Midterm Review

Michael Schatz

October 30, 2024

Applied Comparative Genomics

Class Schedule

~~M~~	~~Oct 14~~	~~Epigenome~~	~~Project Proposal Assigned~~
~~W~~	~~Oct 16~~	~~Single cell~~	
~~M~~	~~Oct 21~~	~~Transformers~~	~~Assignment 5 Assigned~~
~~W~~	~~Oct 23~~	~~Enformer~~	
~~M~~	~~Oct 28~~	~~DL in Genomics~~	~~Preliminary Report Assigned~~
W	Oct 30	Midterm Review	
M	Nov 4	Midterm!	
W	Nov 6	Disease Genomics	
M	Nov 11	Metagenomics	Final Report Assigned
W	Nov 13	No Class (BIODATA24)	
M	Nov 18	Cancer Genomics	
W	Nov 20	Project Presentation 1	
M	Nov 25	Thanksgiving Break	
W	Nov 27	Thanksgiving Break	
M	Dec 2	Project Presentation 2	
W	Dec 4	Project Presentation 3	
M	Dec 16	Project Report Due	

Assignment 5

Assignment 5: Convlutional Neural Networks

Assignment Date: Monday, October 21, 2024
Due Date: Monday, October 28 @ 11:59pm

Assignment Overview

In this assignment you will explore a couple of key aspects of convolutional neural networks such as self-attention and feature encoding as well as explore a pre-trained convolutional neural network for gene expression prediction. For this assignment, we will provide a Jupyter notebook with code for you to use and complete your assignment in.

For this assignment, you will create the environment as follows:

```
mamba create -n asn5 python=3.10 scikit-learn pytorch matplotlib pandas numpy jupyter seaborn kipoiseq logomaker
```

Then activate the environment and install the following package using pip:

```
mamba activate asn5
```

```
pip install enformer-pytorch
```

If you have issues with creating the environment and installing the required packages, you can use Google Colab

You will need to add the following cells to the top of the notebook in Google Colab to install the dependencies:

```
!pip install kipoiseq==0.5.2
!pip install logomaker
!pip install enformer-pytorch
```

Because of the way Google Colab works, you will need to install these everytime you reopen the notebook.

As a reminder, any questions about the assignment should be posted to Piazza.

See the notebook here: Assignment5.ipynb

Packaging

The solutions to the above questions should be submitted as a single PDF document that includes your name, email address, and all relevant code, text, and figures (as needed). If you use ChatGPT for any of the code, also record the prompts used. Submit your solutions by uploading the PDF to GradeScope, and remember to select where in your submission each question/subquestion is. The Entry Code is: Z3J8YV.

If you submit after this time, you will use your late days. Remember, you are only allowed 4 late days for the entire semester!

Resources

- Jupyter notebooks: https://jupyter.org/
- scikit-learn: https://scikit-learn.org/stable/
- pytorch: https://pytorch.org/
- pytorch tutorial: https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html

Preliminary Report
Due Monday November 11

Preliminary Project Report

Assignment Date: October 28, 2024
Due Date: Monday, November 11, 2024 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Monday November 11

The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result (typically a summary of the data you have identified for your project)
- 5+ References to relevant papers and data

The preliminary report must use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. Overleaf is recommended for LaTex submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

Later, you will present your project in class starting the week of November 25. You will also submit your final written report (6-8 pages) of your project by Dec 16

Please use Piazza if you have any questions!

Presentations!

Project Presentations

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 all groups can present in class!

Schedule of Presentations

Slot	Date	Start	Team Name	Team Members	Project Title
1	11/20	3:00	Two single-cells, one big problem	Kevin Meza Landeros, YunZhou Liu	Cluster-based single-cell RNA-seq variant detection
2	11/20	3:12	Team Yuxiang Li	Yuxiang Li	Contrastive Learning Approach to Integrate Single-Cell scRNA-seq and scATAC-seq for Mechanistic Understanding of Gene Regulation
3	11/20	3:24	Team Roujin An	Roujin An	Cell Type-Specific SNP-to-Splicing Variants Mapping Using Deep Learning Models
4	11/20	3:36	Team Miller	Logan Miller	Population-Specific Evolutionary Hotspots in Human Genomes
5	11/20	3:48	Team1D	Ben Miller	Comparative Genomic Analysis of NOD and (Simulated) NOR Mouse Genomes to Identify Variants Associated with Type 1 Diabetes
1	12/2	3:00	Genomic Visionaries	Iason Mihalopoulos, Siam Mohammed	AR/VR Visualization of Individual Genomes with AI-driven Insights
2	12/2	3:12	Silent Codebreakers	Cecelia Zhang, Jiarui Yang	Benchmarking Non-Coding Mutation Analysis Schemes on Cancer Genomes
3	12/2	3:24	Team Table	Oce Bohra, Zoe Rudnick	The emerging contribution of non-coding mutations in glioblastoma development
4	12/2	3:36	Team Brady	Brady Bock	DNN analysis of gut microbiomes to predict colorectal cancer disease state
5	12/2	3:48	Variant Visionaries	Alexandra Gorham, Christine Park, Natalie Vallejo	Benchmarking Non-coding Variant Scoring Tools for Cancer Pathogenicity Prediction
6	12/2	4:00	Human to Plants	Xiaojun Gao, Yujia, Yushan Zou	Evaluation of applicability of ChromHMM for Plants in Chromatin States and Gene Expression
1	12/4	3:00	SE Palmeiras	Caleb Hallinan, Jamie Moore, Rafael dos Santos Peixoto	Evaluating cell-type clustering algorithm's robustness to technical artifacts via synthetic spatial transcriptomics data
2	12/4	3:12	Nuclencoder	Amanda Xu, Angela Yang, Jiamin Li	DNA Cryptography: Digital Signatures for Encryption to Facilitate Safe Data Storage
3	12/4	3:24	Geoguessr	Alex Ostrovsky, Nicole Lauren Brown	Investigating geographic and environmental effects on soil metagenomes by correlating GIS data
4	12/4	3:36	Quetzalli Tlalli	Arshana Welivita, Atticus Colwell	Benchmarking Methods for Inferring the Ethnicity of an Individual from Their Genotype
5	12/4	3:48	Team Barbour	Alexis Barbour	Benchmarking non-coding mutation analysis schemes for evaulating Type 1 Diabetes
6	12/4	4:00	All of Us Team	Levon Galstyan, Nitish Aswani, Talia Haller	Genomic Insights into Sleep Patterns, Disease Outcomes, and Biomarker Associations using the All of Us Dataset

Let me know ASAP on Piazza if you have a *major* conflict

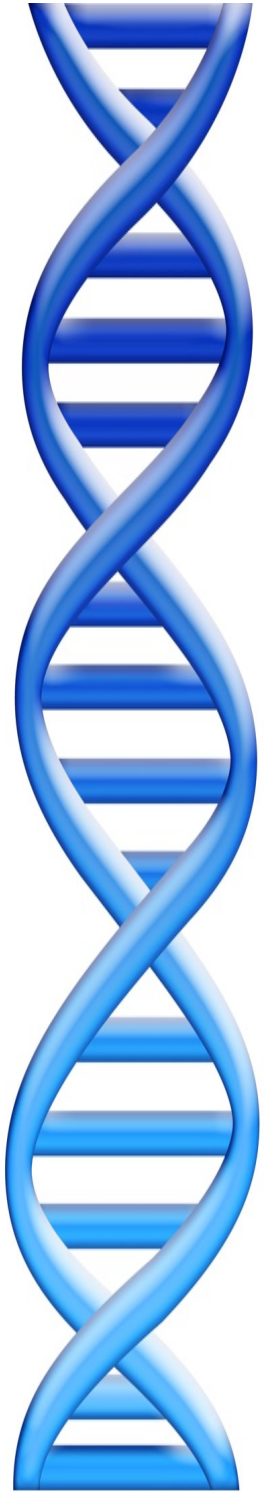
https://github.com/schatzlab/appliedgenomics2024/blob/main/project/presentations.md

