

Biomedicine and the Foundations of Data?

Michael W. Mahoney

(RISELab, ICSI, and Department of Statistics, UC Berkeley)

BD2K: May 2018

(For more info, see: <http://www.stat.berkeley.edu/~mmahoney>)

Insider's vs outsider's views (1 of 2)

Ques: Genetics vs molecular biology vs biochemistry vs biophysics:

- What's the difference?

Insider's vs outsider's views (1 of 2)

Ques: Genetics vs molecular biology vs biochemistry vs biophysics:

- What's the difference?

Answer: *Not much*, (if you are a “methods” person*)

- they are all biology
- you get data from any of those areas, ignoring important domain details, and evaluate your method qua method
- your reviewers evaluate the methods and don't care about the science
- ...

*E.g., one who self-identifies as doing data analysis or machine learning or statistics or theory of algorithms or artificial intelligence or ...

Insider's vs outsider's views (2 of 2)

Ques: Data analysis vs machine learning vs statistics vs theory of algorithms vs artificial intelligence (vs scientific computing vs computational mathematics vs databases ...):

- What's the difference?

Insider's vs outsider's views (2 of 2)

Ques: Data analysis vs machine learning vs statistics vs theory of algorithms vs artificial intelligence (vs scientific computing vs computational mathematics vs databases ...):

- What's the difference?

Answer: *Not much*, (if you are a “science” person*)

- they are all just tools
- you get a tool from any of those areas and bury details in a methods section
- your reviewers evaluate the science and don't care about the methods
- ...

*E.g., one who self identifies as doing genetics or molecular biology or biochemistry or biophysics or ...

BIG data??? MASSIVE data????



NYT, Feb 11, 2012: “The Age of Big Data”

- “What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. ...”

Why are big data big?

- Generate data at different places/times and different resolutions
- Factor of 10 more data is not just more data, but different data

BIG data??? MASSIVE data????

MASSIVE *data*:

- Internet, Customer Transactions, Astronomy/HEP = “Petascale”
- One Petabyte = watching 20 years of movies (HD) = listening to 20,000 years of MP3 (128 kbits/sec) = way too much to browse or comprehend

massive data:

- 10^5 people typed at 10^6 DNA SNPs; 10^6 or 10^9 node social network; etc.

In either case, main issues:

- Memory management issues, e.g., push computation to the data
- Hard to answer even basic questions about what data “looks like”

Thinking about large-scale data



Data generation is **modern version of microscope/telescope**:

- **See things couldn't see before**: e.g., fine-scale movement of people, fine-scale clicks and interests; fine-scale tracking of packages; fine-scale measurements of temperature, chemicals, etc.
- Those inventions ushered new scientific eras and **new understanding of the world** and new technologies to do stuff

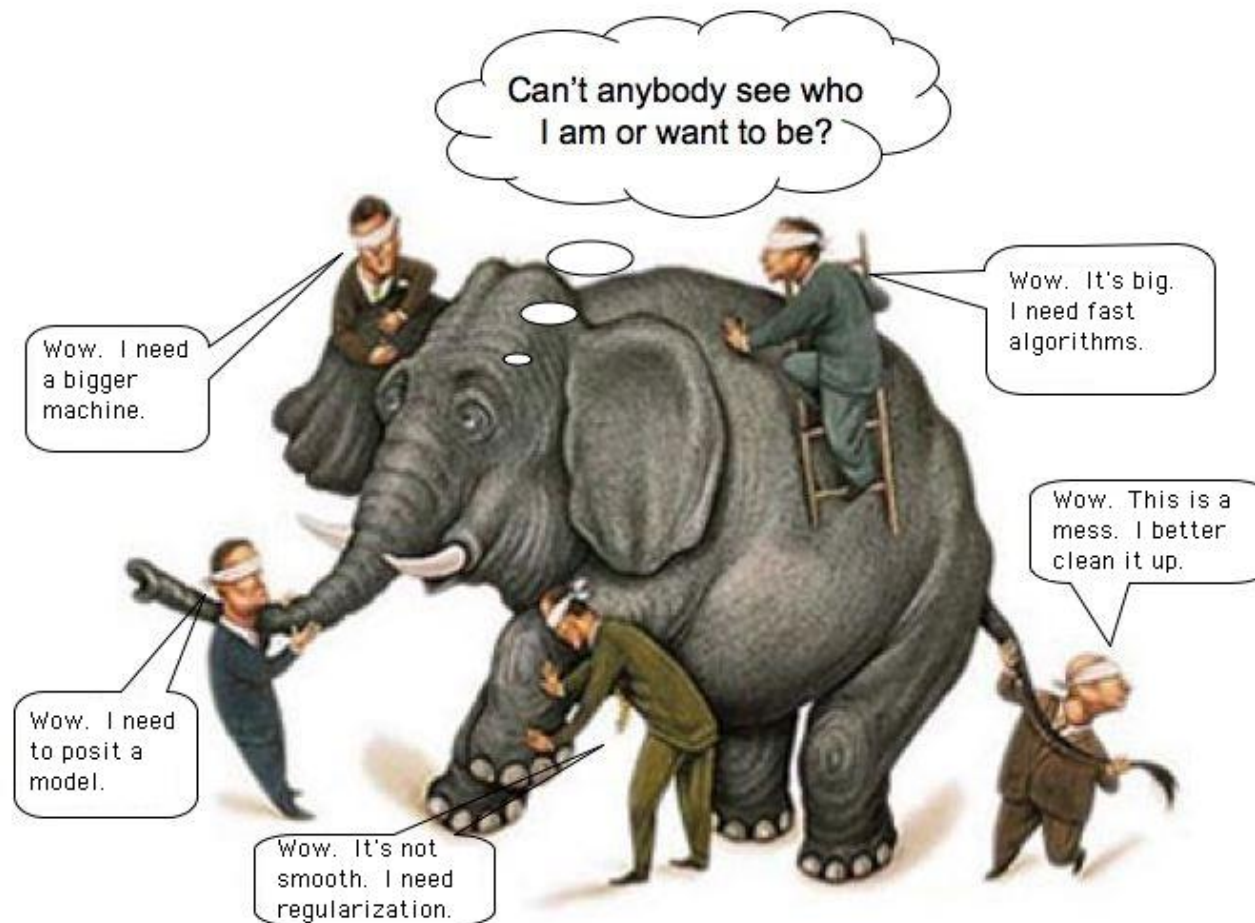
Easy things become hard and **hard things become easy**:

- Easier to see the other side of universe than bottom of ocean
- Means, sums, medians, correlations is easy with small data

Our ability to generate data far exceeds our ability to extract insight from data.



How do we view BIG data?



Algorithmic vs. Statistical Perspectives ...

Lambert (2000), Mahoney (2010)

Computer Scientists

- *Data*: are a **record of everything** that happened.
- *Goal*: process the data to **find interesting patterns** and associations.
- *Methodology*: Develop approximation algorithms under different models of data access since the goal is typically **computationally hard**.

Statisticians (and Natural Scientists)

- *Data*: are a **particular random instantiation** of an underlying process describing unobserved patterns in the world.
- *Goal*: is to **extract information** about the world from noisy data.
- *Methodology*: Make inferences (perhaps about unseen events) by **positing a model** that describes the random variability of the data around the deterministic model.

... are VERY different paradigms

Statistics, natural sciences, scientific computing, etc:

- Problems often involve computation, but the study of *computation per se is secondary*
- Only makes sense to develop algorithms for *well-posed* problems*
- First, write down a model, and think about computation later

Computer science:

- Easier to study *computation per se in discrete settings*, e.g., Turing machines, logic, complexity classes
- Theory of algorithms *divorces computation from data*
- First, run a fast algorithm, and ask what it means later

*Solution exists, is unique, and varies continuously with input data

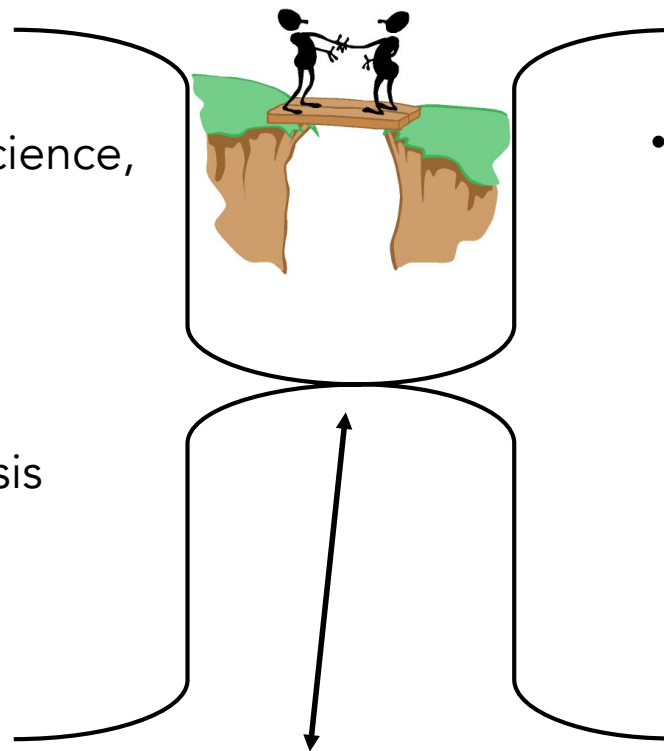
Anecdote 1: Randomized Matrix Algorithms

Mahoney “Algorithmic and Statistical Perspectives on Large-Scale Data Analysis” (2010)

Mahoney “Randomized Algorithms for Matrices and Data” (2011)

Theoretical origins

- theoretical computer science, convex analysis, etc.
- Johnson-Lindenstrauss
- Additive-error algs
- Good worst-case analysis
- No statistical analysis
- No implementations



How to “bridge the gap”?

