Genetic Disease Genomics

Michael Schatz

October 28, 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

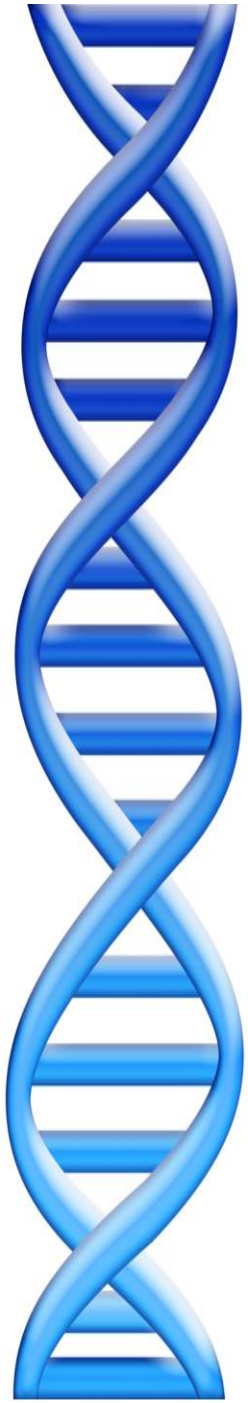
https://github.com/schatzlab/appliedgenomics2024/blob/main/project/presentations.md

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.

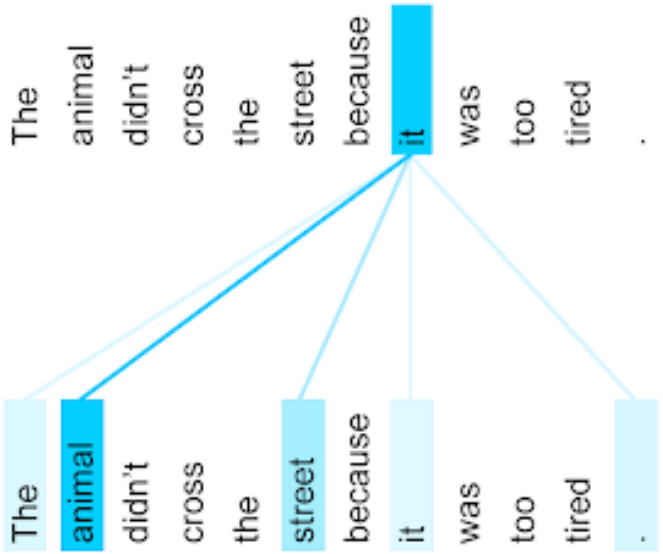


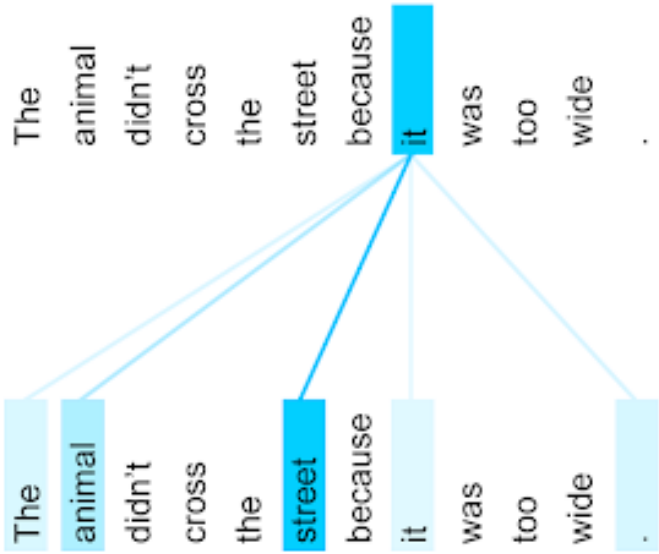
Part 1:
DL in Genomics

Coreference resolution: self attention

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

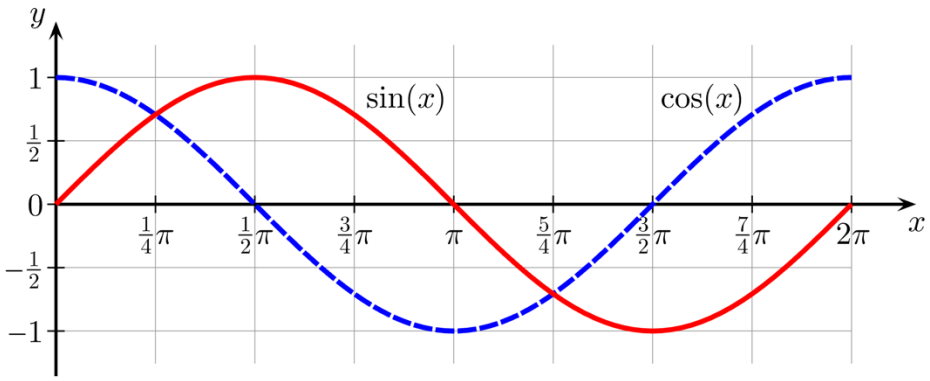
The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

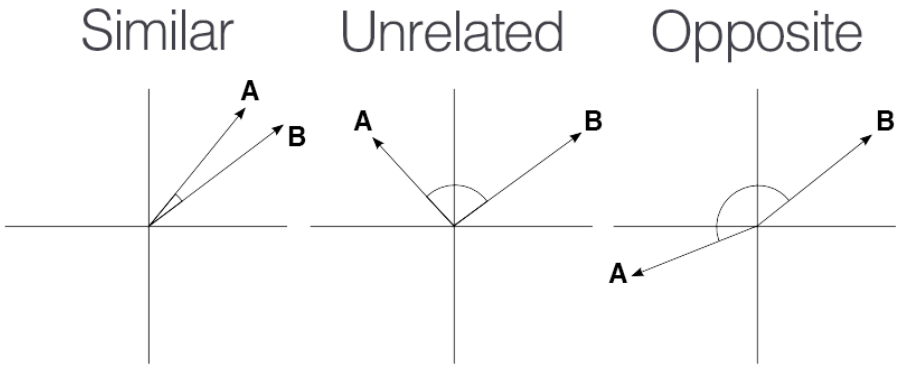




https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

Cosine Similarity





$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}},$$

where A_i and B_i are the ith components of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively.

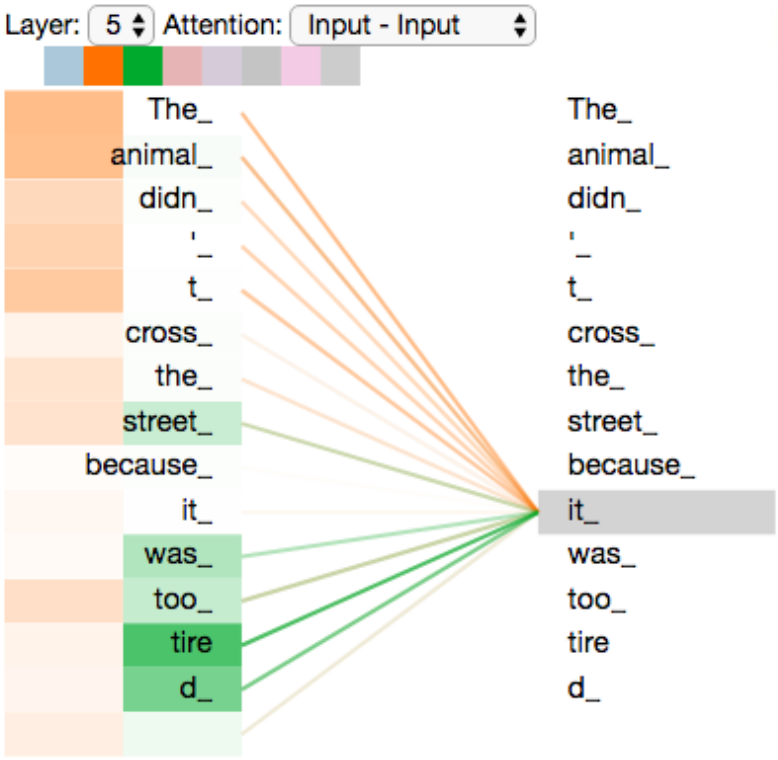
Multi-headed Attention

The orange attention head “knows” “it” refers to animal

The green attention head “knows” “it” refers to tired

Added together we form a new vector representing the concept of “tired animal”

* One head will be trained to look at previous word



https://jalammar.github.io/illustrated-transformer/

Transformer: Multi-headed attention to encode and decode

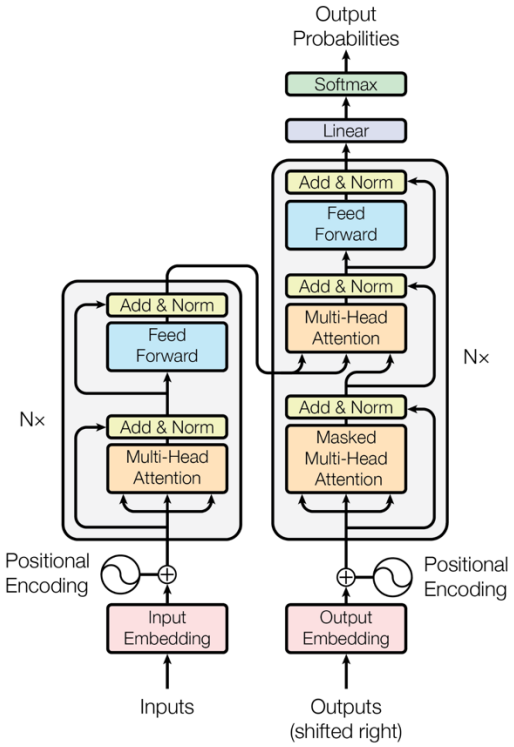


Figure 1: The Transformer - model architecture.

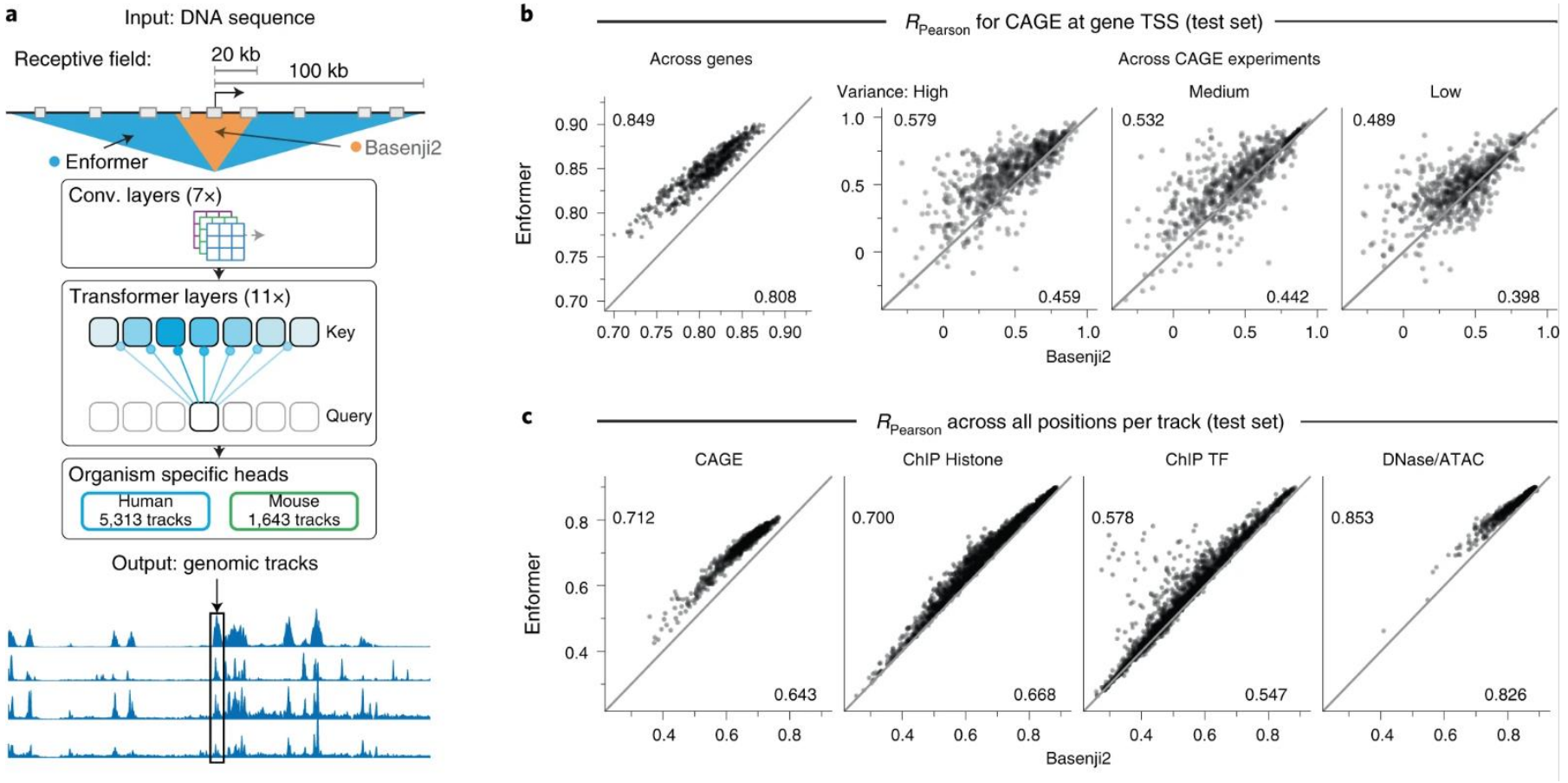
The Transformer uses multi-head attention in three different ways:

"Encoder self-attention": In a self-attention layer all of the keys, values and queries come from the same place, in this case, the output of the previous layer in the encoder. Each position in the encoder can attend to all positions in the previous layer of the encoder.