Presentations!
10 min + 2 min questions

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

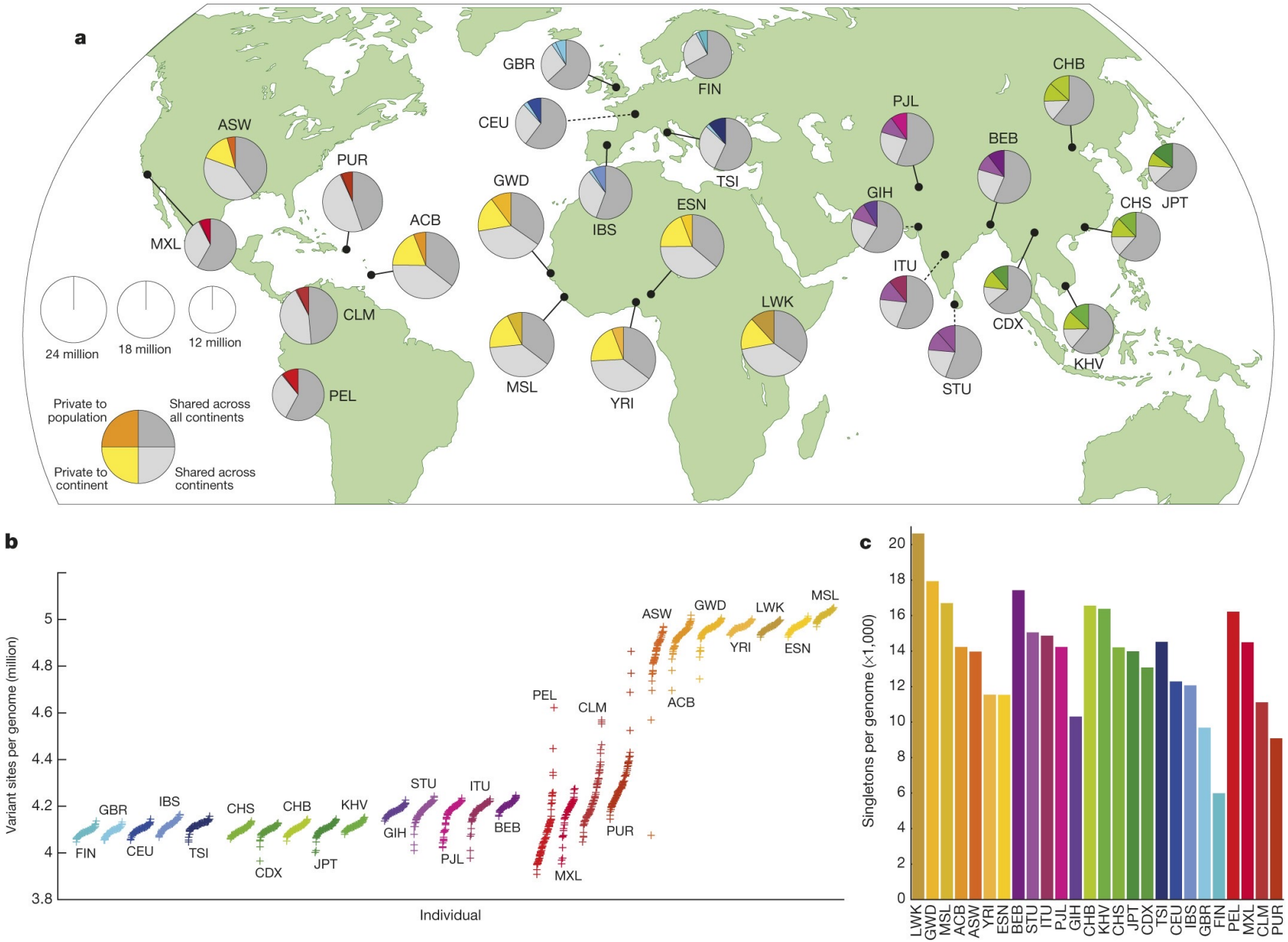
1. Title Slide: Who are you, title, date
2. Intro 1: Whats the big idea???
3. Intro 2: More specifically, what are you trying to learn?
4. Methods 1: What did you try?
5. Methods 2: What is the key idea?
6. Data 1: What data are you looking at?
7. Data 2: Anything notable about the data?
8. Results 1: What did you see!
9. Results 2: How does it compare to other methods/data/ideas?
10. Discussion 1: What did you learn from this study?
11. Discussion 2: What does this mean for the future?
12. Acknowledgements: Who helped you along the way?
13. Thank you!

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.

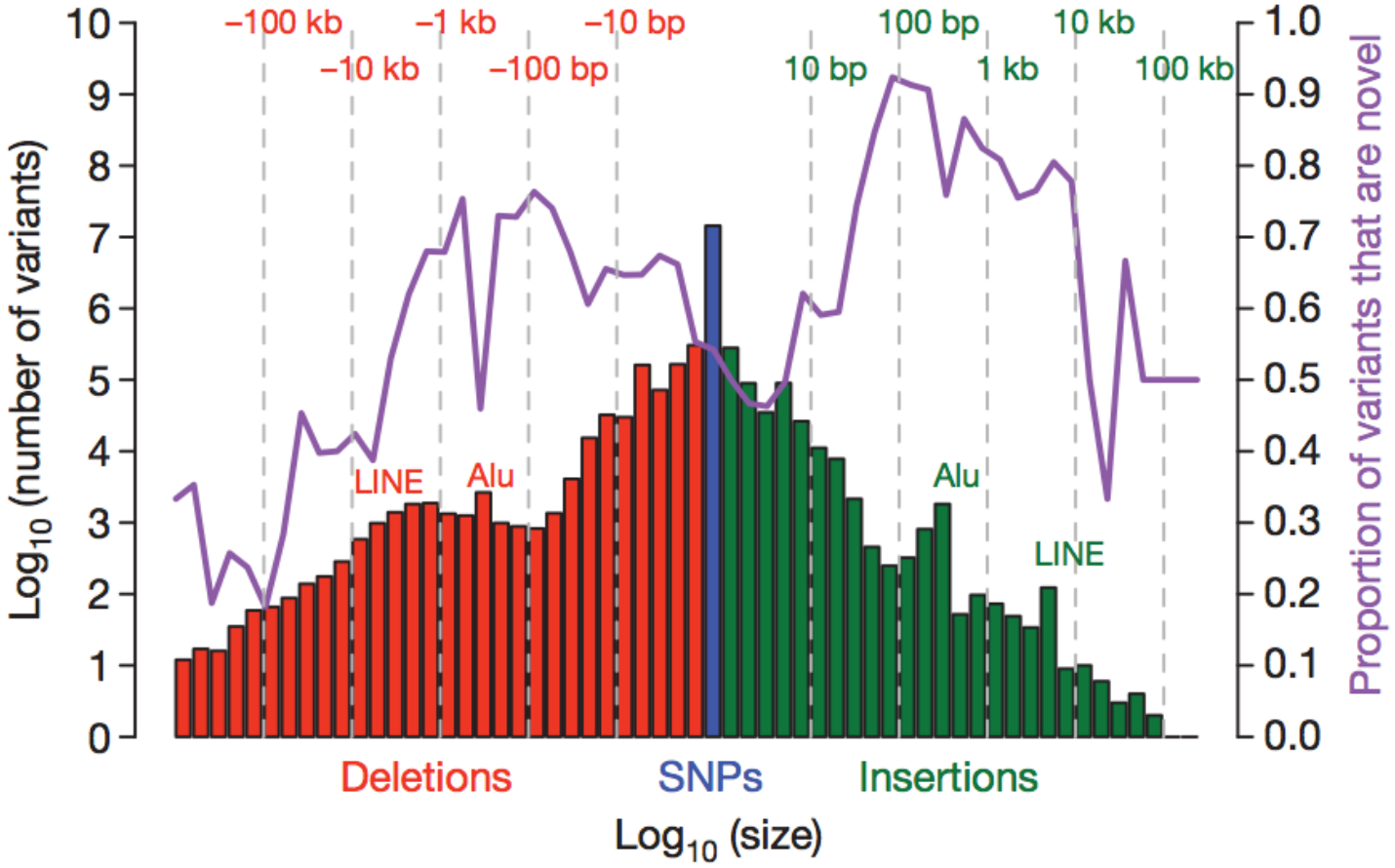


Recap

1000 Genomes Populations



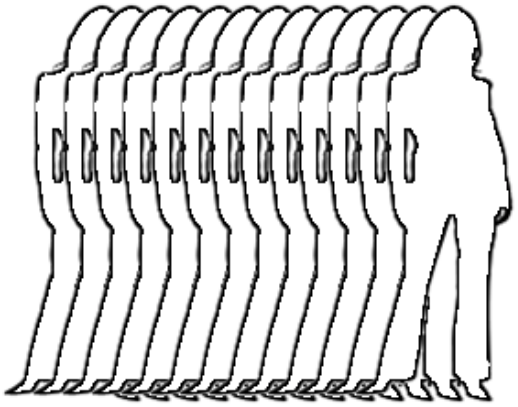
Human Mutation Types



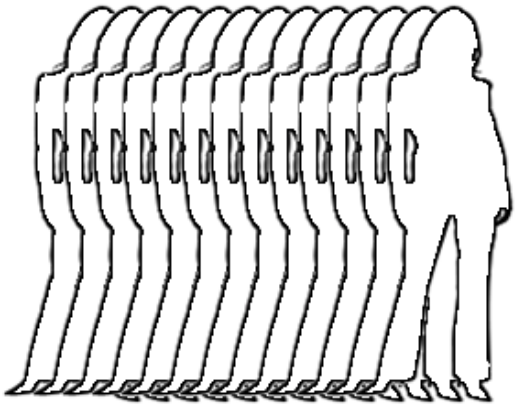
- Mutations follows a "log-normal" frequency distribution
  - Most mutations are SNPs followed by small indels followed by larger events

