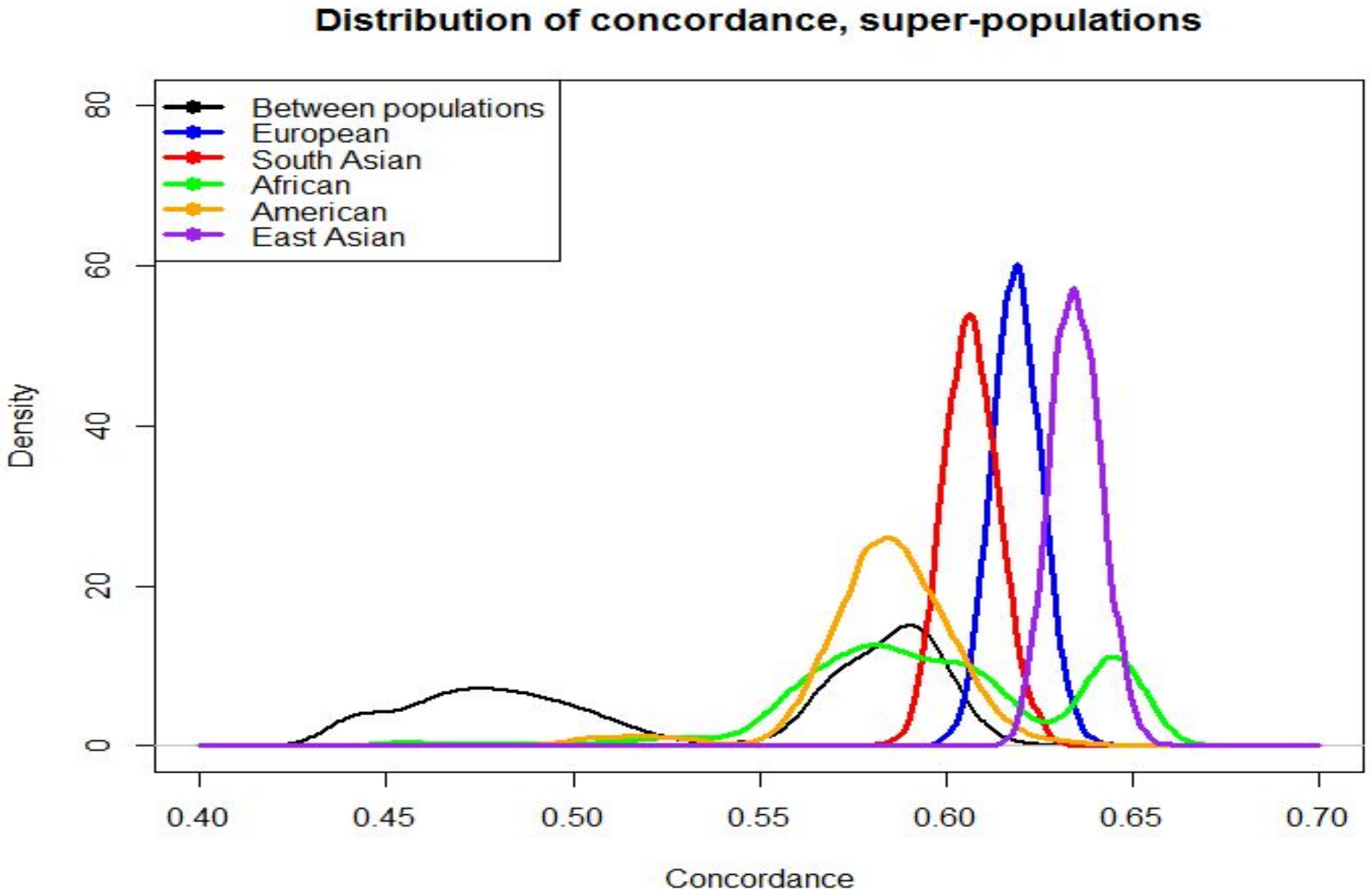
- We compared the HumanQC data with the 1,000 Genomes WGS, Omni 2.5 (OMNI) and Affymetrix 6.0 (AFFY) microarray data.
- Concordance of genotype calls between HumanQC and OMNI, AFFY 6.0 and NGS (using 1000 Genomes Project data) (Genomes Project, Auton et al., 2015)was found to be 99.63%, 99.66% and 99.39% correspondingly when only non-missing bi-allelic calls between both sets are compared (except for the Y chromosome comparison between the HumanQC and 1000 Genomes data, which has a concordance of 95.68%).

