Towards understanding population effect on cancer

Qingyao Huang

Baudis group

A typical human genome

~3 billion base pairs

~20k genes

encode proteins

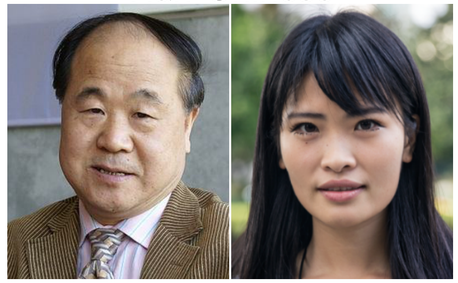
A typical human genome

~3 billion base pairs

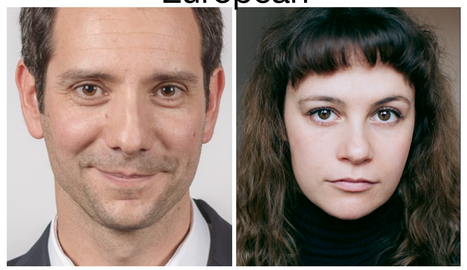
~5 million (germline) variations

Yoruba

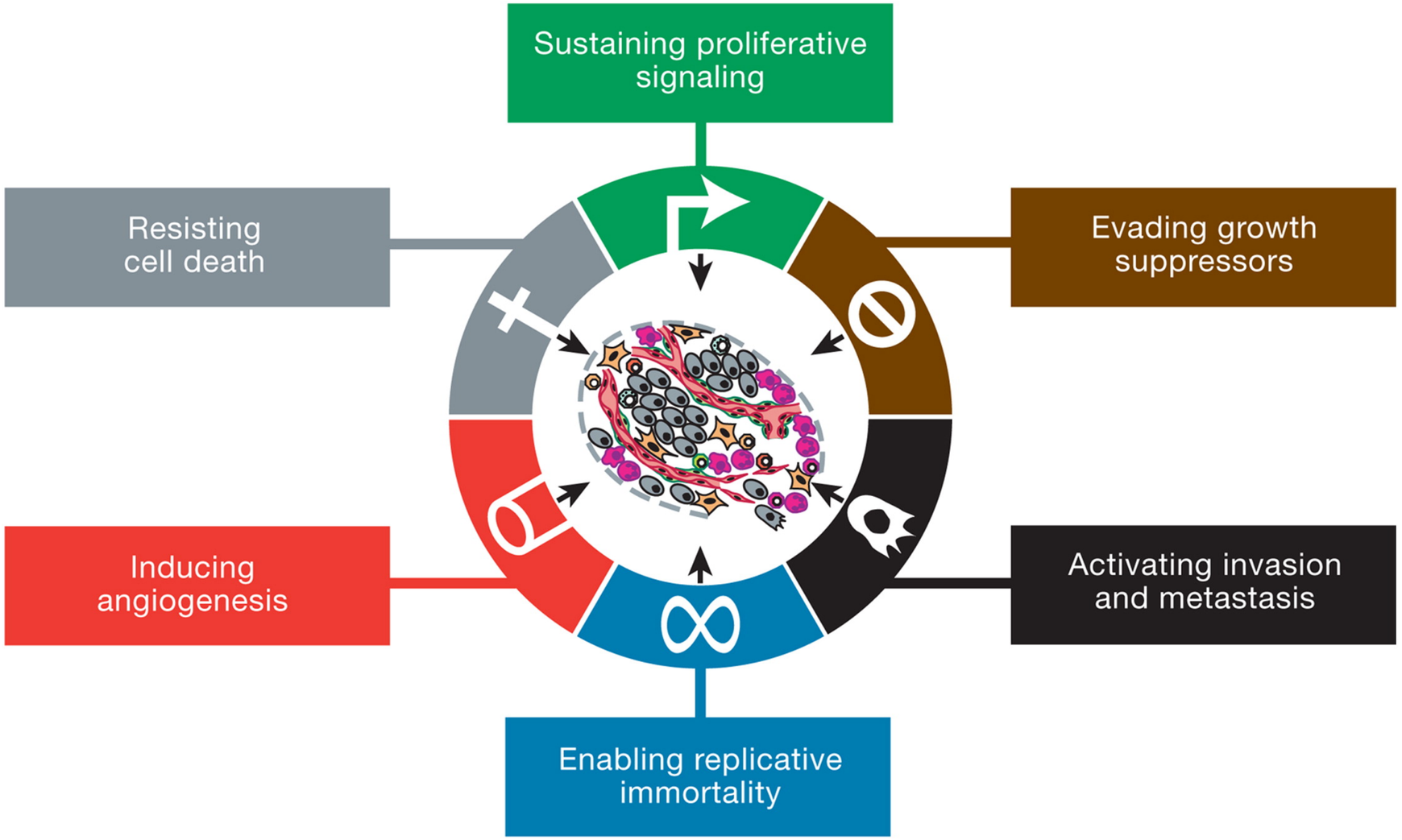
Han-Chinese



European



What happens in cancer?



Hanahan D, Weinberg RA. Cell 2011; 144, 646-74.]

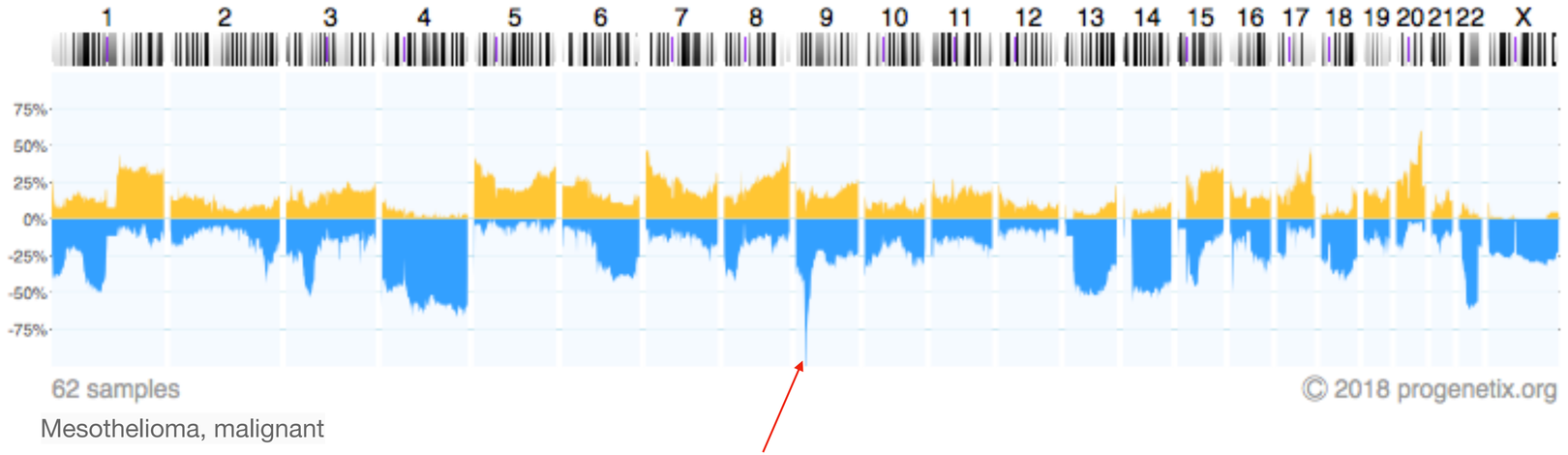
What happens to the genome?