A map of human genome variation from population-scale sequencing
1000 genomes project (2010) *Nature*. doi:10.1038/nature09534

Genome Wide Association (GWAS)



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



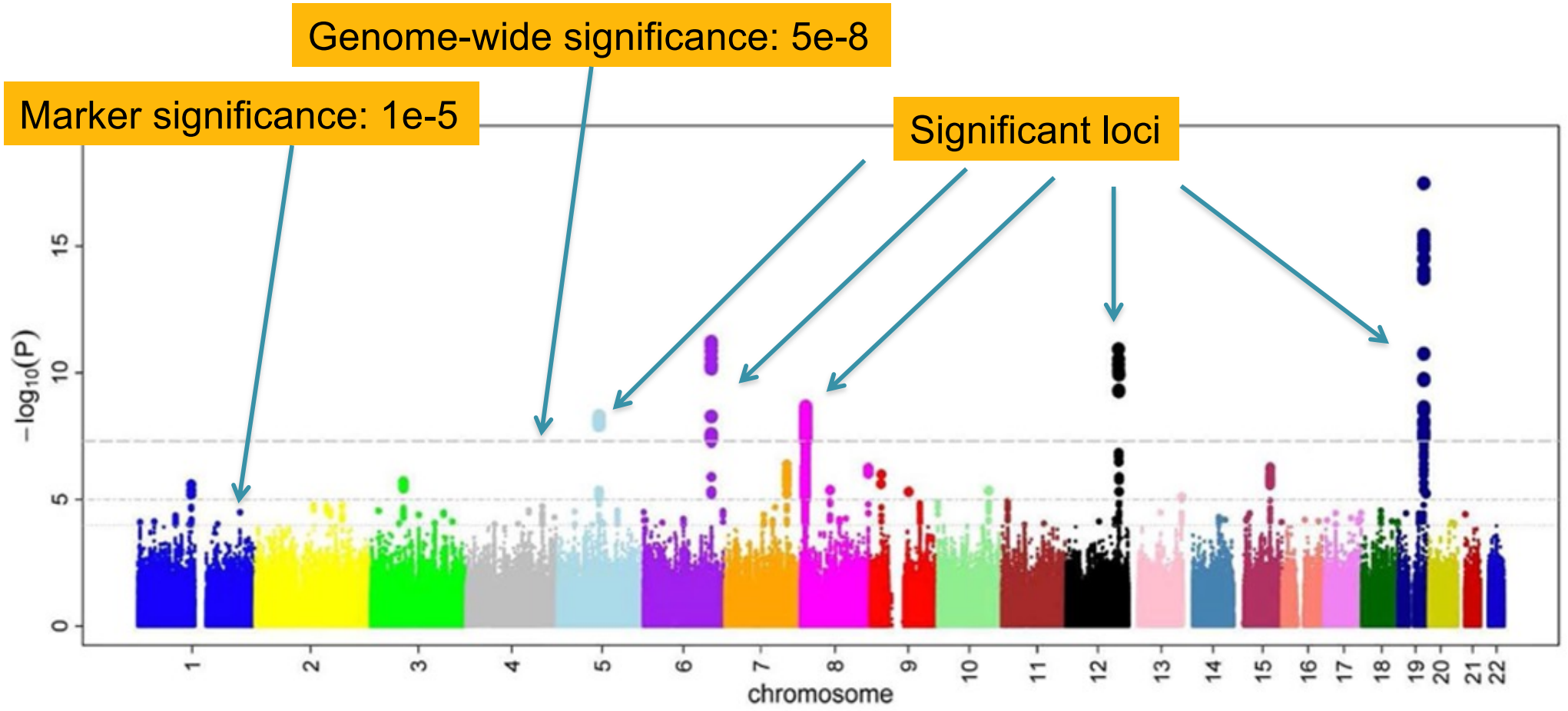
GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1	*SNP2*	*SNP …*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

With a (much) larger population, this might be a significant difference in rate:
25320/60000 =>
p = 5e-7

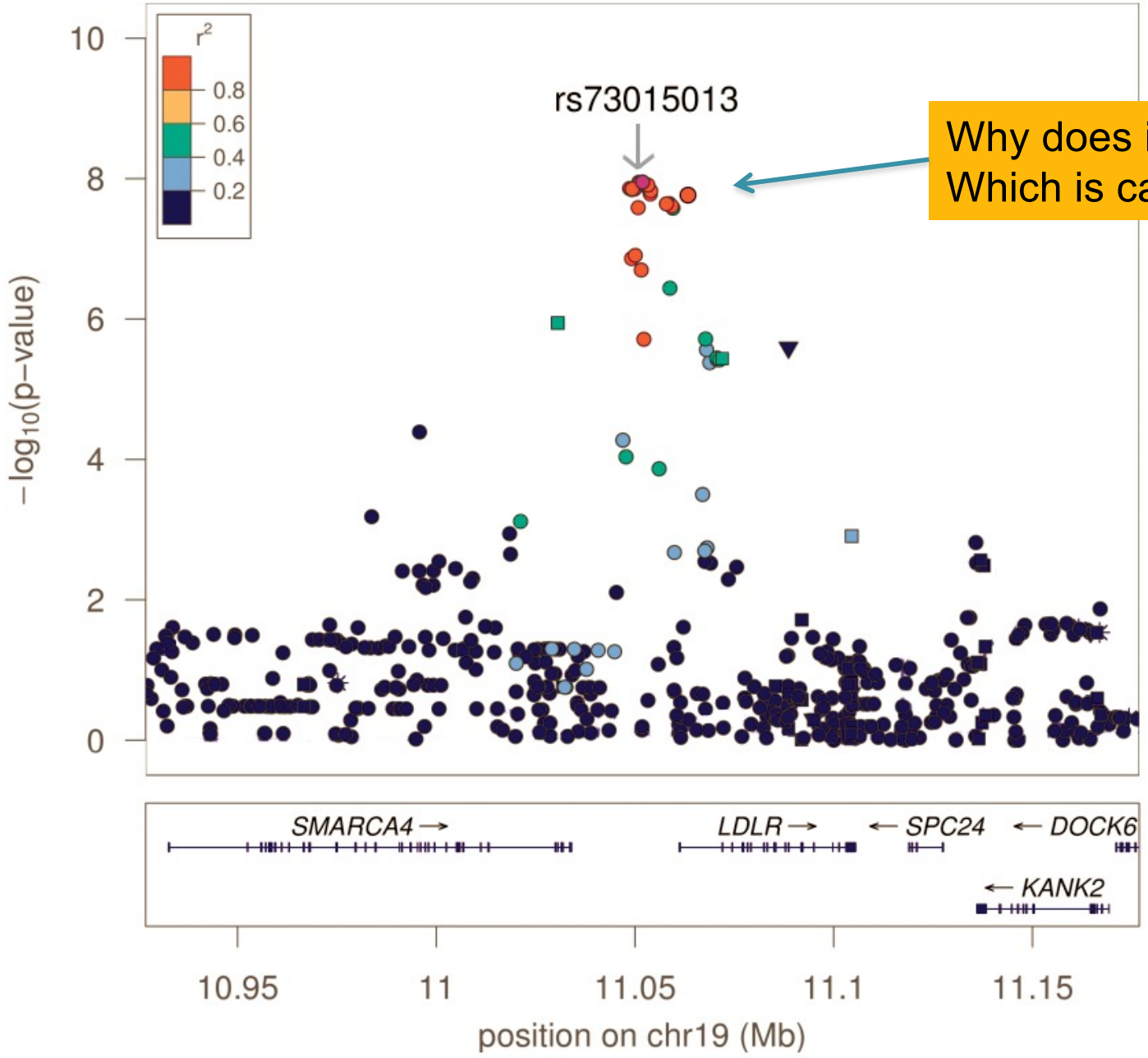
Chi-squared or similar test

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

Regional Association Plot



Why does it look like this?
Which is causal?