"Encoder-decoder attention": The queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence.

"Self-attention in the decoder": Allows each position in the decoder to attend to all positions in the decoder up to and including that position. We need to prevent leftward information flow in the decoder to preserve the auto-regressive property. We implement this inside of scaled dot-product attention by masking out (setting to $-\infty$) all values in the input of the softmax which correspond to illegal connections.

https://www.youtube.com/watch?v=UZDiGooFs54



“Enformer” Effective gene expression prediction from sequence by integrating long-range interactions

Avsec et al. (2021) Nature Methods. https://doi.org/10.1038/s41592-021-01252-x

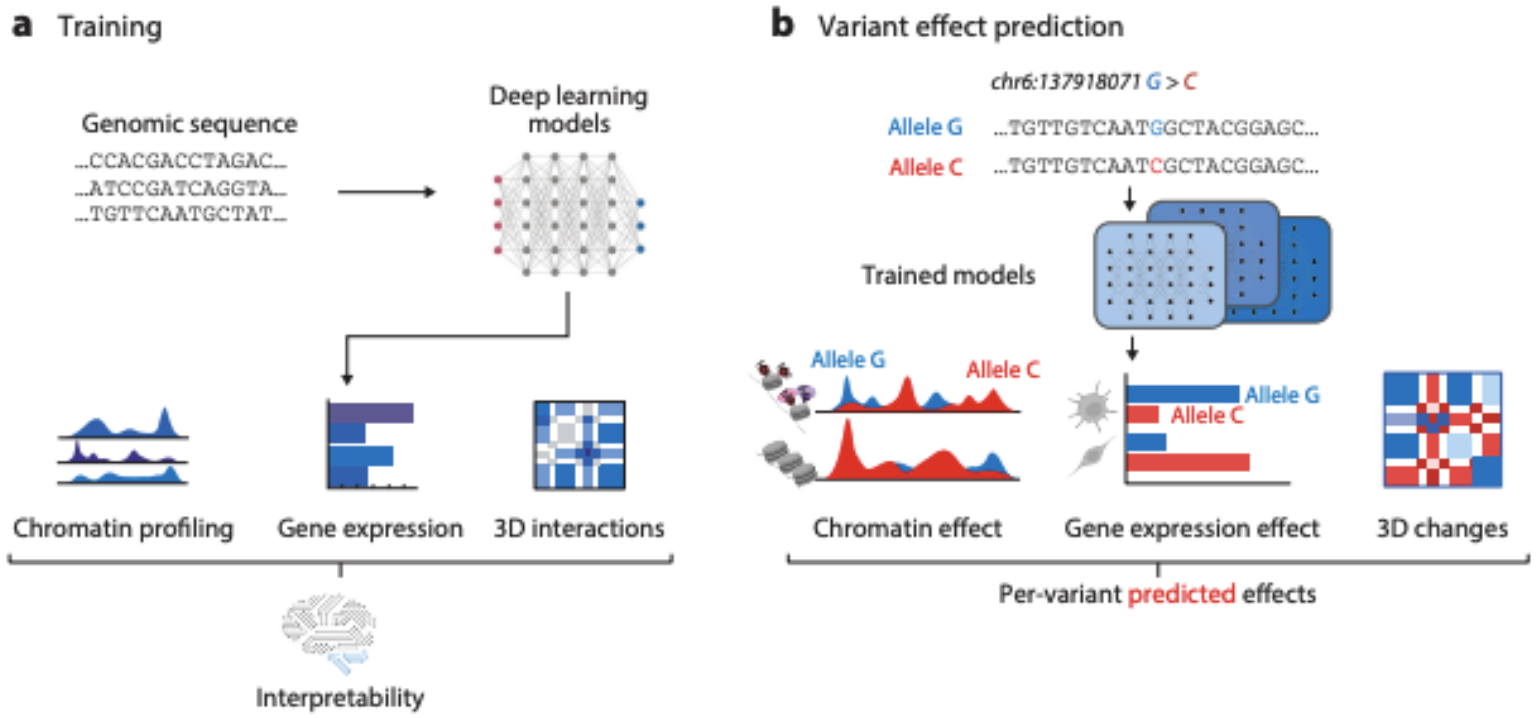


Figure 1

Overview of sequence model prediction tasks and example application. (*a*) Trained deep learning sequence models predict based solely on the DNA sequence as input and are capable of producing tissue- and cell-type-specific chromatin profiling, gene expression, and 3D interaction predictions. Further, interpretation techniques can be used on these models to extract relevant information. (*b*) Variants can be introduced computationally (in silico) into the model through alteration of the sequence, thus enabling the interrogation of millions of variants, including previously unseen variants.

Deep Learning Sequence Models for Transcriptional Regulation

Sokolova et al. (2024) Annual Review of Genomics and Human Genetics.
doi: 10.1146/annurev-genom-021623-024727

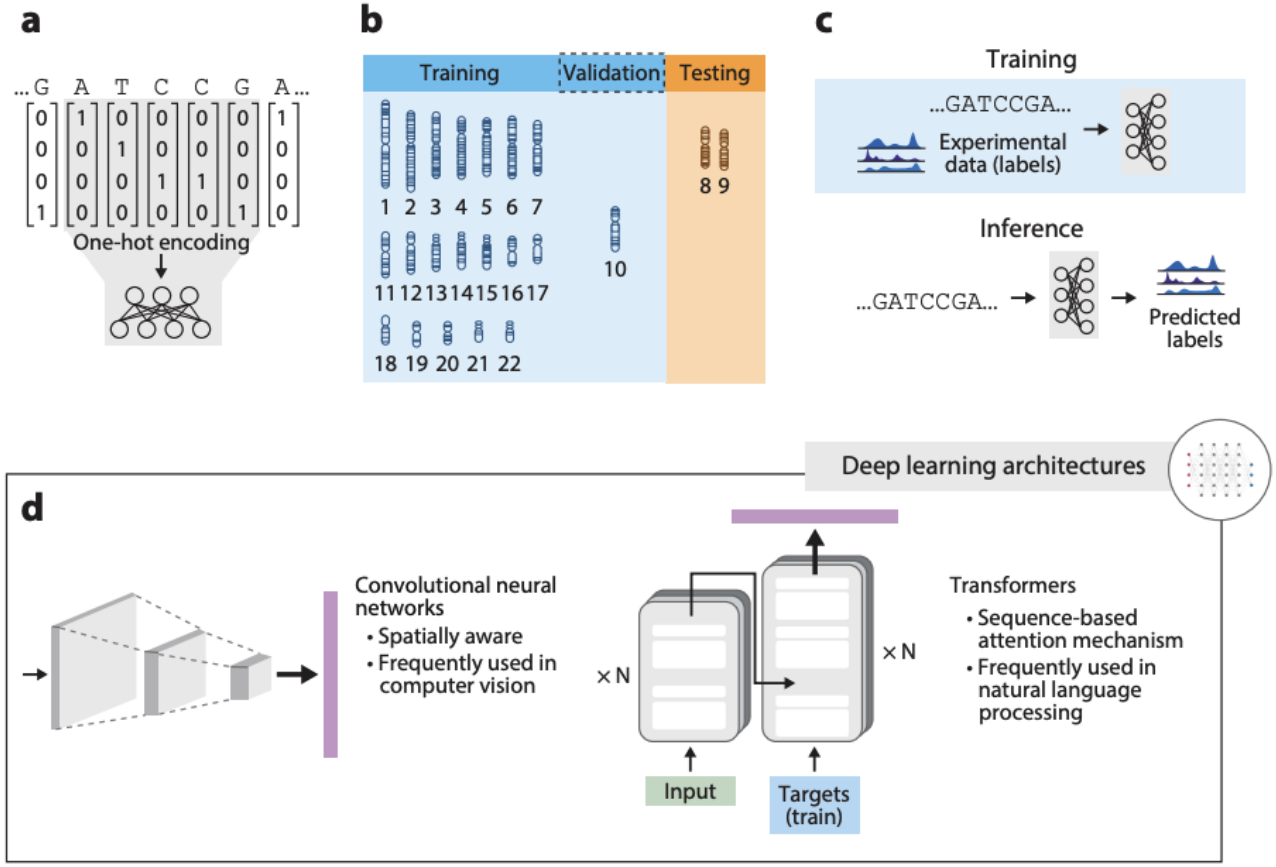


Figure 2

Overview of machine learning concepts. (*a*) To convert DNA sequence into machine-parsable format, the sequence is one-hot encoded: Each nucleotide is converted into a vector of four. (*b*) The dataset is split into training, validation, and testing. The test set is held out from all stages of training and is used to assess the final model. In the context of genomics, one way to split the data is by using chromosomes, for example, using chromosome 10 as a validation set and chromosomes 8 and 9 as a test set (12). (*c*) In supervised training, the targets (labels) are provided together with the sequence, and during inference only the sequence is required. (*d*) Two primary architectures used for the tasks outlined in this review are convolutional neural networks and transformers.