- decouple (implicitly or explicitly) randomization from linear algebra
- importance of statistical leverage scores!

Practical applications

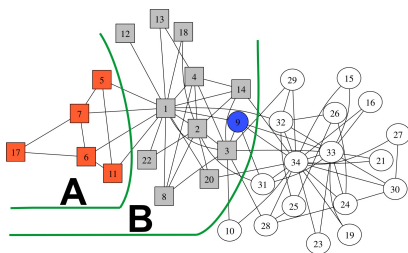
- NLA, ML, statistics, data analysis, **genetics**, etc
- Fast JL transform
- Relative-error algs
- Numerically-stable algs
- Good statistical properties
- *Beats LAPACK & parallel-distributed implementations on terabytes of data*

Anecdote 2:

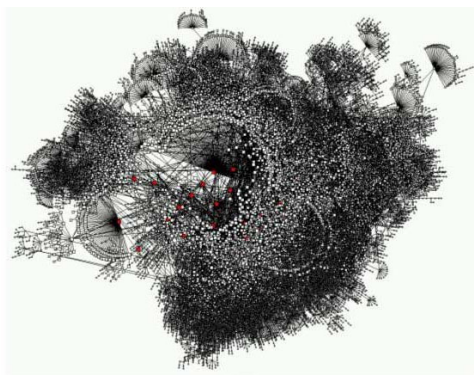
Communities in large informatics graphs

Mahoney “Algorithmic and Statistical Perspectives on Large-Scale Data Analysis” (2010)
Leskovec, Lang, Dasgupta, & Mahoney “Community Structure in Large Networks ...” (2009)

People imagine social networks to look like:



Real social networks actually look like:

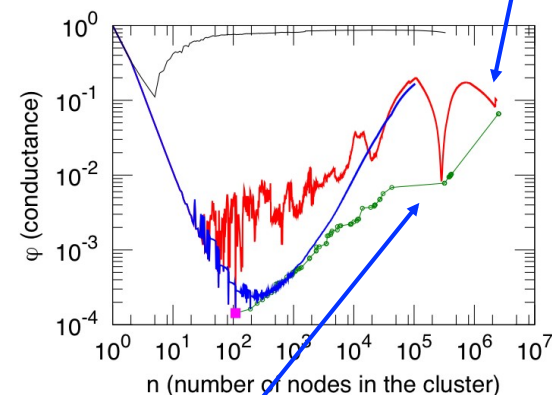


How do we know this plot is “correct”?

- (since computing conductance is intractable)
- Lower Bound Result; Structural Result; Modeling Result; Etc.
- Algorithmic Result (ensemble of sets returned by different approximation algorithms are very different)
- *Statistical Result* (Spectral provides more meaningful communities than flow)

Data are expander-like at large size scales !!!

Size-resolved conductance (degree-weighted expansion) plot looks like:



There do not exist good large clusters in these graphs !!!

Anecdote 3:

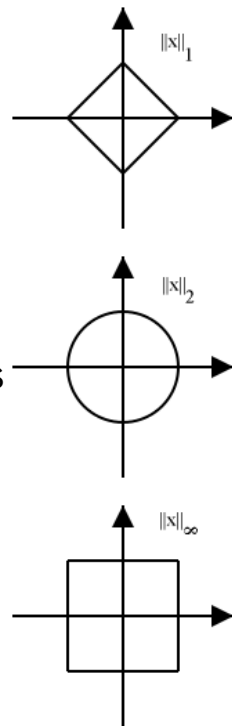
Approx. comp. and implicit regularization

Mahoney “Approximate Computation and Implicit Regularization for Very Large-scale Data Analysis” (2012)

Explicitly-imposed regularization

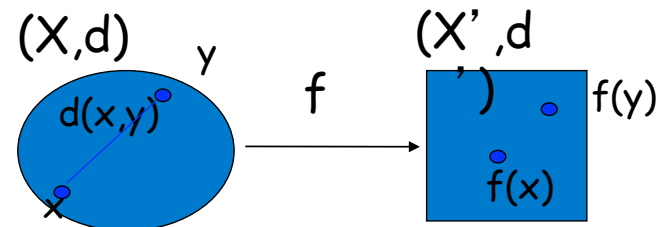
- Traditionally, regularization uses *explicit* norm constraint to make sure solution vector is “small” and not-too-complex

- $\min \|f\| + \lambda \|g(x)\|$



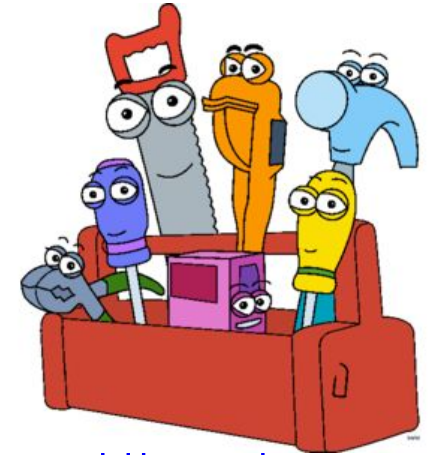
Implicitly-imposed regularization

- Binning, pruning, early stopping, etc.
- Design decisions engineers make
- Approximation algorithms *implicitly* embed data in a “nice” metric/geometric place and then round the solution.



Big question: Can we formalize the notion that/when approximate computation in and of itself can *implicitly* lead to “better” or “more regular” solutions than exact computation? (Short answer: yes!)

Lessons from the anecdotes



We are being forced to **engineer a union between two very different worldviews** on what are fruitful ways to view the data

- in spite of our best efforts *not* to

The forcing function (generation of lots of valuable data) is forcing us to **revisit old methods in a new light**

- often reinventing, but the forcing function makes that acceptable

Given existing forcing functions and disciplinary lines, **many methods and approaches are “undervalued”** for what non-foundational people want

- and it would be good not to loose them

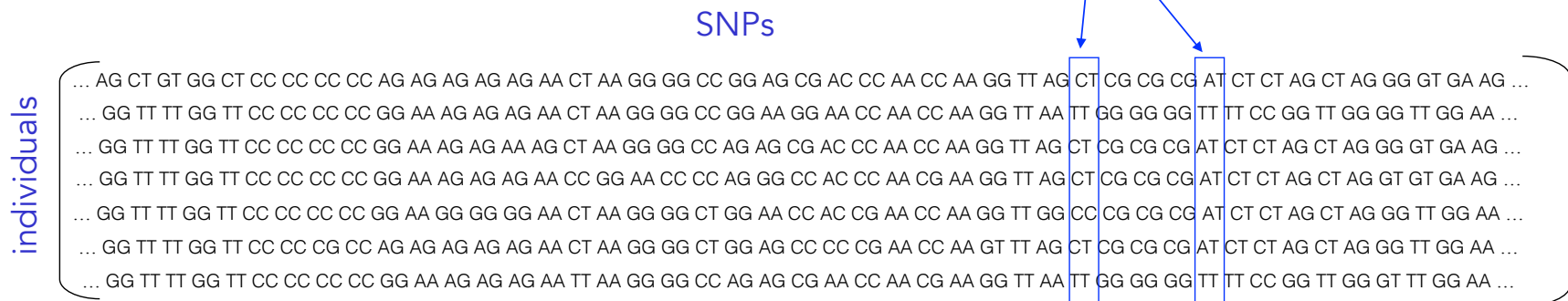
QUESTION: How can we bridge the gap between these two worldviews?

QUESTION: What, if anything, does biomedicine have to offer?

Application in: Human Genetics

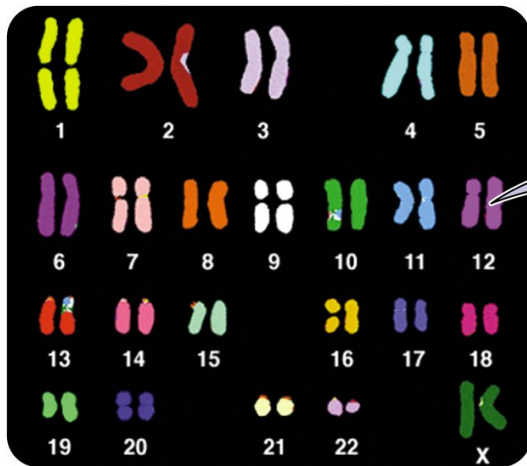
Single Nucleotide Polymorphisms: the most common type of genetic variation in the genome across different individuals.

They are **known** locations at the human genome where **two** alternate nucleotide bases (**alleles**) are observed (out of A, C, G, T).

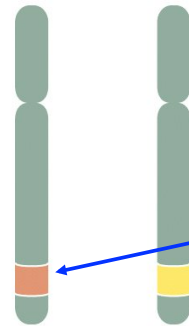


Matrices including thousands of individuals and hundreds of thousands or millions (large for some people, small for other people) if SNPs are available.

This can be written as a “matrix,” assume it’s been preprocessed properly, so let’s call black box matrix algorithms.