Genetic Basis of Autism Spectrum Disorders

Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

U.S. CDC identify around 1 in 68 American children as on the autism spectrum

- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

What is Autism?
https://autisticadvocacy.org/

Autism is NOT caused by vaccines

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associated by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities ranging from lymphoid nodular hyperplasia to aphthoid ulceration. Histology showed patchy chronic inflammation in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls ($p=0.003$), low haemoglobin in four children, and a low serum IgA in four children.

Interpretation We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.

Lancet 1998; **351:** 637–41
See Commentary page

Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology (A J Wakefield FRCS, A Anthony MB, J Linnell PhD, A P Dhillon MRCPath, S E Davies MRCPath) **and the University Departments of Paediatric Gastroenterology** (S H Murch MB, D M Casson MRCP, M Malik MRCP, M A Thomson FRCP, J A Walker-Smith FRCP,), **Child and Adolescent Psychiatry** (M Berelowitz FRCPsych), **Neurology** (P Harvey FRCP), **and Radiology** (A Valentine FRCR), **Royal Free Hospital and School of Medicine, London NW3 2QG, UK**

Correspondence to: Dr A J Wakefield

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for a week, accompanied by their parents.

Clinical investigations

We took histories including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S). Neurological and psychiatric assessments were done by consultant staff (PH, MB) with HMS-4 criteria.[1] Developmental assessments included a review of prospective developmental records from parents, health visitors, and general practitioners. Four children did not undergo psychiatric assessment in hospital; all had been assessed professionally elsewhere, so these assessments were used as the basis for their behavioural diagnosis.

After bowel preparation, ileocolonoscopy was performed by SHM or MAT under sedation with midazolam and pethidine. Paired frozen and formalin-fixed mucosal biopsy samples were taken from the terminal ileum; ascending, transverse, descending, and sigmoid colons, and from the rectum. The procedure was recorded by video or still images, and were compared with images of the previous seven consecutive paediatric colonoscopies (four normal colonoscopies and three on children with ulcerative colitis), in which the physician reported normal appearances in the terminal ileum. Barium follow-through radiography was possible in some cases.

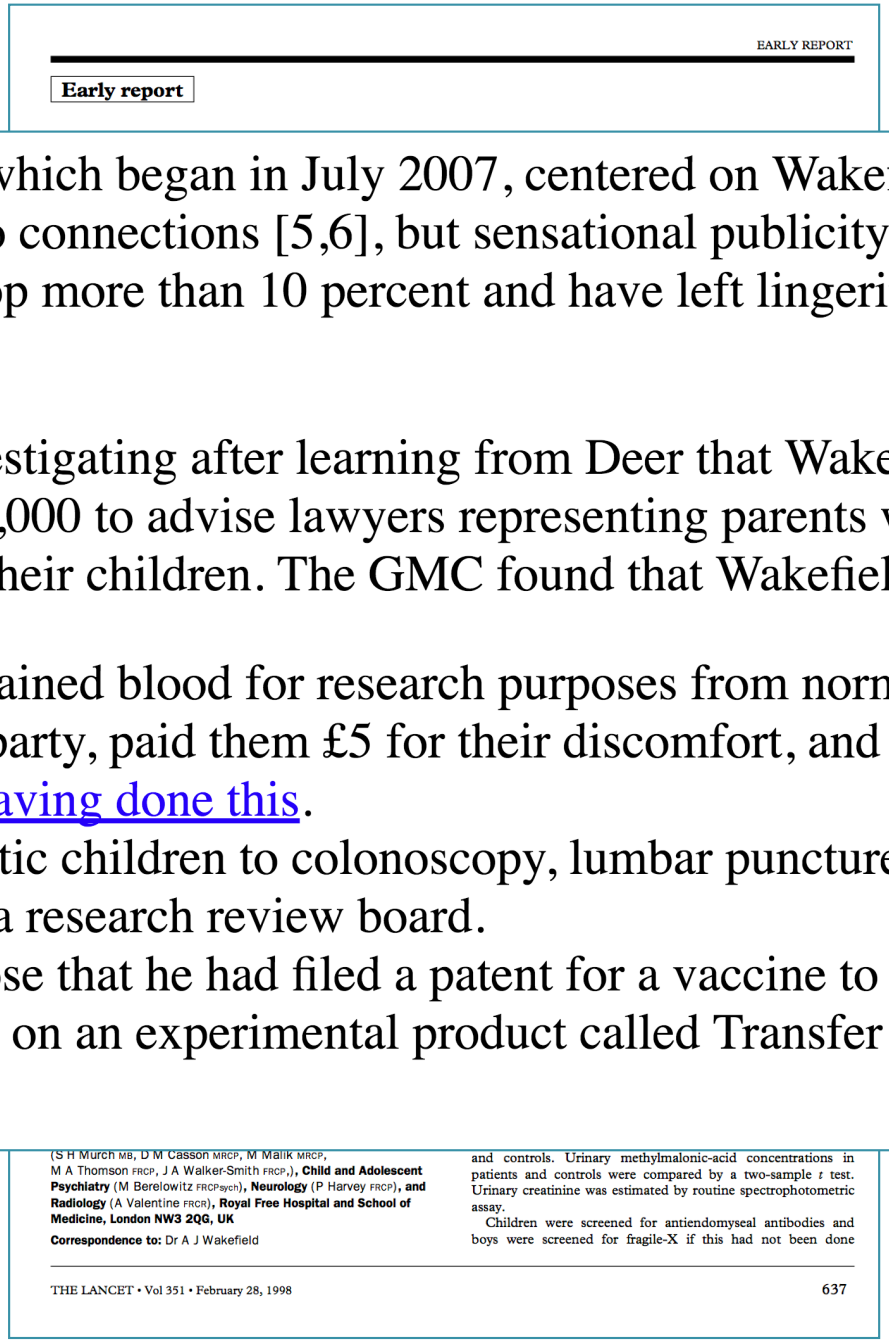
Also under sedation, cerebral magnetic-resonance imaging (MRI), electroencephalography (EEG) including visual, brain stem auditory, and sensory evoked potentials (where compliance made these possible), and lumbar puncture were done.

Laboratory investigations

Thyroid function, serum long-chain fatty acids, and cerebrospinal-fluid lactate were measured to exclude known causes of childhood neurodegenerative disease. Urinary methylmalonic acid was measured in random urine samples from eight of the 12 children and 14 age-matched and sex-matched normal controls, by a modification of a technique described previously.[2] Chromatograms were scanned digitally on computer, to analyse the methylmalonic-acid zones from cases and controls. Urinary methylmalonic-acid concentrations in patients and controls were compared by a two-sample t test. Urinary creatinine was estimated by routine spectrophotometric assay.

Children were screened for antiendomyseal antibodies and boys were screened for fragile-X if this had not been done

Autism is NOT caused by vaccines



EARLY REPORT

Early report

(S H Murch MB, D M Casson MRCP, M Malik MRCP, M A Thomson FRCP, J A Walker-Smith FRCP,), **Child and Adolescent Psychiatry** (M Berelowitz FRCPsych), **Neurology** (P Harvey FRCP), **and Radiology** (A Valentine FRCR), **Royal Free Hospital and School of Medicine, London NW3 2QG, UK**

Correspondence to: Dr A J Wakefield

and controls. Urinary methylmalonic-acid concentrations in patients and controls were compared by a two-sample *t* test. Urinary creatinine was estimated by routine spectrophotometric assay.