Additional (somatic) variation occurs

~5k observed variants

few drive, most hitchhike

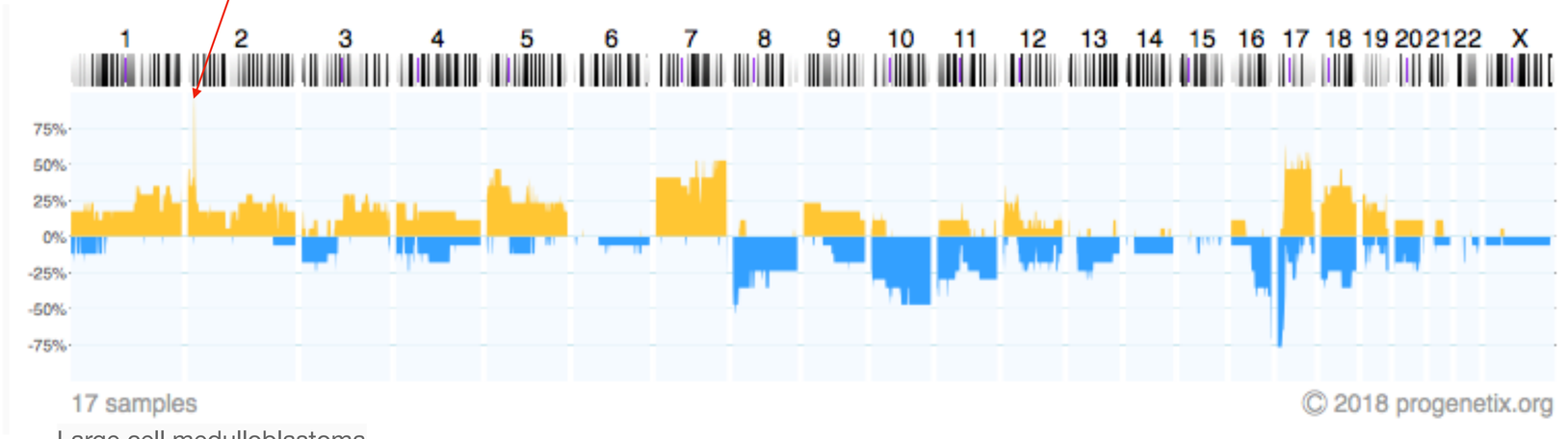
Copy number variation in a cancer genome



Mesothelioma, malignant

CDKN2A/B deletion

MYCN amplification



Large cell medulloblastoma

It can be (partially) inherited as well…

- Early onset: Retinoblastoma

  - congenital mutation in the tumor suppressor RB1

- Later onset: Breast cancer/ Ovarian cancer

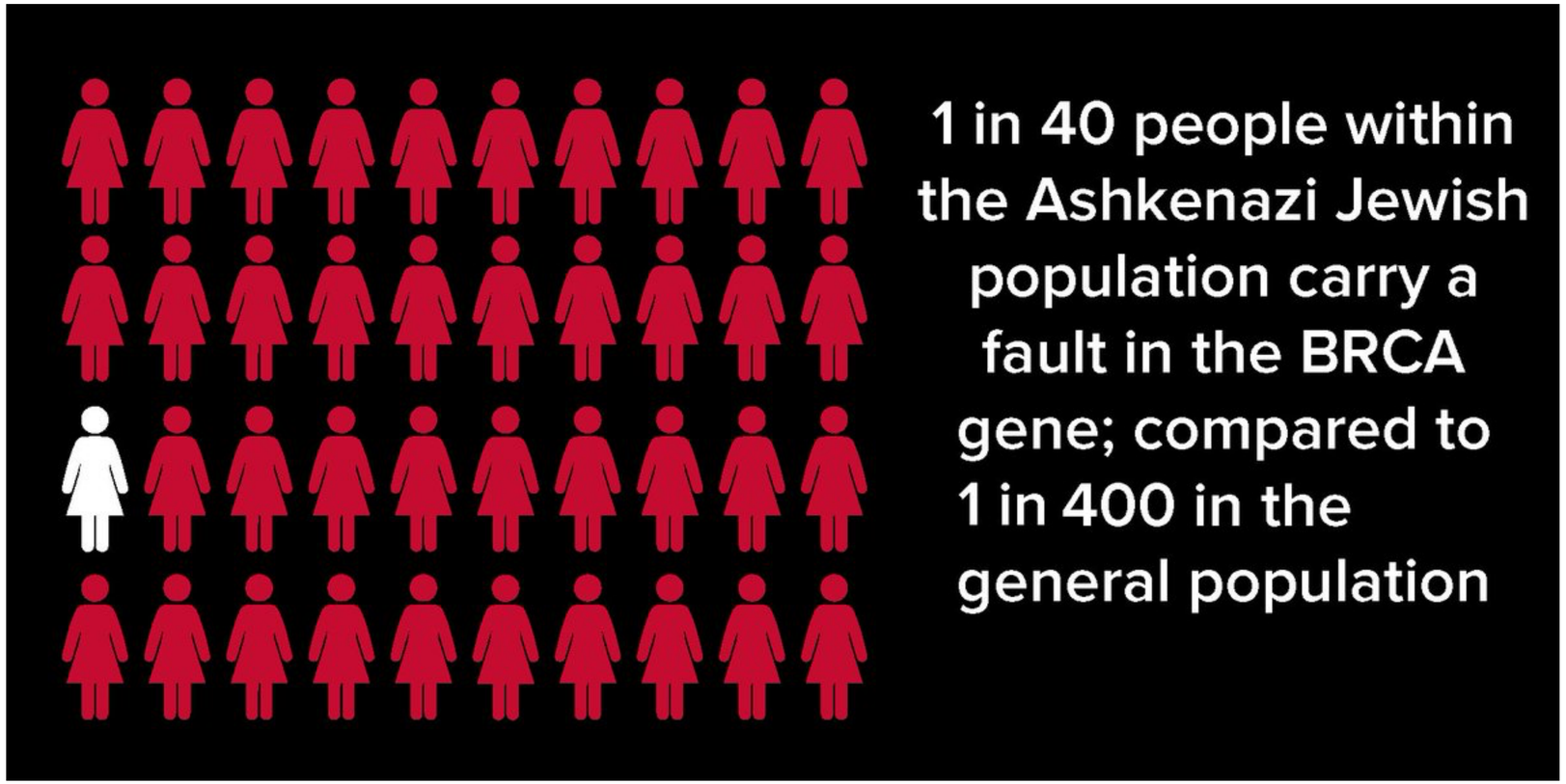
  - mutation in tumor suppressor BRCA1/2 and others…

Breast cancer predisposing genes

Gene	Contribution to Hereditary Breast Cancer
BRCA1	20%–40%
BRCA2	10%–30%
TP53	<1%
PTEN	<1%
Undiscovered genes	30%–70%

adapted from ASCO

BRCA mutation in Ashkenazi Jews



3 founder mutations

BRCA1:

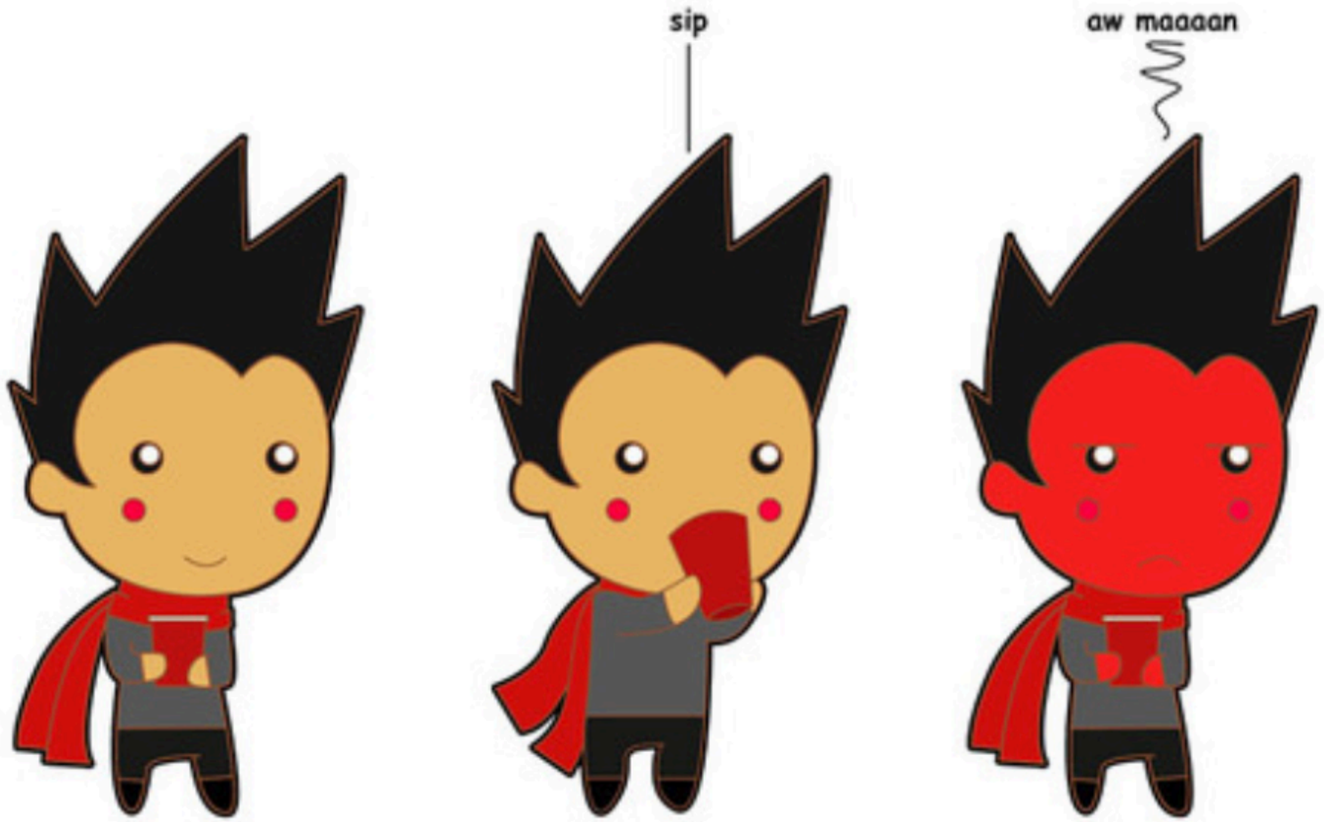
185delAG (~1%)

5382insC (~0.15%)

BRCA2:

6174delT (~1.5%)

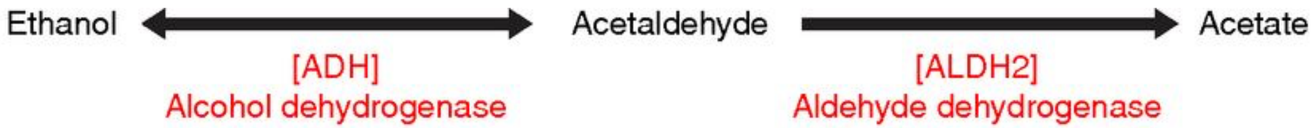
Asia flush

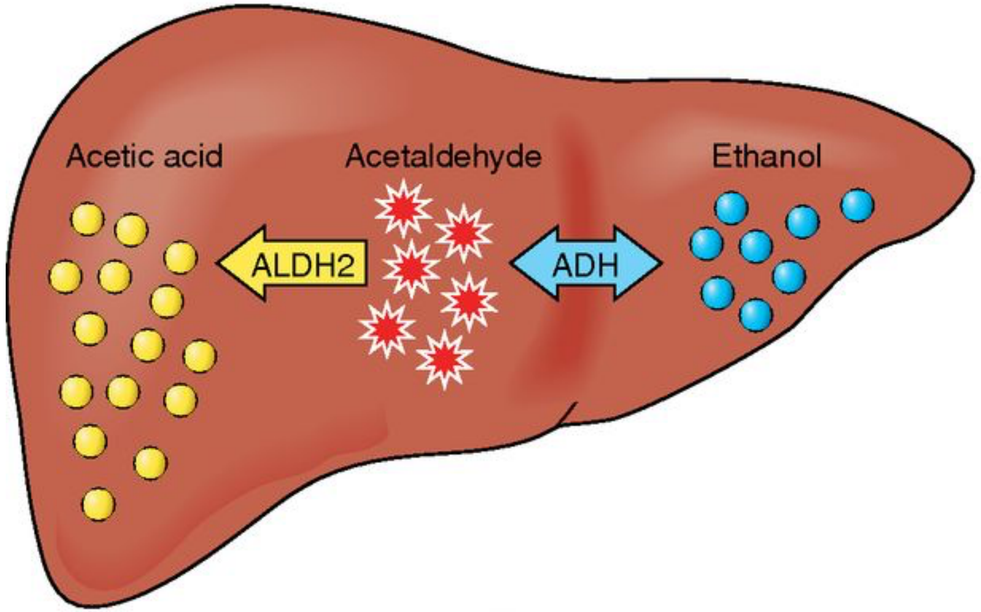


ALDH2*504Lys, enzyme deficiency in 36% East Asians

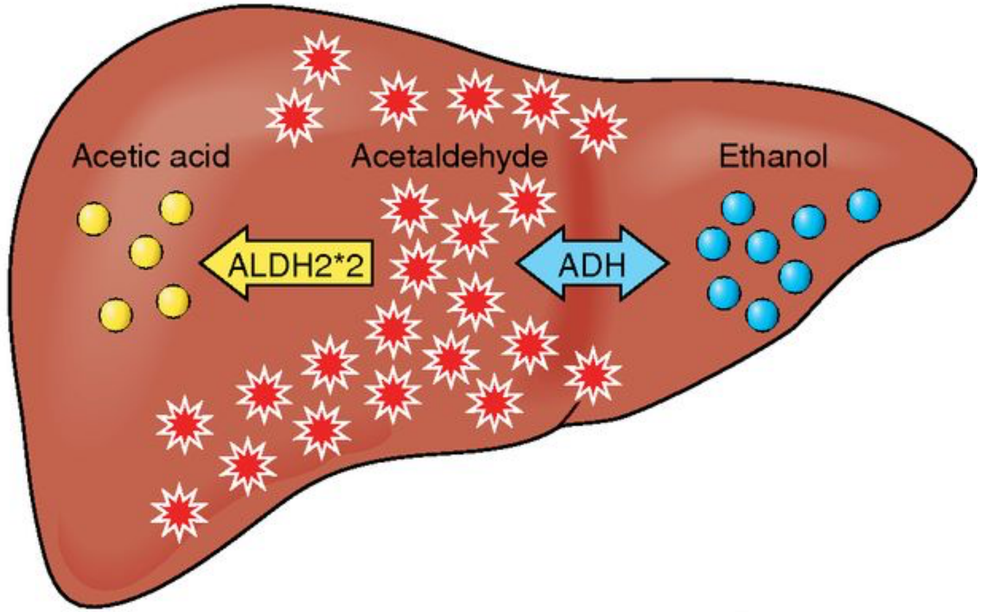
60% higher risk of squamous cell esophageal cancer from alcohol consumption