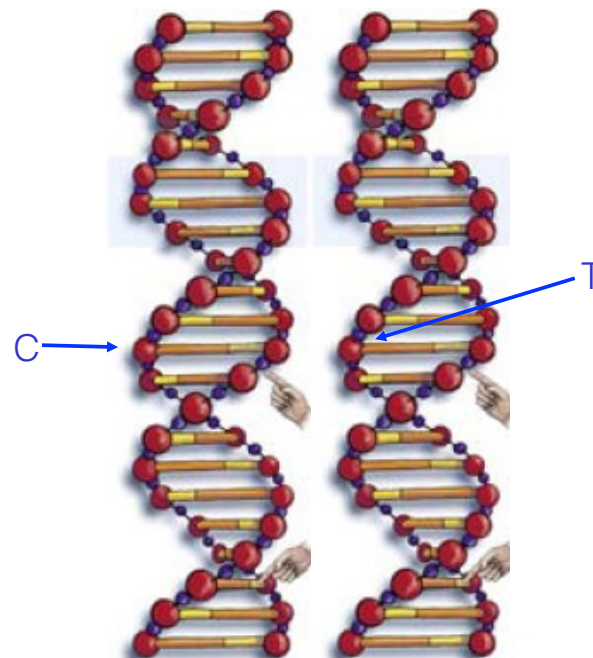


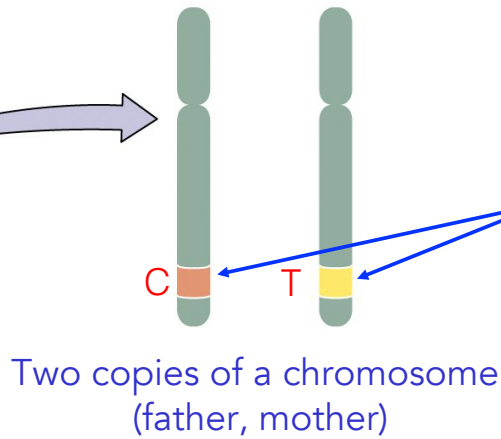
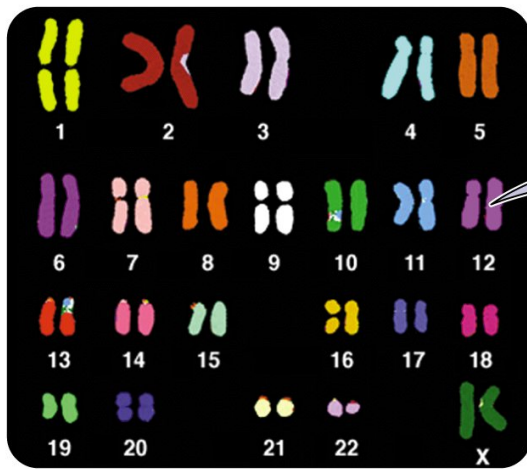
Two copies of a chromosome
(father, mother)



Focus at a specific locus and assay the
observed nucleotide bases (alleles).

SNP: exactly **two alternate** alleles
appear.





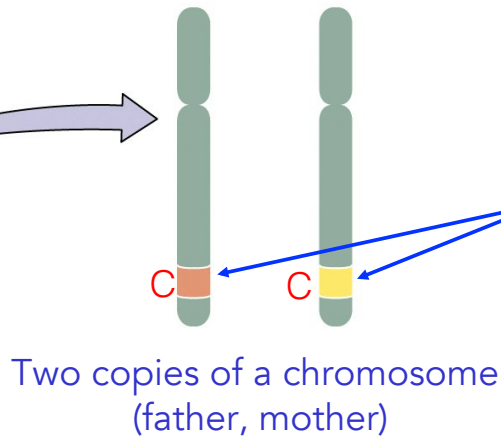
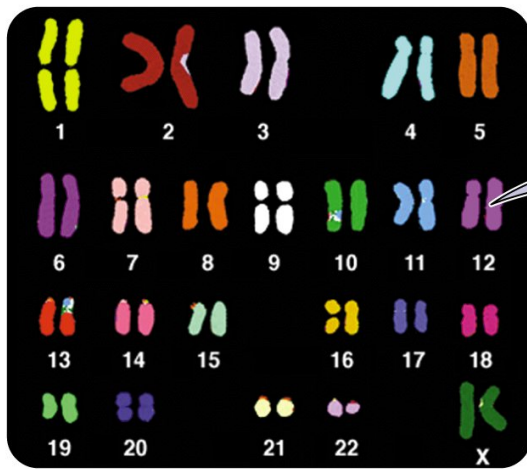
Focus at a specific locus and
assay the observed alleles.
SNP: exactly **two** alternate
alleles appear.

An individual could be:
- Heterozygotic (in our study, CT = TC)

SNPs

individuals

... AG CT GT GG CT CC CC CC CC AG AG AG AG AG AA CT AA GG GG CC GG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CT AA GG GG CC GG AA GG AA CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AA AG CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CC GG AA CC CC AG GG CC AC CC AA CG AA GG TT AG CT CG CG CG AT CT CT AG CT AG GT GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA GG GG GG AA CT AA GG GG CT GG AA CC AC CG AA CC AA GG TT GG CC CG CG CG AT CT CT AG CT AG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CG CC AG AG AG AG AG AA CT AA GG GG CT GG AG CC CC CG AA CC AA GT TT AG CT CG CG CG AT CT CT AG CT AG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA TT AA GG GG CC AG AG CG AA CC AA CG AA GG TT AA TT GG GG GG TT TT CC GG TT GG GT TT GG AA ...



Focus at a specific locus and
assay the observed alleles.
SNP: exactly **two** alternate
alleles appear.

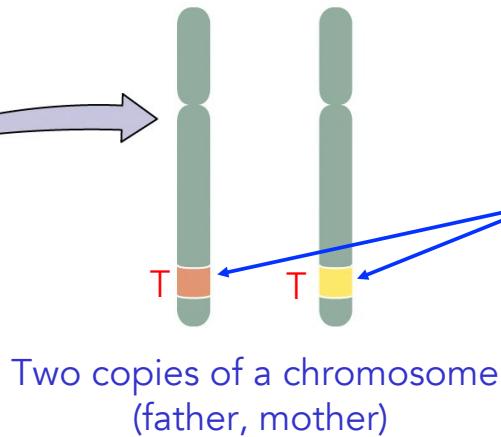
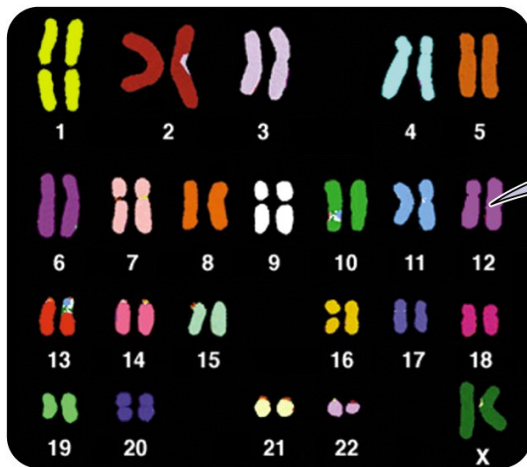
An individual could be:

- Heterozygotic (in our studies, CT = TC)
- Homozygotic at the first allele, e.g., C

SNPs

individuals

... AG CT GT GG CT CC CC CC CC AG AG AG AG AA CT AA GG GG CC GG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CT AA GG GG CC GG AA GG AA CC AA CC AA GG TT AA TT GG GG GG TT TT CC GG TT GG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AA AG CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AA CC GG AA CC CC AG GG CC AC CC AA CG AA GG TT AG CT CG CG CG AT CT CT AG CT AG GT GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA GG GG GG AA CT AA GG GG CT GG AA CC AC CG AA CC AA GG TT GG **CC** CG CG CG AT CT CT AG CT AG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CG CC AG AG AG AG AA CT AA GG GG CT GG AG CC CC CG AA CC AA GT TT AG CT CG CG CG AT CT CT AG CT AG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA TT AA GG GG CC AG AG CG AA CC AA CG AA GG TT AA TT GG GG GG TT TT CC GG TT GG GT TT GG AA ...



Focus at a specific locus and
assay the observed alleles.
SNP: exactly **two** alternate
alleles appear.

An individual could be:

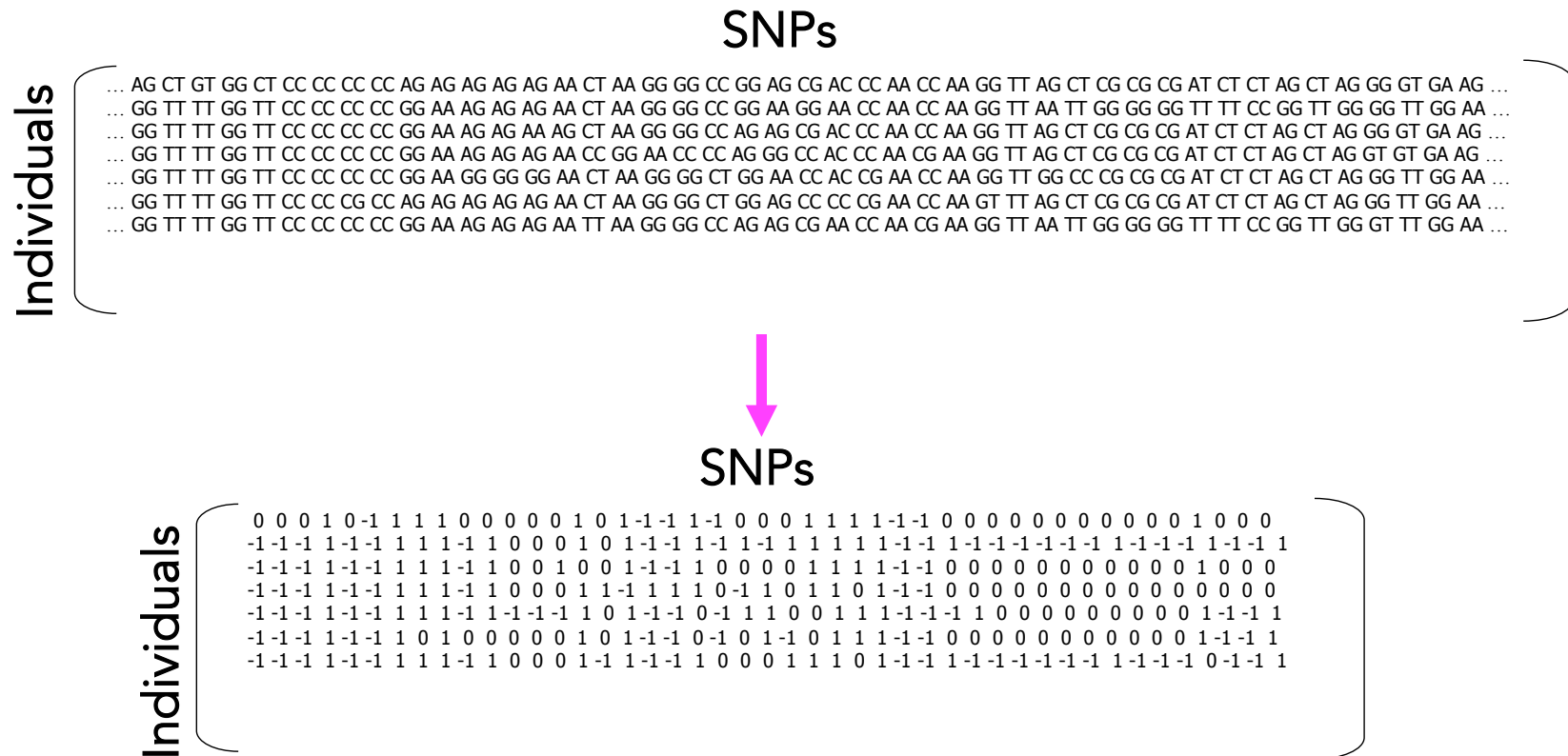
- Heterozygotic (in our studies, CT = TC) → Encode as 0
- Homozygotic at the first allele, e.g., C → Encode as +1
- Homozygotic at the second allele, e.g., T → Encode as -1

SNPs

individuals

... AG CT GT GG CT CC CC CC CC AG AG AG AG AG AA CT AA GG GG CC GG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CT AA GG GG CC GG AA GG AA CC AA CC AA GG TT AA TT GG GG GG TT TT CC GG TT GG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AA AG CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CC GG AA CC CC AG GG CC AC CC AA CG AA GG TT AG CT CG CG CG AT CT CT AG CT AG GT GT GA AG ...
 ... GG TT TT GG TT CC CC CC CC GG AA GG GG GG AA CT AA GG GG CT GG AA CC AC CG AA CC AA GG TT GG CC CG CG CG AT CT CT AG CT AG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CG CC AG AG AG AG AG AA CT AA GG GG CT GG AG CC CC CG AA CC AA GT TT AG CT CG CG CG AT CT CT AG CT AG GG TT GG AA ...
 ... GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA TT AA GG GG CC AG AG CG AA CC AA CG AA GG TT AA **TT** CG GG GG TT TT CC GG TT GG GT TT GG AA ...

Our SNP data as a matrix



We are quite similar, but we are different ...

The average genome (~2x3 billion base pairs) contains:

- 3-4 million single nucleotide variations, compared to the reference sequence (Single Nucleotide Polymorphisms – SNPs)
- ~0.4 million small insertions or deletions 'indels' (1-100bp)
- ~5,000 larger insertions or deletions (>100bp)

Variation across all (~23,000) genes - the 'exome'

- ~18,000 variant
- ~8-9,000 functional variant
- ~95% of variants are common
- ~500-1000 genes with new mutation
- ~100-200 knock-out mutations

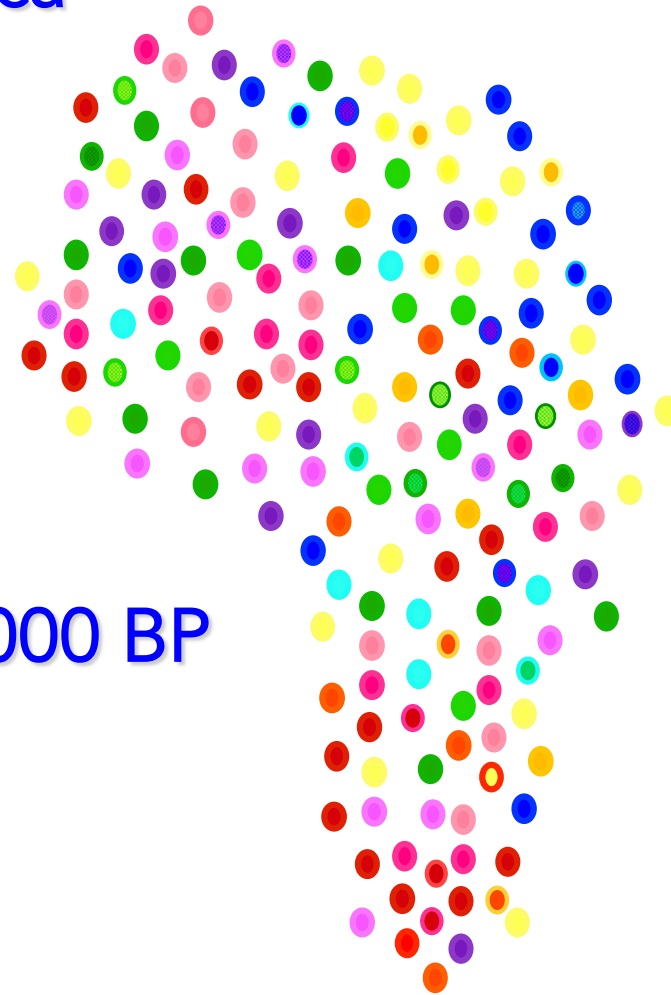
Genetic variation shaped by evolutionary forces

- Mutation
- Genetic drift
- Population structure (inbreeding, mating patterns, etc.)
- Gene flow and admixture
- Natural selection



Great application domain to stress test novel methods ...