Children were screened for antiendomyseal antibodies and boys were screened for fragile-X if this had not been done

THE LANCET • Vol 351 • February 28, 1998 637

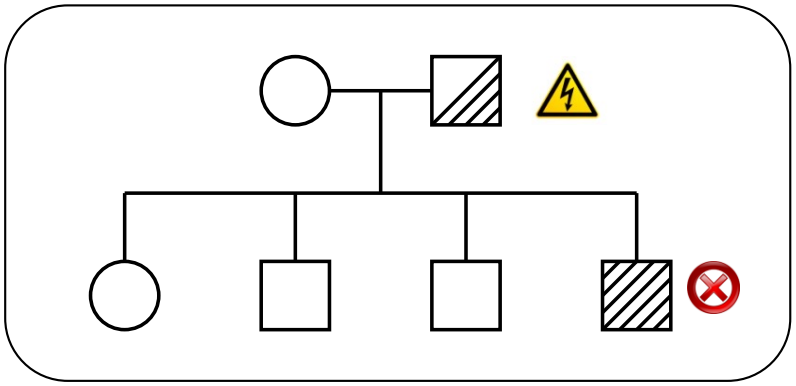
The GMC hearings, which began in July 2007, centered on Wakefield's 1998 report. Many studies have found no connections [5,6], but sensational publicity caused immunization rates in the UK to drop more than 10 percent and have left lingering doubts among parents worldwide.

The GMC began investigating after learning from Deer that Wakefield had failed to declare he had been paid £55,000 to advise lawyers representing parents who believed that the vaccine had harmed their children. The GMC found that Wakefield had:

- Improperly obtained blood for research purposes from normal children attending his son's birthday party, paid them £5 for their discomfort, and later joked during a lecture about having done this.
- Subjected autistic children to colonoscopy, lumbar punctures, and other tests without approval from a research review board.
- Failed to disclose that he had filed a patent for a vaccine to compete with the MMR
- Starting a child on an experimental product called Transfer Factor, which he planned to market.

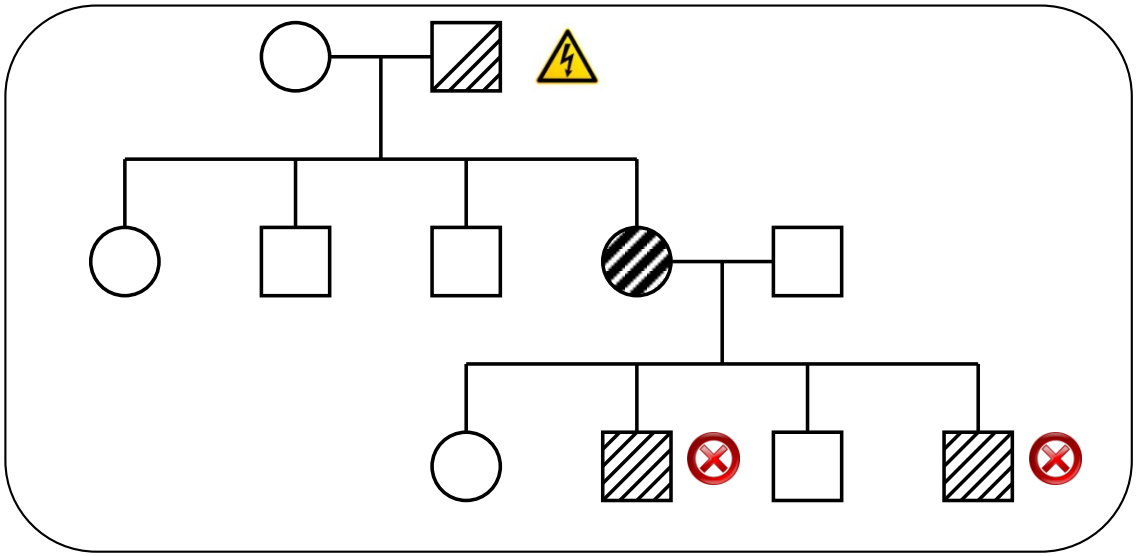
Unified Model of Autism

Sporadic Autism



De novo mutations of high penetrance contributes to autism, especially in low risk families with no history of autism.

Familial Autism



Legend


Sporadic mutation

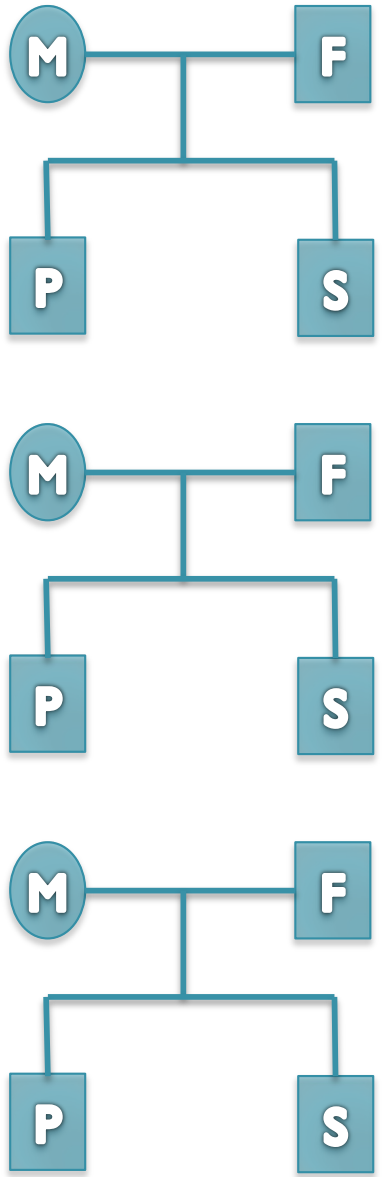

Fails to procreate

A unified genetic theory for sporadic and inherited autism
Zhao *et al.* (2007) *PNAS. 104(31)12831-12836.*

Searching for the genetic risk factors

Search Strategy

- Thousands of families identified from a dozen hospitals around the United States
- Large scale genome sequencing of "simplex" families: mother, father, affected child, unaffected sibling
- Unaffected siblings provide a natural control for environmental factors

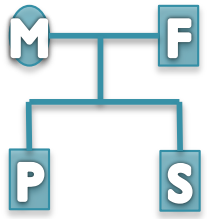


Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?

SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

De novo mutation discovery and validation

De novo mutations:
Sequences not inherited from your parents.




```
Reference:    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father(1):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Father(2):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Mother(1):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother(2):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Sibling(1):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling(2):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Proband(1):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(2):   ...TCAAATCCTTTTAAT****AAGAGCTGACA...
```