Platform Comparison	Number of Common Variants	Number of Variants in HumanQC Only	Number of Variants in the Comparison Platform Only
HumanQC vs Affymetrix	3290	10365	33
HumanQC vs Omni (Illumina)	9166	4489	159
HumanQC vs 1000 Genomes	12820	835	908
CPM HumanQC vs CPM CES	761	12821	114862

ПОХОЖЕСТЬ ОБРАЗЦОВ






Distribution of concordance, super-populations
Between populations
European
South Asian
African
American
East Asian
Density
0
20
40
60
80
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Concordance

ПОХОЖЕСТЬ СУПЕР-ПОПУЛЯЦИЙ

	AFR	AMR	EAS	EUR	SAS
AFR	0.559928	0.417711	0.431412	0.390875	0.417188
AMR		0.55099	0.543173	0.558732	0.550021
EAS			0.60714	0.539716	0.555402
EUR				0.597023	0.567043
SAS					0.576606

1. Concordance of Africans vs. “all other super populations” is lower (0.39-0.43)
2. Concordance values inside same population are higher (0.55-0.61)
3. Other pairings are similar in concordance with within-population concordance (0.54-0.57)

Comparison with HumanQC	OMNI	AFFY 6.0 (WES)	1KG no Y	1KG only Y
Concordance, %	99.63% (97.38%)	99.66% (99.91%)	99.39% (98.63%)	95.68% (99.86%)
shared markers	7,781 (708)	2,526 (205)	10,096 (1,177)	47 (3)
non-missing genotype calls	4,806,200 (465,156)	1,637,639 (132,796)	5,071,268 (591,461)	11,458 (732)
matching samples	657	652	503	252
missing genotype calls	305,917 (29,971)	9,313 (864)	7,020 (570)	386 (24)
missing genotype calls, %	5.98% (6.44%)	0.56% (0.64%)	0.13% (0.09%)	3.26%
non-missing mismatches	17,782 (11,384)	5,607 (119)	30,829 (8,116)	485 (1)
only one allele matches	2,061 (96)	1,789 (116)	12,901 (1,623)	(1)
% one allele matches out of all genotype mismatches	11.5% (0.8%)	31.9% (97.5%)	41.8% (20%)	(100%)

БЯКИ- 30 ТОЧЕК

Chromosome: Position	Marker name	Most discordant genotype calls between HumanQC and			Average genotype call score
		1KG	OMNI	AFFY	
chr4:69512637	rs4148271	451	596		0.915051523
chr7:87133470	rs17064	442	560		0.92561961
chr6:29712759*	exm-rs2844845	395	48	520	0.959497865
chr19:41354533	rs1801272	489			0.662827184
chr8:145639681	rs1871534	464			0.871873314
chr6:32411846	exm-rs2239802	352	437		0.83848067
chr13:20901724*	rs1335873	323	34	425	0.849843877
chr15:74710485	rs2072649	410			0.949527

КОЭФФИЦИЕНТ РОДСТВА

Degree of relationship	Relationship	Coefficient of relationship (r)	Kinship coefficient
0	identical twins; clones	100%	0.5
1	parent-offspring	50%	0.25
2	full siblings	50%	0.25
2	3/4 siblings or sibling-cousins	37.50%	0.1875
2	grandparent-grandchild	25%	0.125
2	half siblings	25%	0.125
3	aunt/uncle-nephew/niece	25%	0.125
4	double first cousins	25%	0.125
3	great grandparent-great grandchild	12.50%	0.0625
4	first cousins	12.50%	0.0625
6	quadruple second cousins	12.50%	0.0625
6	triple second cousins	9.38%	0.0469
4	half-first cousins	6.25%	0.03125
5	first cousins once removed	6.25%	0.03125
6	double second cousins	6.25%	0.03125
6	second cousins	3.13%	0.01565
8	third cousins	0.78%	0.0039
10	fourth cousins	0.2%	0.001

СРАВНЕНИЕ ГЕНОМНОГО РОДСТВА С ЗАПИСАННЫМ

- 2,208 pairs of individuals.
- Two of the recorded pairs of siblings in 1000 Genomes database (NA20344/NA20334 and NA20344/NA20336) have suspiciously weak similarity (kinship of 0. 0148 and -0.0081), while the pair NA20334/NA20336 have kinship consistent with siblings (0.2251).
- http://www.internationalgenome.org/data-portal/sample/NA20344.

