

Genomic Futures

Michael Schatz

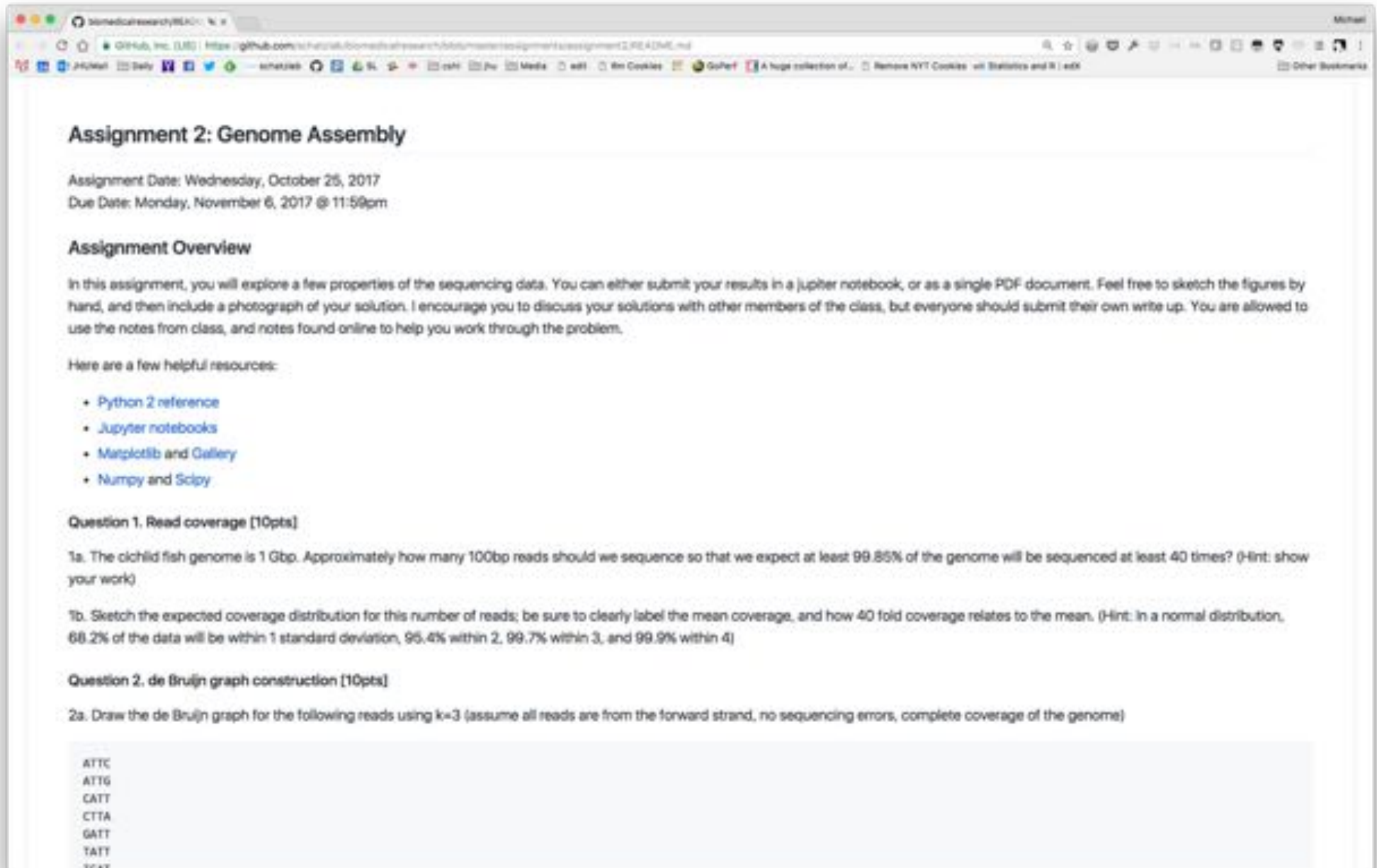
November 27 – Lecture 23

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research



HW 2 Review



The screenshot shows a web browser window with the address bar displaying a GitHub repository URL. The page title is "Assignment 2: Genome Assembly". Below the title, the assignment date is "Wednesday, October 25, 2017" and the due date is "Monday, November 6, 2017 @ 11:59pm". The "Assignment Overview" section states that students will explore sequencing data properties and can submit results as a Jupyter notebook or a PDF. It encourages discussion but requires individual submissions. A list of helpful resources includes Python 2 reference, Jupyter notebooks, Matplotlib and Gallery, and Numpy and Scipy. The "Question 1. Read coverage [10pts]" section contains two sub-questions: 1a. asks for the number of 100bp reads needed for 99.85% coverage of a 1 Gbp genome at least 40 times; 1b. asks for a sketch of the expected coverage distribution. The "Question 2. de Bruijn graph construction [10pts]" section contains sub-question 2a, which asks for a de Bruijn graph construction for a set of reads using k=3.

Assignment 2: Genome Assembly

Assignment Date: Wednesday, October 25, 2017
Due Date: Monday, November 6, 2017 @ 11:59pm

Assignment Overview

In this assignment, you will explore a few properties of the sequencing data. You can either submit your results in a jupyter notebook, or as a single PDF document. Feel free to sketch the figures by hand, and then include a photograph of your solution. I encourage you to discuss your solutions with other members of the class, but everyone should submit their own write up. You are allowed to use the notes from class, and notes found online to help you work through the problem.

Here are a few helpful resources:

- [Python 2 reference](#)
- [Jupyter notebooks](#)
- [Matplotlib and Gallery](#)
- [Numpy and Scipy](#)

Question 1. Read coverage [10pts]

1a. The cichlid fish genome is 1 Gbp. Approximately how many 100bp reads should we sequence so that we expect at least 99.85% of the genome will be sequenced at least 40 times? (Hint: show your work)

1b. Sketch the expected coverage distribution for this number of reads; be sure to clearly label the mean coverage, and how 40 fold coverage relates to the mean. (Hint: in a normal distribution, 68.2% of the data will be within 1 standard deviation, 95.4% within 2, 99.7% within 3, and 99.9% within 4)

Question 2. de Bruijn graph construction [10pts]

2a. Draw the de Bruijn graph for the following reads using $k=3$ (assume all reads are from the forward strand, no sequencing errors, complete coverage of the genome)

```
ATTC
ATTG
CATT
CTTA
GATT
TATT
TCAT
```

Project Presentations

Project Presentations

Recommended outline for your talk (1 minute per slide):

0. Title Slide: Who are you, title, date, mentor
1. Intro 1: Whats the big idea???
2. Intro 2: More specifically, what are you trying to learn?
3. Methods 1: What did you try?
4. Methods 2: What is the key idea?
5. Data 1: What data are you looking at?
6. Data 2: Anything notable about the data?
7. Results 1: What did you see!
8. Results 2: Does it work?
9. Discussion: What did you learn from this study & what is your future work?
10. Acknowledgements: Who helped you along the way?

I strongly discourage you from trying to give a live demo as they are too unpredictable for a short talk. If you have running software you want to show, use a "cooking show" approach, where you have screen shots of the important steps.

Presentations will be a total of 12 minutes: 10 minutes for the presentation, followed by 2 minutes for questions. We will strictly keep to the schedule to ensure that everyone can present in class.

Schedule of Presentations

Day	Time	Student	Title
We 11/29	3:00-3:12	Tatiana Romer	Establishing an improved Genome and Methylome for Manduca sexta
We 11/29	3:12-3:24	Sophie Shoemaker	AutoPhaser: Hedging our bets for phasing x-ray crystallography using molecular replacement
We 11/29	3:24-3:36	George Botev	Fast Twister: Fast Haplotype Phasing
-			
Mo 12/4	3:00-3:12	Syed Usman Enam	Regional mapping of RNA Binding Proteins on mRNA transcripts
Mo 12/4	3:12-3:24	Sun Jay Yoo	High-Performance Software Pipeline for Single-Molecule Tracking Data Analysis
Mo 12/4	3:24-3:36	David Li	Mixture Models for Visual Concept Detection
-			
We 12/6	3:00-3:12	Elisabeth Wood	Inferring ethnicity from whole genome single nucleotide polymorphisms
We 12/6	3:12-3:24	Amanda Laine	In situ detection of DNA repair alterations in cancer
We 12/6	3:24-3:36	Cathleen Nguyen	Implementation of Precision Medicine in the Clinical Management of Oncological Patients
We 12/6	3:36-3:48	Kiki Chang	Extending annotations in Comprehensive Human Expressed Sequences

Final Report

Project Final Report

Due Date: Wed, Dec 20, 2017 @ 11:59pm

Each student should email a PDF of your final project report (6-10 pages) to "jhubiomedicalresearch at gmail dot com" by 11:59pm on Wednesday December 20, 2017

The report should have:

- Title of your project
- Name and email addresses of you, your mentor, and anyone else that you worked closely with
- 1 paragraph Abstract summarizing the project
- ~1 page of Introduction
- ~1-2 page of Methods that you are using (at least 1 figure of the method/workflow)
- ~1-2 page of Results, describing the data evaluated and your results (plan for 3 - 4 figures showing data and results)
- ~1 page of Discussion (what you have seen or expect to see)
- 1 paragraph of Acknowledgements
- 5 to 20 References to relevant papers and data

The report can use any style, although clearly mark each section (Title, Name, Abstract, Introduction, Methods, Results, Discussion, Acknowledgements, References). You can use the [Bioinformatics Templates](#) if you wish but they are not required. Feel free to reuse text from your preliminary report if it is still appropriate.

Please use Piazza if you have any general questions!

The human genome (2001)



"Without doubt, this is the most important, most wonderful achievement by humankind."



Human disease genes

Gerardo Jimenez-Sanchez*, Barton Childs* & David Valle*†

* Department of Pediatrics, McKusick-Nathans Institute of Genetic Medicine, and † Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

The complete human genome sequence will facilitate the identification of all genes that contribute to disease. We propose that the functional classification of disease genes and their products will reveal general principles of human disease. We have determined functional categories for nearly 1,000 documented disease genes, and found striking correlations between the function of the gene product and features of disease, such as age of onset and mode of inheritance. As knowledge of disease genes grows, including those contributing to complex traits, more sophisticated analyses will be possible; their results will yield a deeper understanding of disease and an enhanced integration of medicine with biology.

To test the proposal that classifying disease genes and their products according to function will provide general insight into disease processes^{1,2}, we have compiled and classified a list of disease genes. To assemble the list, we began with 269 genes identified in a survey of the 7th edition of *Metabolic and Molecular Bases of Inherited Disease*². We then searched the 'morbid map' and allelic variants listed in the *Online Mendelian Inheritance in Man*³ (OMIM), an online resource documenting human diseases and their associated genes

(www.ncbi.nlm.nih.gov), and increased the total disease gene set to 923. This sample included genes that cause monogenic disease (97% of the sample) and genes that increase susceptibility for complex traits. We excluded genes associated only with somatic genetic disease (such as non-inherited forms of cancer) or the mitochondrial genome.