Sokolova et al. (2024) Annual Review of Genomics and Human Genetics.
doi: 10.1146/annurev-genom-021623-024727

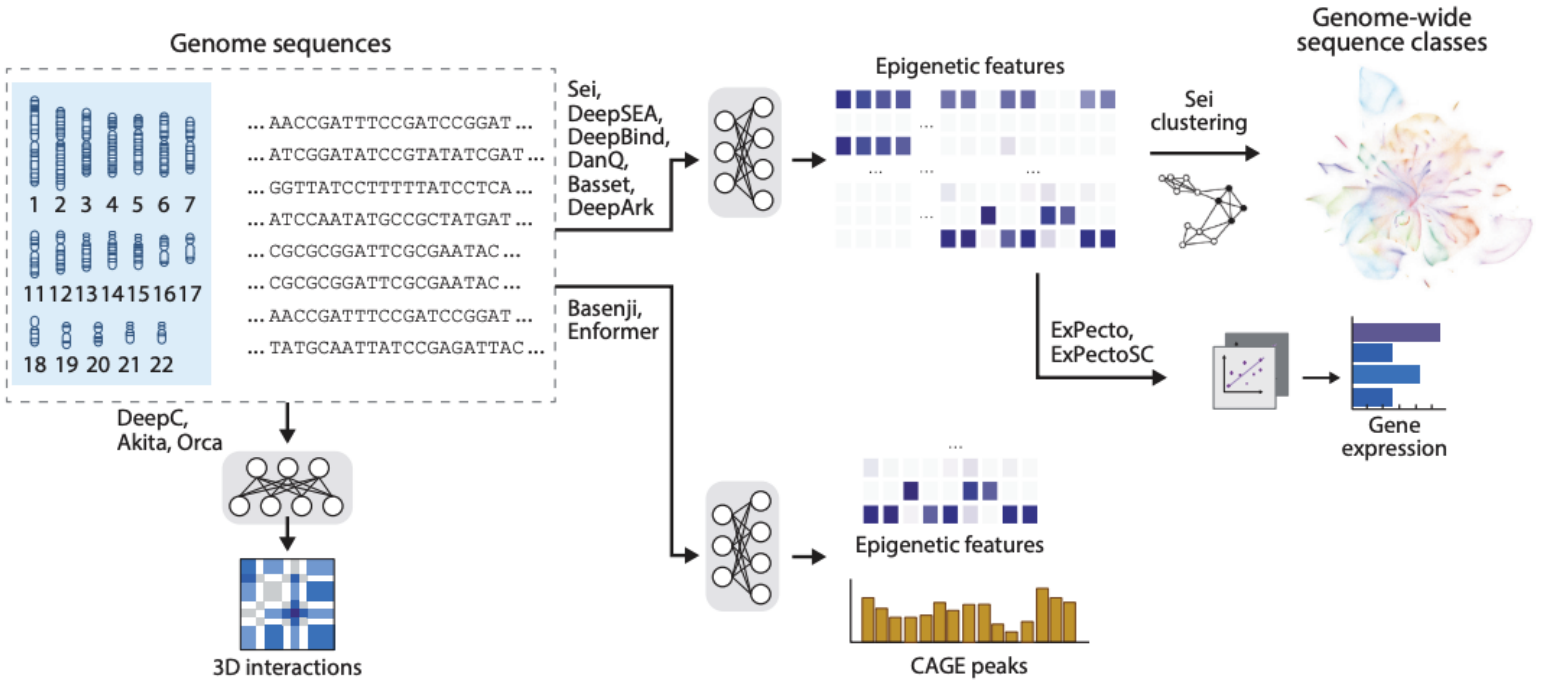


Figure 3

Overview of sequence-based deep learning models and how they work together. Once trained, these models require only the genomic sequence input. For example, the Sei architecture scans the entire genome to identify predicted chromatin profiling patterns and provides genome-wide sequence classes. ExPecto and ExPectoSC use predicted epigenetic marks in a modular framework to generate gene expression predictions, while Basenji and Enformer concurrently predict epigenetic features and cap analysis of gene expression (CAGE) peaks. Akita and Orca take in genomic sequences to predict 3D interaction maps; notably, Orca is capable of predicting Hi-C maps for entire chromosomes in a single pass.

Sokolova et al. (2024) Annual Review of Genomics and Human Genetics.
doi: 10.1146/annurev-genom-021623-024727

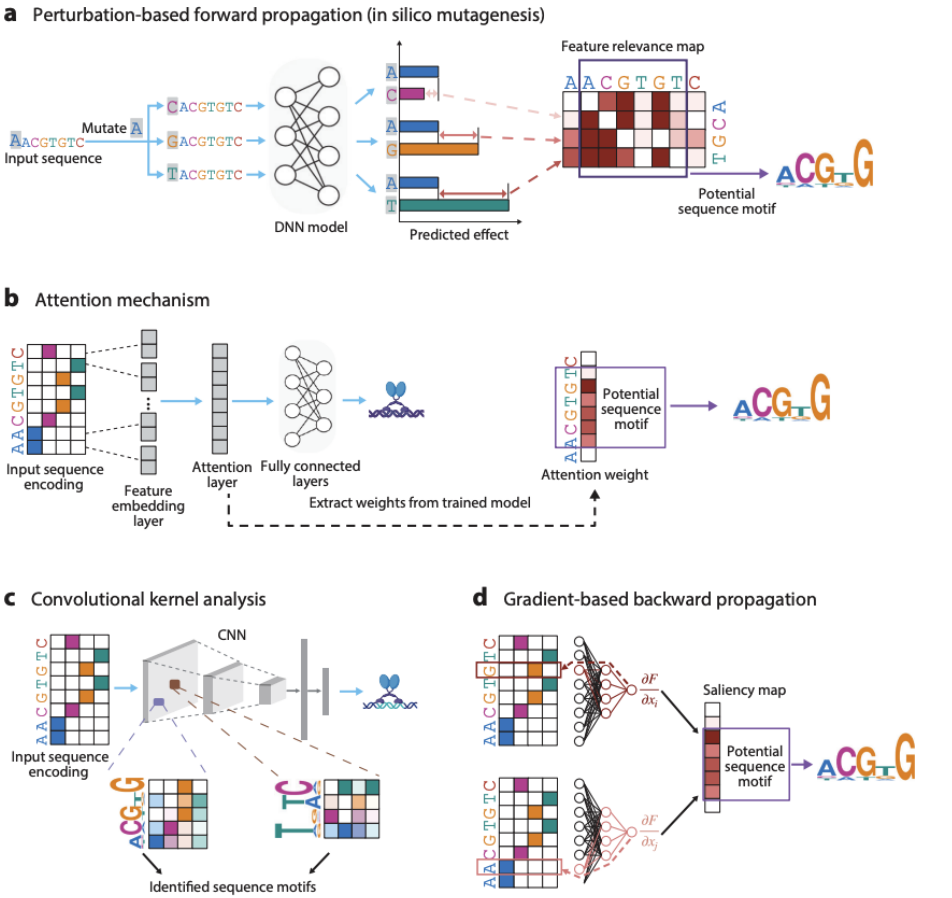


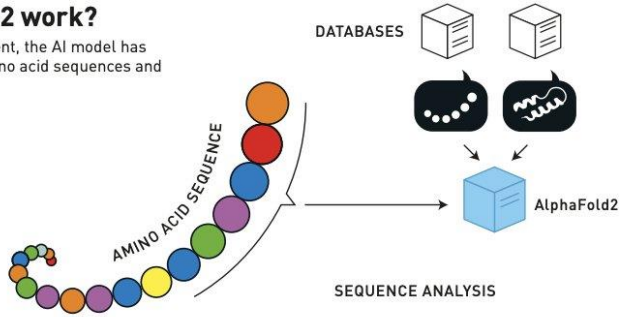
Figure 4

Overview of deep learning interpretation approaches for transcriptional regulation. (*a*) Perturbation-based forward propagation determines the relevance of a feature via mutating each nucleotide to all three possible alternatives. Feature relevance is computed as the corresponding difference in model predictions, such that positions particularly sensitive to perturbations can be prioritized for subsequent motif analysis. (*b*) Attention mechanism assigns weights to parts of the sequence based on their relevance to the sequence-to-activity prediction. Sequence elements of higher attention weights can be prioritized for motif identification. (*c*) Convolutional kernel analysis utilizes the convolutional layer to identify motif or motif-like sequence patterns, scaling and transforming the weights of a trained CNN into a position weight matrix. (*d*) Gradient-based backward propagation determines feature relevance by computing the partial derivative of trained DNN function with respect to the input, producing a gradient vector that reflects the extent to which the prediction of the DNN would change from perturbations of each input nucleotide. Abbreviations: CNN, convolutional neural network; DNN, deep neural network.

Sokolova et al. (2024) Annual Review of Genomics and Human Genetics.
doi: 10.1146/annurev-genom-021623-024727

How does AlphaFold2 work?

As part of AlphaFold2's development, the AI model has been trained on all the known amino acid sequences and determined protein structures.



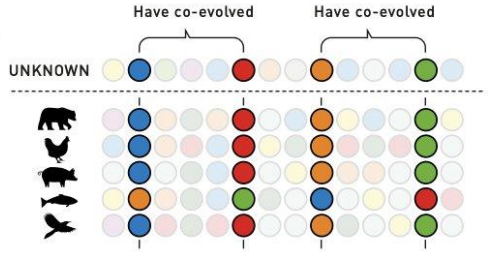
1. DATA ENTRY AND DATABASE SEARCHES

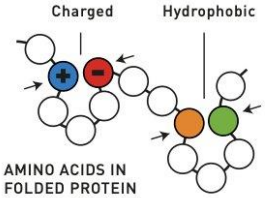
An amino acid sequence with unknown structure is fed into AlphaFold2, which searches databases for similar amino acid sequences and protein structures.

2. SEQUENCE ANALYSIS

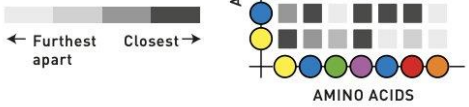
The AI model aligns all the similar amino acid sequences – often from different species – and investigates which parts have been preserved during evolution.

In the next step, AlphaFold2 explores which amino acids could interact with each other in the three-dimensional protein structure. Interacting amino acids co-evolve. If one is charged, the other has the opposite charge, so they are attracted to each other. If one is replaced by a water-repellent (hydrophobic) amino acid, the other also becomes hydrophobic.





Using this analysis, AlphaFold2 produces a distance map that estimates how close amino acids are to each other in the structure.

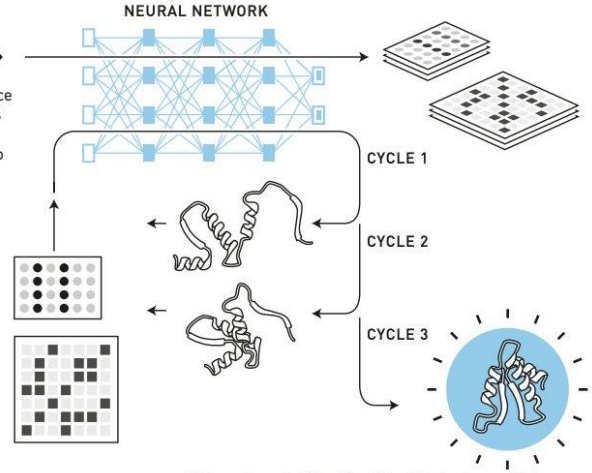


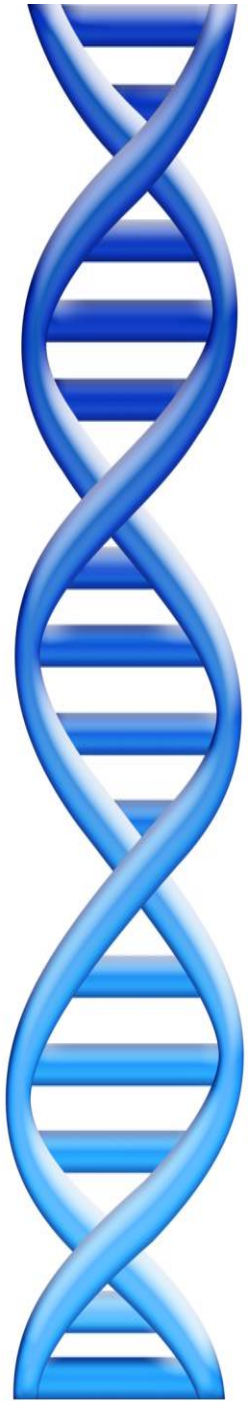
3. AI ANALYSIS

Using an iterative process, AlphaFold2 refines the sequence analysis and distance map. The AI model uses neural networks called transformers, which have a great capacity to identify important elements to focus on. Data about other protein structures – if they were found in step 1 – is also utilised.

4. HYPOTHETICAL STRUCTURE

AlphaFold2 puts together a puzzle of all the amino acids and tests pathways to produce a hypothetical protein structure. This is re-run through step 3. After three cycles, AlphaFold2 arrives at a particular structure. The AI model calculates the probability that different parts of this structure correspond to reality.