Asia flush





Normal Alcohol Metabolsm



Accumulation of Acetaldehyde in ALDH2*2

ALDH2*504Lys, enzyme deficiency in 36% East Asians

60% higher risk of squamous cell esophageal cancer from alcohol consumption

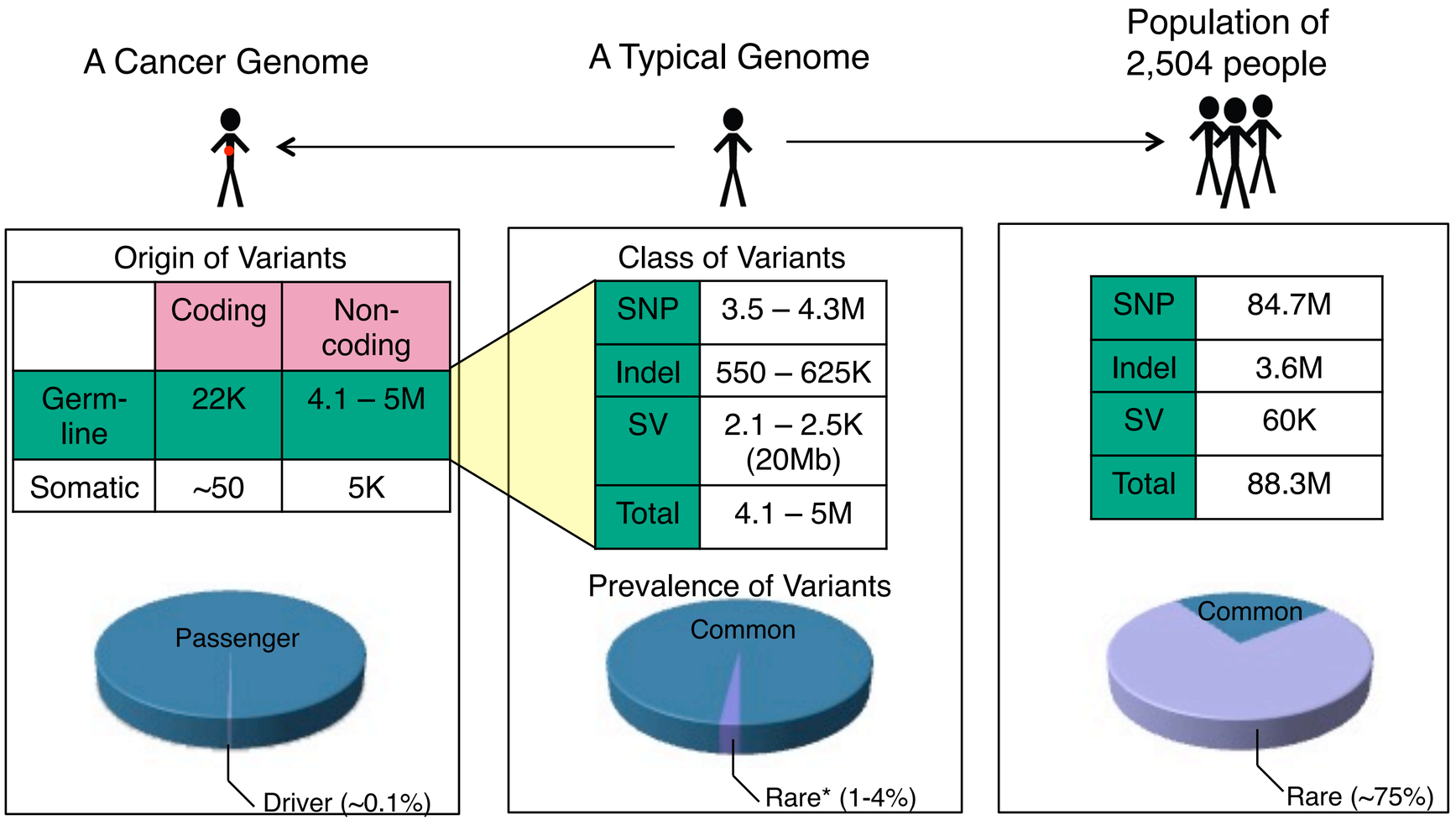
Personalised Medicine in Cancer

the **individual genetic background** and **somatic variants** in tumor

currently mostly check for "**actionable**" mutations

need interpretation of **non-standard** variants

What else in a cancer genome?



* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Goal

Find population-specific cancer-predisposing germline variations

Goal

Find population-specific cancer-predisposing germline variations

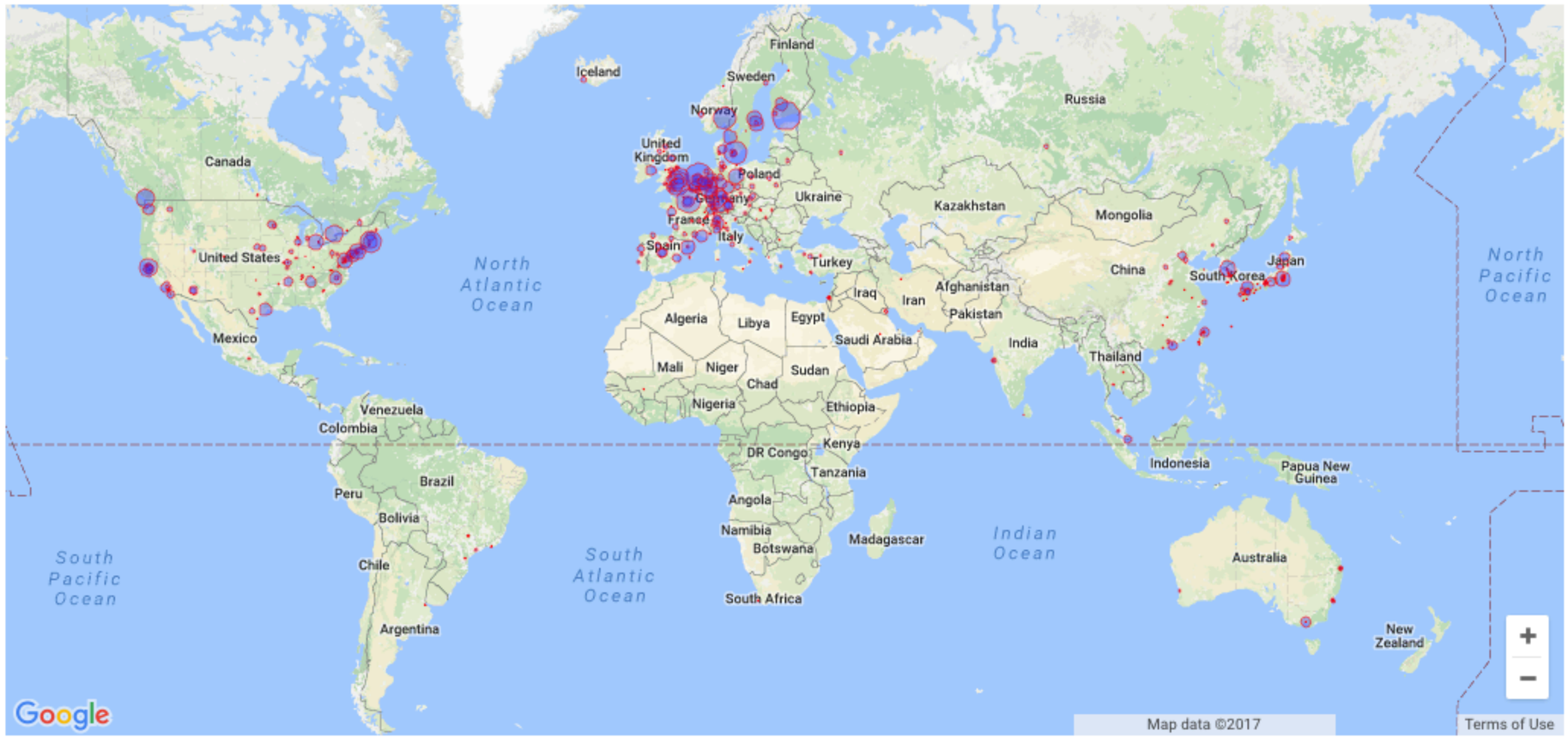
Need to classify population

Sources for ethnicity information

- Geographical location of research

Sources for ethnicity information

- Geographical location of research



The map displays the geographic distribution (by corresponding author) of the 96703 genomic array, 36747 chromosomal CGH and 6274 whole genome/exome based cancer genome datasets. The numbers are derived from the 2993 I 2993 publications registered in the Progenetix database.