4bp heterozygous deletion at chr15:93524061 CHD2

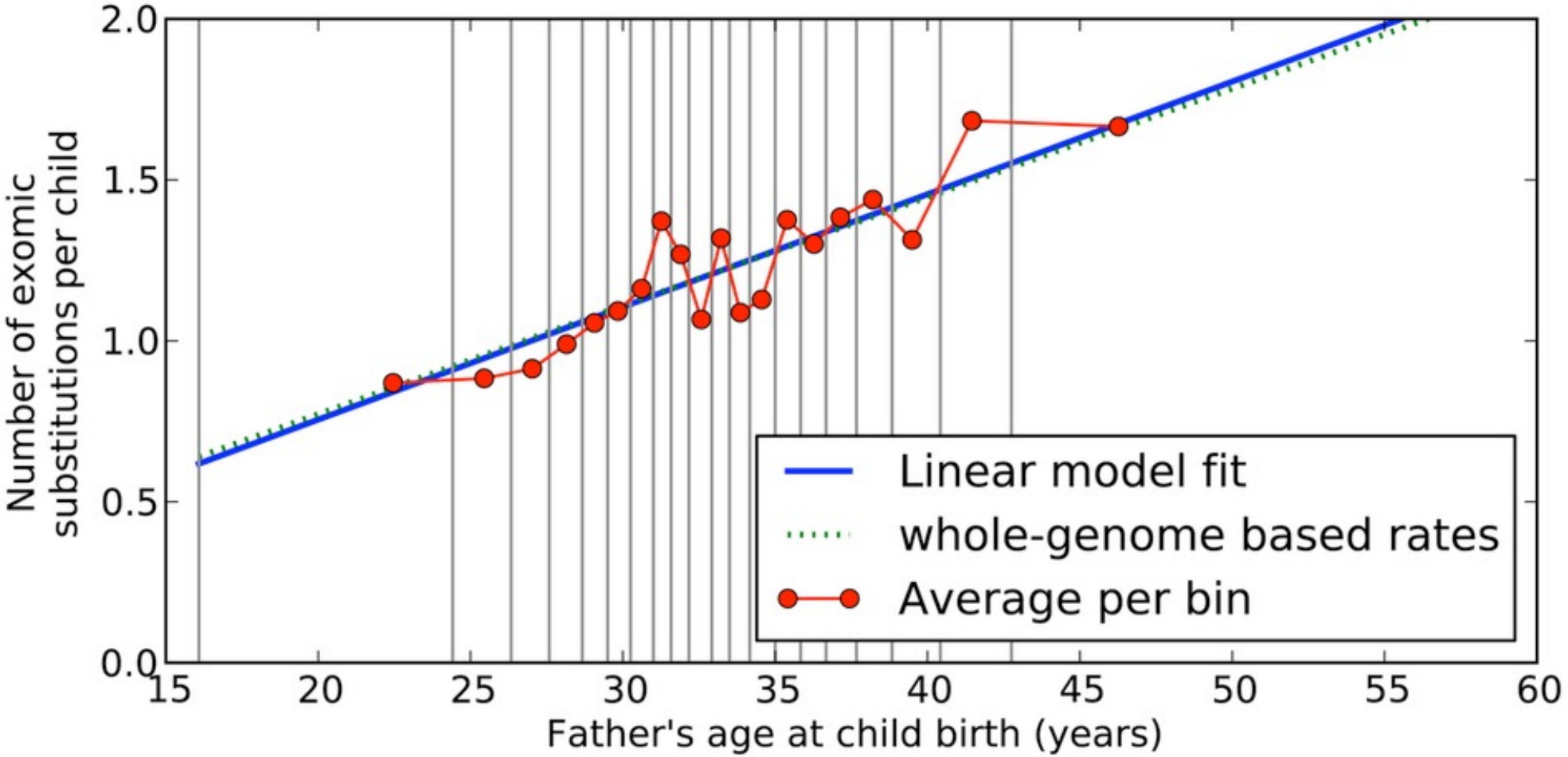
De novo Genetics of Autism

- In 593 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMPR
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

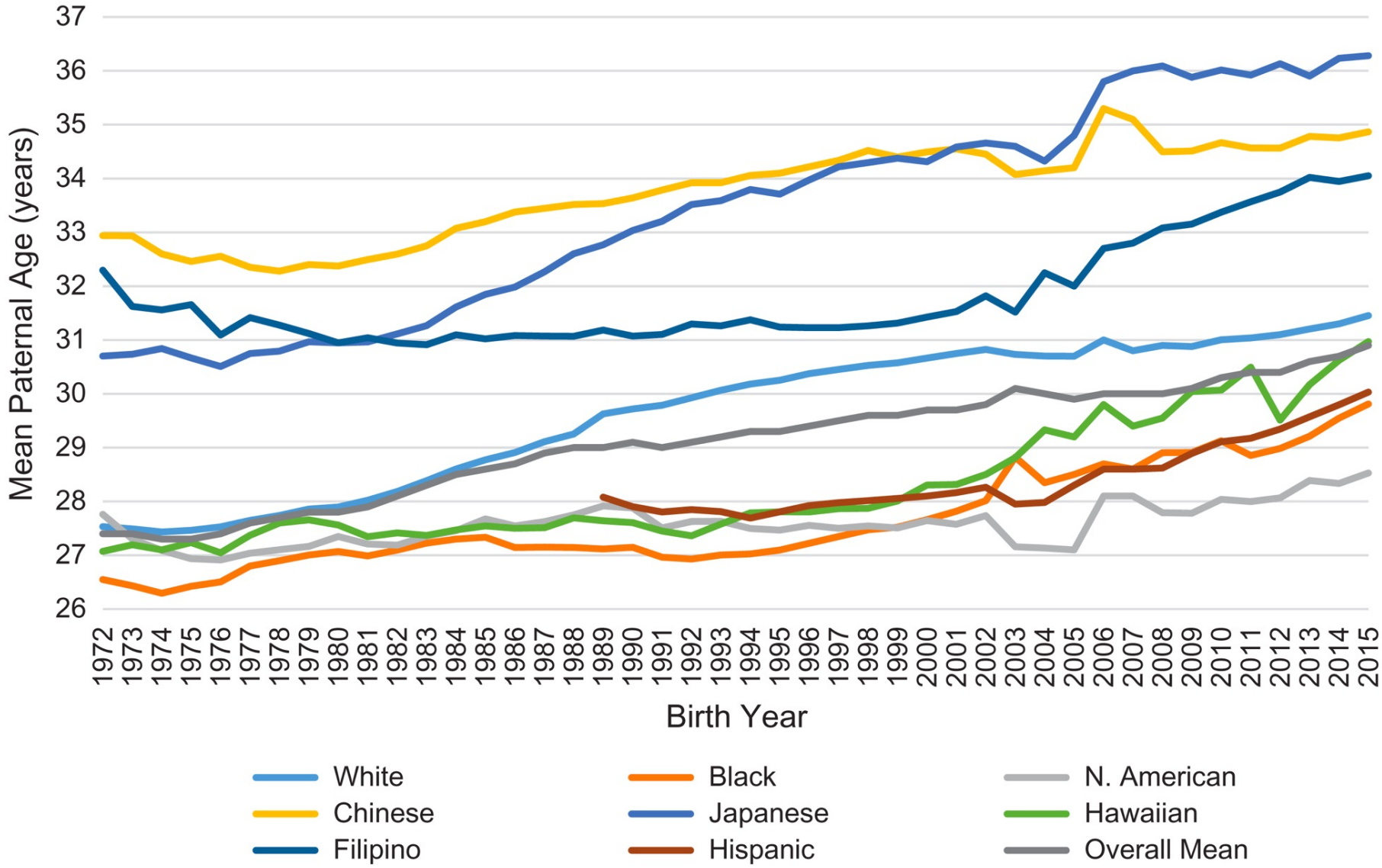
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi *et al* (2014) Nature Methods doi:10.1038/nmeth.3069

De novo Mutations in Men



The contribution of de novo coding mutations to autism spectrum disorder
Iossifov *et al* (2014) *Nature*. doi:10.1038/nature13908

Age of Fatherhood



The age of fathers in the USA is rising: an analysis of 168 867 480 births from 1972 to 2015
Khandwala *et al* (2017) *Human Reproduction. https://doi.org/10.1093/humrep/dex267*