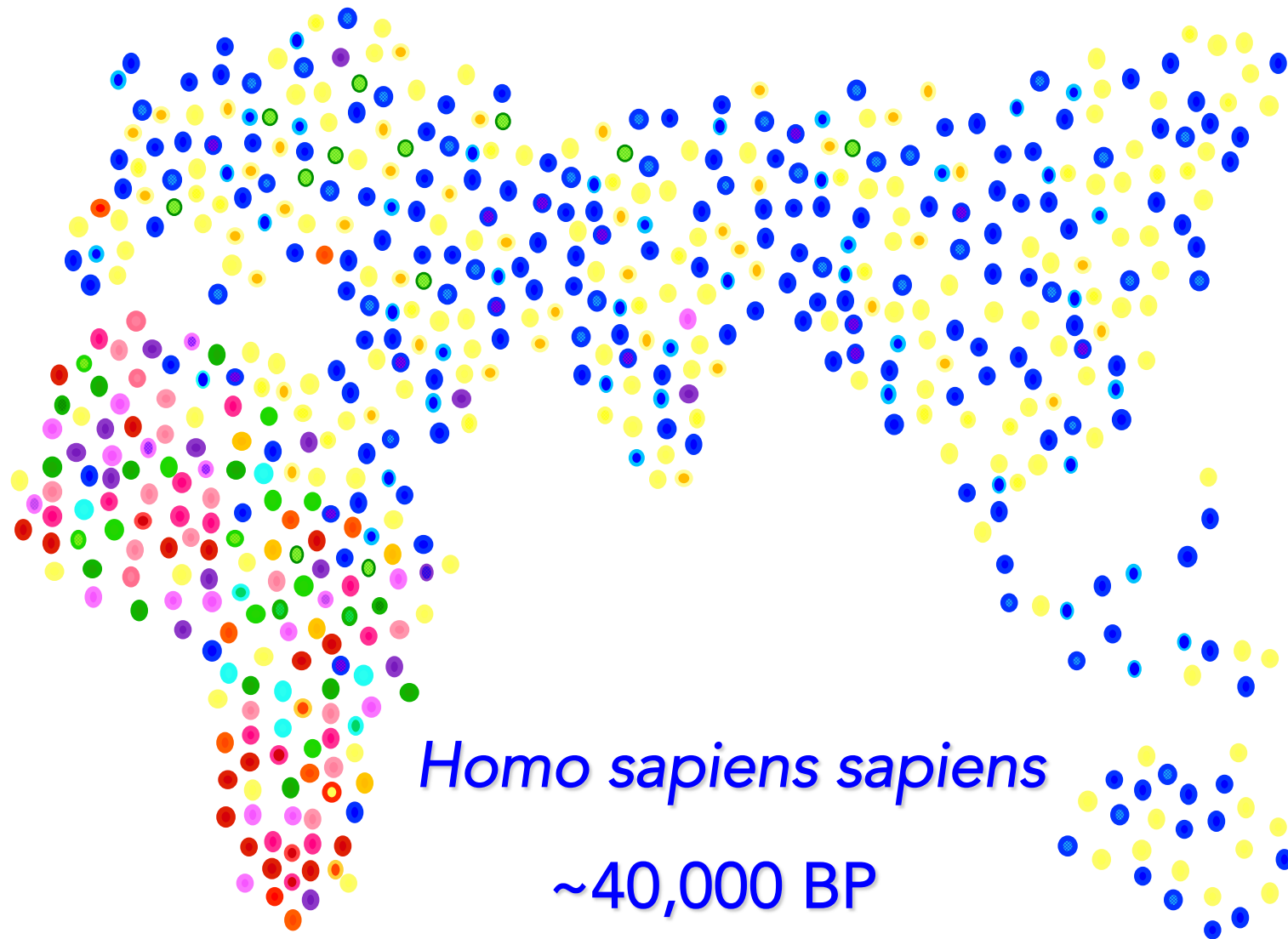
Early *Homo sapiens sapiens* in Africa



150,000 to 100,000 BP

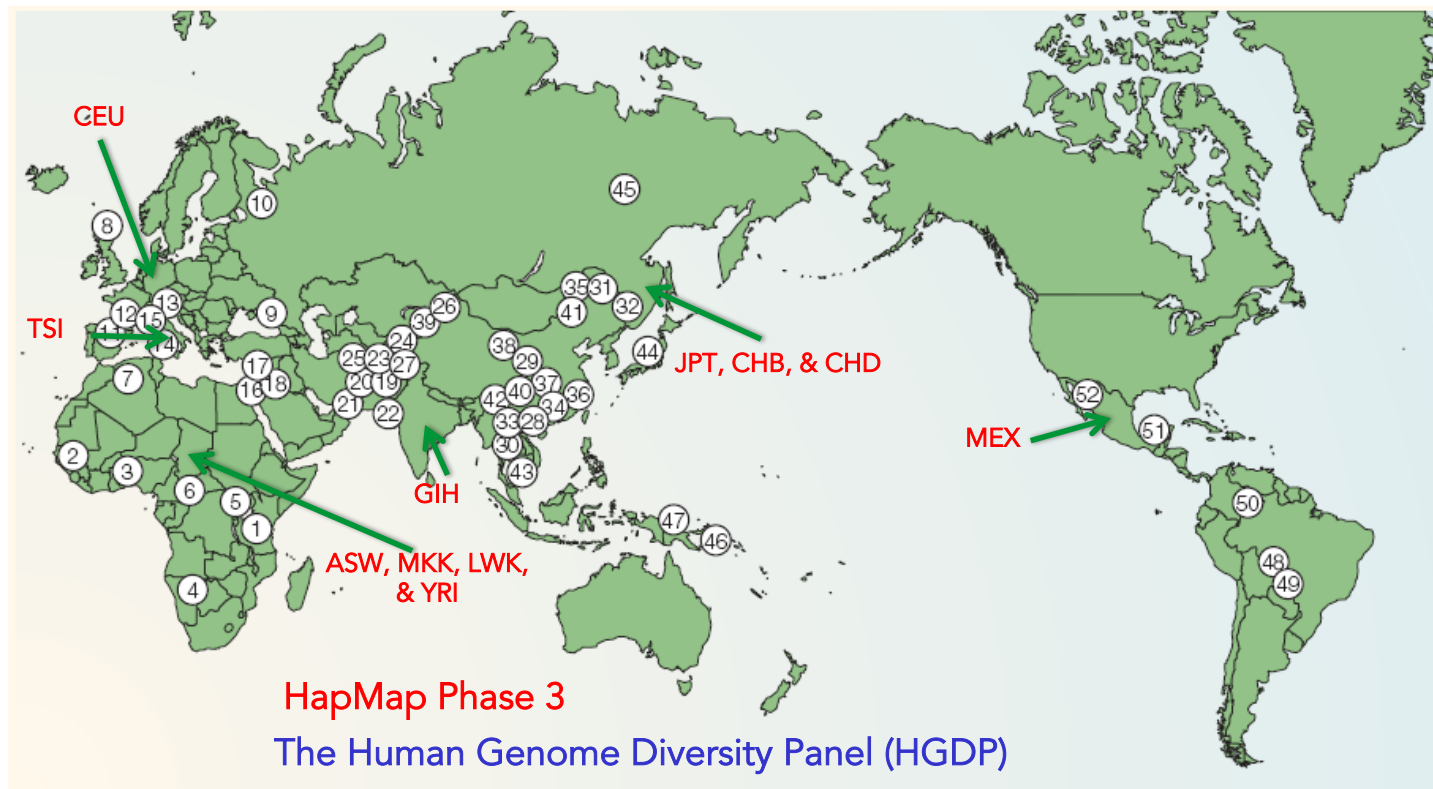


Homo sapiens sapiens
Colonizing south west Asia
~100,000 BP



Homo sapiens sapiens

~40,000 BP



Africans

1 Bantu
2 Mandenka
3 Yoruba
4 San
5 Mbuti pygmy
6 Biaka
7 Mozabite

Europeans

8 Orcadian
9 Adygei
10 Russian
11 Basque
12 French
13 North Italian
14 Sardinian
15 Tuscan

Western Asians

16 Bedouin
17 Druze
18 Palestinian

Central and Southern Asians

19 Balochi
20 Brahui
21 Makrani
22 Sindhi
23 Pathan
24 Burusho
25 Hazara
26 Uygur
27 Kalash

Eastern Asians

28 Han (S. China)
29 Han (N. China)
30 Dai
31 Daur
32 Hezhen
33 Lahu
34 Miao
35 Oroqen
36 She
37 Tujia
38 Tu
39 Xibo
40 Yi
41 Mongola
42 Naxi
43 Cambodian
44 Japanese
45 Yakut

Oceanians

46 Melanesian
47 Papuan

Native Americans

48 Karitiana
49 Surui
50 Colombian
51 Maya
52 Pima

Cavalli-Sforza (2005) *Nat Genet Rev*

Rosenberg et al. (2002) *Science*

Li et al. (2008) *Science*

The International HapMap Consortium
(2003, 2005, 2007) *Nature*

HGDP data

- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data

- 1,207 samples
- 11 populations

Apply SVD/PCA on the
(joint) HGDP and HapMap
Phase 3 data.

Matrix dimensions:

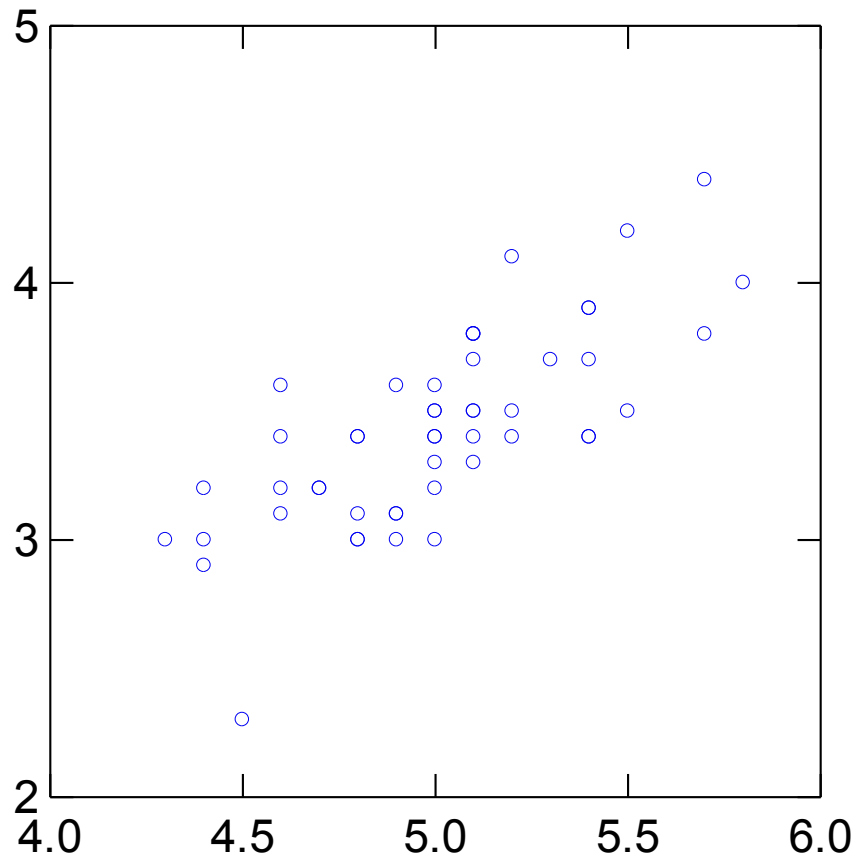
2,240 subjects (rows)

447,143 SNPs (columns)