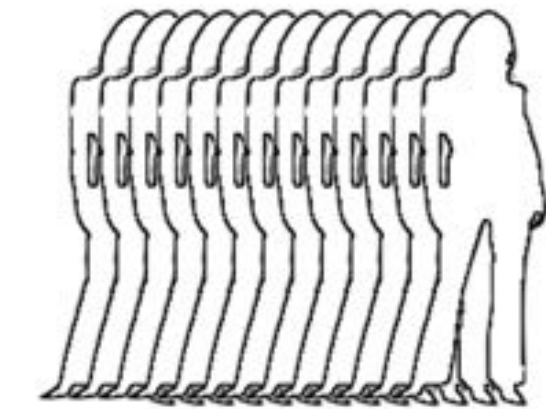
Functional classification

We categorized each disease gene according to the function of its

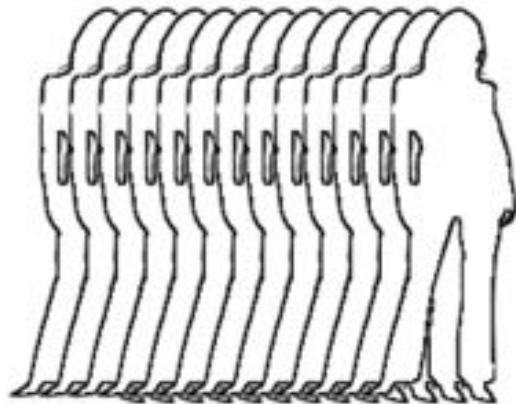
Human disease genes

Jimenez-Sanchez, G., Childs, B. & Valle, D. (2001) *Nature* 409, 853–855

Genome Wide Association (GWAS)



GC CC GG GC CC GC GC
GG CC GC GG GC GG



GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1

Cases

Count of G:
2104 of 4000

Frequency of G:
52.6%

Controls

Count of G:
2676 of 6000

Frequency of G:
44.6%

SNP2

Cases

Count of G:
1648 of 4000

Frequency of G:
41.2%

Controls

Count of G:
2532 of 6000

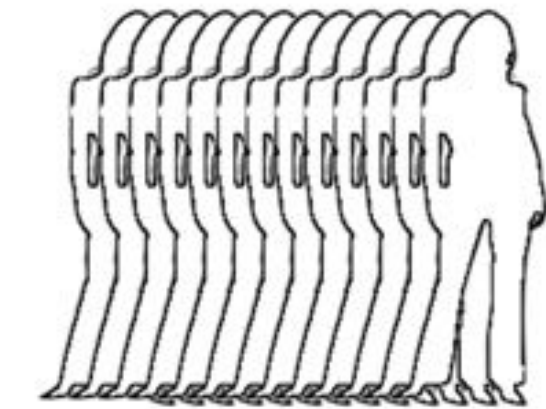
Frequency of G:
42.2%

SNP...

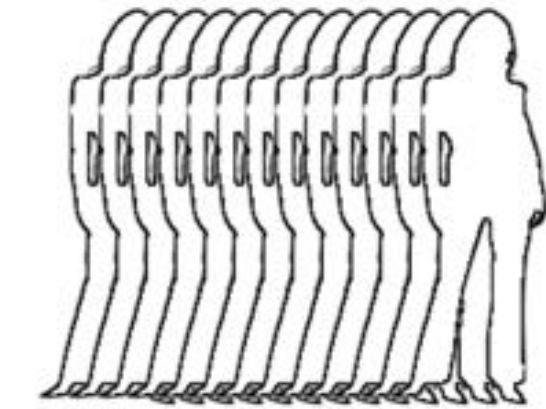
*Repeat for all
SNPs*

Are these significant
differences in frequencies?

Genome Wide Association (GWAS)



GC CC GG GC CC GC GC
GG CC GC GG GC GG



GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1

Cases

Count of G:
2104 of 4000

Frequency of G:
52.6%

Controls

Count of G:
2676 of 6000

Frequency of G:
44.6%

P-value:
 $5.0 \cdot 10^{-15}$

SNP2

Cases

Count of G:
1648 of 4000

Frequency of G:
41.2%

Controls

Count of G:
2532 of 6000

Frequency of G:
42.2%

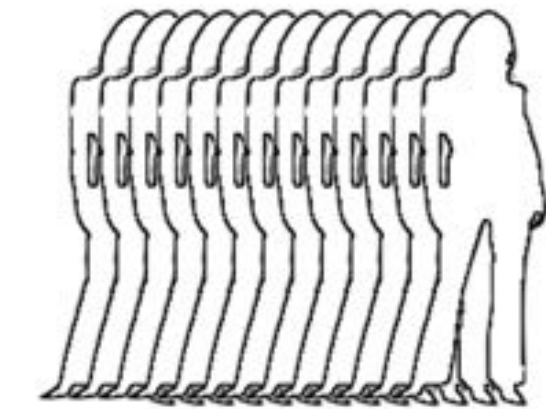
P-value:
0.33

SNP...

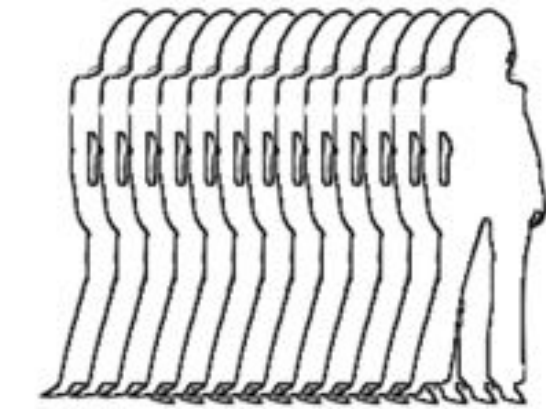
*Repeat for all
SNPs*

Chi-squared or
similar test

Genome Wide Association (GWAS)



GC CC GG GC CC GC GC
GG CC GC GG GC GG



GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1

Cases

Count of G:
2104 of 4000

Frequency of G:
52.6%

Controls

Count of G:
2676 of 6000

Frequency of G:
44.6%

P-value:
 $5.0 \cdot 10^{-15}$

SNP2

Cases

Count of G:
1648 of 4000

Frequency of G:
41.2%

Controls

Count of G:
2532 of 6000

Frequency of G:
42.2%

P-value:
0.33

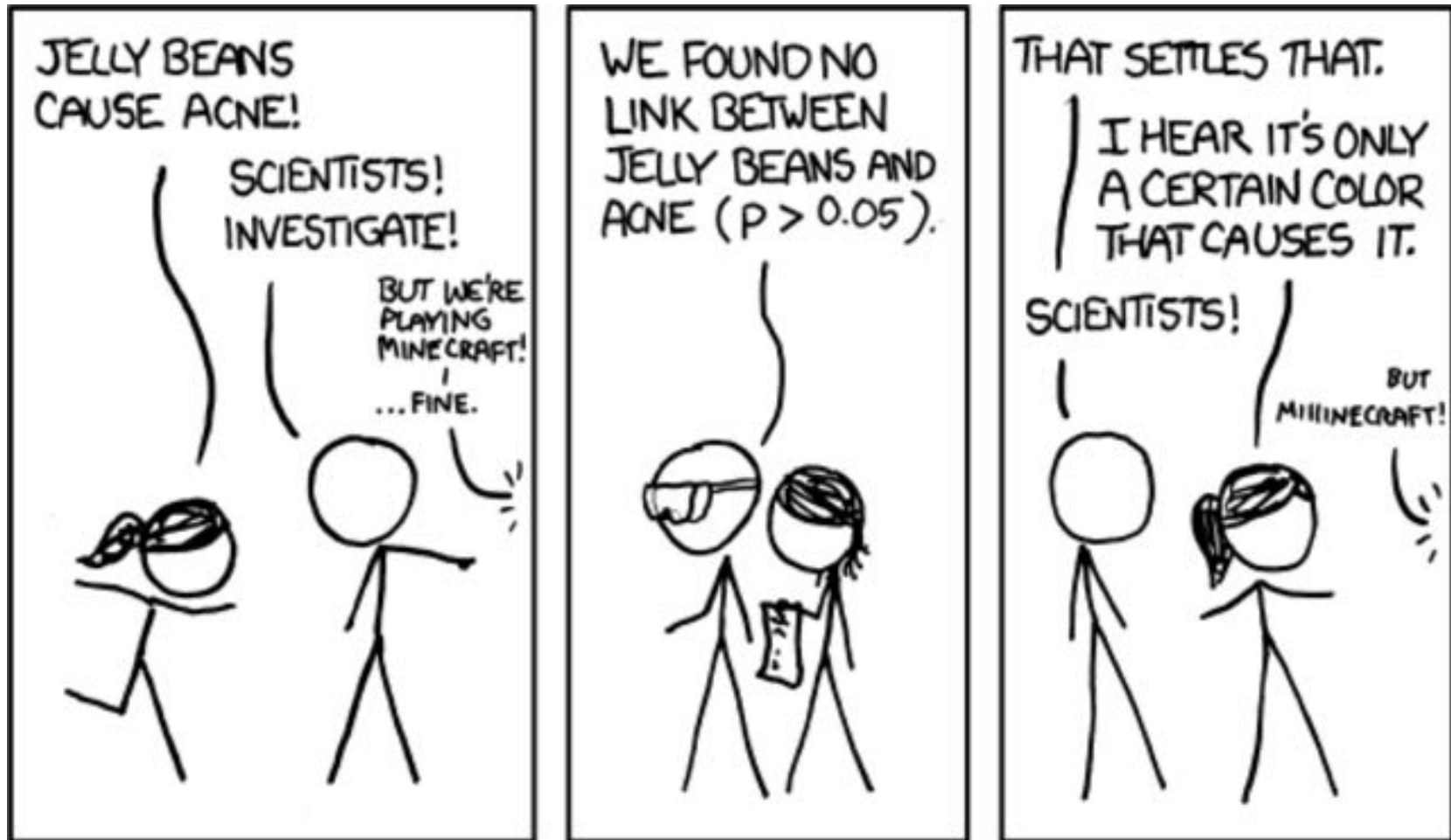
SNP...