Sources for ethnicity information

- Geographical location of research

Sources for ethnicity information

- Geographical location of research
- Self-report metadata

Sources for ethnicity information

- Geographical location of research
- Self-report metadata ⟶ Inaccurate, often incomplete

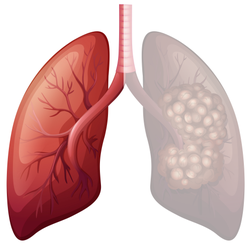
Sources for ethnicity information

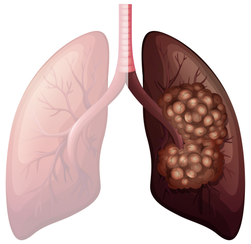
- Geographical location of research
- Self-report metadata ⟶ Inaccurate, often incomplete
- Cancer genome profile (way to go)

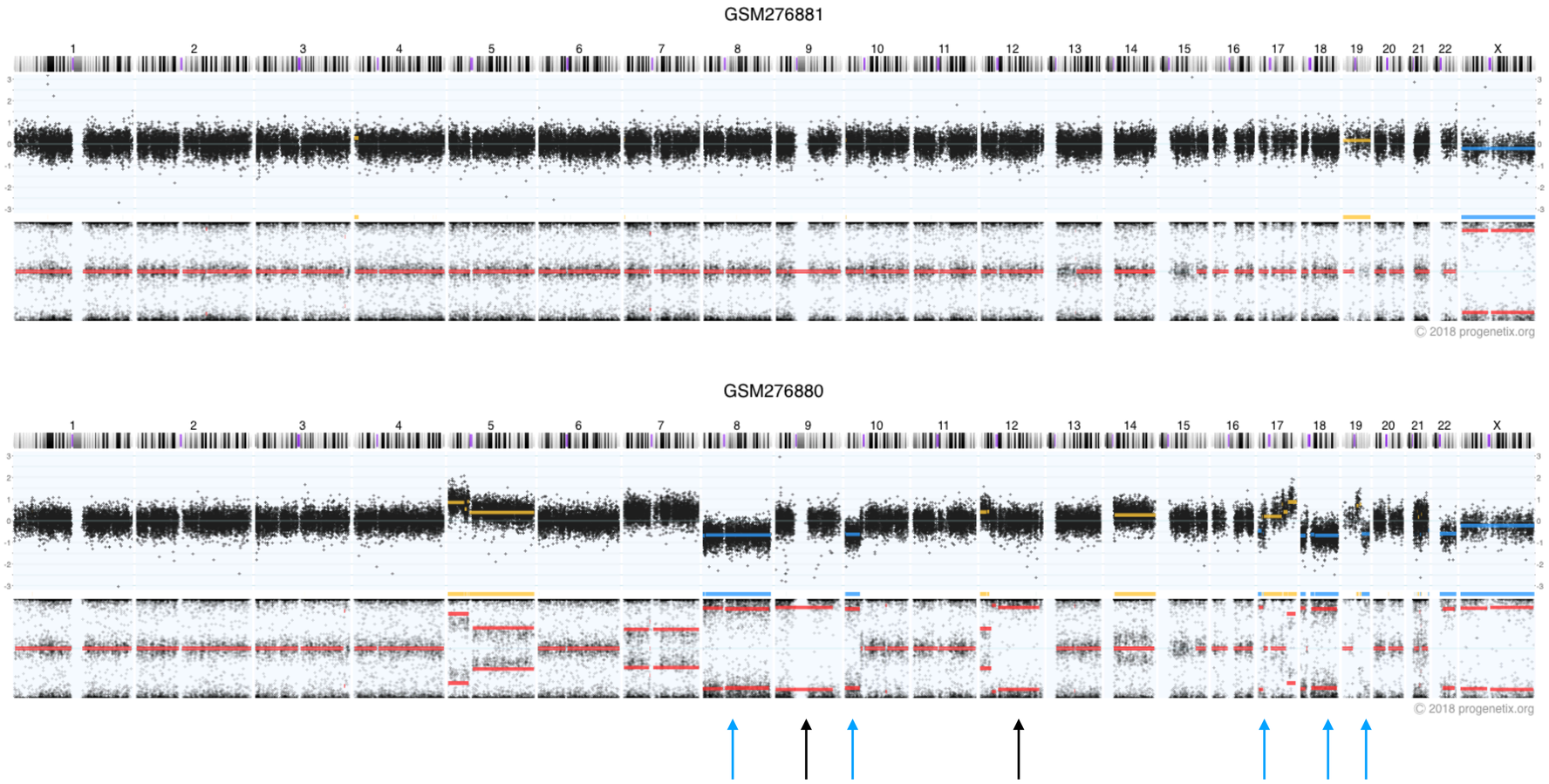
Sources for ethnicity information

- Geographical location of research
- Self-report metadata ⟶ Inaccurate, often incomplete
- Cancer genome profile (way to go)
    - Somatic variation

Noisy cancer genome







Copy DEL

Copy neutral LOH

18.4%

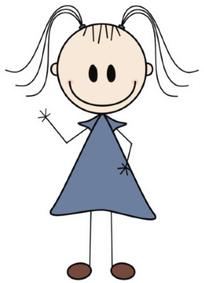
Sources for ethnicity information

- Geographical location of research
- Self-report metadata ⟶ Inaccurate, often incomplete
- Cancer genome profile (way to go)
  - Somatic variation
  - Diverse platforms

Diverse SNP platforms

AACGTCTAGGATCGACTAG

Platform1



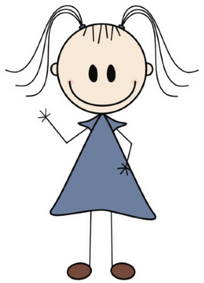
AACGTCTAGGATCGACTAG

Platform2

Diverse SNP platforms

AACGTCTAGGATCGACTAG

Platform1



AACGTCTAGGATCGACTAG

Platform2

Linked SNP markers carry same information.

Overlap between platforms is rare.

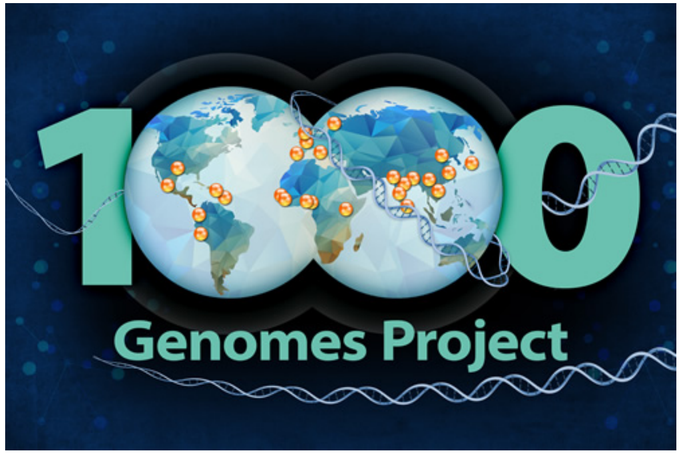
(16-18% shared between 3 main genetic testing companies)

23andMe

ancestry

myHERITAGE

Reference data



- 1000 Genomes Project
- public catalogue of human variation and genotype data
- sequencing data
- **2,504 individuals** from **26 populations (5 continental groups)**

Admixture analysis

Allele Frequency per SNP

A(ncestry)	A1	A2	A3	A4	A5	A6
A	30%	1%	20%	40%	20%	10%
B	70%	99%	80%	60%	80%	90%

...

Ancestry fraction per individual

	A1	A2	A3	A4	A5	A6
Individual1	85%	0	10%	0	0	5%

...