КОЕФФИЦИЕНТ РОДСТВА

Relatedness	Median Kinship	SAMPLE SIZE	Theoretical kinship	MIN KIN	MAX KIN
Siblings	0.2354	9	0.25	-0.0081	0.3029
Parent-Child	0.2441	221	0.25	0.1712	0.2620
Second Order	0.1107	9	0.125-0.1875	0.0714	0.1475
Unrelated	-0.1300	1679	<0.001	-0.3074	0.0443

NA20344

Sample NA20344

Data portal beta

NA20344 details

Sex:	Female
Biosample ID:	SAME124404
Search Coriell:	NA20344

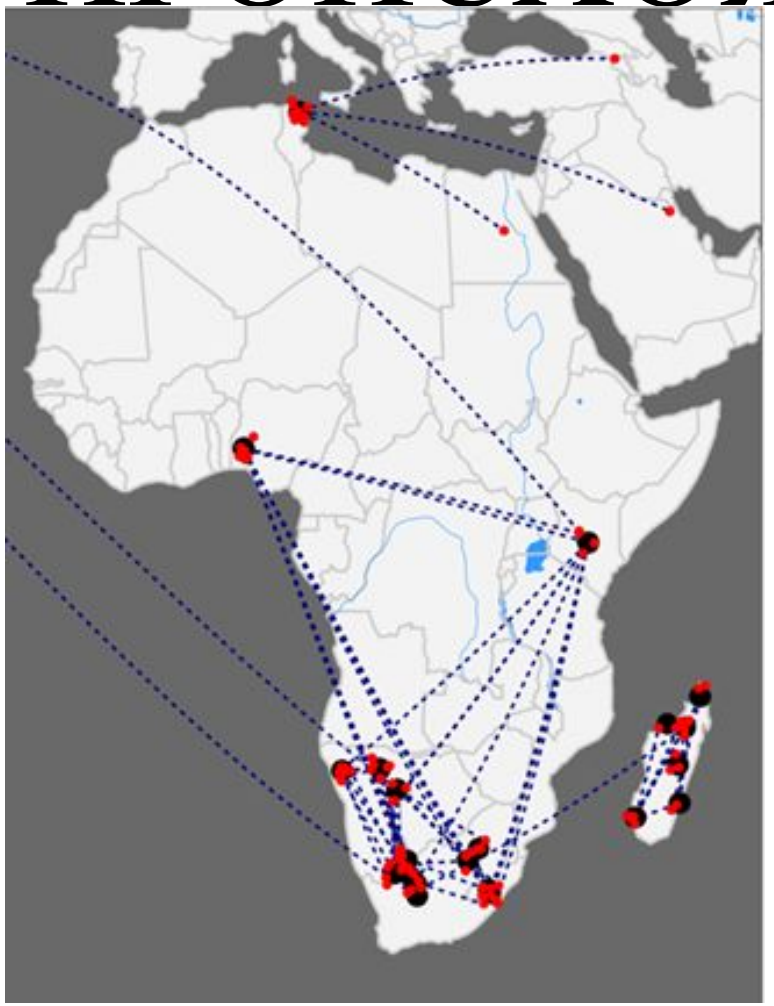
ASW population

Population:	African-American SW
Code:	ASW
Description:	African Ancestry in Southwest US
Superpopulation:	African
Superpopulation code:	AFR

RelatedSamples

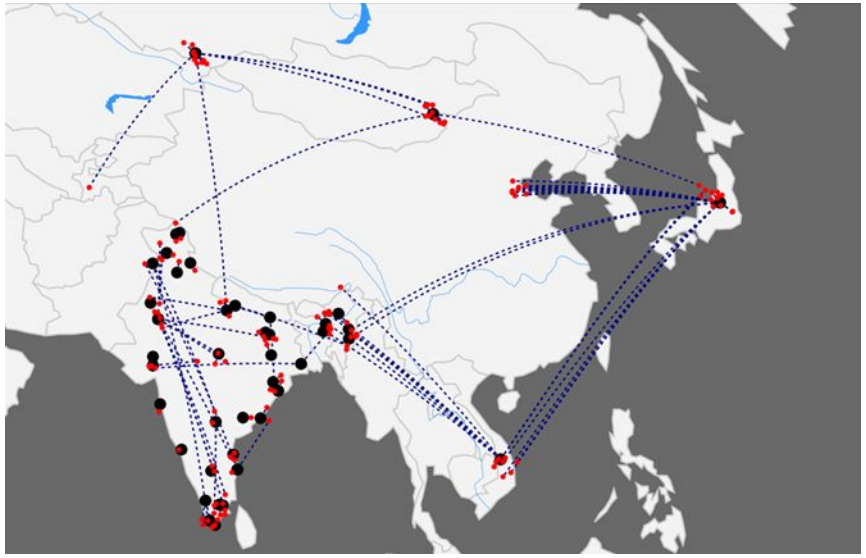
Sibling	NA20334
Sibling	NA20344
Child	NA20345
Sibling	NA20349
Second Order	NA20350