Part II:
(Healthy) Humans

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

ARTICLE

OPEN

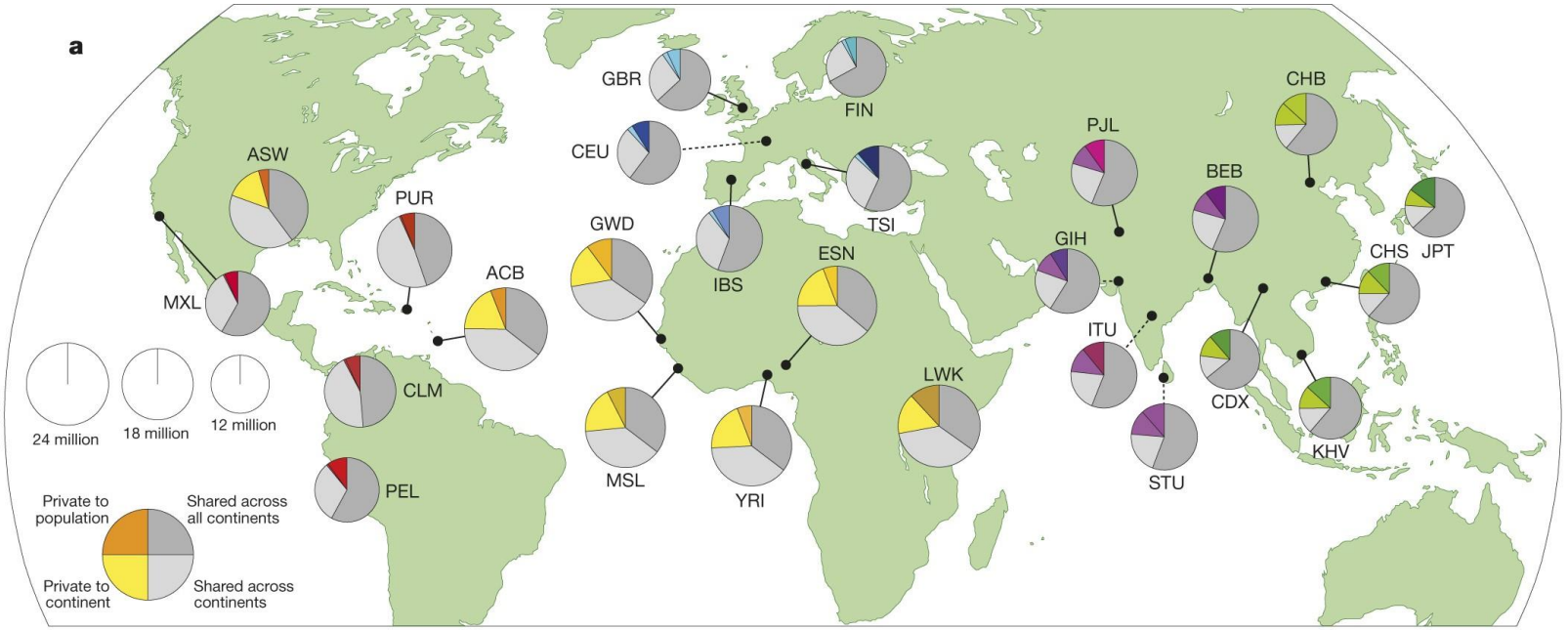
doi:10.1038/nature15393

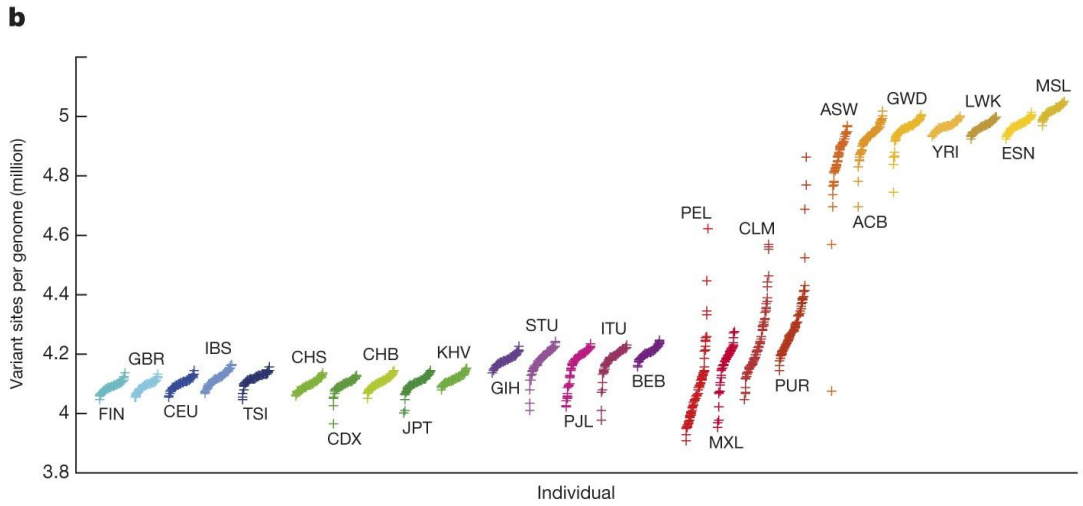
A global reference for human genetic variation

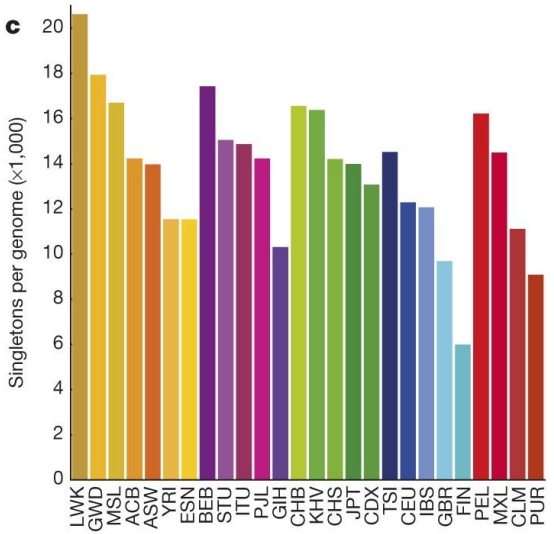
The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

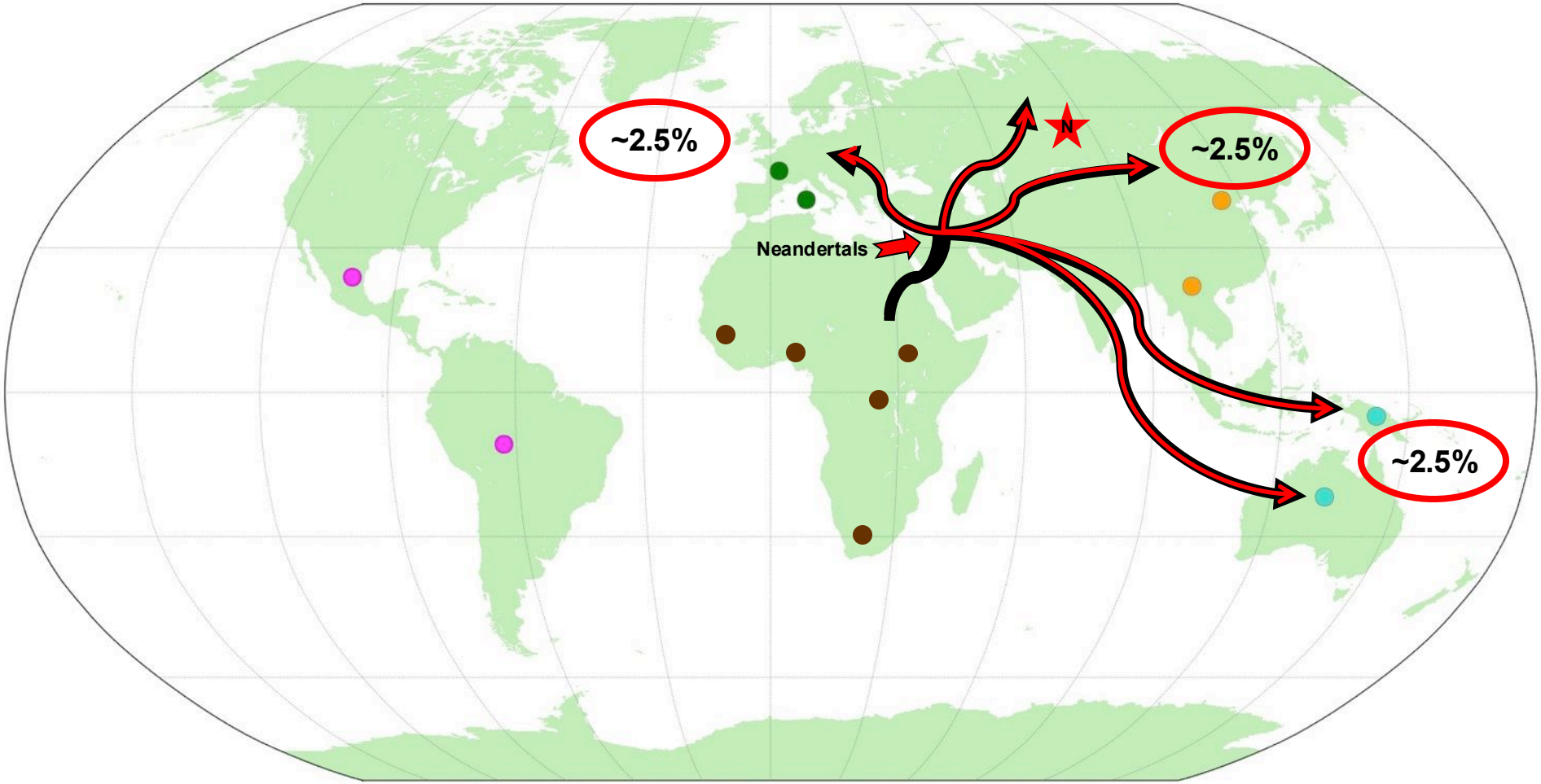
1000 Genomes Populations





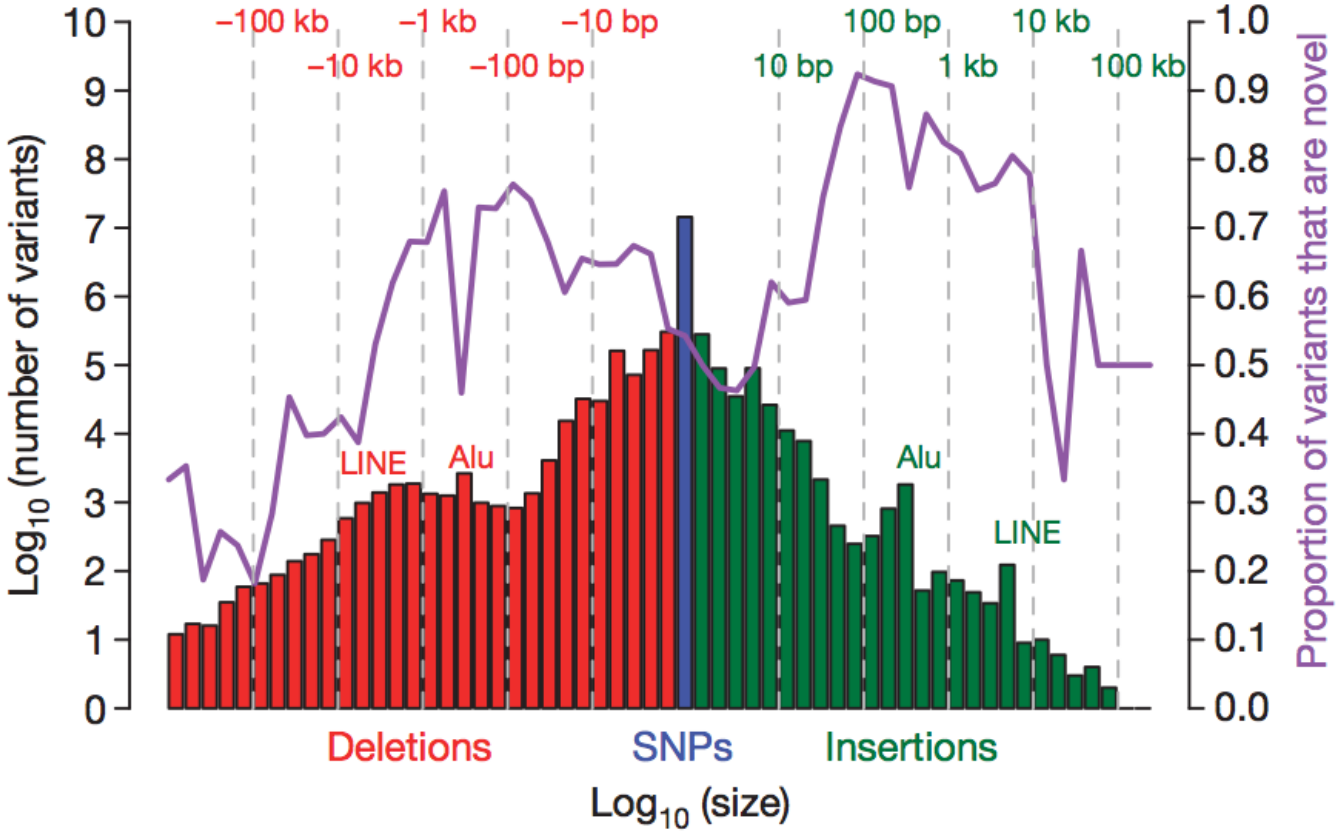


Human Migration Patterns



As modern humans migrated out of Africa, they apparently interbred with Neanderthal's so we see their alleles across the rest of the world and carry about 2.5% of their genome with us!