*Repeat for all
SNPs*

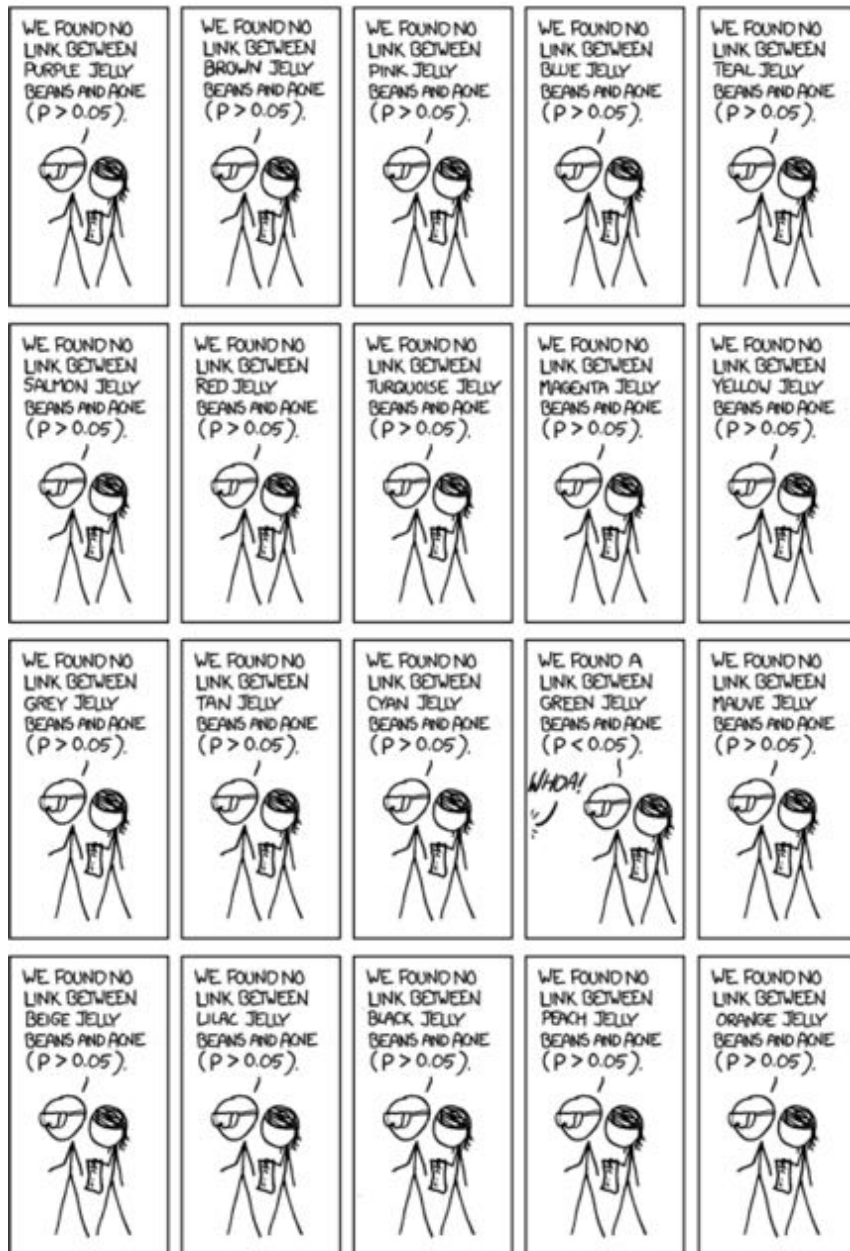
With a (much) larger
population, this might
be a significant
difference:
 $25320/60000 \Rightarrow$
 $p = 5e-7$

Chi-squared or
similar test

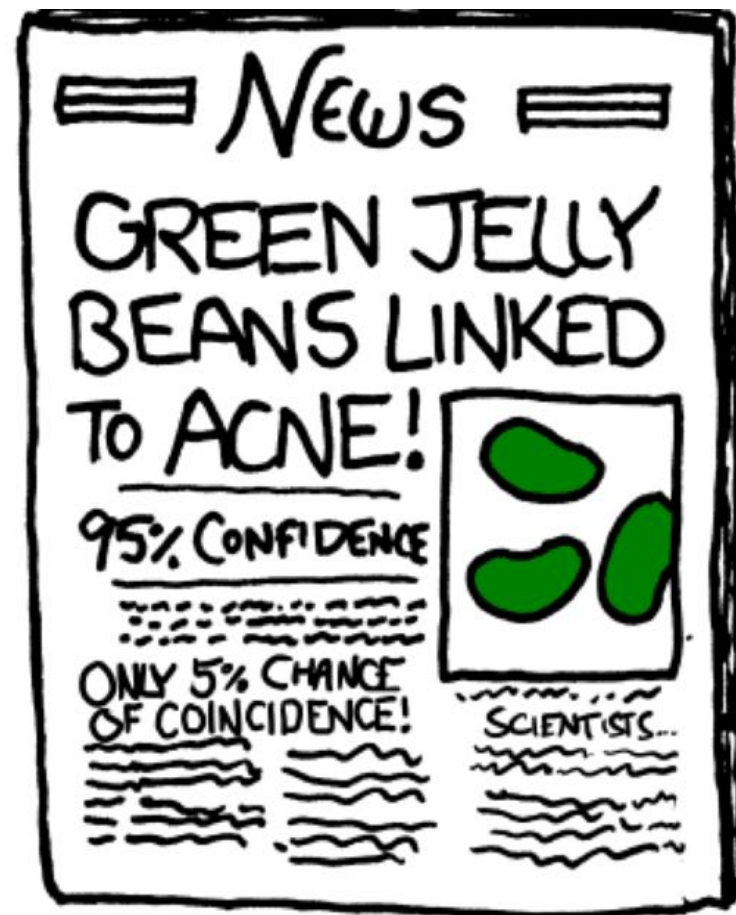
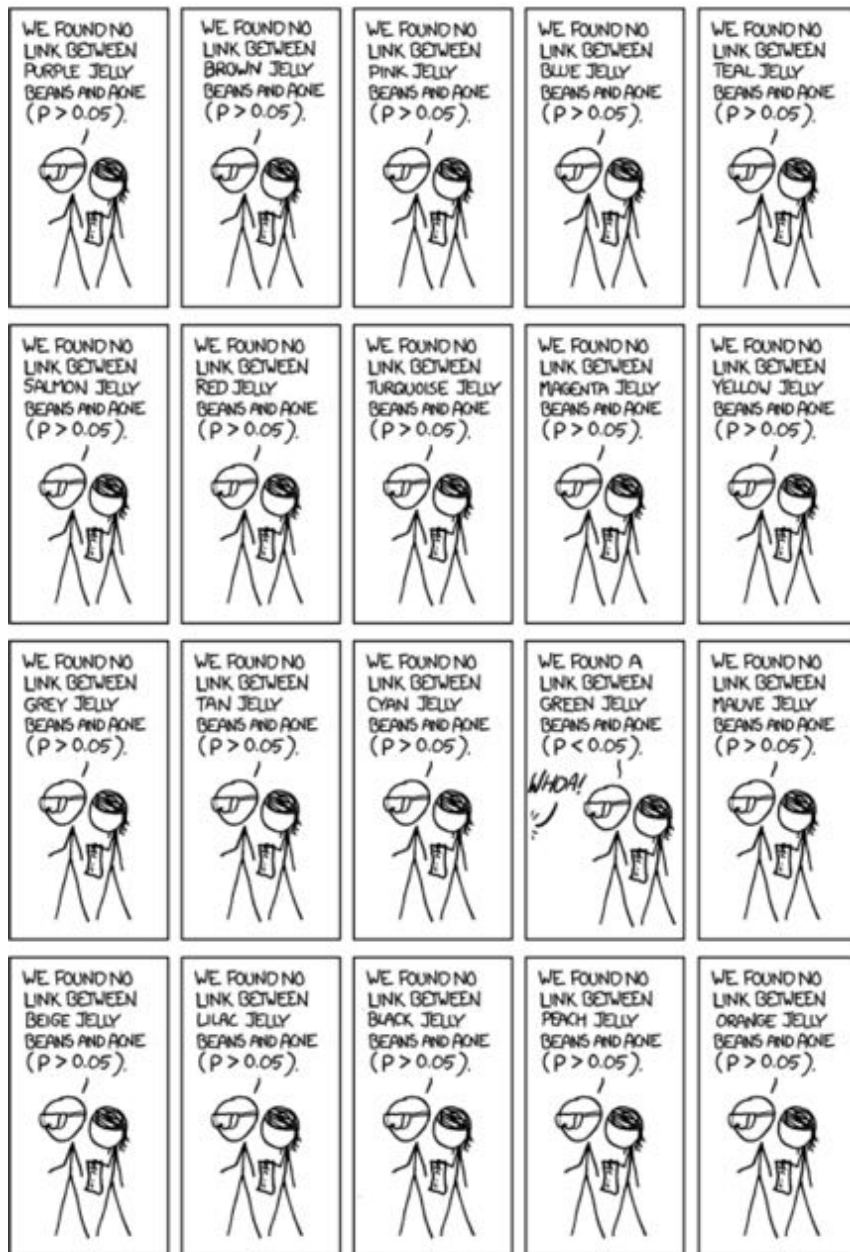
The curse of multiple testing