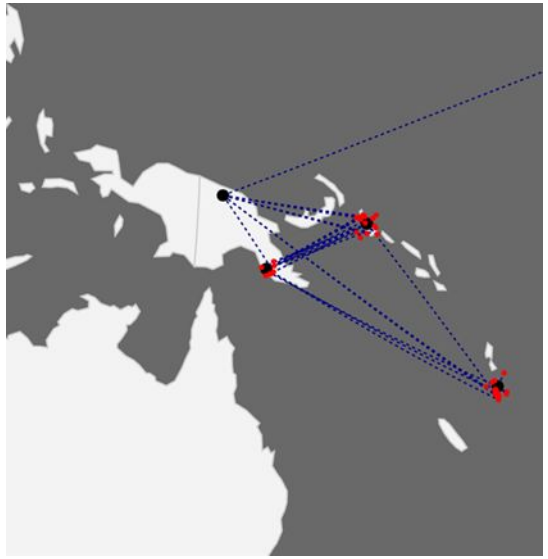
ТОЧНОСТЬ ПРОИСХОЖДЕНИЯ



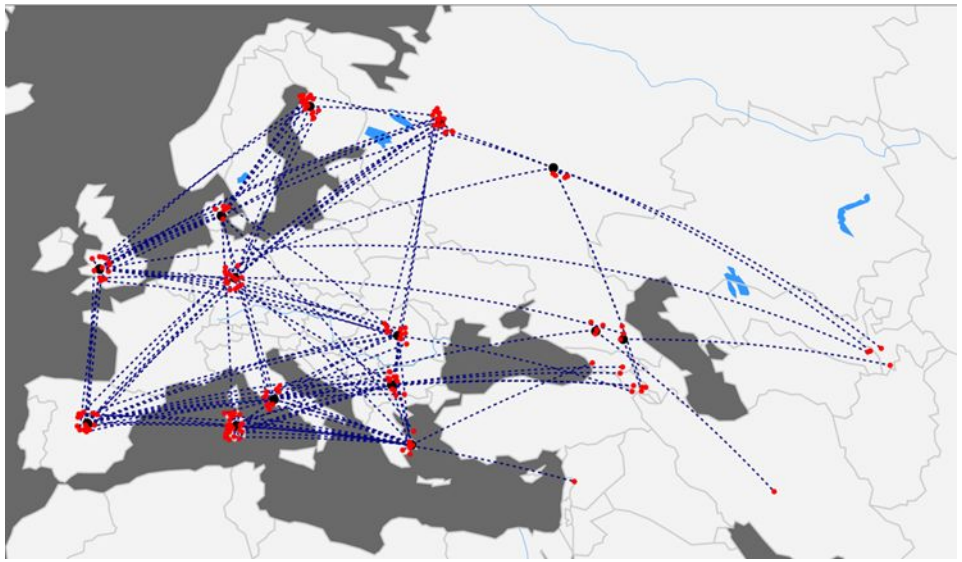
- Из 150К маркеров чипа Geno2.0 в нашем распоряжении только 1900
- Как точно мы можем предсказать?