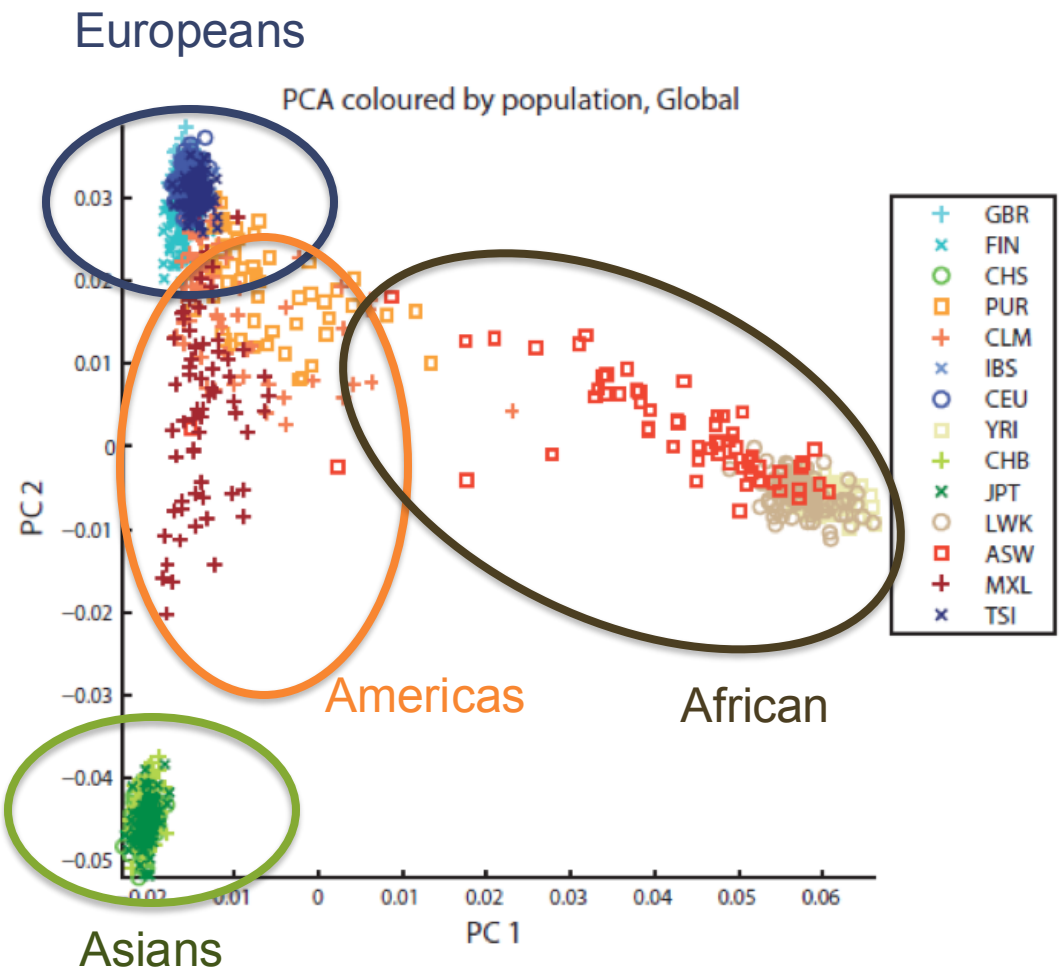
Human Mutation Types



- 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

Variation across populations

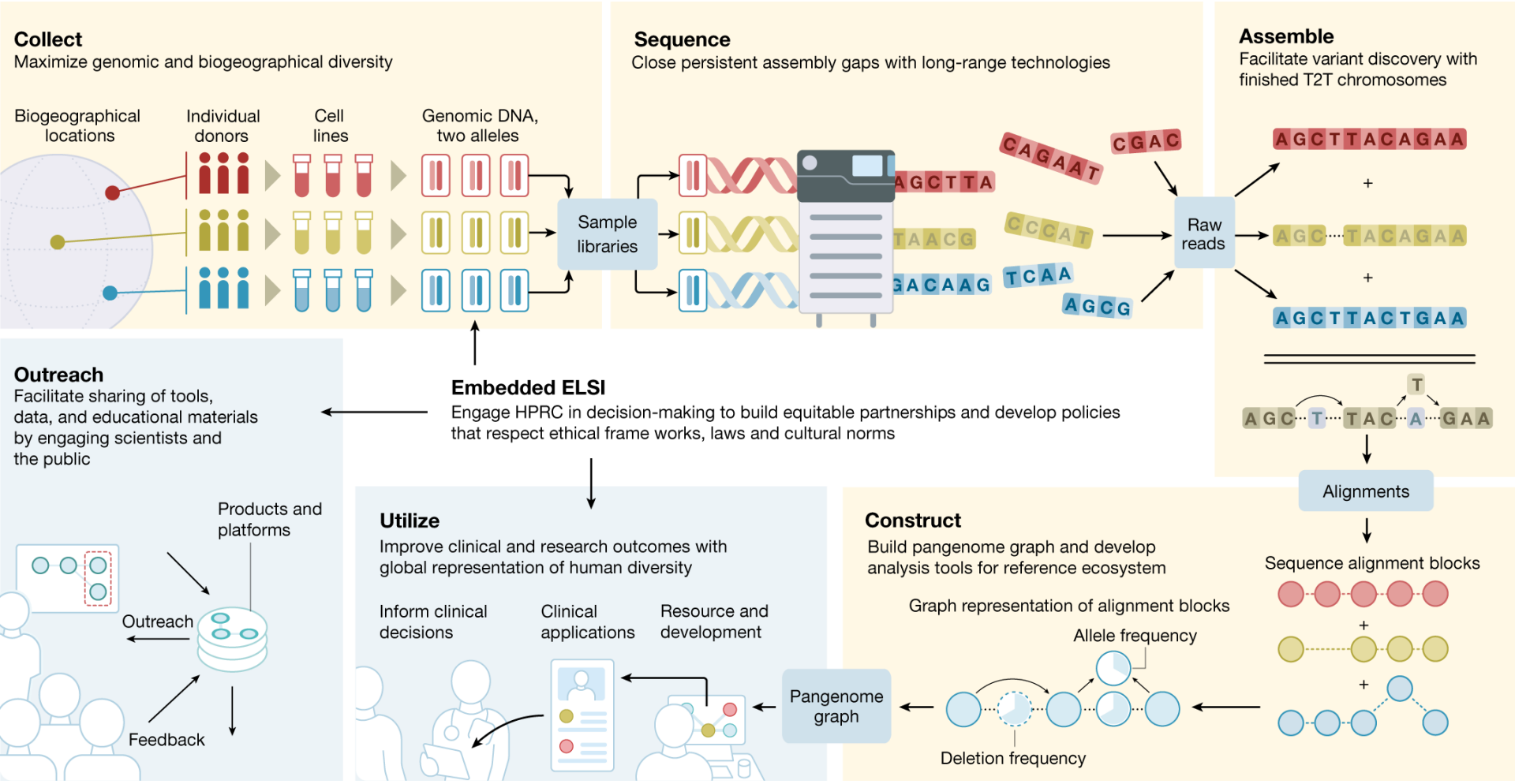


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8
EUR	CEU-GBR	17	14.3
CON	AFR-EUR	348	52.2
CON	AFR-ASN	317	52.6
CON	ASN-EUR	190	53.4

Table S12A Summary of sites showing high levels of population differentiation

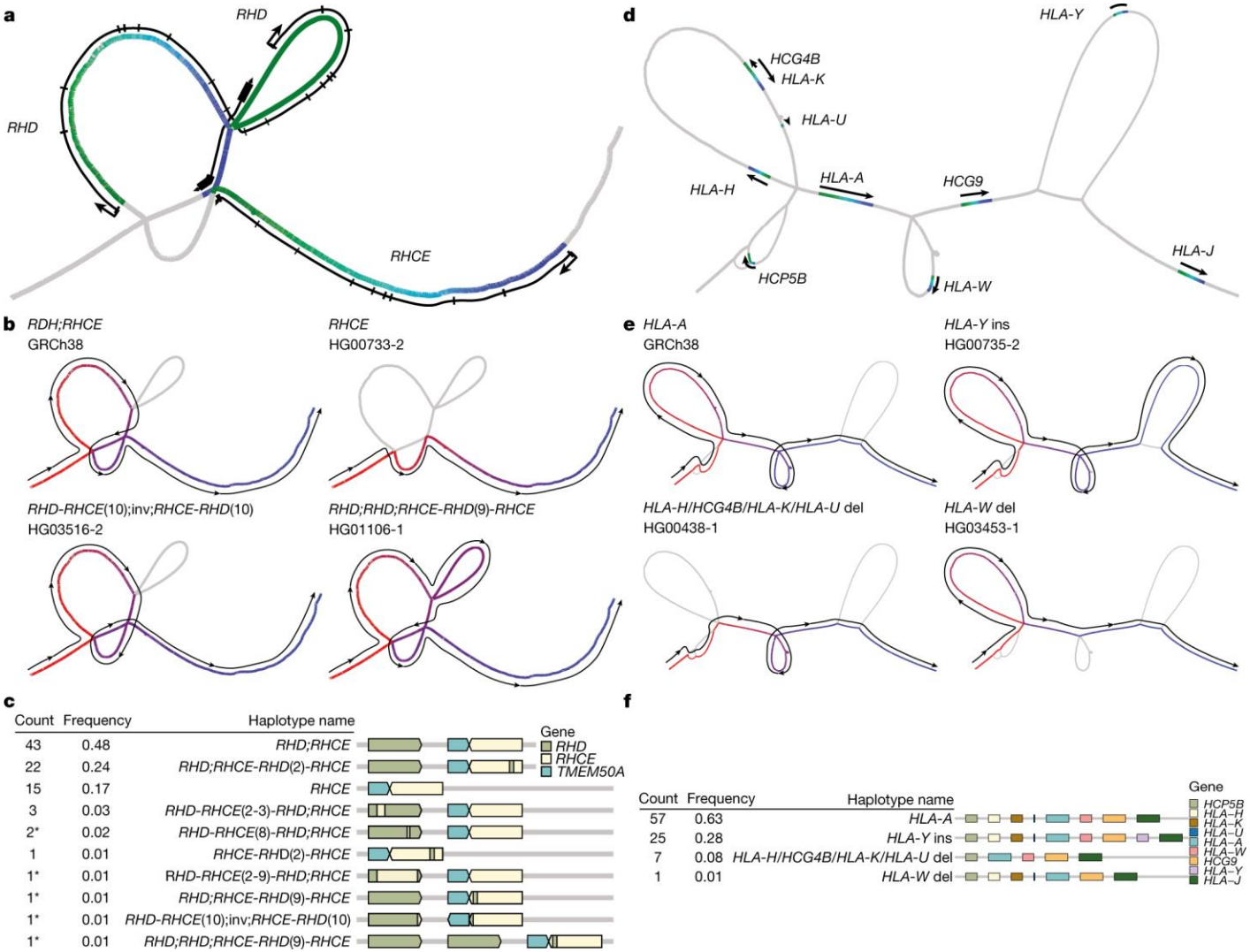
- Not a single variant 100% unique to a given population
- 17% of low-frequency variants (.5-5% pop. freq) observed in a single ancestry group
- 50% of rare variants (<.5%) observed in a single population

Variation across populations



The Human Pangenome Project: a global resource to map genomic diversity
Wang et al (2022) Nature. https://doi.org/10.1038/s41586-022-04601-8