Dense matrix:

over one billion entries

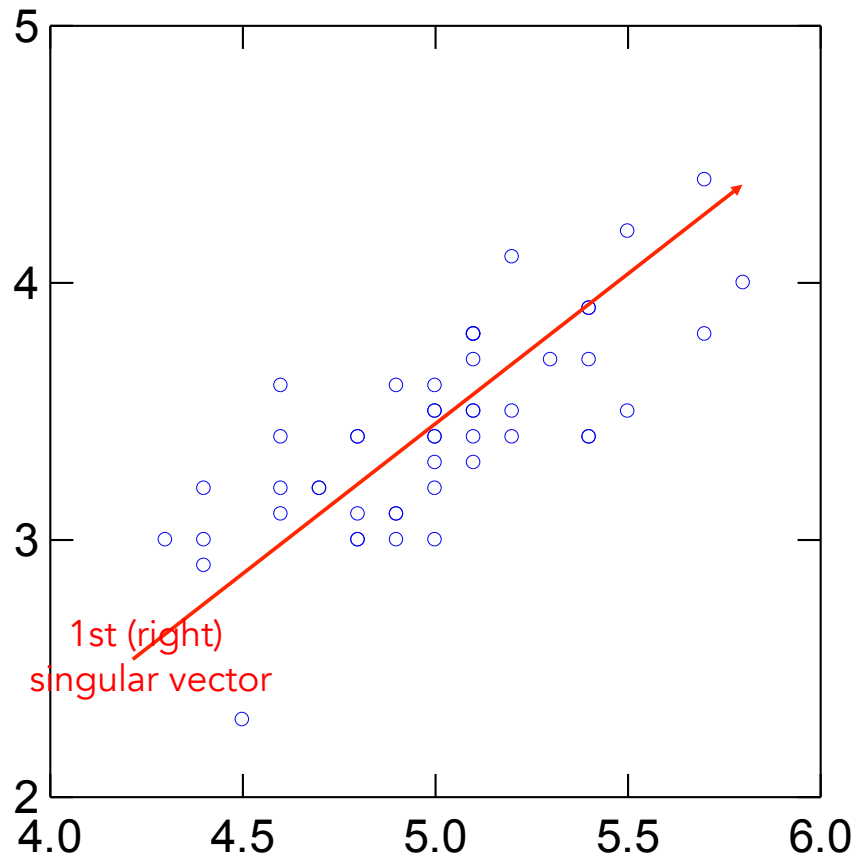
The Singular Value Decomposition (SVD)



Let the blue circles represent m data points in a 2-D Euclidean space.

Then, the SVD of the m -by-2 matrix of the data will return ...

The Singular Value Decomposition (SVD)



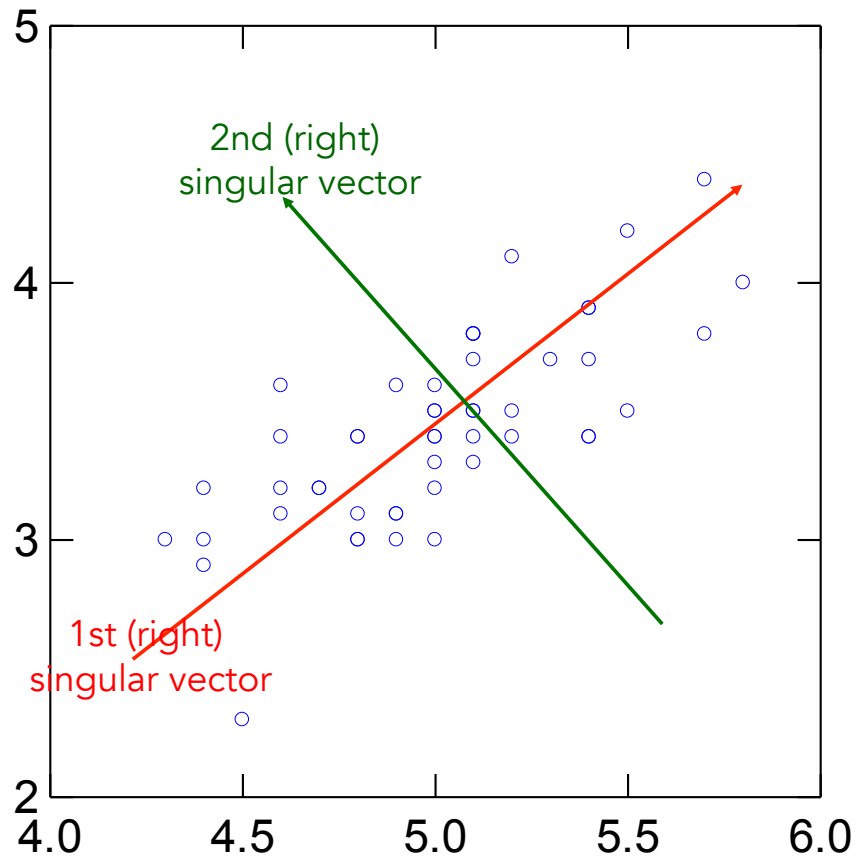
Let the blue circles represent m data points in a 2-D Euclidean space.

Then, the SVD of the m -by-2 matrix of the data will return ...

1st (right) singular vector:

direction of maximal variance,

The Singular Value Decomposition (SVD)



Let the blue circles represent m data points in a 2-D Euclidean space.

Then, the SVD of the m -by-2 matrix of the data will return ...

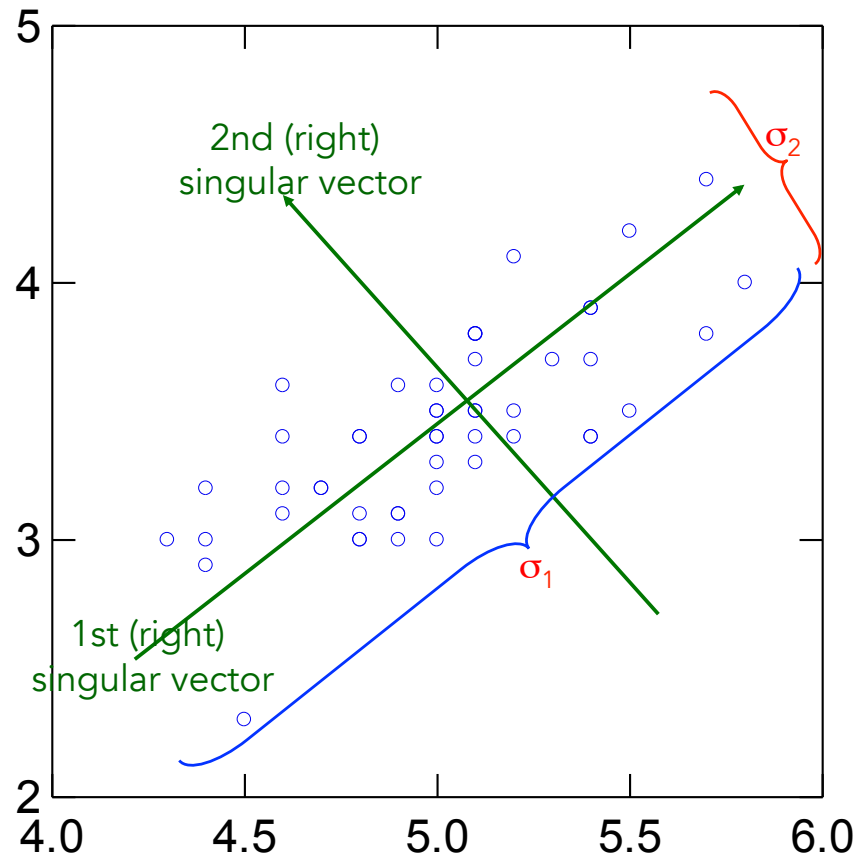
1st (right) singular vector:

direction of maximal variance,

2nd (right) singular vector:

direction of maximal variance, after removing the projection of the data along the first singular vector.

Singular values



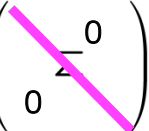
σ_1 : measures how much of the data variance is explained by the first singular vector.

σ_2 : measures how much of the data variance is explained by the second singular vector.

Principal Components Analysis (PCA) is done via the computation of the Singular Value Decomposition (SVD) of a (mean-centered) covariance matrix.

Typically, a small constant number (say k) of the top singular vectors and values are kept.

SVD: formal definition

$$\begin{pmatrix} A \\ m \times n \end{pmatrix} = \begin{pmatrix} U \\ m \times \rho \end{pmatrix} \cdot \begin{pmatrix} \text{ } & 0 \\ 0 & \Sigma \end{pmatrix} \cdot \begin{pmatrix} V \\ \rho \times n \end{pmatrix}^T$$


ρ : rank of A

U (V): orthogonal matrix containing the left (right) singular vectors of A .

Σ : diagonal matrix containing the singular values of A .

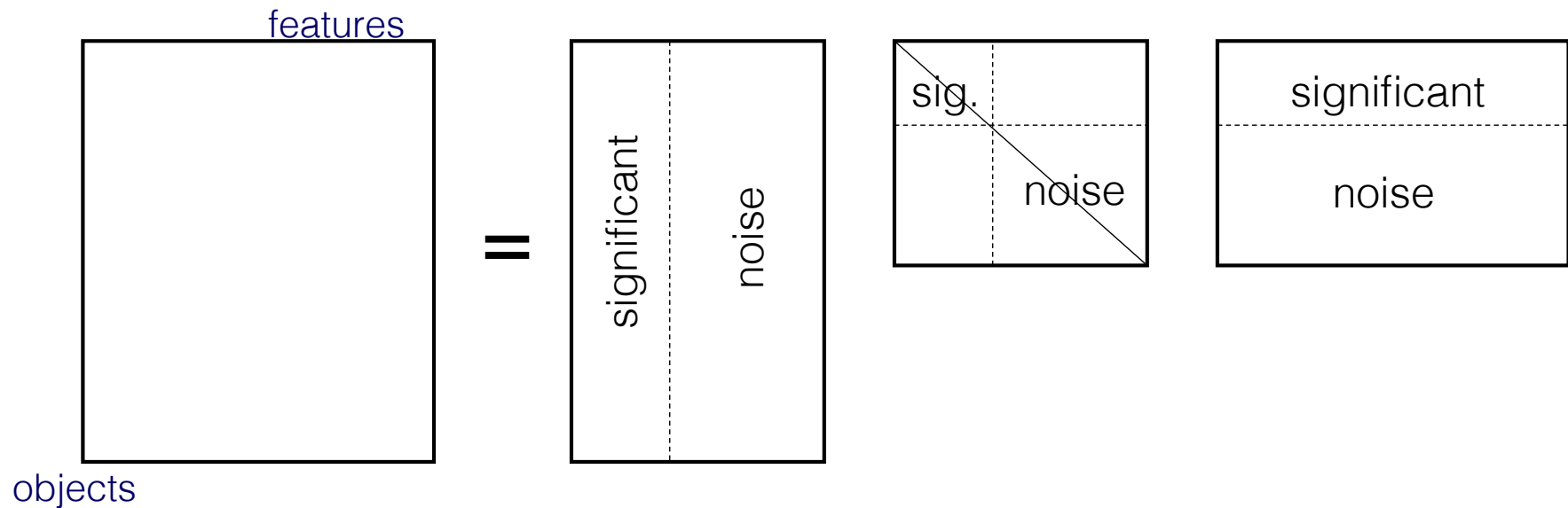
Let $\sigma_1, \sigma_2, \dots, \sigma_\rho$ be the entries of Σ .