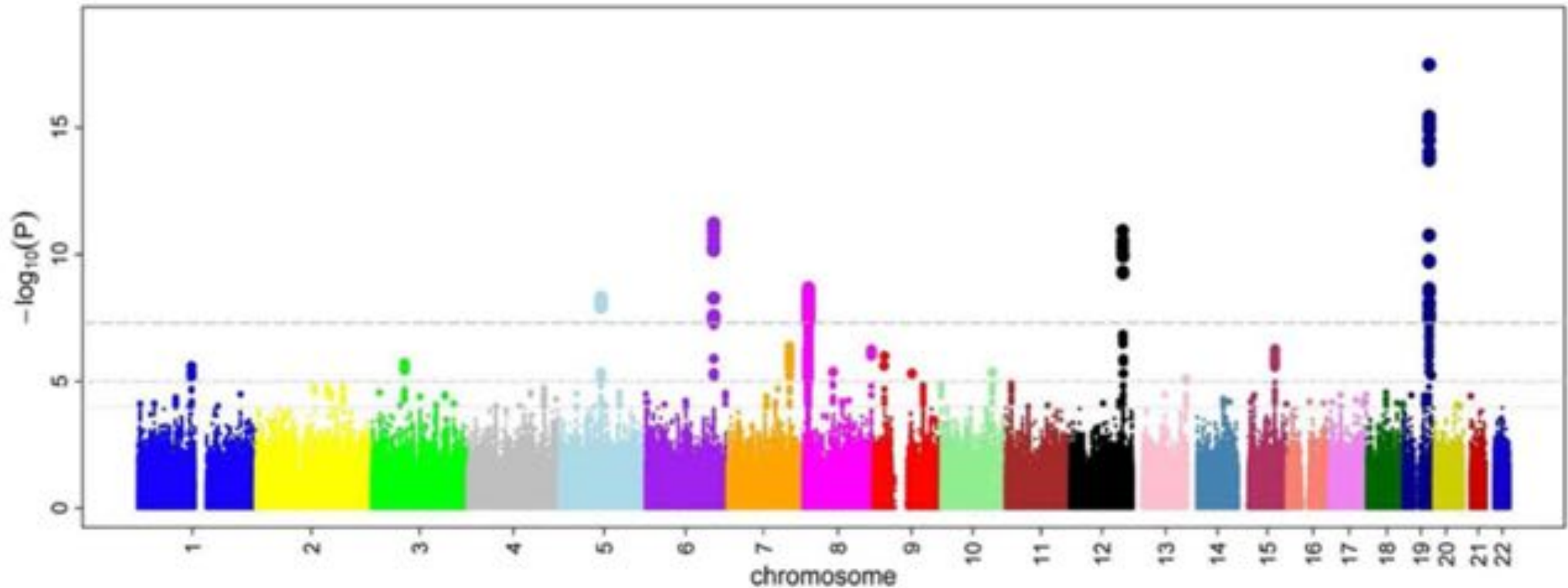
The curse of multiple testing



The curse of multiple testing



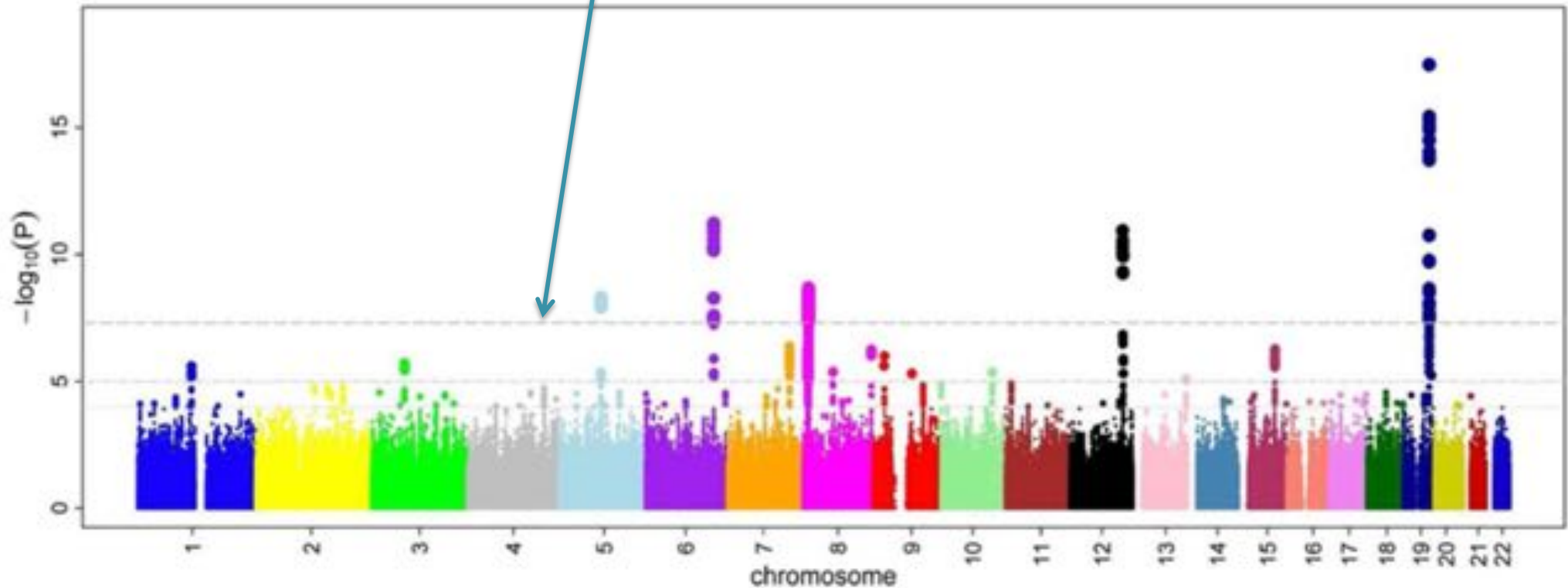
Manhattan Plot



Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

Manhattan Plot

Genome-wide significance: $5e-8$
“Bonferroni correction”

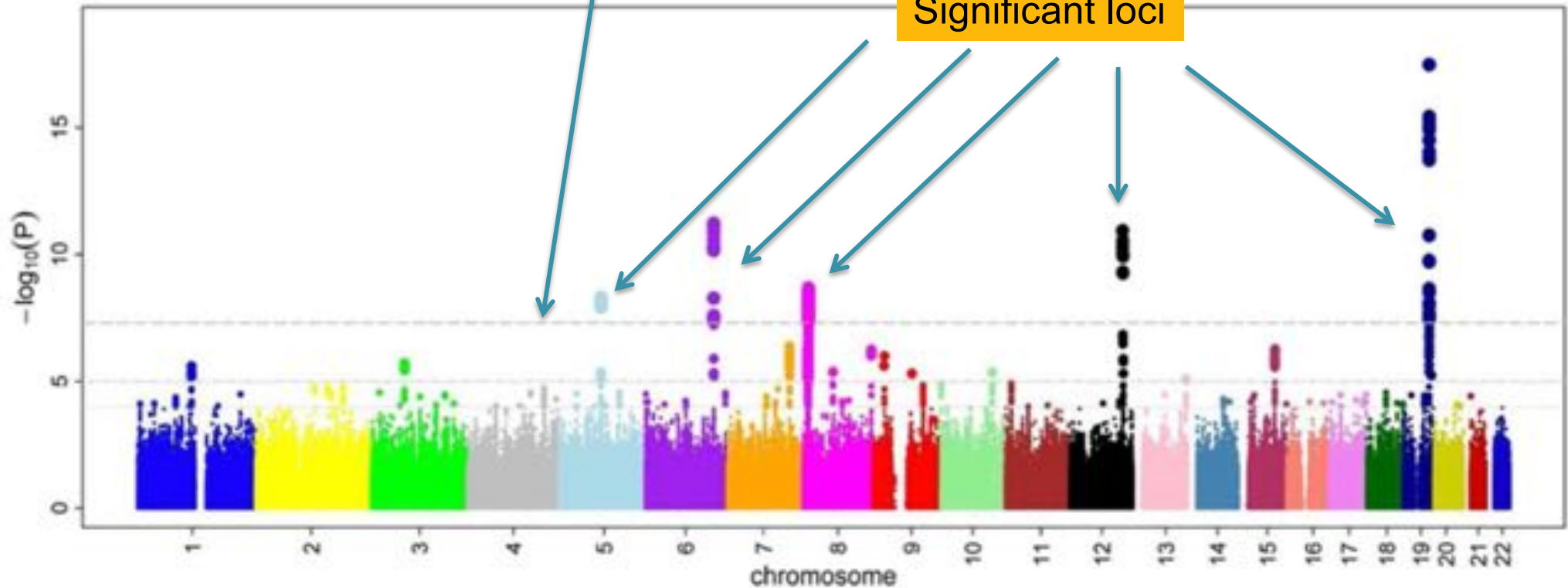


Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

Manhattan Plot

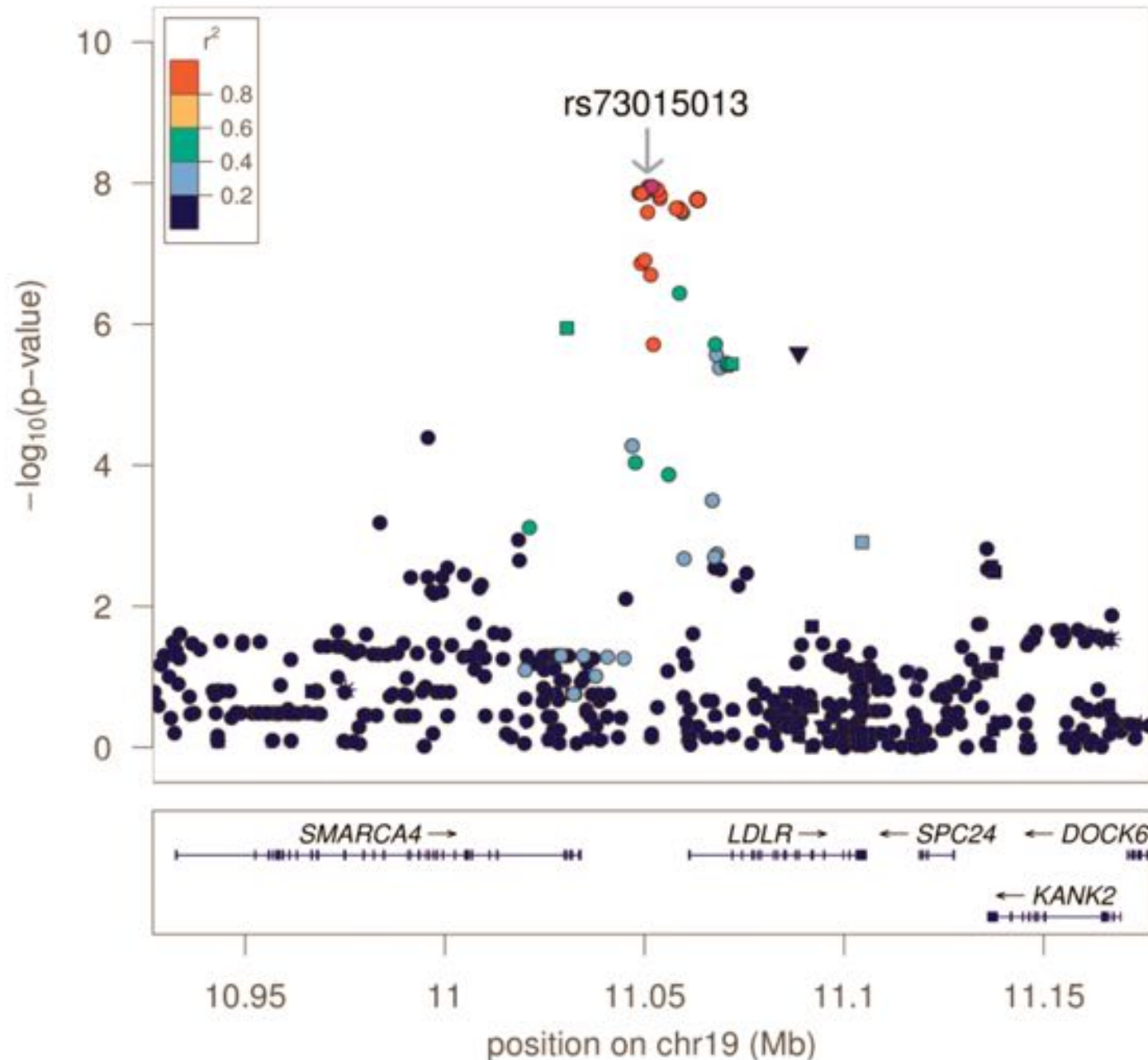
Genome-wide significance: $5e-8$
“Bonferroni correction”