Variation across populations



A draft human pangenome reference

Liao et al (2023) Nature. https://doi.org/10.1038/s41586-023-05896-x

A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes

Daniel G. MacArthur,[1,2*] Suganthi Balasubramanian,[3,4] Adam Frankish,[1] Ni Huang,[1] James Morris,[1] Klaudia Walter,[1] Luke Jostins,[1] Lukas Habegger,[3,4] Joseph K. Pickrell,[5] Stephen B. Montgomery,[6,7] Cornelis A. Albers,[1,8] Zhengdong D. Zhang,[9] Donald F. Conrad,[10] Gerton Lunter,[11] Hancheng Zheng,[12] Qasim Ayub,[1] Mark A. DePristo,[13] Eric Banks,[13] Min Hu,[1] Robert E. Handsaker,[13,14] Jeffrey A. Rosenfeld,[15] Menachem Fromer,[13] Mike Jin,[3] Xinmeng Jasmine Mu,[3,4] Ekta Khurana,[3,4] Kai Ye,[16] Mike Kay,[1] Gary Ian Saunders,[1] Marie-Marthe Suner,[1] Toby Hunt,[1] If H. A. Barnes,[1] Clara Amid,[1,17] Denise R. Carvalho-Silva,[1] Alexandra H. Bignell,[1] Catherine Snow,[1] Bryndis Yngvadottir,[1] Suzannah Bumpstead,[1] David N. Cooper,[18] Yali Xue,[1] Irene Gallego Romero,[1,5] 1000 Genomes Project Consortium, Jun Wang,[12] Yingrui Li,[12] Richard A. Gibbs,[19] Steven A. McCarroll,[13,14] Emmanouil T. Dermitzakis,[7] Jonathan K. Pritchard,[5,20] Jeffrey C. Barrett,[1] Jennifer Harrow,[1] Matthew E. Hurles,[1] Mark B. Gerstein,[3,4,21]† Chris Tyler-Smith[1]†

Genome-sequencing studies indicate that all humans carry many genetic variants predicted to cause loss of function (LoF) of protein-coding genes, suggesting unexpected redundancy in the human genome. Here we apply stringent filters to 2951 putative LoF variants obtained from 185 human genomes to determine their true prevalence and properties. We estimate that human genomes typically contain ~100 genuine LoF variants with ~20 genes completely inactivated. We identify rare and likely deleterious LoF alleles, including 26 known and 21 predicted severe disease–causing variants, as well as common LoF variants in nonessential genes. We describe functional and evolutionary differences between LoF-tolerant and recessive disease genes and a method for using these differences to prioritize candidate genes found in clinical sequencing studies.

(2012) Science. doi: 10.1126/science.1215040

Homozygous LoF Mutations

LETTER

doi:10.1038/nature22034

Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity

Danish Saleheen[1,2]*, Pradeep Natarajan[3,4]*, Irina M. Armean[4,5], Wei Zhao[1], Asif Rasheed[2], Sumeet A. Khetarpal[6], Hong-Hee Won[7], Konrad J. Karczewski[4,5], Anne H. O'Donnell-Luria[4,5,8], Kaitlin E. Samocha[4,5], Benjamin Weisburd[4,5], Namrata Gupta[4], Mozzam Zaidi[2], Maria Samuel[2], Atif Imran[2], Shahid Abbas[9], Faisal Majeed[2], Madiha Ishaq[2], Saba Akhtar[2], Kevin Trindade[6], Megan Mucksavage[6], Nadeem Qamar[10], Khan Shah Zaman[10], Zia Yaqoob[10], Tahir Saghir[10], Syed Nadeem Hasan Rizvi[10], Anis Memon[10], Nadeem Hayyat Mallick[11], Mohammad Ishaq[12], Syed Zahed Rasheed[12], Fazal-ur-Rehman Memon[13], Khalid Mahmood[14], Naveeduddin Ahmed[15], Ron Do[16,17], Ronald M. Krauss[18], Daniel G. MacArthur[4,5], Stacey Gabriel[4], Eric S. Lander[4], Mark J. Daly[4,5], Philippe Frossard[2]§, John Danesh[19,20]§, Daniel J. Rader[6,21]§ & Sekar Kathiresan[3,4]§

A major goal of biomedicine is to understand the function of every gene in the human genome[1]. Loss-of-function mutations can disrupt both copies of a given gene in humans and phenotypic analysis of such 'human knockouts' can provide insight into gene function. Consanguineous unions are more likely to result in offspring carrying homozygous loss-of-function mutations. In Pakistan, consanguinity rates are notably high[2]. Here we sequence the protein-coding regions of 10,503 adult participants in the Pakistan Risk of Myocardial Infarction Study (PROMIS), designed to understand the determinants of cardiometabolic diseases in individuals from South Asia[3]. We identified individuals carrying homozygous predicted loss-of-function (pLoF) mutations, and performed phenotypic analysis involving more than 200 biochemical and disease traits. We enumerated 49,138 rare (<1% minor allele frequency) pLoF mutations. These pLoF mutations are estimated to knock out 1,317 genes, each in at least one participant. Homozygosity for pLoF mutations at *PLA2G7* was associated with absent enzymatic activity of soluble lipoprotein-associated phospholipase A2; at *CYP2F1*, with higher plasma interleukin-8 concentrations; at *TREH*, with lower concentrations of apoB-containing lipoprotein subfractions; at either *A3GALT2* or *NRG4*, with markedly reduced plasma insulin C-peptide concentrations; and at *SLC9A3R1*, with mediators of calcium and phosphate signalling. Heterozygous deficiency of *APOC3* has been shown to protect against coronary heart disease[4,5]; we identified *APOC3* homozygous pLoF carriers in our cohort. We recruited these human knockouts and challenged them with an oral fat load. Compared with family members lacking the mutation, individuals with *APOC3* knocked out displayed marked blunting of the usual post-prandial rise in plasma triglycerides. Overall, these observations provide a roadmap for a 'human knockout project', a systematic effort to understand the phenotypic consequences of complete disruption of genes in humans.

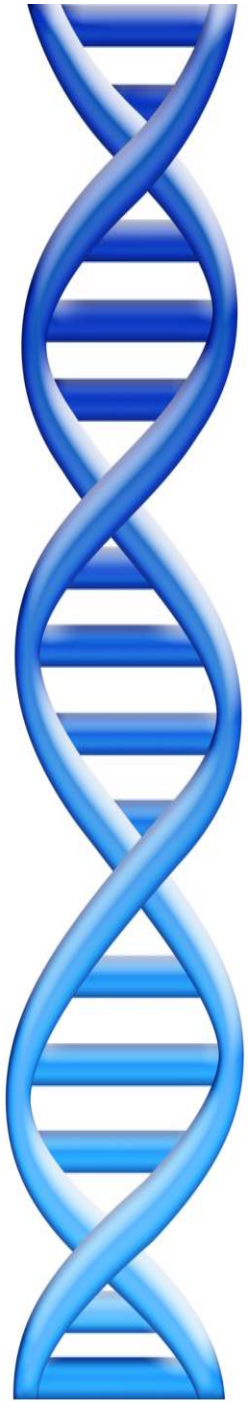
Across all participants (Table 1), exome sequencing yielded 1,639,223 exonic and splice-site sequence variants in 19,026 autosomal genes that passed initial quality control metrics. Of these, 57,137 mutations across 14,345 autosomal genes were annotated as pLoF mutations (that is, nonsense, frameshift, or canonical splice-site mutations predicted to inactivate a gene). To increase the probability that mutations are correctly annotated as pLoF by automated algorithms, we removed nonsense and frameshift mutations occurring within the last 5% of the transcript and within exons flanked by non-canonical splice sites, splice-site mutations at small (<15 bp) introns, at non-canonical splice sites, and where the purported pLoF allele is observed across primates. Common pLoF alleles are less likely to exert strong functional effects as they are less constrained by purifying selection; thus, we define pLoF mutations in the rest of the manuscript as variants with a minor allele frequency (MAF) of <1% and passing the aforementioned bioinformatic filters. Applying these criteria, we generated a set of 49,138 pLoF mutations across 13,074 autosomal genes. The site-frequency spectrum for these pLoF mutations revealed that the majority was seen only in one or a few individuals (Extended Data Fig. 1).

Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated owing to pLoF mutations. A full listing of all 1,317 genes knocked out, the number of knockout participants for each gene, and the specific pLoF mutation(s) are provided in Supplementary Table 1. 891 (67.7%) of the genes were knocked out only in one participant (Fig. 1a). Nearly 1 in 5 of the participants that were sequenced (1,843 individuals, 17.5%) had at least one gene knocked out by a homozygous pLoF mutation. 1,504 of these 1,843 individuals (81.6%) were homozygous pLoF carriers for just one gene, but the minority of participants had more than one gene knocked out and one participant had six genes with homozygous pLoF genotypes.

We compared the coefficient of inbreeding (*F* coefficient) in PROMIS participants with that of 15,249 individuals from outbred populations of European or African American ancestry. The *F* coefficient estimates the excess homozygosity compared with an outbred ancestor. PROMIS participants had a fourfold higher median inbreeding coefficient compared to outbred populations (0.016 versus 0.0041; $P < 2 \times 10^{-16}$) (Fig. 1b). Additionally, those in PROMIS who reported that their parents were closely related had even higher median inbreeding coefficients than

(2017) Nature. doi:10.1038/nature22034

- Homozygous LoF mutations are rare in most people, but enriched in people born from consanguineous relationships
- Sequence the exomes of many such people, find their homozygous LoFs, relate to 200 biochemical or disease traits
- A "natural" experiment to understand what genes do: people with both copies of APOC3 disabled can clear fat from their bloodstream much faster than others, suggests we should develop compounds to prevent heart attacks



Part III:

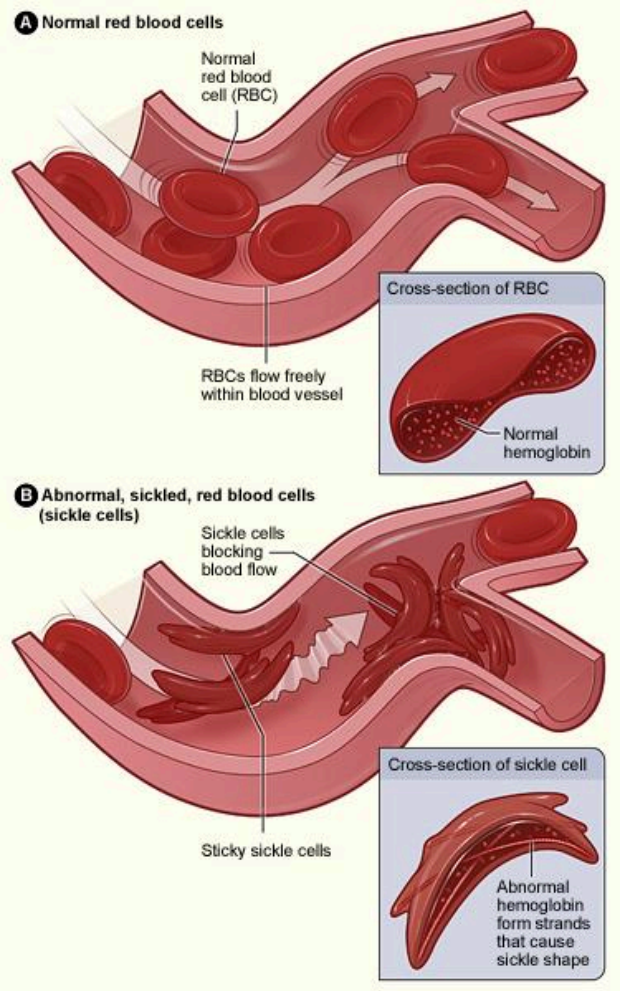
Pre-genome Genetic Medicine

Sickle Cell Anaemia

- Sickle-cell anaemia (SCA) is an abnormality in the oxygen-carrying protein haemoglobin (hemoglobin S) found in red blood cells. First modern clinical description in 1910s

- **The genetic basis of sickle cell disease is an A-to-T transversion in the sixth codon of the HBB gene**.