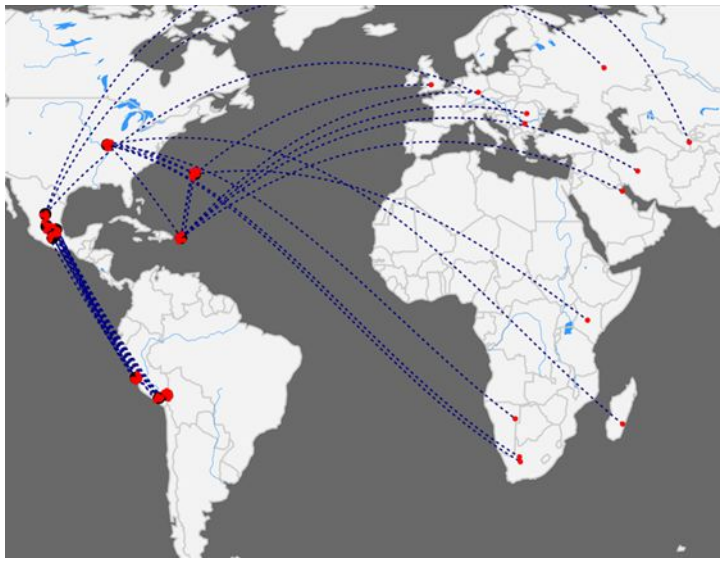
АЗИЯ И ОКЕАНИЯ





ЕВРОПА И АМЕРИКА



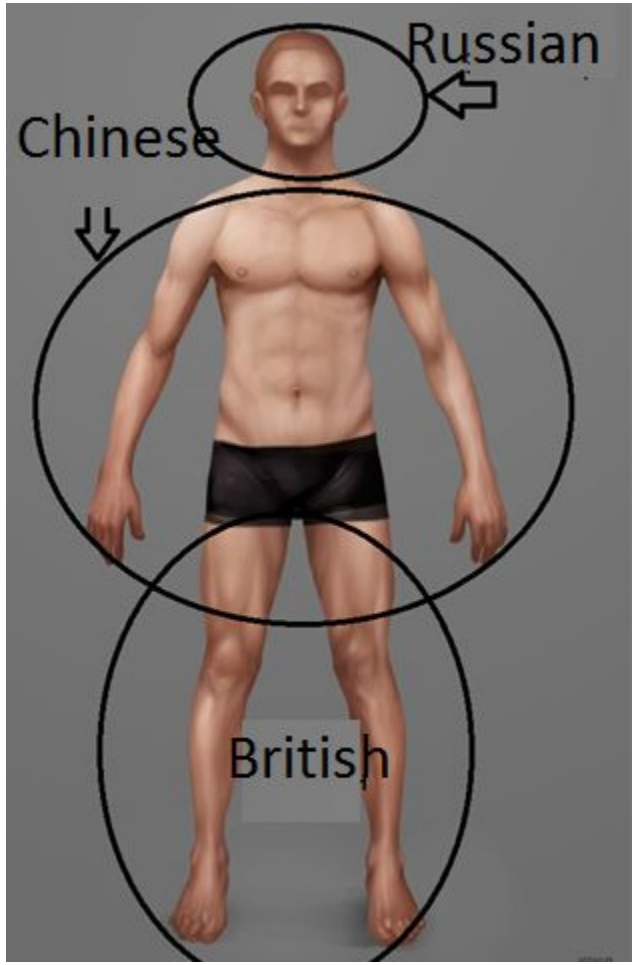


ПОПУЛЯЦИИ

POPULATION CODE	NUMBER OF SAMPLES	POPULATION
ASW	90	Americans of African Ancestry in SW USA
CEU	88	Utah Residents (CEPH) with Northern and Western Ancestry
CHB	38	Han Chinese in Beijing, China
GIH	77	Gujarati Indian from Houston, Texas
JPT	45	Japanese in Tokyo, Japan
MXL	82	Mexican Ancestry from Los Angeles, USA
PUR	72	Puerto Ricans from Puerto Rico
TSI	83	Toscani in Italy
YRI	88	Yoruba in Ibadan, Nigeria

reAdmix

- ReAdmix developed to treat individuals of mixed origin and represents an individual as a linear combination of admixture vectors of reference populations
  - 30%British+10%Russian+60%Chinese
  - $P=a_1rs_1 + a_2rs_2 +... + a_prs_p+error$



HOW IT WORKS

- We assume right away that the given ancient proportions contain error
- Start with a guess population
- Add/remove populations to achieve optimal fit
- Conditional optimization (such as "*I know that there was a Jewish ancestor somewhere in my pedigree*")

READMIX APPROACH

Aim: to find the smallest subset of modern populations whose combined admixture components are similar to those of the individual within a small tolerance margin.

Let $R=\{r_i\}_{i=1}$ be the set of **modern populations**, where each $r_i=(r_{i,1}, ..., r_{i,K})$ and K is the **dimension** (K=9).

We seek two sets $S=(s_1,...,s_p)$ and $A=(a_1,...,a_p)$, where s_i are the indices of modern populations in the approximation

$$P = a_1 r_{s(1)} + a_2 r_{s(2)} + \cdots + a_p r_{s(p)}$$

where a_i are the coefficients of modern of test vector T

The algorithm consists of three phases:

1. Iteratively **build** the first candidate solution and improve it.
2. Generate the predefined number M of additional candidate solutions randomly and apply the **Differential Evolution (DEEP)**.
3. **Identify** the populations that have stable membership in the solution across the set, that is, are part of solution in at least **75% of cases**.

PHASE 1. BUILD AND IMPROVE THE INITIAL SOLUTION SET

Find the population vector with the highest affinity score.
Append this population to the solution set.
Calculate the weight of the population vector to be proportional to the maximal possible.