Admixture analysis

Allele Frequency per SNP

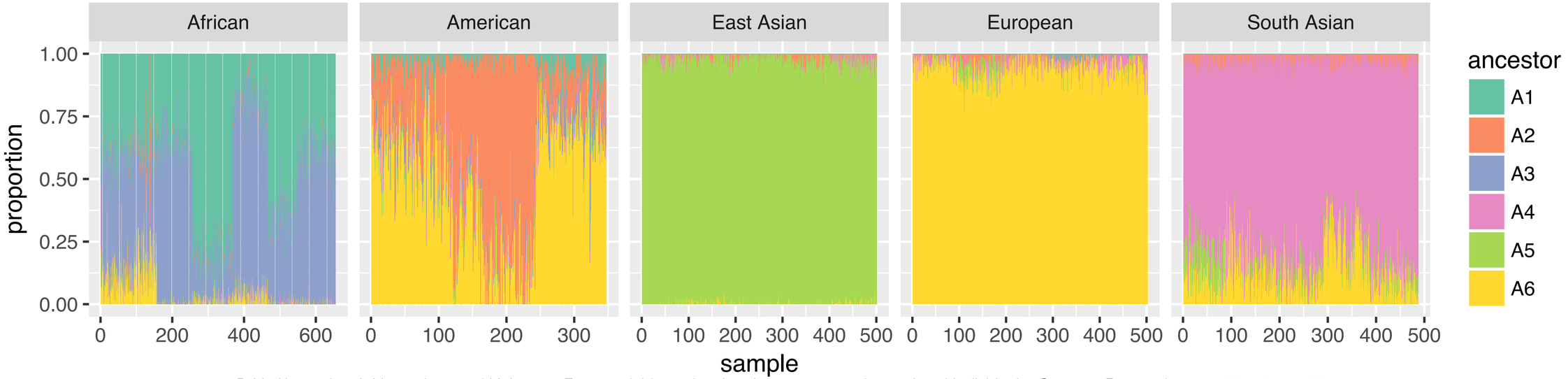
A(ncestry)	A1	A2	A3	A4	A5	A6
A	30%	1%	20%	40%	20%	10%
B	70%	99%	80%	60%	80%	90%

...

Ancestry fraction per individual

	A1	A2	A3	A4	A5	A6
Individual1	85%	0	10%	0	0	5%

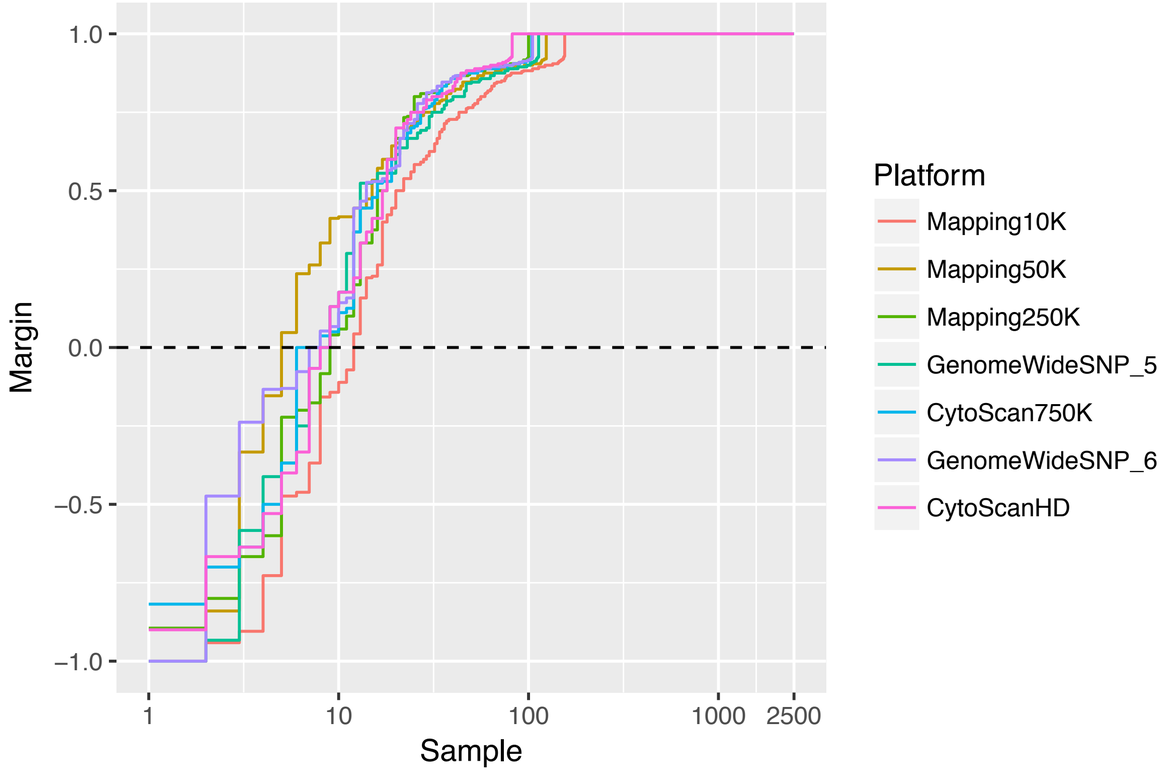
...



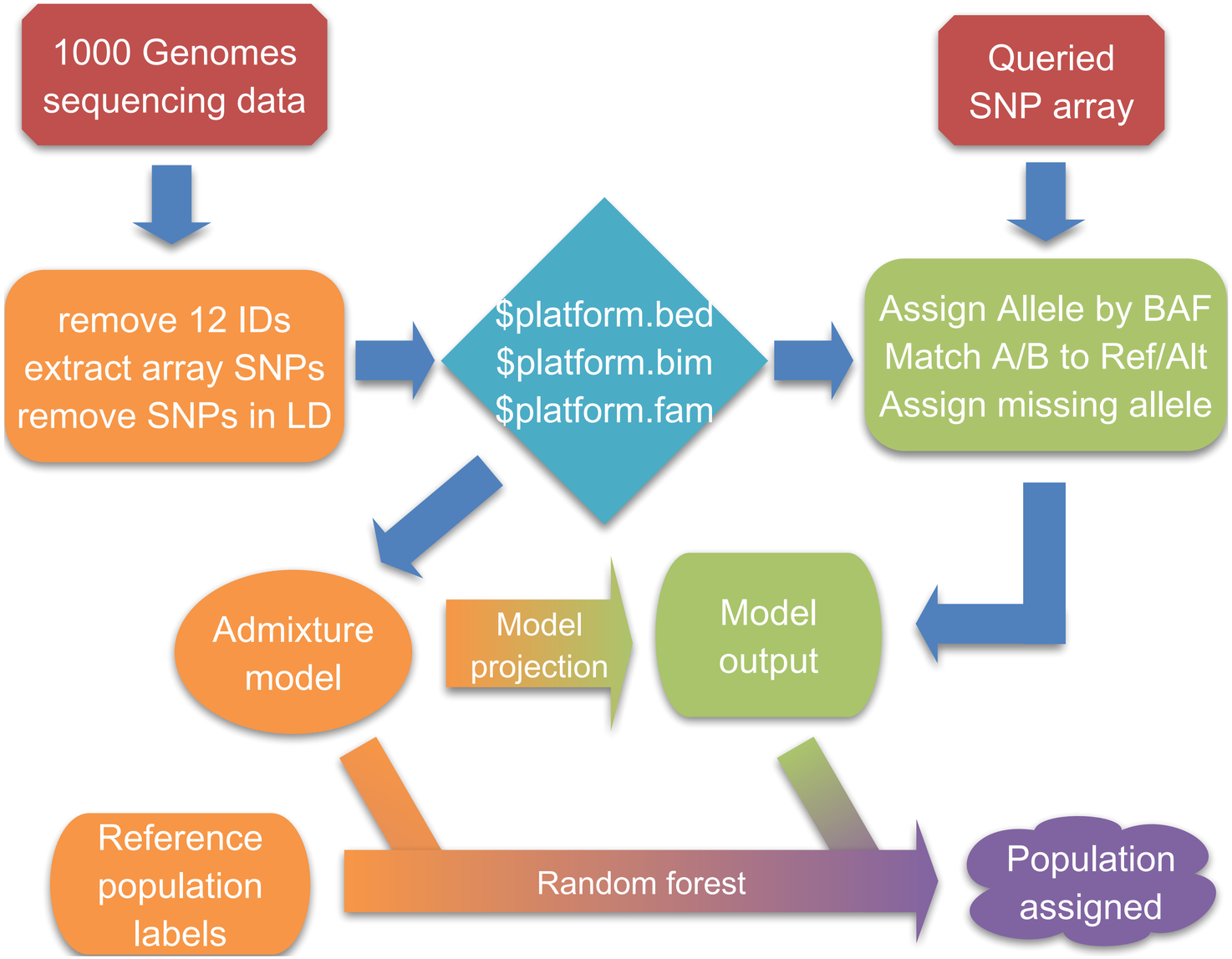
D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