Needles in stacks of needles

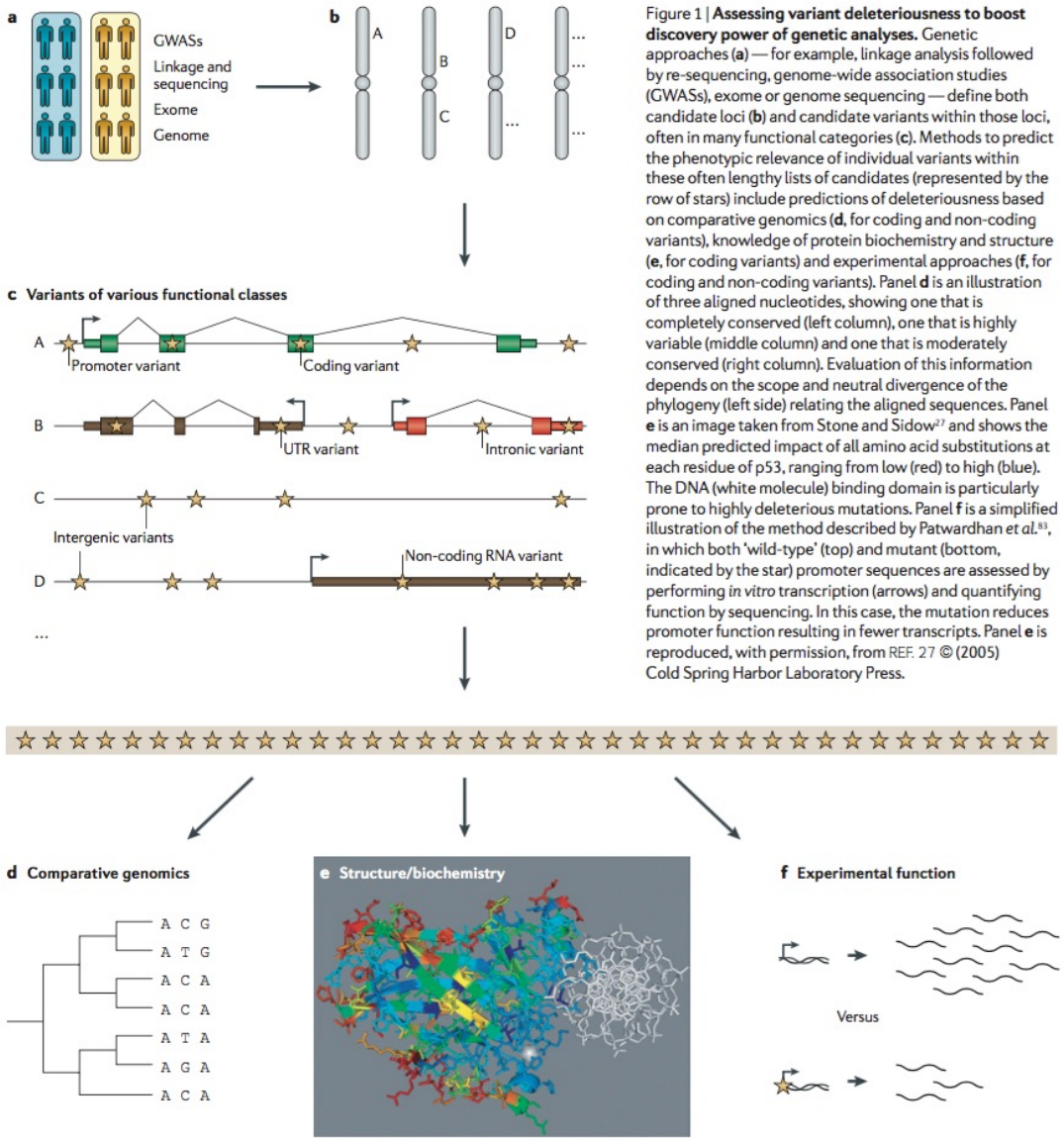


Figure 1 | **Assessing variant deleteriousness to boost discovery power of genetic analyses.** Genetic approaches (**a**) — for example, linkage analysis followed by re-sequencing, genome-wide association studies (GWASs), exome or genome sequencing — define both candidate loci (**b**) and candidate variants within those loci, often in many functional categories (**c**). Methods to predict the phenotypic relevance of individual variants within these often lengthy lists of candidates (represented by the row of stars) include predictions of deleteriousness based on comparative genomics (**d**, for coding and non-coding variants), knowledge of protein biochemistry and structure (**e**, for coding variants) and experimental approaches (**f**, for coding and non-coding variants). Panel **d** is an illustration of three aligned nucleotides, showing one that is completely conserved (left column), one that is highly variable (middle column) and one that is moderately conserved (right column). Evaluation of this information depends on the scope and neutral divergence of the phylogeny (left side) relating the aligned sequences. Panel **e** is an image taken from Stone and Sidow[27] and shows the median predicted impact of all amino acid substitutions at each residue of p53, ranging from low (red) to high (blue). The DNA (white molecule) binding domain is particularly prone to highly deleterious mutations. Panel **f** is a simplified illustration of the method described by Patwardhan *et al.*[83], in which both 'wild-type' (top) and mutant (bottom, indicated by the star) promoter sequences are assessed by performing *in vitro* transcription (arrows) and quantifying function by sequencing. In this case, the mutation reduces promoter function resulting in fewer transcripts. Panel **e** is reproduced, with permission, from REF. 27 © (2005) Cold Spring Harbor Laboratory Press.

Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data
Cooper & Shendure (2011) Nature Reviews Genetics.

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

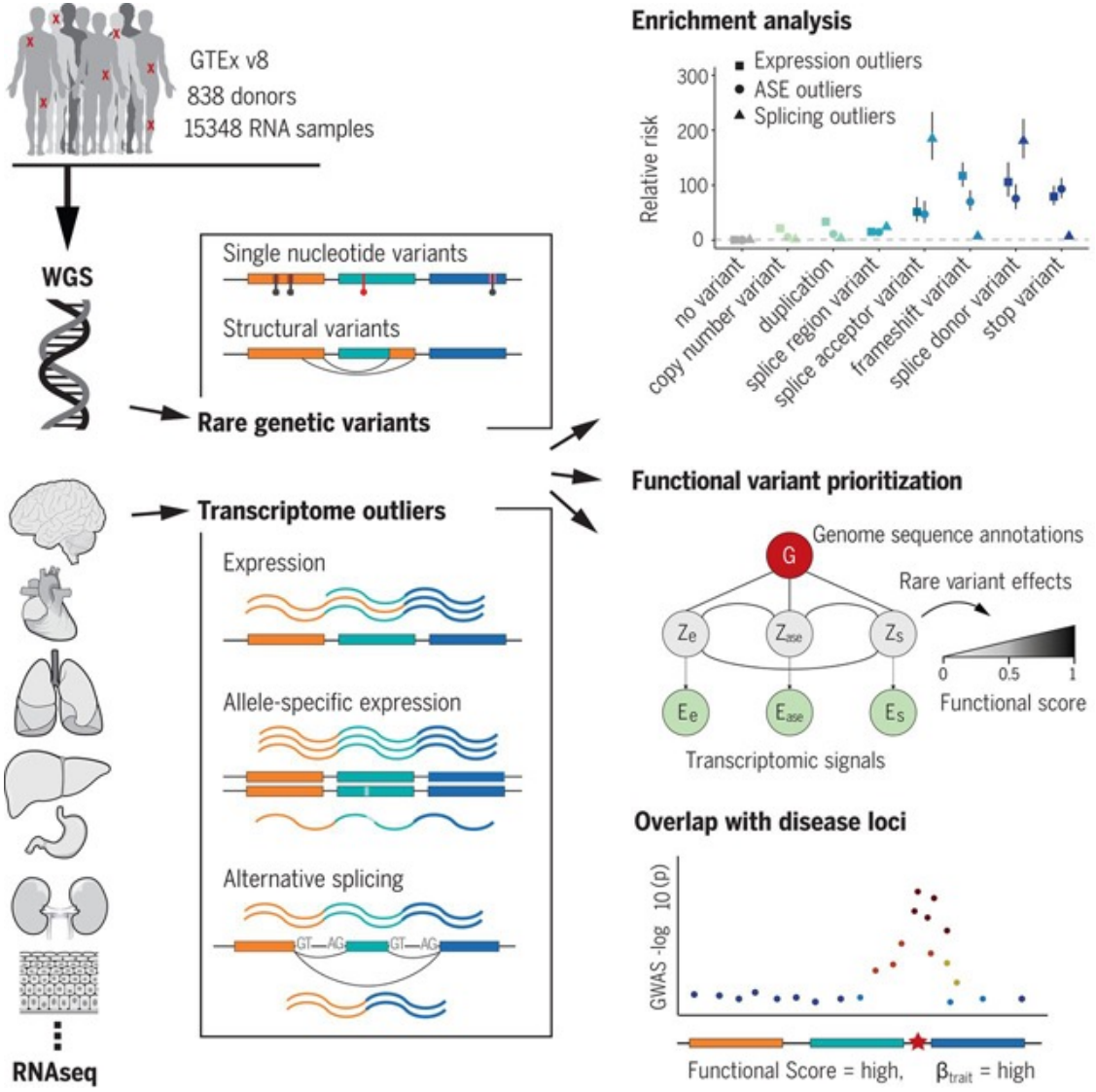
Current methods for annotating and interpreting human genetic variation tend to exploit a single information type (for example, conservation) and/or are restricted in scope (for example, to missense changes). Here we describe Combined Annotation–Dependent Depletion (CADD), a method for objectively integrating many diverse annotations into a single measure (C score) for each variant. We implement CADD as a support vector machine trained to differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants. We precompute C scores for all 8.6 billion possible human single-nucleotide variants and enable scoring of short insertions-deletions. C scores correlate with allelic diversity, annotations of functionality, pathogenicity, disease severity, experimentally measured regulatory effects and complex trait associations, and they highly rank known pathogenic variants within individual genomes. The ability of CADD to prioritize functional, deleterious and pathogenic variants across many functional categories, effect sizes and genetic architectures is unmatched by any current single-annotation method.

comparable, making it difficult to evaluate the relative importance of distinct variant categories or annotations. Third, annotation methods trained on known pathogenic mutations are subject to major ascertainment biases and may not be generalizable. Fourth, it is a major practical challenge to obtain, let alone to objectively evaluate or combine, the existing panoply of partially correlated and partially overlapping annotations; this challenge will only increase in size as large-scale projects such as the Encyclopedia of DNA Elements (ENCODE)[11] continually increase the amount of relevant data available. The net result of these limitations is that many potentially relevant annotations are ignored, while the annotations that are used are applied and combined in *ad hoc* and subjective ways that undermine their usefulness.