Significant loci

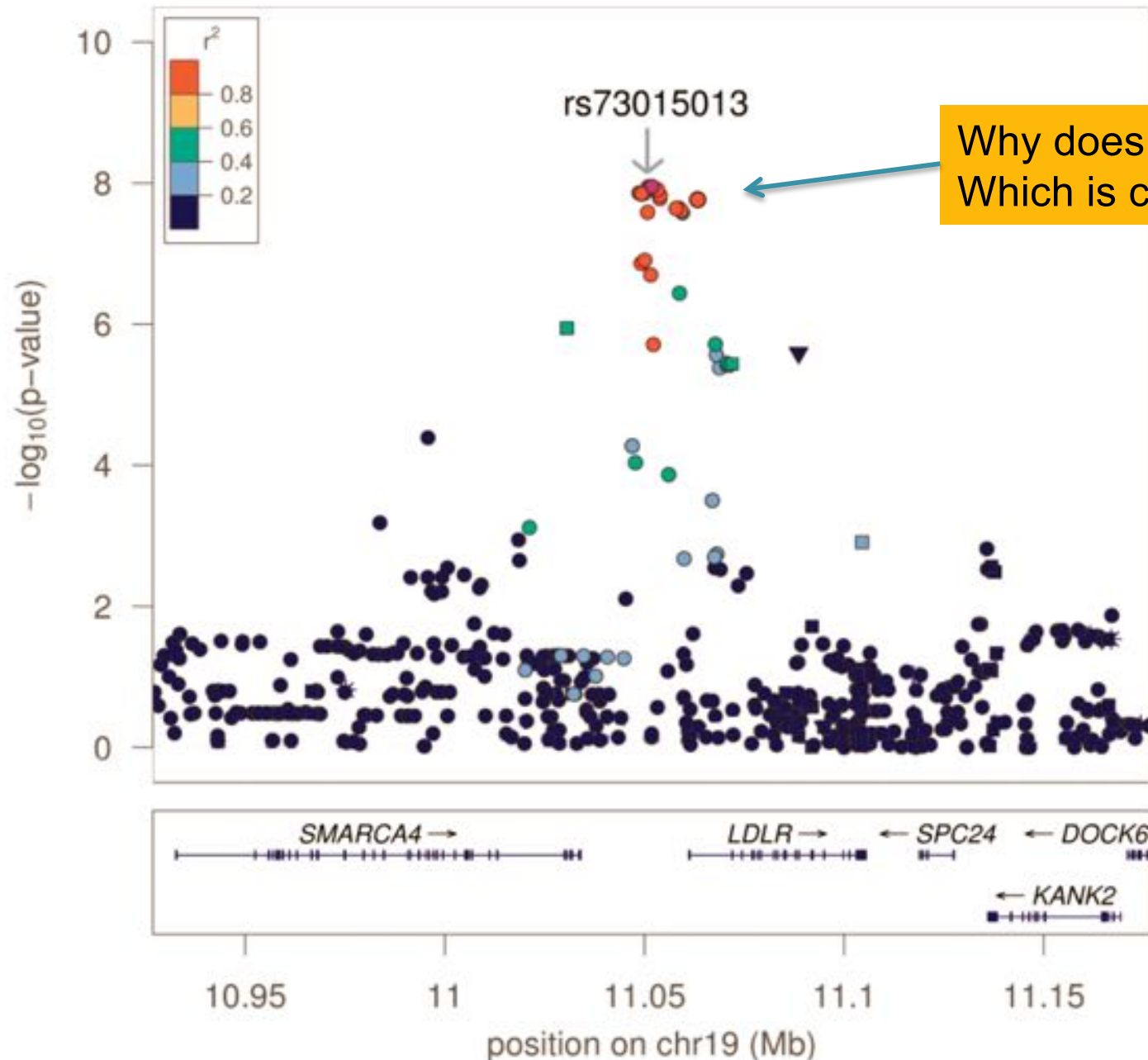


Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

Regional Association Plot



Regional Association Plot



OMIM

Secure <https://www.omim.org>

About Statistics Downloads Contact Us MIMmatch Donate Help

50 YEARS OMIM
Human Genetics Knowledge for the World

OMIM®
Online Mendelian Inheritance in Man®
An Online Catalog of Human Genes and Genetic Disorders
Updated April 7, 2017

Search OMIM for clinical features, phenotypes, genes, and more...

Advanced Search : OMIM, Clinical Synopses, Gene Map | Search History
Need help? : Example Searches, OMIM Search Help, OMIM Tutorial
Mirror site : mirror.omim.org

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

[Make a donation!](#)

[Follow us on Twitter](#)

NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions.
OMIM® and Online Mendelian Inheritance in Man® are registered trademarks of the Johns Hopkins University.
Copyright© 1966-2017 Johns Hopkins University.

- For many different diseases and phenotypes, lists what are all of the known genetic associations
- Has records for nearly all genes, ~5k different conditions with known molecular basis, ~1k with unknown basis, ~1k with questionable basis
- Started at JHU 50 years ago 😊

GWAS in Crisis

NEWS FEATURE PERSONAL GENOMES

NATURE Vol 456/6 November 2008



The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

If you want to predict how tall your children might one day be, a good bet would be to look in the mirror, and at your mate. Studies going back almost a century have estimated that height is 80–90% heritable. So if 29 centimetres separate the tallest 5% of a population from the shortest, then genetics would account for as many as 27 of them.

This year, three groups of researchers scoured the genomes of huge populations (the largest study looked at more than 30,000 people) for genetic variants associated with the height differences. More than 40 turned up.

But there was a problem: the variants had tiny effects. Altogether, they accounted for little more than 5% of height's heritability — just 6 centimetres by the calculations above.



Even though these genome-wide association studies (GWAS) turned up dozens of variants, they did "very little of the prediction that you would do just by asking people how tall their parents are", says Joel Hirschhorn at the Broad Institute in Cambridge, Massachusetts, who led one of the studies.

Height isn't the only trait in which genes have gone missing, nor is it the most important. Studies looking at similarities between identical and fraternal twins estimate heritability at more than 90% for autism and more than 80% for schizophrenia. And genetics makes a major contribution to disorders such as obesity, diabetes and heart disease. GWAS, one of the most celebrated techniques of the past five years, promised to deliver many of the genes involved (see "Where's the reward?", page 20). And to some extent they have, identifying more than 400 genetic variants that

contribute to a variety of traits and common diseases. But even when dozens of genes have been linked to a trait, both the individual and cumulative effects are disappointingly small and nowhere near enough to explain earlier estimates of heritability. "It is the big topic in the genetics of common disease right now," says Francis Collins, former head of the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. The unexpected results left researchers at a point "where we all had to scratch our heads and say, 'Huh!'", he says.

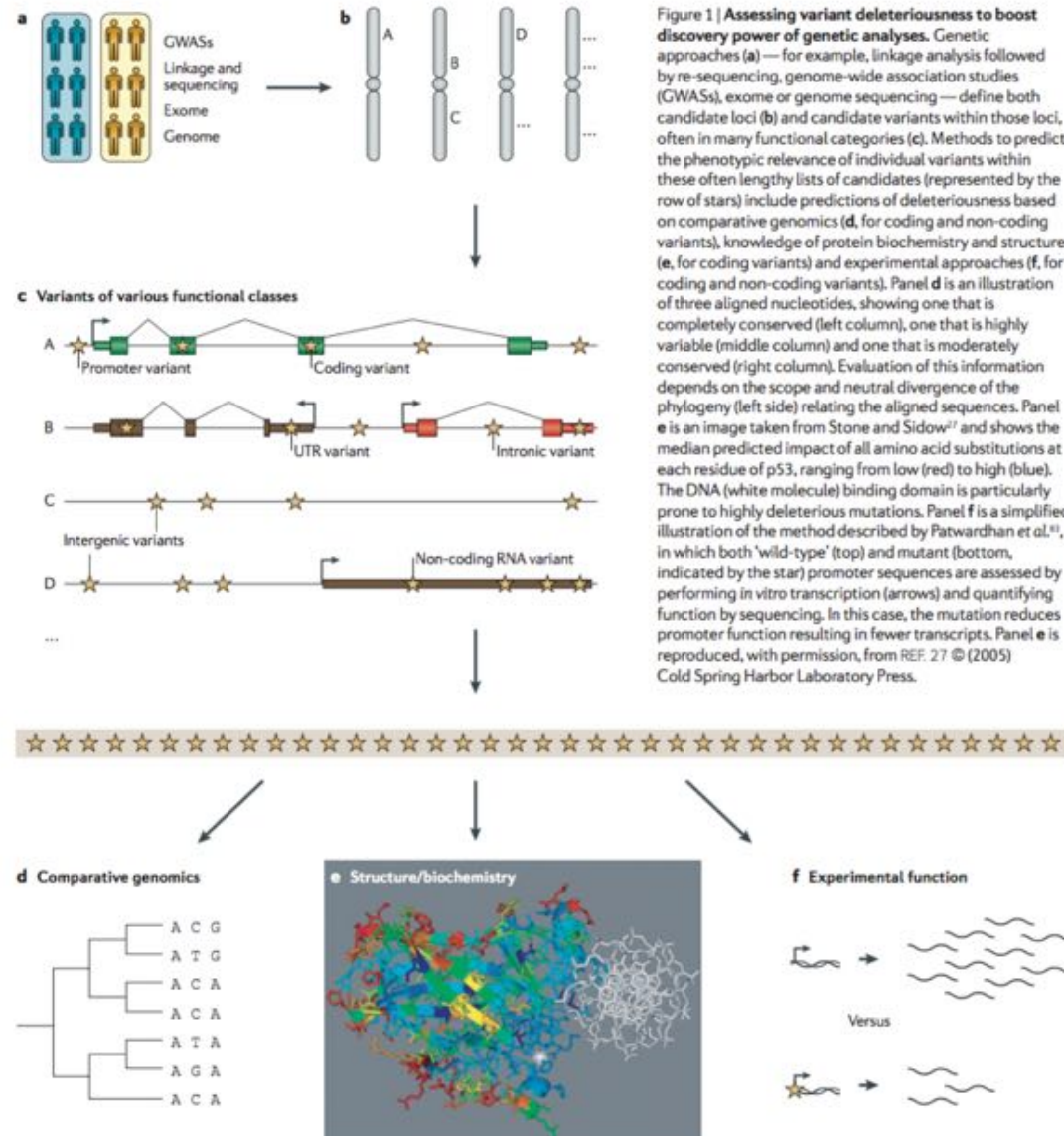
Although flummoxed by this missing heritability, geneticists remain optimistic that they can find more of it. "These are very early days, and there are things that are doable in the next year or two that may well explain another sizeable chunk of heritability," says Hirschhorn. So where might it be hiding?