- The mutation was actually found in the protein sequence first in the 1950s! Occurs when a person inherits two abnormal copies of the haemoglobin gene, one from each parent. Interestingly, heterozygous patients also incur a resistance to malaria infection, contributing to its prevalence in Africa where malaria infections remain a major disease



OMIM: SICKLE CELL ANEMIA

https://www.omim.org/entry/603903

Huntington's Disease

A polymorphic DNA marker genetically linked to Huntington's disease

James F. Gusella*, Nancy S. Wexler†‖, P. Michael Conneally†, Susan L. Naylor§, Mary Anne Anderson*, Rudolph E. Tanzi*, Paul C. Watkins*¶, Kathleen Ottina*, Margaret R. Wallace‡, Alan Y. Sakaguchi§, Anne B. Young‖, Ira Shoulson‖, Ernesto Bonilla‖ & Joseph B. Martin*

* Neurology Department and Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA
† Hereditary Disease Foundation, 9701 Wilshire Blvd, Beverley Hills, California 90212, USA
‡ Department of Medical Genetics, Indiana University Medical Center, Indianapolis, Indiana 46223, USA
§ Department of Human Genetics, Roswell Park Memorial Institute, Buffalo, New York 14263, USA
‖ Venezuela Collaborative Huntington's Disease Project*

Family studies show that the Huntington's disease gene is linked to a polymorphic DNA marker that maps to human chromosome 4. The chromosomal localization of the Huntington's disease gene is the first step in using recombinant DNA technology to identify the primary genetic defect in this disorder.

Gusella et al (1983) Nature. doi:10.1038/306234a0

Huntington's Disease

A polymorphic D
to H

James F. Gusella*, Nan
Mary Anne Anderson*,
Margaret R. Wallace
Er

* Neurology Department and Genetics Unit, M
† Hereditary Disease
‡ Department of Medical Ge
§ Department of Human
‖ Ve

Family studies show that the Huntingt
chromosome 4. The chromosomal loca
DNA technology to identify the primar

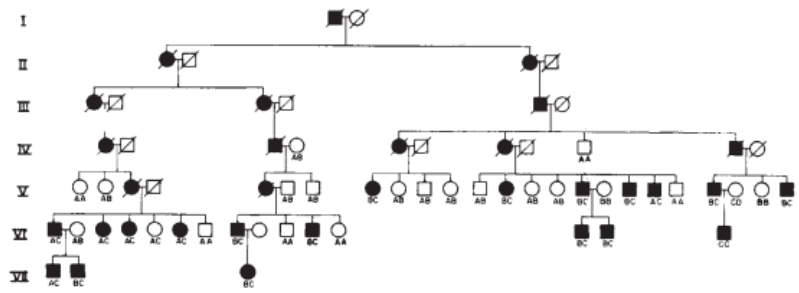


Fig. 2 Pedigree of the Venezuelan Huntington's disease family. This pedigree represents a small part of a much larger pedigree that will be described in detail elsewhere. Permanent EBV-transformed lymphoblastoid cell lines were established from blood samples of these individuals (unpublished data). DNA prepared from the lymphoblastoid lines will used to determine the phenotype of each individual at the G8 locus as described in Fig. 3. The data were analysed for linkage to the Huntington's disease gene using the program LIPED[17] with a correction for the late age of onset[5]. Because of the high frequency of the Huntington's disease gene in this population some of the spouses of affected individuals have also descended from identified Huntington's disease gene carriers. In none of these cases, however, was the unaffected individual at significantly greater risk for Huntington's disease than a member of the general population. Although a number of younger at-risk individuals were also analysed as part of this study, for the sake of these family members the data are not shown due to their predictive nature. The data are available upon request if confidentiality can be assured.

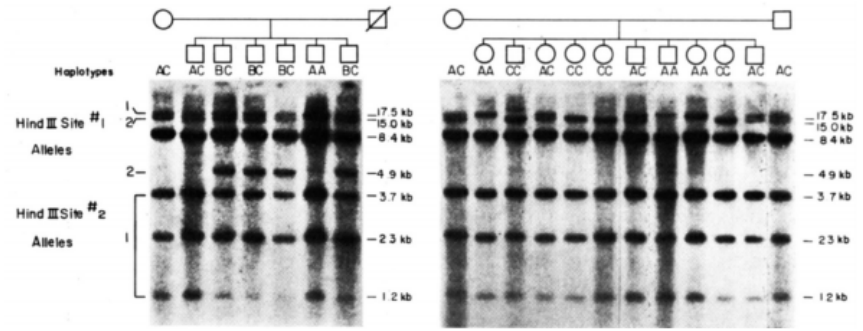


Fig. 3 Hybridization of the G8 Probe to *Hin*dIII-digested human genomic DNA.
Methods: DNA was prepared as described[23] from lymphoblastoid cell lines derived from members of two nuclear families. 5 μg of each DNA was digested to completion with 20 units of *Hin*dIII in a volume of 30 μl using the buffer recommended by the supplier. The DNAs were fractionated on a 1% horizontal agarose gel in TBE buffer (89 mM Tris, *p*H 8, 89 mM Na borate, 2 mM Na EDTA) for 18 h. *Hin*dIII-digested λC1857 DNA was loaded in a separate lane as a size marker. The gels were stained with ethidium bromide (0.5 μg ml^{-1}) for 30 min and the DNA was visualized with UV light. The gels were incubated for 45 min in 1 M NaOH with gentle shaking and for two successive 20 min periods in 1 M Tris, *p*H 7.6, 1.5 M NaCl. DNA from the gel was transferred in 20×SSC (3 M NaCl, 0.3 M Na citrate) by capillary action to a positively charged nylon membrane. After overnight transfer, agarose clinging to the filters was removed by washing in 3×SSC and the filters were air dried and baked for 2 h under vacuum at 80 °C. Baked filters were prehybridized in 500 ml 6×SSC, 1×Denhardt's solution (0.02% bovine serum albumin, 0.02% polyvinyl pyrollidone, 0.02% Ficoll), 0.3% SDS and 100 μg ml^{-1} denatured salmon sperm DNA at 65 °C for 18 h. Prehybridized filters were washed extensively at room temperature in 3×SSC until no evidence of SDS remained. Excess liquid was removed from the filters by blotting on Whatman 3MM paper and damp filters were placed individually in heat-sealable plastic bags. 5 ml of hybridization solution (6×SSC, 1×Denhardt's solution, 0.1% SDS, 100 μg ml^{-1} denatured salmon sperm DNA) containing approximately 5×10^6 c.p.m. of nick-translated G8 DNA (specific activity ~2×10^8 c.p.m. μg^{-1})[24] was added to each bag which was then sealed and placed at 65 °C for 24–48 h. Filters were removed from the bags and washed at 65 °C for 30 min each in 3×SSC, 2×SSC, 1×SSC and 0.3×SSC. The filters were dried and exposed to X-ray film (Kodak XR-5) at −70 °C with a Dupont Cronex intensifying screen for 1 to 4 days. The haplotypes observed in each individual were determined from the alleles seen for each *Hin*dIII RFLP (site 1 and 2) as explained in Fig. 4.

Gusella et al (1983) Nature. doi:10.1038/306234a0

Huntington's Disease

Cell, Vol. 72, 971–983, March 26, 1993, Copyright © 1993 by Cell Press

A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes

The Huntington's Disease Collaborative Research Group*

Summary

The Huntington's disease (HD) gene has been mapped in 4p16.3 but has eluded identification. We have used haplotype analysis of linkage disequilibrium to spotlight a small segment of 4p16.3 as the likely location of the defect. A new gene, IT15, isolated using cloned trapped exons from the target area contains a polymorphic trinucleotide repeat that is expanded and unstable on *HD* chromosomes. A $(CAG)_n$ repeat longer than the normal range was observed on *HD* chromosomes from all 75 disease families examined, comprising a variety of ethnic backgrounds and 4p16.3 haplotypes. The $(CAG)_n$ repeat appears to be located within the coding sequence of a predicted ~348 kd protein that is widely expressed but unrelated to any known gene. Thus, the *HD* mutation involves an unstable DNA segment, similar to those described in fragile X syndrome, spino-bulbar muscular atrophy, and myotonic dystrophy, acting in the context of a novel 4p16.3 gene to produce a dominant phenotype.

Introduction

Huntington's disease (HD) is a progressive neurodegenerative disorder characterized by motor disturbance, cognitive loss, and psychiatric manifestations (Martin and Gusella, 1986). It is inherited in an autosomal dominant fashion and affects ~1 in 10,000 individuals in most populations of European origin (Harper et al., 1991). The hallmark of HD is a distinctive choreic movement disorder that typically has a subtle, insidious onset in the fourth to fifth decade of life and gradually worsens over a course of 10 to 20 years until death. Occasionally, HD is expressed in juveniles, typically manifesting with more severe symptoms including rigidity and a more rapid course. Juvenile onset of HD is associated with a preponderance of paternal transmission of the disease allele. The neuropathology of HD also displays a distinctive pattern, with selective loss of neurons that is most severe in the caudate and putamen. The biochemical basis for neuronal death in HD has not yet been explained, and there is consequently no treatment effective in delaying or preventing the onset and progression of this devastating disorder.

The genetic defect causing HD was assigned to chromosome 4 in 1983 in one of the first successful linkage analyses using polymorphic DNA markers in humans (Gusella

Huntington's Disease

Cell, Vol. 72, 971–983, March 26, 1993, Copyright © 1993 by Cell Press

A Novel Gene Containing a Trinucleotide Repe That Is Expanded and Unstable on Huntington's Disease Chromosomes

The Huntington's Disease Collaborative Research Group*

Summary

The Huntington's disease (HD) gene has been mapped in 4p16.3 but has eluded identification. We have used haplotype analysis of linkage disequilibrium to spotlight a small segment of 4p16.3 as the likely location of the defect. A new gene, IT15, isolated using cloned trapped exons from the target area contains a polymorphic trinucleotide repeat that is expanded and unstable on *HD* chromosomes. A $(CAG)_n$ repeat longer than the normal range was observed on *HD* chromosomes from all 75 disease families examined, comprising a variety of ethnic backgrounds and 4p16.3 haplotypes. The $(CAG)_n$ repeat appears to be located within the coding sequence of a predicted ~348 kd protein that is widely expressed but unrelated to any known gene. Thus, the *HD* mutation involves an unstable DNA segment, similar to those described in fragile X syndrome, spino-bulbar muscular atrophy, and myotonic dystrophy, acting in the context of a novel 4p16.3 gene to produce a dominant phenotype.

Introduction

Huntington's disease (HD) is a progre
ative disorder characterized by moto
tive loss, and psychiatric manifestati
sella, 1986). It is inherited in an a
fashion and affects ~1 in 10,000 indiv
lations of European origin (Harper et
mark of HD is a distinctive choreic
that typically has a subtle, insidious
fifth decade of life and gradually wo
of 10 to 20 years until death. Occ
pressed in juveniles, typically manif
vere symptoms including rigidity and
Juvenile onset of HD is associated w
of paternal transmission of the disea
pathology of HD also displays a dist
selective loss of neurons that is most s
and putamen. The biochemical basi
in HD has not yet been explained,
quently no treatment effective in de
the onset and progression of this de

The genetic defect causing HD was
some 4 in 1983 in one of the first succ
ses using polymorphic DNA markers

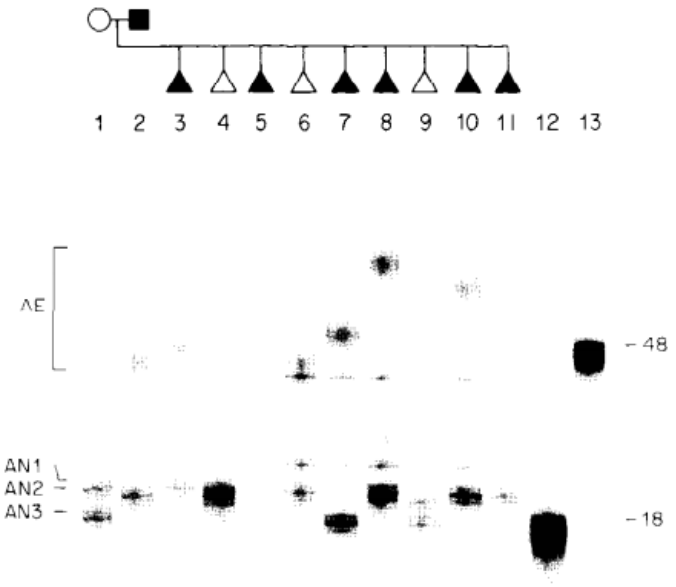


Figure 6. PCR Analysis of the $(CAG)_n$ Repeat in a Venezuelan HD Sibship with Some Offspring Displaying Juvenile Onset

Results of PCR analysis of a sibship in the Venezuelan HD pedigree are shown. Affected individuals are represented by closed symbols. Progeny are shown as triangles, and the birth order of some individuals has been changed for confidentiality. AN1, AN2, and AN3 mark the positions of the allelic products from normal chromosomes. AE marks the range of PCR products from the *HD* chromosome. The intensity of background constant bands, which represent a useful reference for comparison of the above PCR products, varies with slight differences in PCR conditions. The PCR products from cosmids L191F1 and GUS72-2130 are loaded in lanes 12 and 13 and have 18 and 48 CAG repeats, respectively.