Population classification by platform

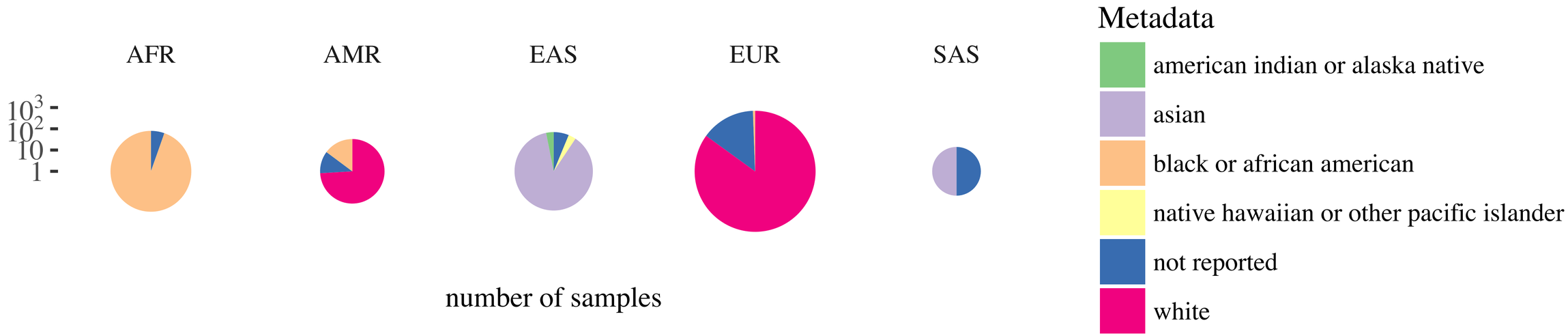
- Target data from 9 genotyping SNP arrays
- extract SNPs from array platforms
- train a population admixture model
- perform random forest classification
- Margin > 0 = correct assignment



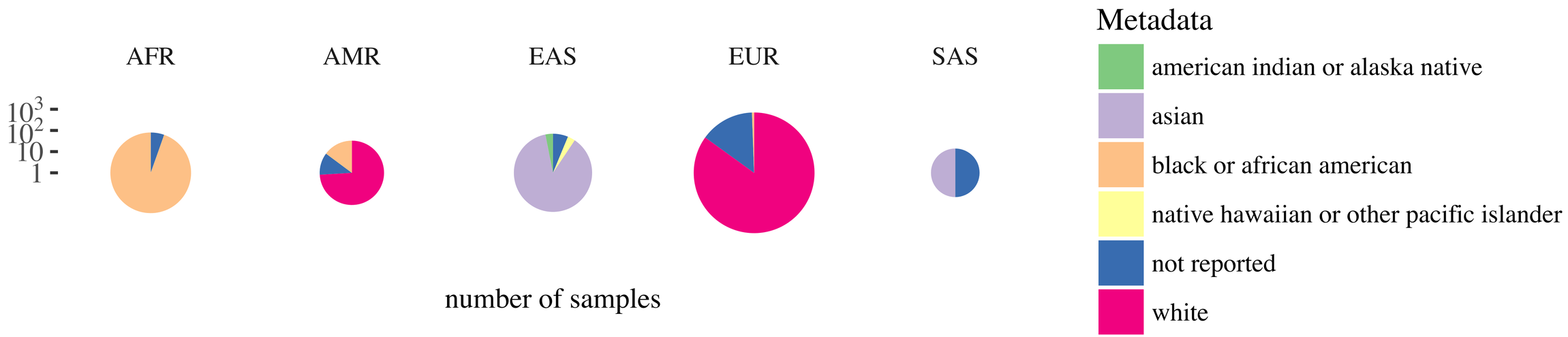
Pipeline


1000 Genomes sequencing data
Queried SNP array
remove 12 IDs
extract array SNPs
remove SNPs in LD
$platform.bed
$platform.bim
$platform.fam
Assign Allele by BAF
Match A/B to Ref/Alt
Assign missing allele
Admixture model
Model projection
Model output
Reference population labels
Random forest
Population assigned

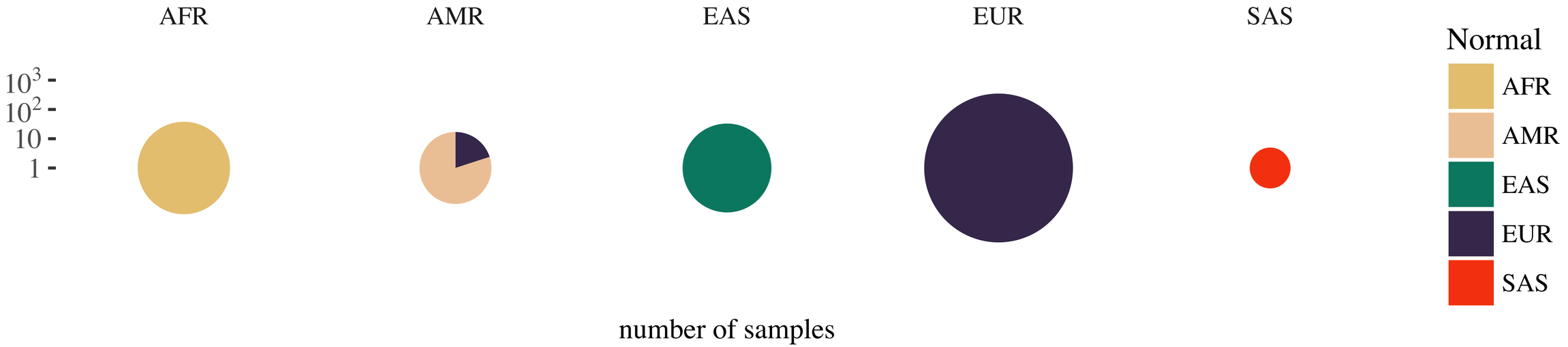
Composition of metadata groups with predicted population groups from TCGA