ILLUSTRATION BY D. J. HARRIS

"Three groups of researchers scoured the genomes of huge populations (>30,000 people) for genetic variants associated with the height differences. More than 40 turned up. **But there was a problem: the variants had tiny effects.** Altogether, they accounted for little more than 5% of height's heritability"

- **Rare, moderately penetrant or common, weakly penetrant variants?**
- **CNVs and SVs?**
- **Epistasis (multiple genes working together)?**
- **Epigenetic effects, especially in utero?**

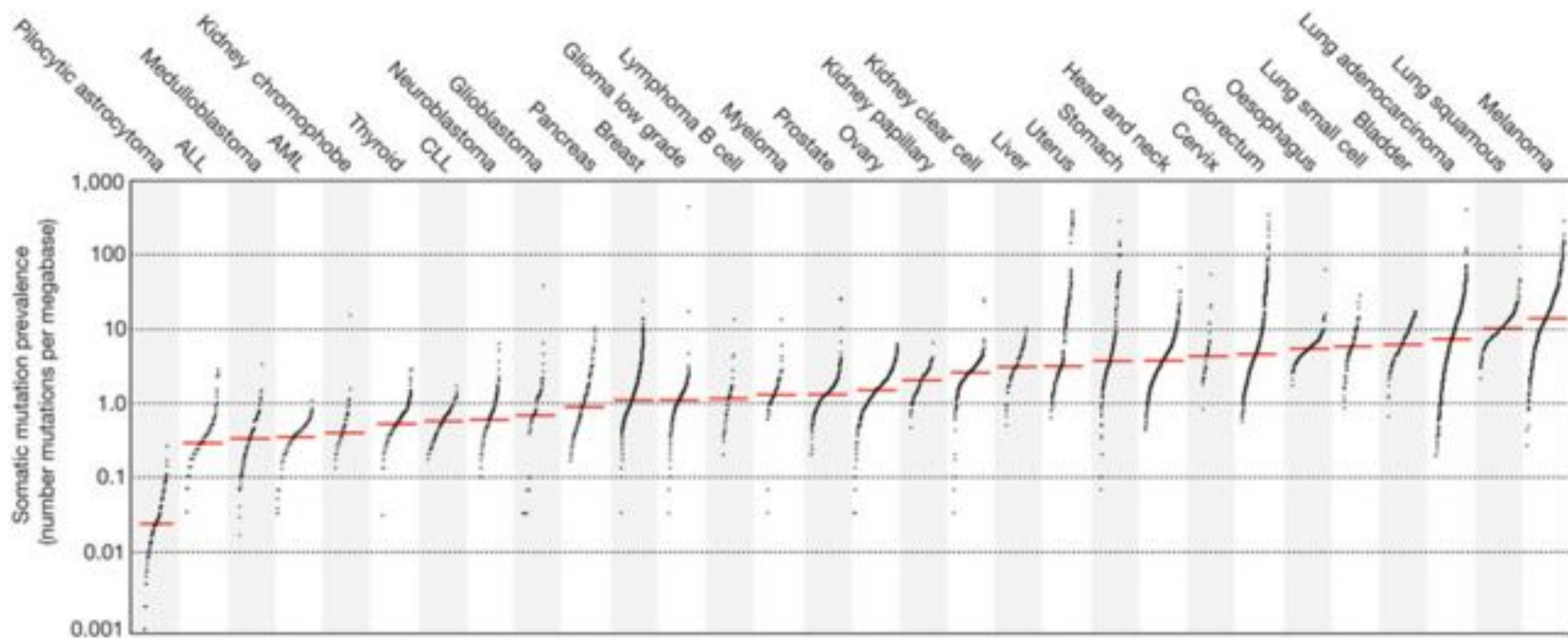
Needles in stacks of needles



Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data

Cooper & Shendure (2011) Nature Reviews Genetics.