Affinity

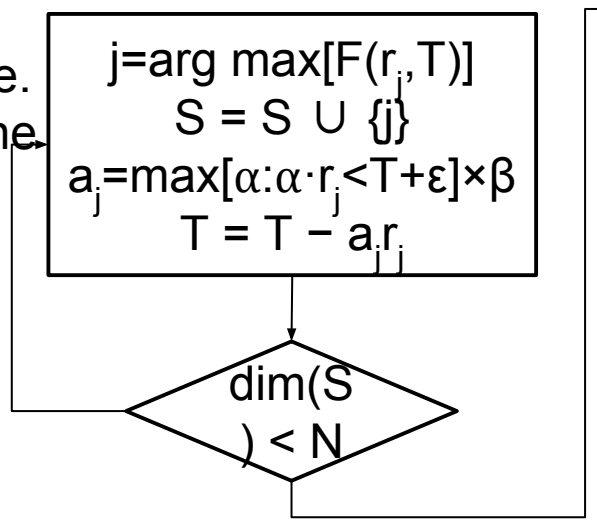
$$F(P,T) = arg\min_{\alpha} L(d(\alpha))$$

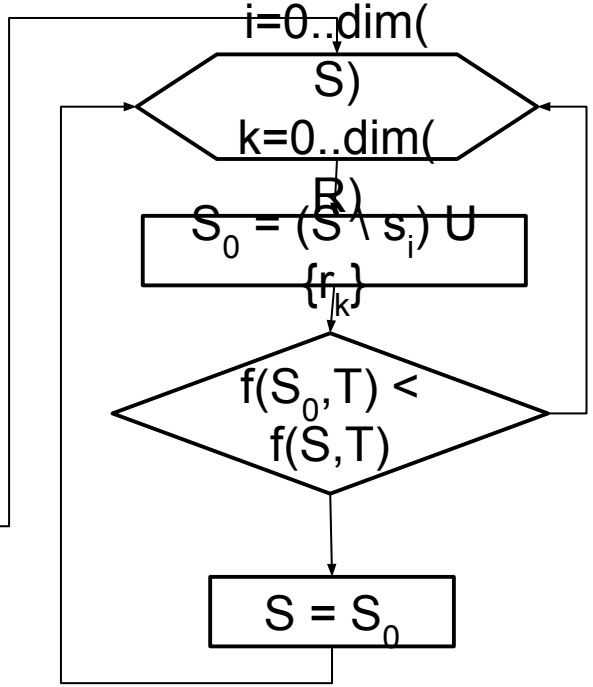
minimizes the **loss function**

$$L(d) = \sum_{i=1}^{K} d_i^2 + \sum_{i:d_i<-\varepsilon} (1 + 2|d_i|)$$

where

$$d = T - \alpha P$$





S - individual's ancestry

N – desired size of the solution

T – test vector

R – the set of modern populations

For all populations **x** in the current solution and for all **y** outside the solution, **replace** x with y, if it reduces the error.

SOLUTION BY GLOBAL STOCHASTIC AND LOCAL SEARCH

Objective function

$$f(S,T) = \min_{A=(a_1,a_2,\ldots,a_p)} \max_{k=1\ldots K} |P - T|,$$

wher $P = a_1 r_{s(1)} + a_2 r_{s(2)} + \cdots + a_p r_{s(p)}$ is the approximation of T
e

$S_1, S_2, \ldots, S_M$

$S_2, \ldots, S_M$ - are generated randomly

1. **Differential Evolution**

G_{max} number of iterations

2. **Local search** over all populations

close to the preliminary solution

This step selects between **related populations** (e.g. Belorussian, Russian, and Ukrainian) that could have been misplaced in previous steps

DE: optimization method used for multidimensional real valued functions.
Good for
Treatment of noisy problems (Storn and Price, 1997)

PHASE 3. AVERAGING

Reliable estimate:

- the **populations** that are part of solution in at least **75% of cases**,
- their **averaged** estimates of **admixture proportion**.

S_1, S_2, ... , S_{M-1}, S_M — the set of candidate solutions.

S_1:$(\mathbf{a_1,r_1})_1$,$(a_2,r_2)_1$,..,$(\mathbf{a_{p-1},r_{p-1}})_1$,$(a_p,r_p)_1$

S_2:$(a_1,r_1)_2$,$(\mathbf{a_2,r_2})_2$,..,$(a_{p-1},r_{p-1})_2$,$(a_p,r_p)_2$

..