Exact computation of the SVD takes $O(\min\{mn^2, m^2n\})$ time.

The top k left/right singular vectors/values can be computed faster using iterative methods.

Rank- k approximations via the SVD

$$A = U \Sigma V^T$$



Rank- k approximations (A_k)

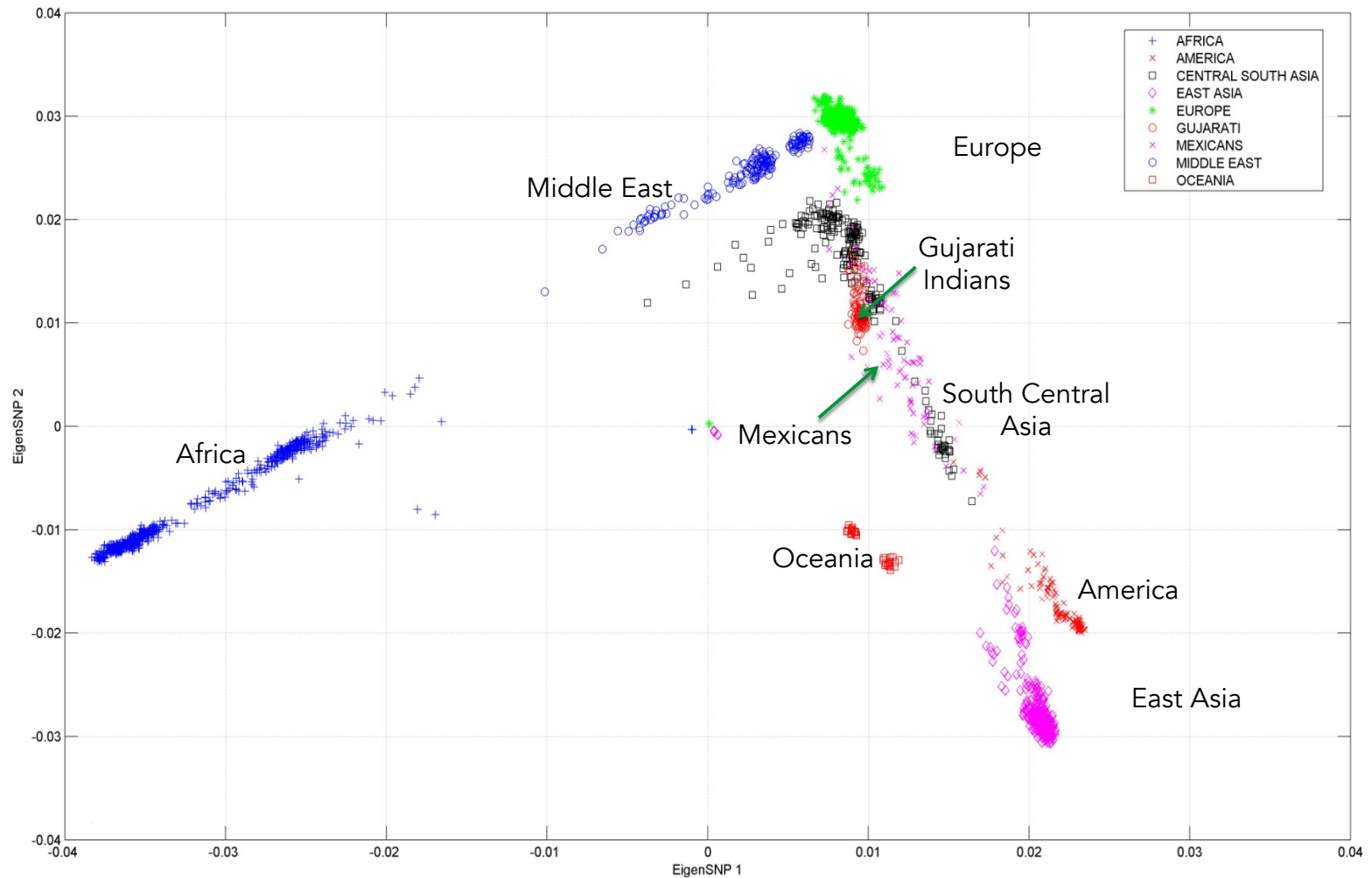
$$\begin{pmatrix} A_k \\ m \times n \end{pmatrix} = \begin{pmatrix} U_k \\ m \times k \end{pmatrix} \cdot \begin{pmatrix} \Sigma_k \\ k \times k \end{pmatrix} \cdot \begin{pmatrix} V_k^T \\ k \times n \end{pmatrix}$$

U_k (V_k): orthogonal matrix containing the top k left (right) singular vectors of A .
 Σ_k : diagonal matrix containing the top k singular values of A .

PCA (Principal Components Analysis) essentially amounts to the computation of the SVD of a mean-centered covariance matrix.

SVD is the algorithmic tool behind MultiDimensional Scaling (MDS). Factor Analysis, etc.

Paschou, et al (2010) J Med Genet

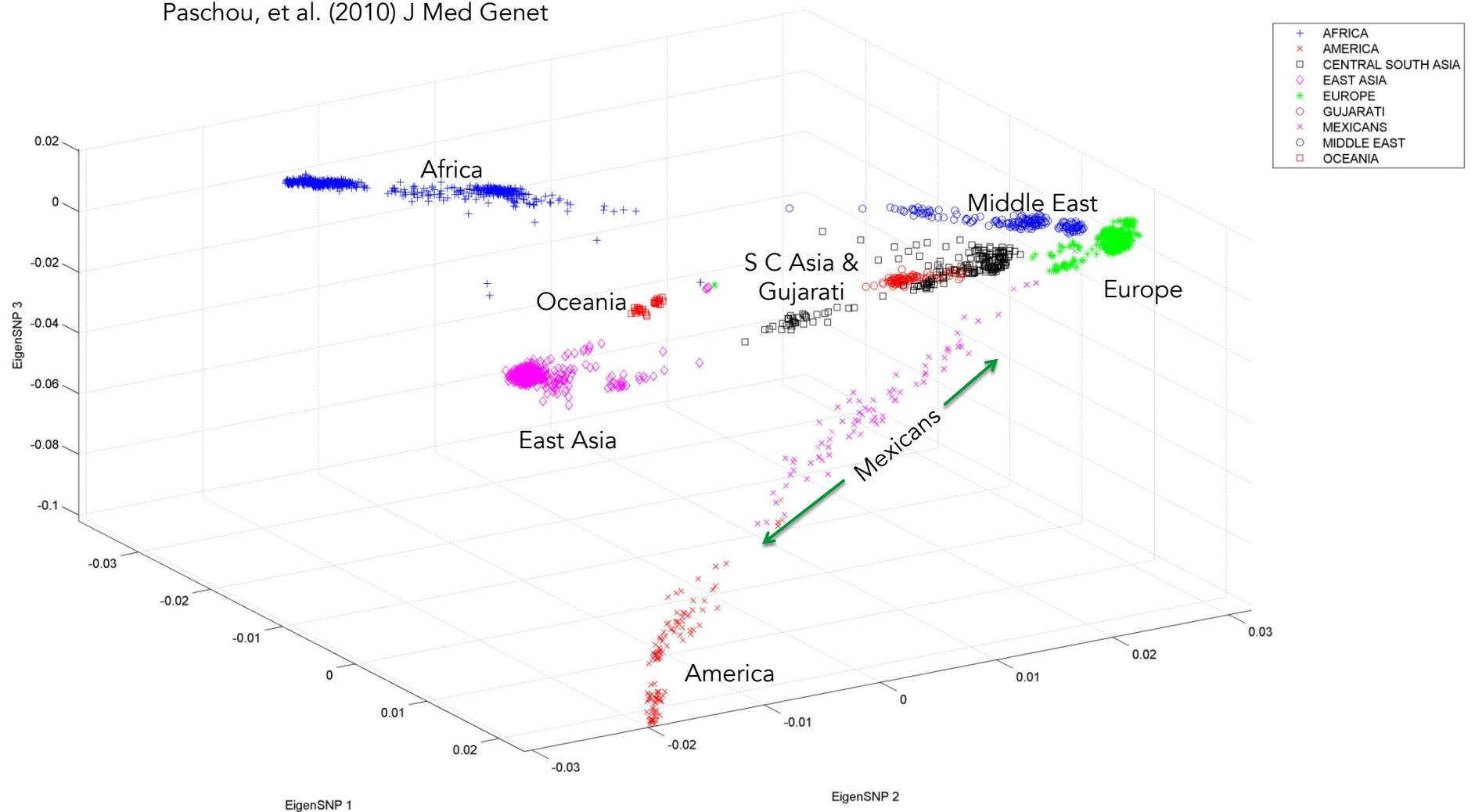


- Top two Principal Components (PCs or eigenSNPs)

(Lin and Altman (2005) *Am J Hum Genet*)

- The figure renders visual support to the “out-of-Africa” hypothesis.
- Mexican population seems out of place: we move to the top three PCs.