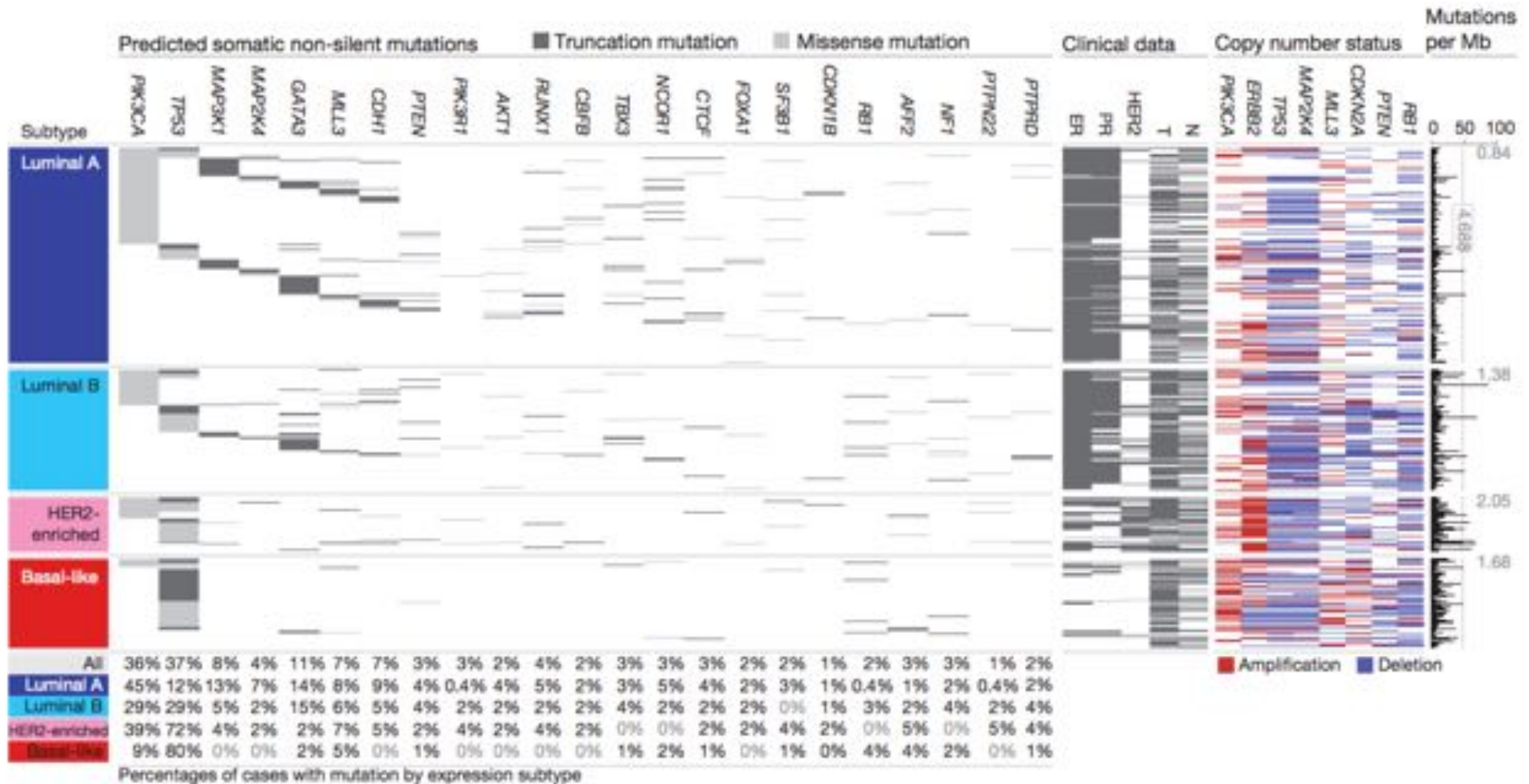
Somatic Mutations In Cancer



Signatures of mutational processes in human cancer

Alexandrov et al (2013) *Nature*. doi:10.1038/nature12477

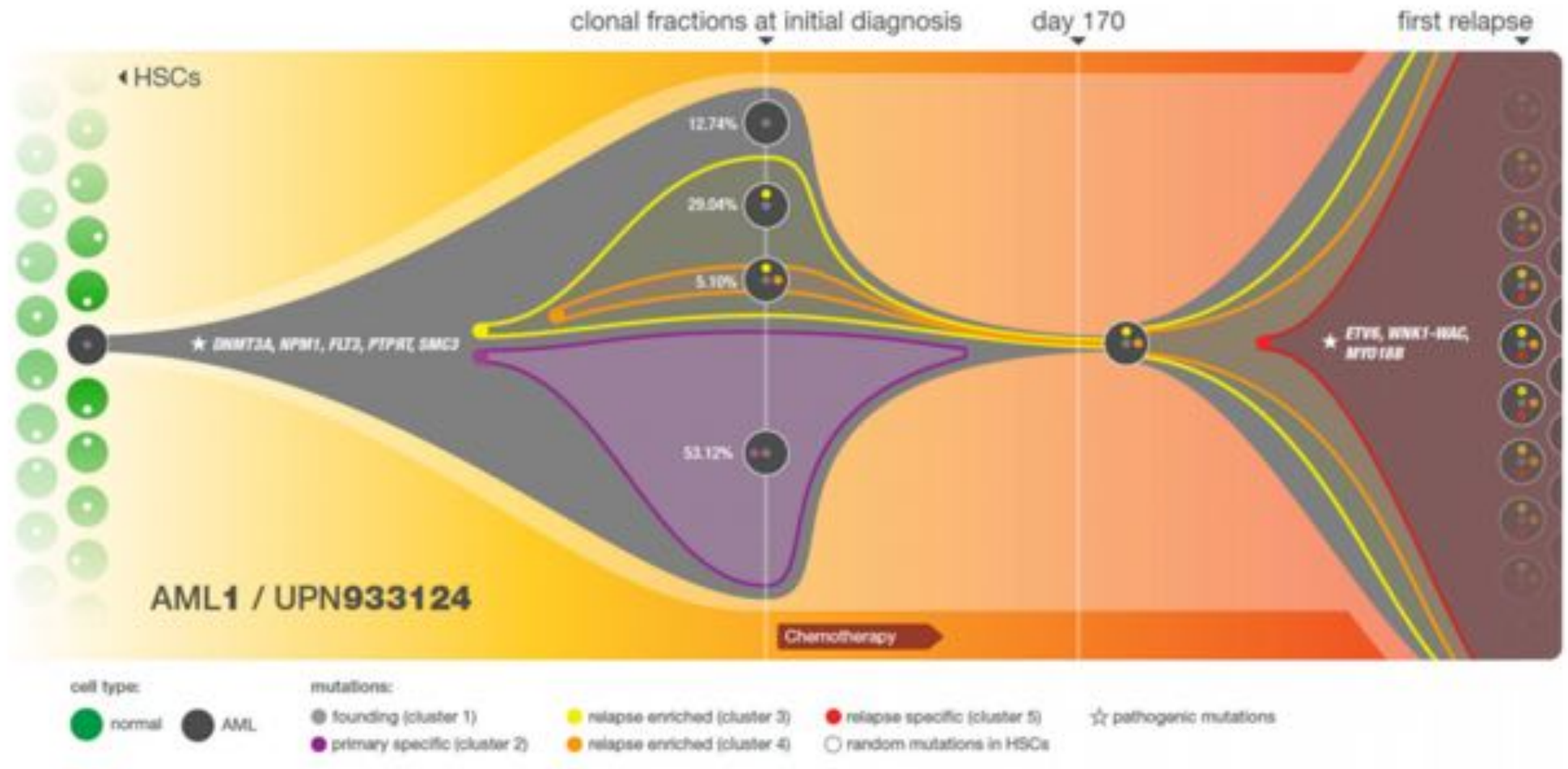
Mutations in Breast Cancer



Comprehensive molecular portraits of human breast tumours

Cancer Genome Atlas Network (2012) Nature. doi:10.1038/nature11412

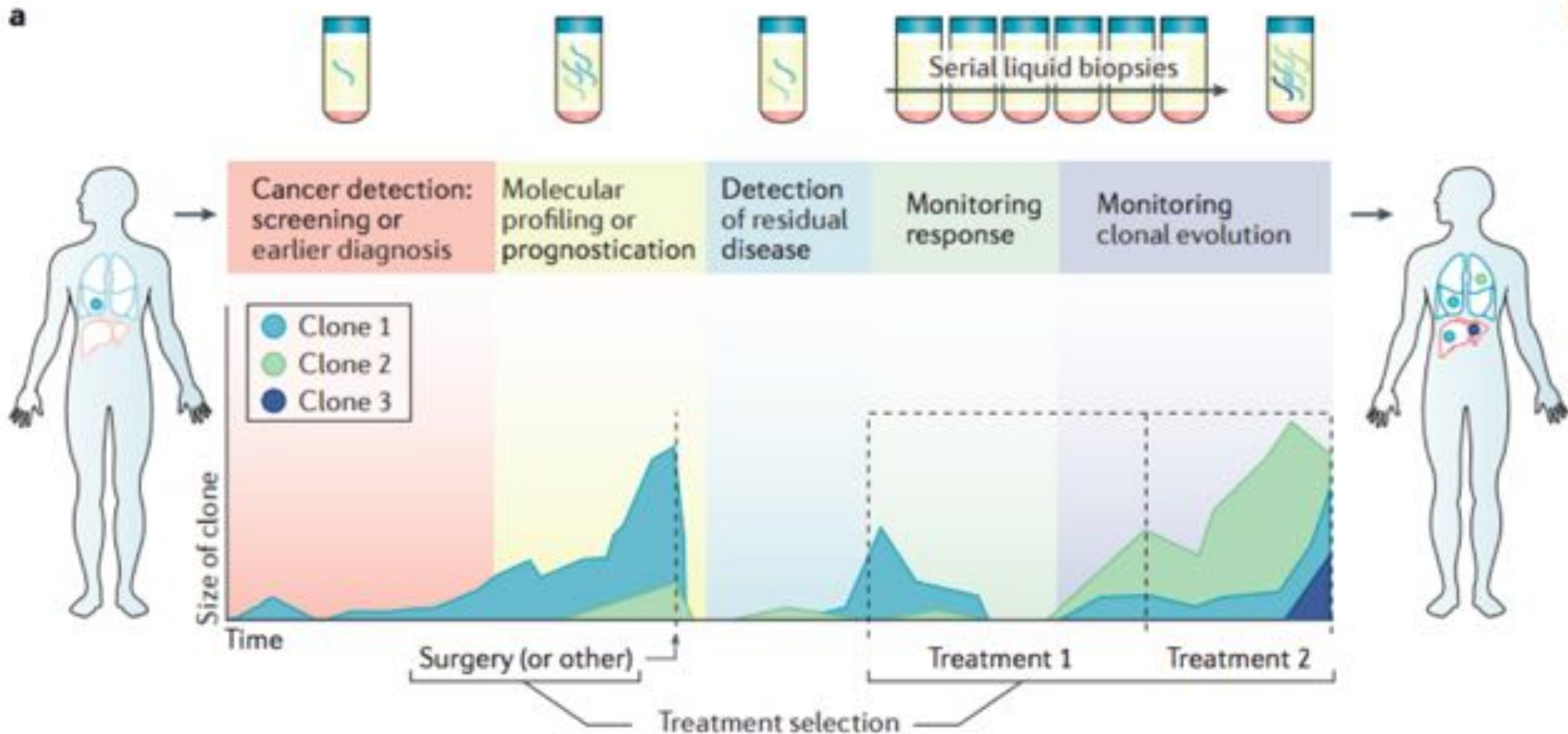
Tumor Heterogeneity and Treatment



Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing

Ding et al (2012) Nature. doi:10.1038/nature10738

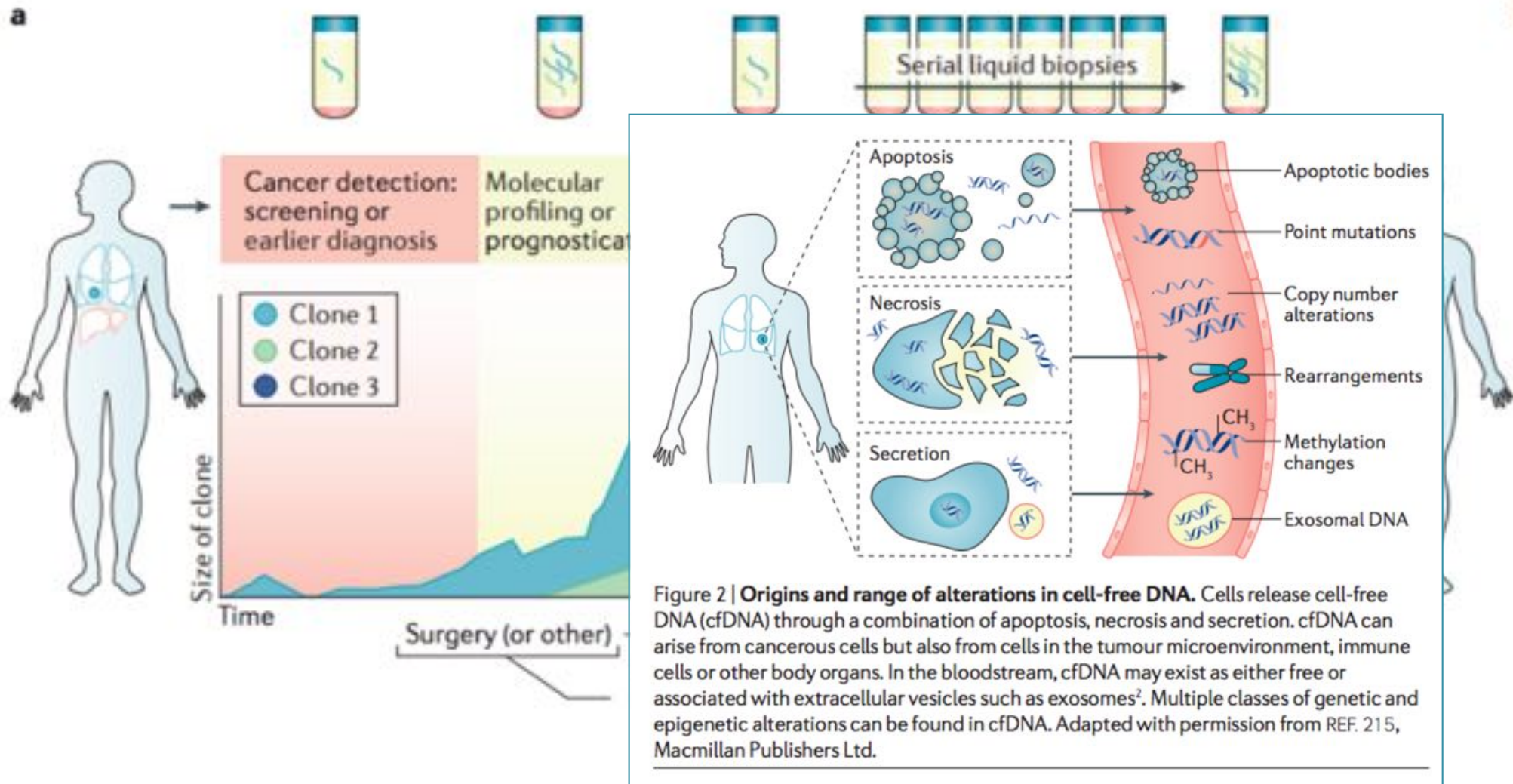
Liquid Biopsies



Liquid biopsies come of age: towards implementation of circulating tumour DNA

Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7

Liquid Biopsies



Liquid biopsies come of age: towards implementation of circulating tumour DNA

Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7

Your second genome?



***Human body:
~10 trillion cells***

***Microbiome
~100 trillion cells***

***Human brain:
~3.3 lbs***

***Total mass:
~3.3 lbs***

Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans

Sender et al (2016) Cell. <http://doi.org/10.1016/j.cell.2016.01.013>

Microbes and Human Health

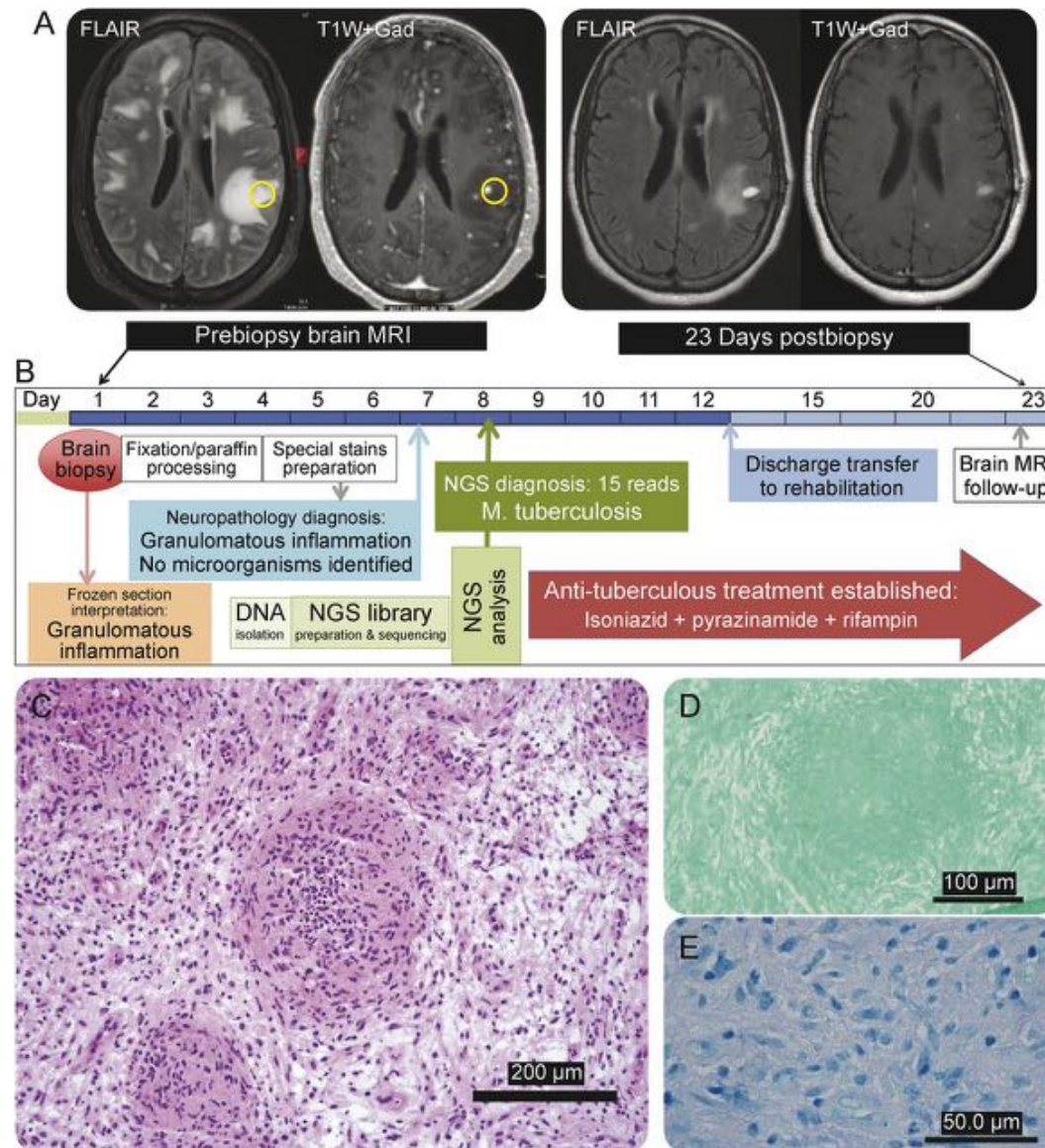


“MICROBE DIET Mice fed microbes from obese people tend to gain fat. Microbes from lean people protect mice from excessive weight gain, even when animals eat a high-fat, low-fiber diet.”

Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice

Ridaura et al (2013) Science. doi: 10.1126/science.1241214

Diagnosing Brain Infections with NGS



Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system
Salzberg et al (2016) Neurol Neuroimmunol Neuroinflamm dx/doi.org/10.1212/NXI.0000000000000251



Identifying Personal Genomes by Surname Inference

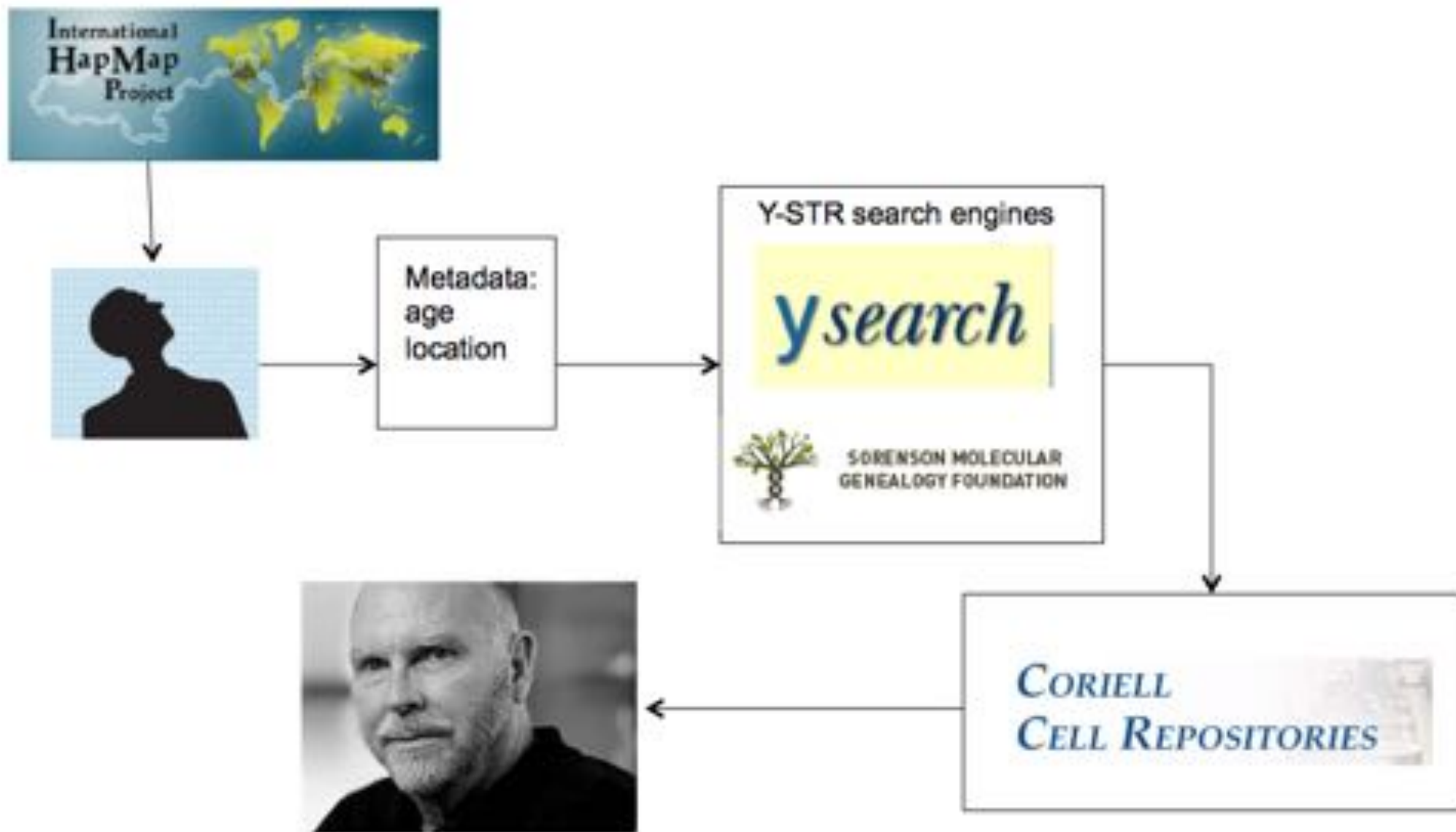
Melissa Gymrek *et al.*

Science **339**, 321 (2013);

DOI: 10.1126/science.1229566



Surname Inference Overview



The risks of big data?

Predicting Social Security numbers from public data

Alessandro Acquisti¹ and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master

File and the widespread accessibility of personal data from multiple sources, such as data brokers or professional working sites. Our results highlight the unexpected sequences of the complex interactions among data sources in modern information economies and the risks associated with information revelation in

identity theft | online social networks | privacy | statistics

In modern information economies, sensitive personal data are in plain sight amid transactions that rely on their unhindered circulation. Such is the case with Social Security numbers in the United States: Created as identifiers for tracking individual earnings (1), they have turned into authentication devices (2), becoming one of the most often sought by identity thieves. The Social Security Administration (SSA), which issues them, has kept SSNs confidential (3), coordinating with the Federal Reserve to keep their public exposure (4).^{*} After embarrassing data breaches, sector entities also have attempted to strengthen their consumers' and employees' data (7).[†] However, we have already left the barn: We demonstrate that

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within

publish on social networking sites (10). Using this method, we identified with a single attempt the first 5 digits for 44% of DMF records of deceased individuals born in the U.S. from 1989 to 2003 and the complete SSNs with <1,000 attempts (making SSNs akin to 3-digit financial PINs) for 8.5% of those records. Extrapolating to the U.S. living population, this would imply the potential identification of millions of SSNs for individuals whose birth data were available. Such findings highlight the hidden privacy costs of widespread information dissemination and the complex interactions among multiple data sources in modern information economies (11), underscoring the role of public records as breeder documents (12) of more sensitive data.

Keywords:

SEE COMMENTARY

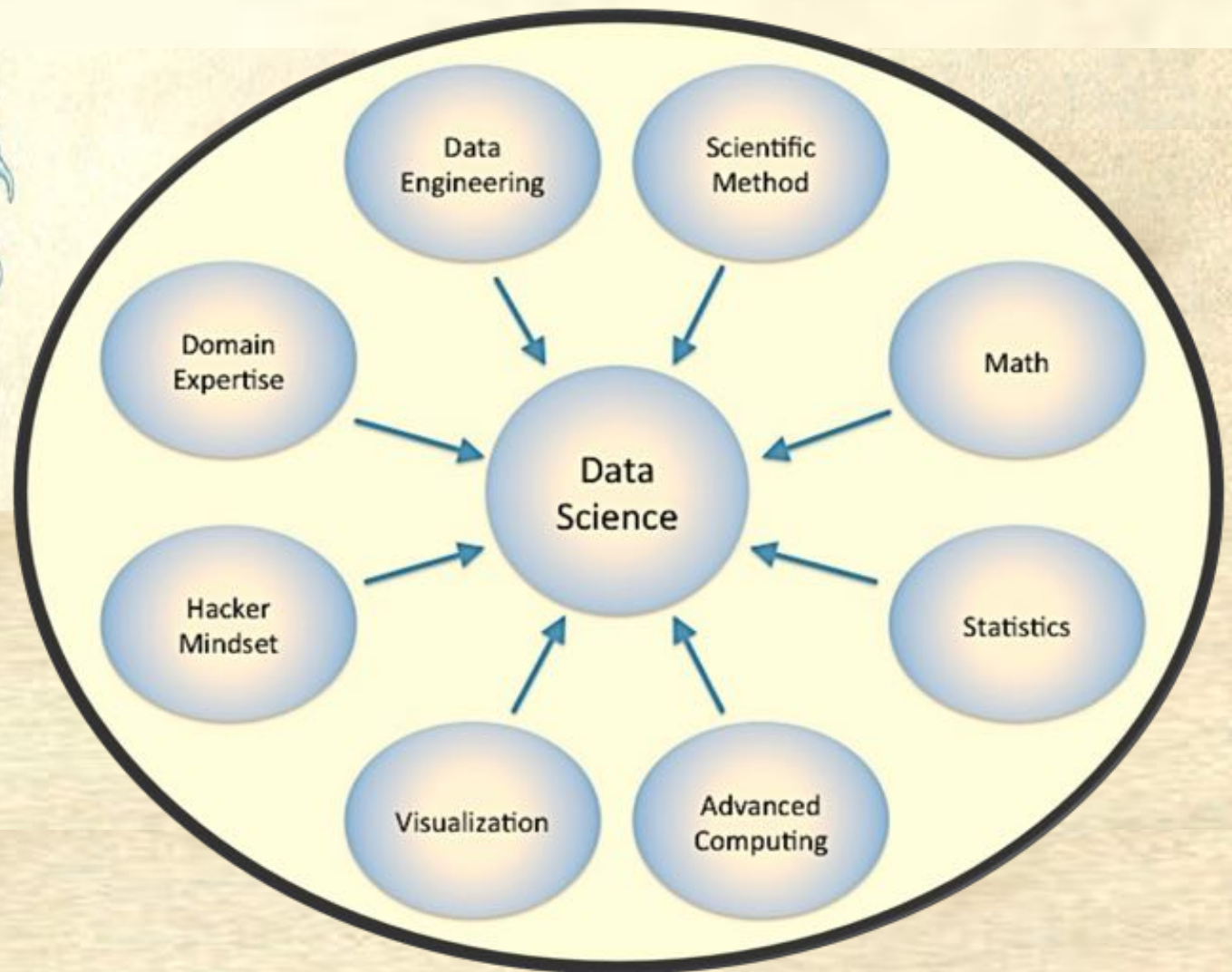
Genomic Futures?



The rise of a digital immune system

Schatz & Phillippy (2012) GigaScience 1:4

Who is a Data Scientist?



http://en.wikipedia.org/wiki/Data_science

Topics for next time?

- 01. Intro
- 02. Programming And Stats
- 03. Alan Yuille
- 04. Intro to Programming
- 05. Exponential Distribution
- 06. TJ Ha
- 07. Human Genome
- 08. Ulrich Mueller
- 09. Rachel Green
- 10. Steven Salzberg
- 11. Alexis Battle
- 12. Genome Sequencing

- 13. Assembly Algorithms
- 14. Carl Wu
- 15. Variant Calling
- 16. Human Evolution
- 17. Human Disease
- 18. Grad Student Panel
- 19. Jennifer Doudna
- 20. Ben Langmead
- 21. Andy Feinberg
- 22. Rong Li
- 23. Wrap Up
- 24-26. Presentations