Here we describe a general framework, Combined Annotation–Dependent Depletion (CADD), for integrating diverse genome annotations and scoring any possible human single-nucleotide variant (SNV) or small insertion-deletion (indel) event. The basis of CADD is to contrast the annotations of fixed or nearly fixed derived alleles in humans with those of simulated variants. Deleterious variants—that is, variants that reduce organismal fitness—are depleted by natural selection in fixed but not simulated variation. CADD therefore

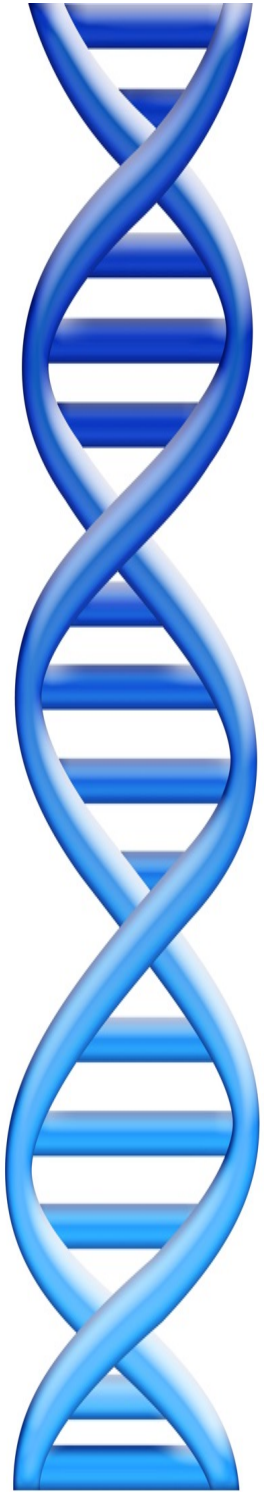
CADD Key Idea: Evaluate amino acid substitutions AND allele frequencies in 1000 genomes project AND ENCODE regions AND … (63 annotations total :)

Nature Genetics 46, 310–315 (2014) doi:10.1038/ng.2892



Watershed Key Idea: Identify rare variants that are associated with major changes in gene expression using a synthesis of conservation and other annotations

Transcriptomic signatures across human tissues identify functional rare genetic variation
Ferraro et al. (2020) Science doi: 10.1126/science.aaz5900



Exam

What to study?

Closed book / no electronics in class exam [75 min]:

- One sheet of paper for notes (double sided)
- Recommend powerpoint / gslides / keynote but hand-written is okay
- Bring a pen / pencil!

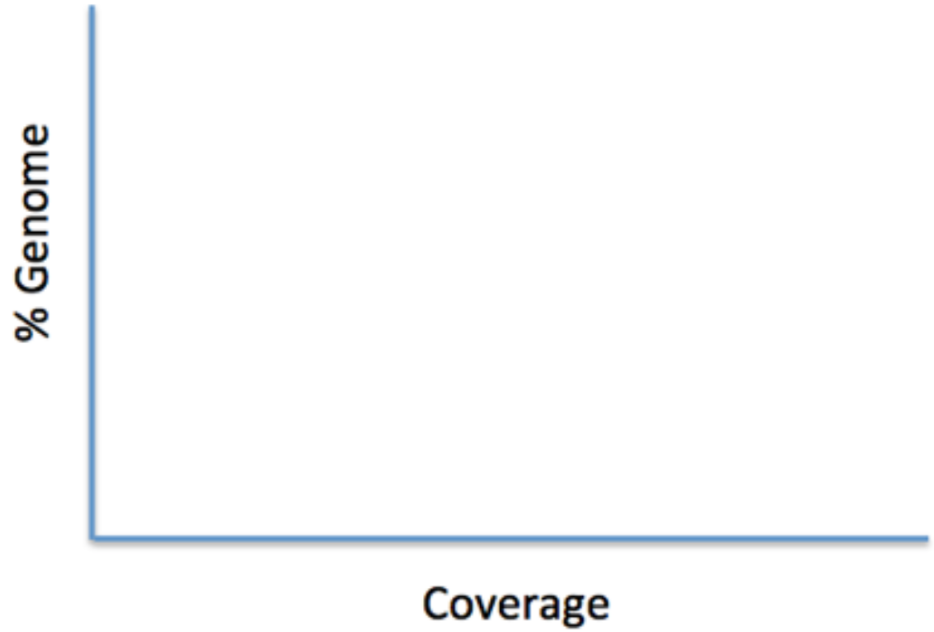
1. **Homework assignments**
   - Make sure you understand the key concepts, algorithms, data structures, and methods involved
   - Don't worry about running software or writing code, but you should know what are the major software tools and how they work

2. **Lectures [with videos on canvas!]**
   - Make sure you understand the key concepts, algorithms, data structures, and methods involved :)
   - No "complicated" numerical computing, but you should be able to sketch a distribution, approximate values, & other simple calculations

3. **Readings**
   - These are to provide more details and provide additional perspectives
   - You are not responsible for plots/content only presented in the readings
   - Of course, you are free to read/watch/discuss additional resources

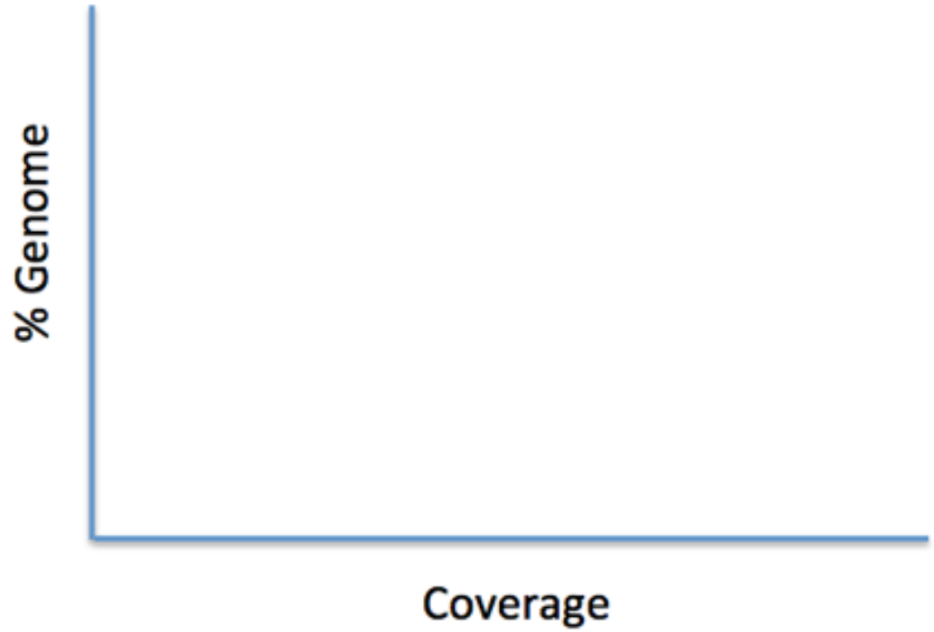
Sample Question

Q3. The Maryland blue crab genome is 1 Gbp in size. 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? 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 is within 1 standard deviation of the mean, 95.4% within 2, 99.7% within 3, and 99.9% within 4)



Sample Question

Q3. The Maryland blue crab genome is 1 Gbp in size. 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? 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 is within 1 standard deviation of the mean, 95.4% within 2, 99.7% within 3, and 99.9% within 4)



Also expect a few multiple choice, short answer, and longer essay questions

Exam Topics

Genomics

- Genomics Technologies
  - Illumina, PacBio, Nanopore
- Coverage, Kmers, Motifs
- Genome Assembly: GRCh38, T2T
- Whole Genome Alignment, Dotplot
- Read mapping
- Variant Identification
- Gene Finding
- Genome Annotation
- BLAST
- RNA-seq, Volcano Plots
- Single-cell analysis
- Methyl-seq, Chip-Seq, Hi-C

Quantitative Techniques

- Normal, Poisson, Binomial,
- P-value, E-value
- de Bruijn and overlap graphs
- Quality Values (Phred Scale)
- Full text indexing & BWT
- Seed & Extend
- Differential Expression
- Expectation Maximization
- PCA / t-SNE / UMAP
- One-hot encoding
- Logistic Regression, Random Forest
- Convolutional Neural Networks
- Transformers / Attention

What is the goal? What is the approach? What are the key challenges?

How did we explore these topics in the homeworks and lectures?