S_{M-1}:$(a_1,r_1)_{M-1}$,$(a_2,r_2)_{M-1}$,:..,$(a_{p-1},r_{p-1})_{M-1}$,$(\mathbf{a_p,r_p})_{M-1}$

S_M:$(a_1,r_1)_M$,$(\mathbf{a_2,r_2})_M$,..,$(a_{p-1},r_{p-1})_M$,$(a_p,r_p)_M$

$S=(s_1,...,s_p)$

$A=(a_1,...,a_p)$

Final solution:

r belongs to **final solution** if:

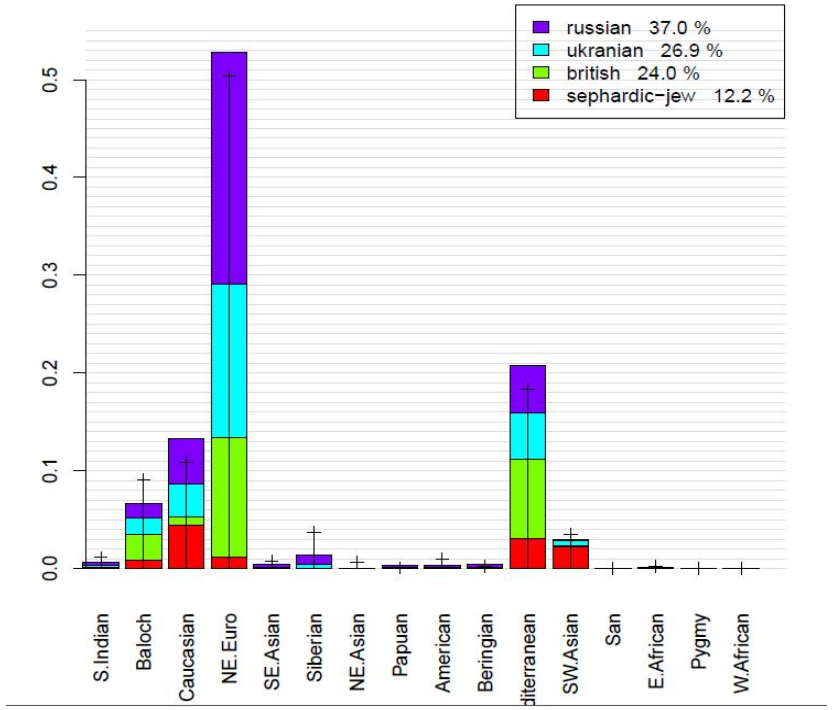
$r=r_{1,1}=r_{2,M-1}=...=r_{p,M}$ — is the same **modern population**

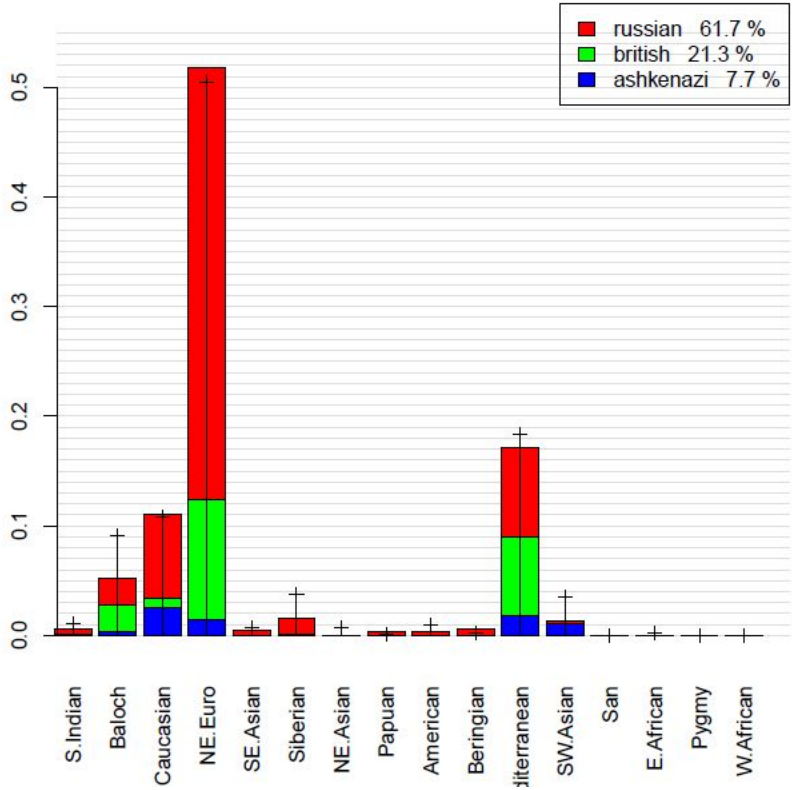
and

is present in ***L*** > 75% solutions

and $a=(1/L)(a_{1,1}+a_{2,M-1}+..+a_{p,M})$

EXAMPLE


russian 37.0 %
ukranian 26.9 %
british 24.0 %
sephardic-jew 12.2 %
0.0
0.1
0.2
0.3
0.4
0.5
S.Indian
Baloch
Caucasian
NE.Euro
SE.Asian
Siberian
NE.Asian
Papuan
American
Beringian
SW.Asian
San
E.African
Pygmy
W.African


russian 61.7 %
british 21.3 %
ashkenazi 7.7 %
0.0
0.1
0.2
0.3
0.4
0.5
S.Indian
Baloch
Caucasian
NE.Euro
SE.Asian
Siberian
NE.Asian
Papuan
American
Beringian
SW.Asian
San
E.African
Pygmy
W.African

BENCHMARK-1

Take sample that were validated by GPS_1

Then we used ReAdmix in (a) unconditional and (b) conditional with incorrect guess mode.