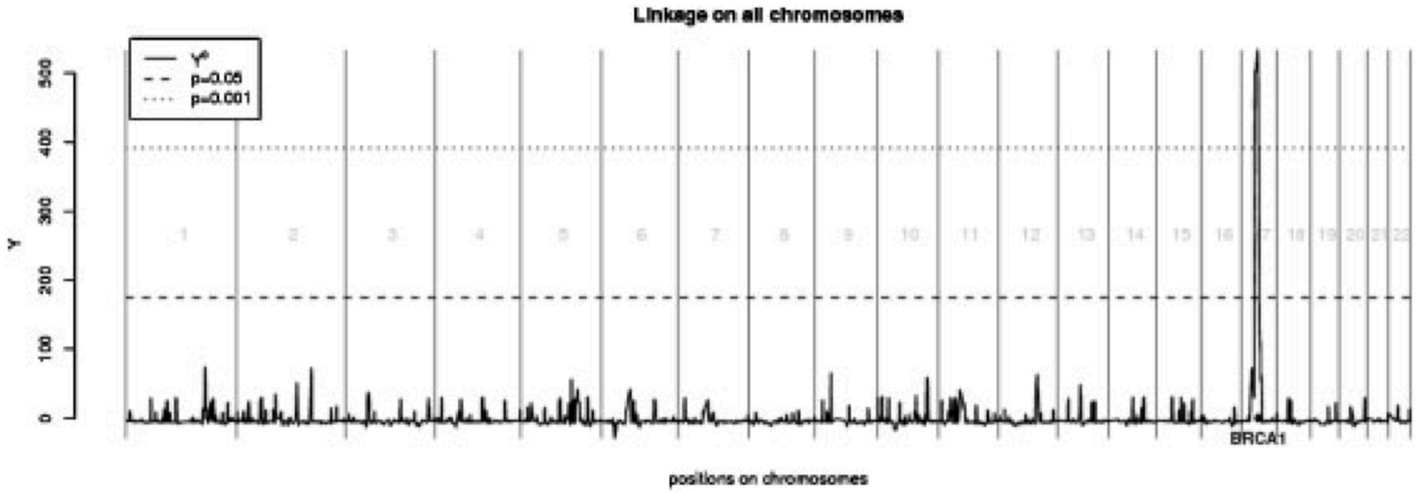
Composition of metadata groups with predicted population groups from TCGA



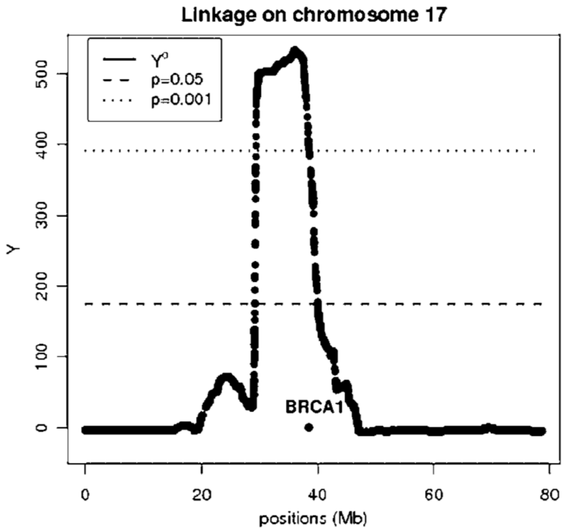
Predicted population group of paired tumor sample from TCGA data



Outlook



Albrechtsen et al. Genetic Epidemiology 2009



- Haplotype phasing and linkage mapping
- CNV pattern search

7 unrelated Danish breast/ovarian against HapMap control

outbred enough, but reasonably similar background

Thank you!

The Baudis group:

Prof. Michael Baudis

Bo Gao

Paula Carrio Cordo

Rahel Paloots

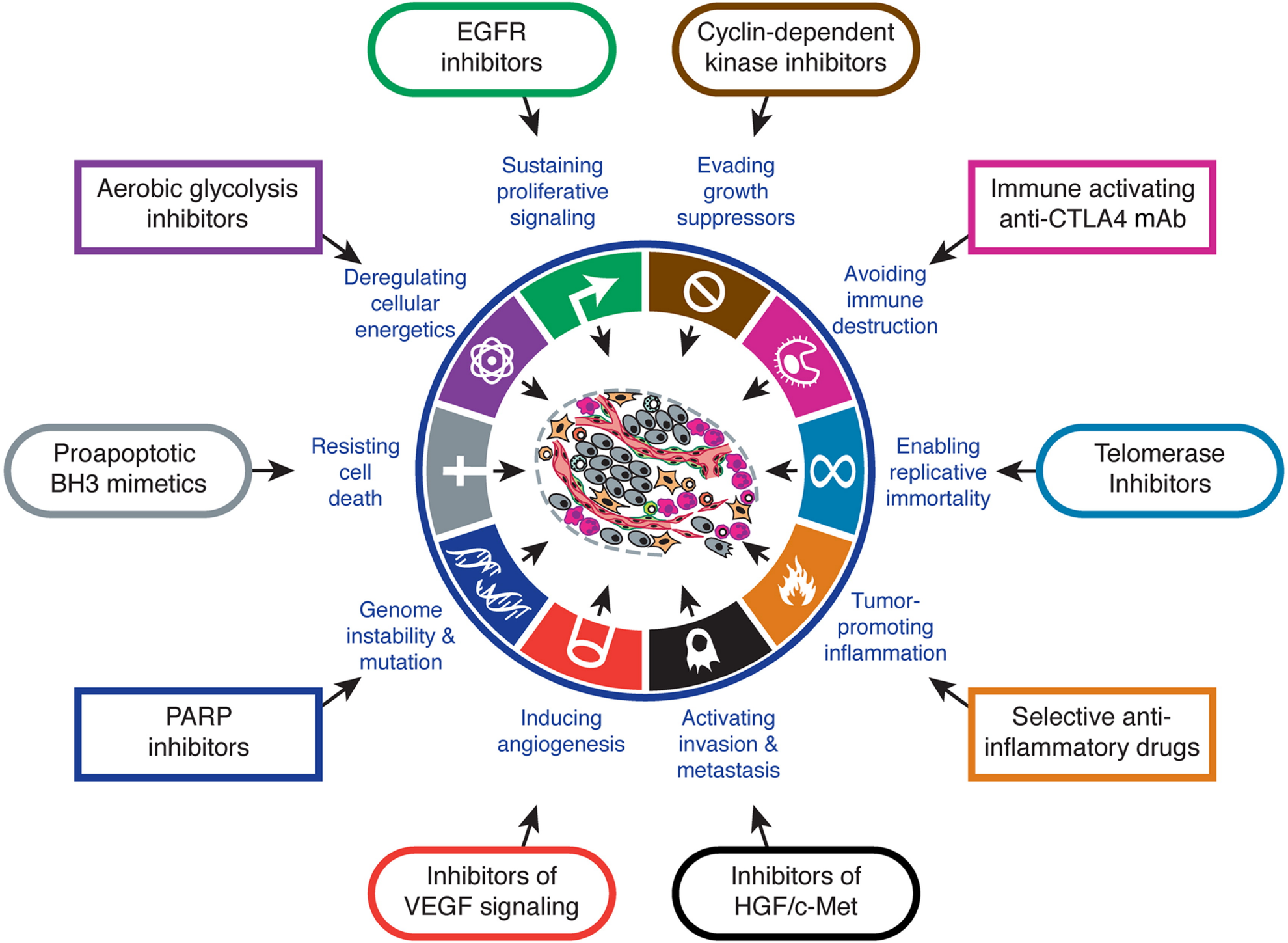
Zurich Bioinformatics Seminar

SIB Swiss Institute of Bioinformatics

Universität Zürich[UZH]

arrayMap

elixir


EGFR inhibitors
Cyclin-dependent kinase inhibitors
Sustaining proliferative signaling
Evading growth suppressors
Aerobic glycolysis inhibitors
Immune activating anti-CTLA4 mAb
Deregulating cellular energetics
Avoiding immune destruction
Proapoptotic BH3 mimetics
Resisting cell death
Enabling replicative immortality
Telomerase Inhibitors
Genome instability & mutation
Tumor-promoting inflammation
PARP inhibitors
Selective anti-inflammatory drugs
Inducing angiogenesis
Activating invasion & metastasis
Inhibitors of VEGF signaling
Inhibitors of HGF/c-Met