Paschou, et al. (2010) J Med Genet



- **Not altogether satisfactory:** the principal components are linear combinations of all SNPs, and – of course – can not be assayed!
- Can we find **actual SNPs** that capture the information in the singular vectors?
 - Relatedly, can we compute them and/or the truncated SVD “efficiently.”

Two related issues with eigen-analysis

Computing large SVDs: computational time

- In [commodity hardware](#) (e.g., a 4GB RAM, dual-core laptop), using MatLab 7.0 (R14), the computation of the SVD of the dense 2,240-by-447,143 matrix [A takes ca 20 minutes](#).
- Computing this SVD is not a one-liner, since we can not load the whole matrix in RAM (runs out-of-memory in MatLab).
- Instead, compute the SVD of AA^T .
- In a similar experiment, compute **1,200 SVDs** on matrices of dimensions (approx.) 1,200-by-450,000 (roughly, a full leave-one-out cross-validation experiment) (DLP2010)

Selecting *actual columns* that “capture the structure” of the top PCs

- Combinatorial optimization problem; hard even for small matrices.
- Often called the Column Subset Selection Problem (CSSP).
- Not clear that such “good” columns even exist.
- Avoid “reification” problem of “interpreting” singular vectors!
- (Solvable in “random projection time” with CX/CUR decompositions! (PNAS, MD09))

CUR matrix decompositions

Mahoney and Drineas "CUR Matrix Decompositions for Improved Data Analysis" (PNAS, 2009)

Mahoney, "Randomized Algorithms for Matrices and Data," FnTML, 2011

Drineas and Mahoney, "RandNLA: Randomized Numerical Linear Algebra," CACM, 2016

Goal. Solve the following problem:

"While very efficient basis vectors, the (singular) vectors themselves are completely artificial and do not correspond to actual (DNA expression) profiles. . . . Thus, it would be interesting to try to find basis vectors for all experiment vectors, using actual experiment vectors and not artificial bases that offer little insight." Kuruvilla et al. (2002)

Theorem:

Given an arbitrary matrix, call a black box that I won't describe.

- You get a small number of actual columns/rows that are only marginally worse than the truncated PCA/SVD.
- The black box runs faster than computing a truncated PCA/SVD for arbitrary input.
- It's very robust to heuristic modifications.

Corollary:

We can use the same methods to approximate the PCA/SVD.

CUR matrix decompositions and RandNLA

Mahoney and Drineas "CUR Matrix Decompositions for Improved Data Analysis" (PNAS, 2009)

Mahoney, "Randomized Algorithms for Matrices and Data," FnTML, 2011

Drineas and Mahoney, "RandNLA: Randomized Numerical Linear Algebra," CACM, 2016

One of many methods from **Randomized (Numerical) Linear Algebra (RandNLA)**:

- Interdisciplinary research area
- Exploits randomization as a computational resource to develop improved algorithms for large-scale linear algebra problems

Qua **methods**,
RandNLA:

- Roots in theoretical computer science (TCS)
- Deep connections to mathematics and applied mathematics
- Statistical interpretation that complement bootstrapping, etc.

Qua **applications**,
RandNLA:

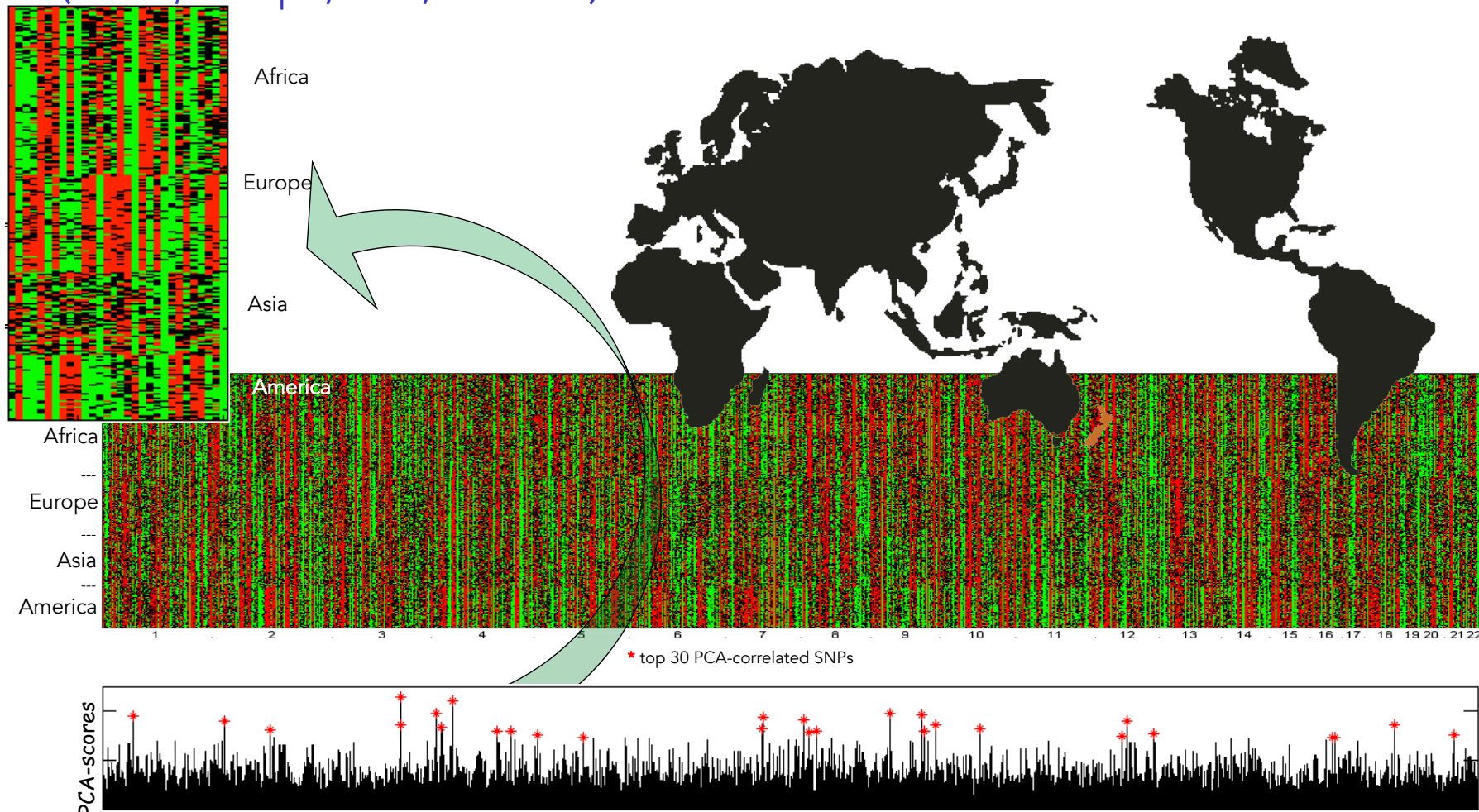
- Vital new tool for machine learning, statistics, and data analysis
- Solved state-of-the-art problems in genetics, astronomy, mass spec imaging, etc.

Qua **implementations**,
RandNLA:

- Outperform highly-optimized software (LAPACK)
- Scalability in parallel and distributed environments
- Terabyte-scale PCA in Spark/Alchemist

Promises sound algorithmic and statistical foundation for modern large-scale data analysis.

Selecting PCA SNPs for individual assignment to four continents (Africa, Europe, Asia, America)



- Data analysis and machine learning and statistics and theory of algorithms and scientific computing ... and genetics and astronomy and mass spectrometry and ... likes this---but each for different reasons!

- Good “hydrogen atom” for methods development!

Mahoney and Drineas (2009) PNAS
 Paschou et al (2007; 2008) PLoS Genetics
 Paschou et al (2010) J Med Genet
 Drineas et al (2010) PLoS One
 Javed et al (2011) Annals Hum Genet

Bioinformatics: a cautionary tale?

- How did/does bioinformatics relate to computer science, statistics, and applied mathematics, “technically” and “sociologically”?
- How did NIH choose to fund graduate students and postdocs in the budget expansion of the 90s?
- What effect did this have on the number of American/foreign going into biomedical research?
- How will the pay structure of biomedical researchers effect which cs/stats “data scientists” engage you in your efforts?
- What effect does med schools deciding not to do joint faculty hires with cs departments have on bioinformatics and big biomedical data?
- How is this Big Biomedical Data phenomenon similar to and different than the Bioinformatics experience?

Big changes in the past ... and future

Consider the creation of:

- Modern Physics
- Computer Science
- Molecular Biology
- OR and Management Science
- Transistors and Microelectronics
- Biotechnology

These were driven by *new measurement techniques* and *technological advances*, but they led to:

- big new (academic and applied) questions
- new perspectives on the world
- lots of downstream applications

We are in the middle of a similarly big shift!



QUESTION: What, if anything, does biomedicine have to offer?