(a) 94% of the samples were mapped to their reported ethnic group, and average distance to the correct location was 54±13 miles.

(b) with randomly chosen incorrect guess, 92% of samples was mapped to the reported ethnic group, with average distance to the correct location equal to 65±17 miles.

In all cases ReAdmix correctly identified the cases as un-mixed

BENCHMARK-2: SIMULATED 500 50:50 MIXTURES

Testing mode	Percent of at least one correctly predicted origin	Percent of completely correct predictions	Average distance to correct population, miles
Unconditional	78%	56%	324±46
Conditional on the equal weights, populations unknown	91%	74%	176±33
Conditional on one of the populations, weights unknown	NA	71%	104±16

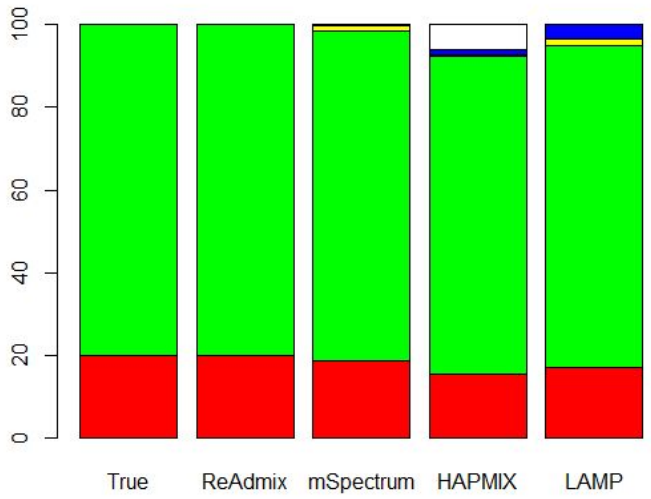
BENCHMARK-2: SIMULATED 500 50:25:25 MIXTURES

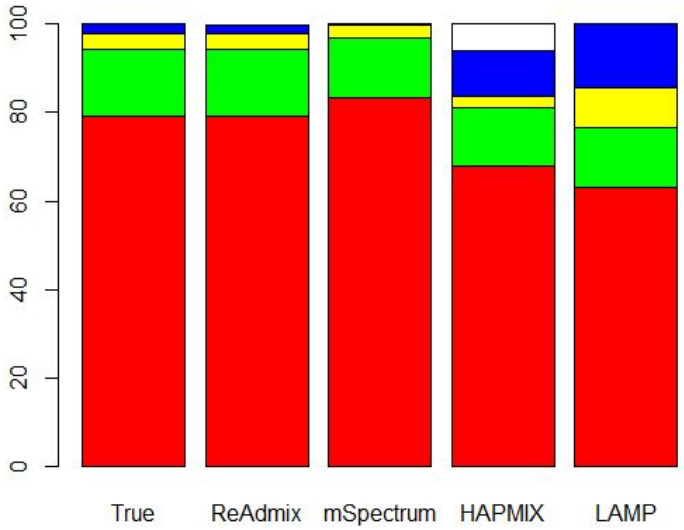
Testing mode	Percent of at least one correctly predicted origin	Percent of completely correct predictions	Average distance to correct population, miles
Unconditional	88%	21%	346±34
Conditional on the major population, weights unknown	73%	27%	222±26
Conditional on one of the minor populations, weights unknown	78%	33%	234±30

http://chcb.saban-chla.usc.edu/gps/

SOHN ET (2012) AL BENCHMARK

- 4 components
- 2 comp





4-dim space: European, African, Native American and East Asian

Color coding: red-European, green-African, yellow- Native American, blue-East Asian, and white- unassigned

РЕАДМИКС

POPULATION	AVERAGE NUMBER OF ETHNIC ASSIGNMENTS PER INDIVIDUAL	WEIGHT OF THE MOST SIGNIFICANT ETHNIC ASSIGNMENT
PUR	1.78	0.59
CEU	1.58	0.67
MXL	1.39	0.65
ASW	1.28	0.76
TSI	1.25	0.74
GIH	1.18	0.82
CHB	1.13	0.90
YRI	1.01	0.99