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.

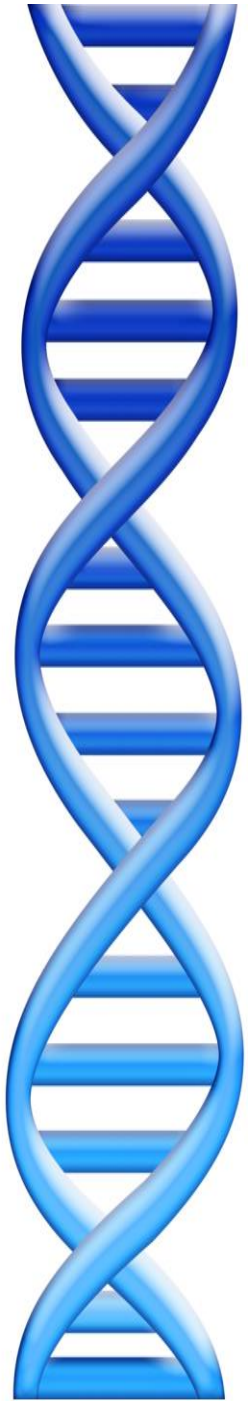
To test the proposal that classifying disease genes and their products according to function will provide general insight into disease processes[1,2], we have compiled and classified a list of disease genes. To assemble the list, we began with 269 genes identified in a survey of the 7th edition of *Metabolic 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

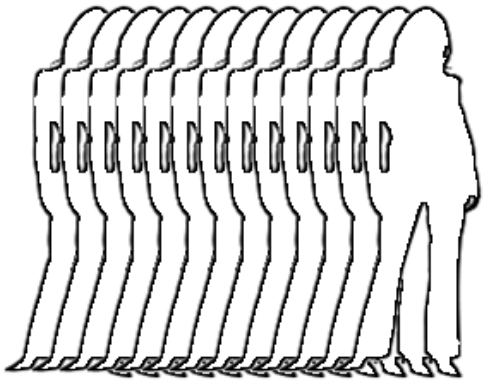
Jimenez-Sanchez, G., Childs, B. & Valle, D. (2001) Nature 409, 853–855



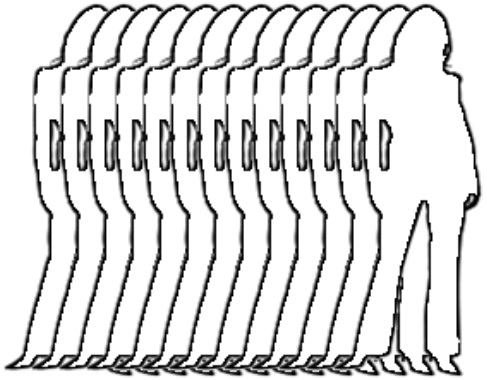
Part IV:

Post-genome Inherited Diseases

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%	

Are these significant differences in frequencies?

Pearson's Chi-squared test

The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

where

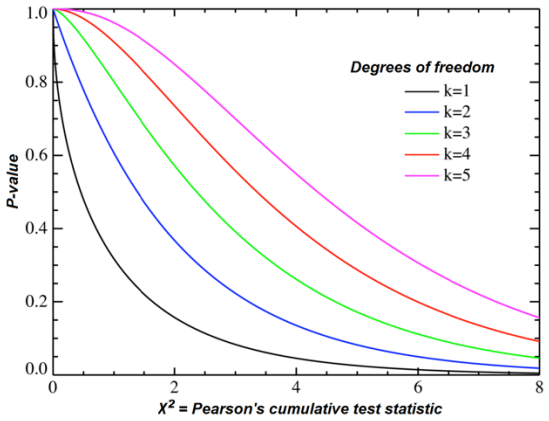
χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.

O_i = the number of observations of type *i*.

N = total number of observations

$E_i = Np_i$ = the expected (theoretical) frequency of type *i*, asserted by the null hypothesis that the fraction of type *i* in the population is p_i

n = the number of cells in the table.



$$P(\chi^2_P(\{p_i\}) > T) \sim C \int_{\sum_{i=1}^{m-1} y_i^2 > T} \left\{ \prod_{i=1}^{m-1} dy_i \right\} \prod_{i=1}^{m-1} \exp\left[-\frac{1}{2} \left(\sum_{i=1}^{m-1} y_i^2 \right) \right]$$

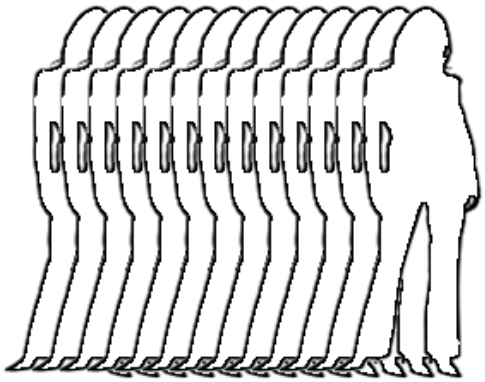
	has G	**Not G**	***Marginal Row Totals***
Cases	2104 (1912) [19.28]	1896 (2088) [17.66]	4000
Controls	2676 (2868) [12.85]	3324 (3132) [11.77]	6000
Marginal Column Totals	4780	5220	10000 (Grand Total)

Cases/hasG expected: 4000 * (4780/10000) = 1912 expected

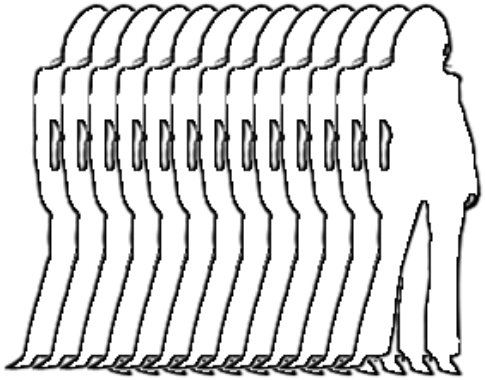
Cases/hasG squared deviation: $(2104 - 1912)^2 / 1912 = 19.28$ deviation

The chi-square statistic is 19.28+17.66+12.85+11.77 = 61.56. The p-value is 5e-15

Genome Wide Association (GWAS)



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

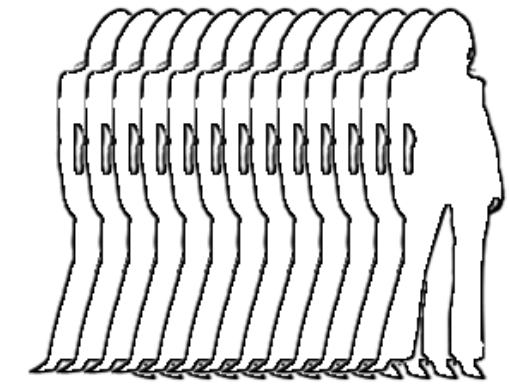


GC CC GC GC GG CC CC
CC GC GC GG GC GG

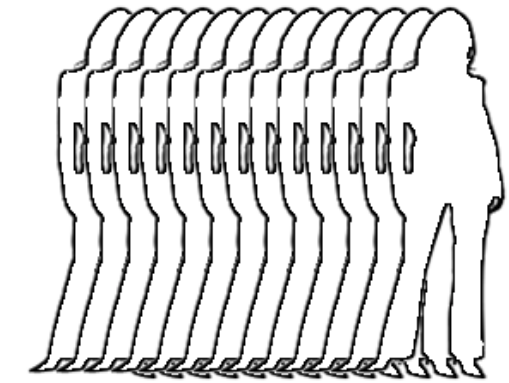
SNP1	*SNP2*	*SNP …*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:**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

Chi-squared or similar test

Genome Wide Association (GWAS)



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



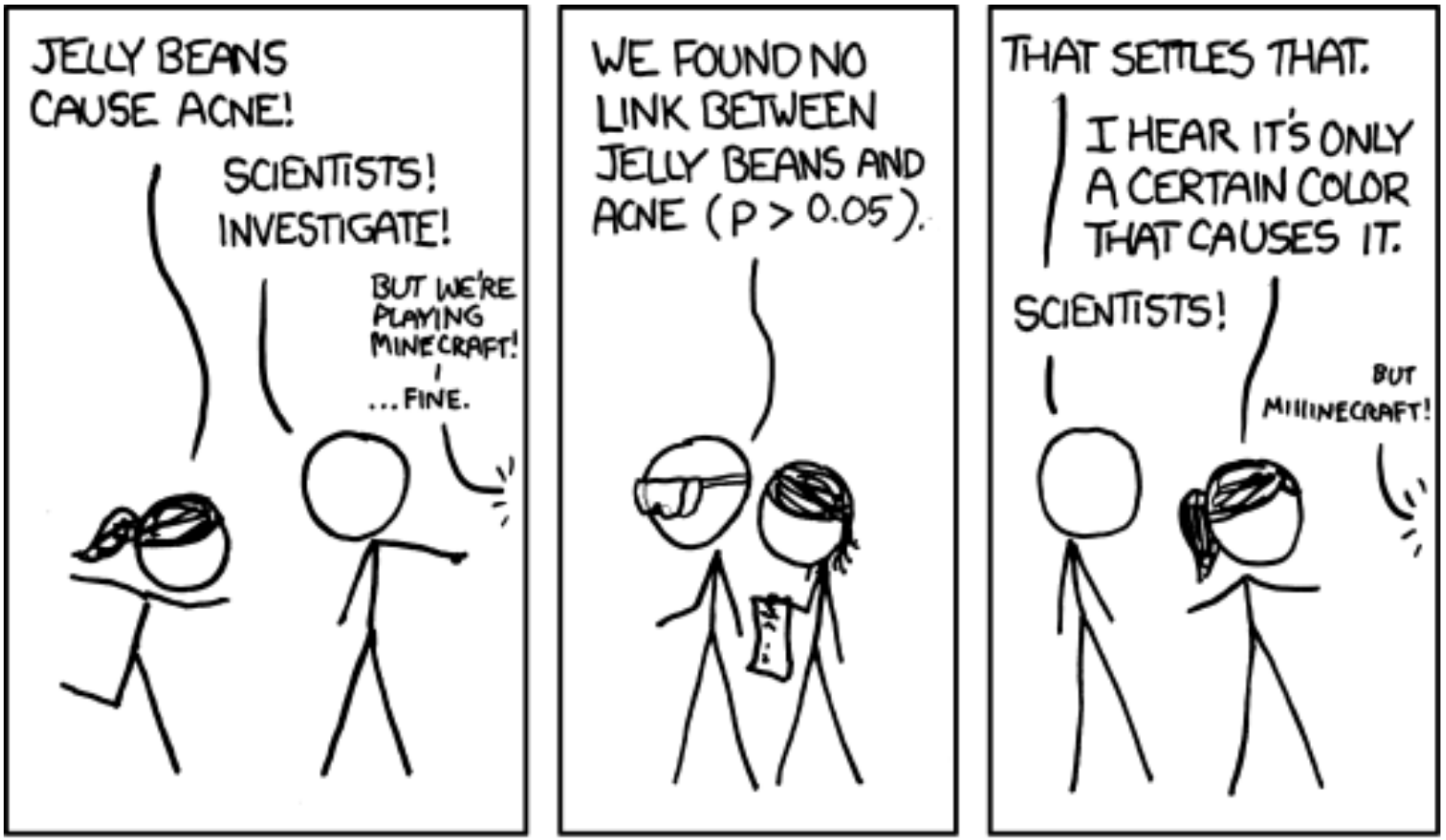
GC CC GC GC GG CC CC
CC GC GC GG GC GG

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

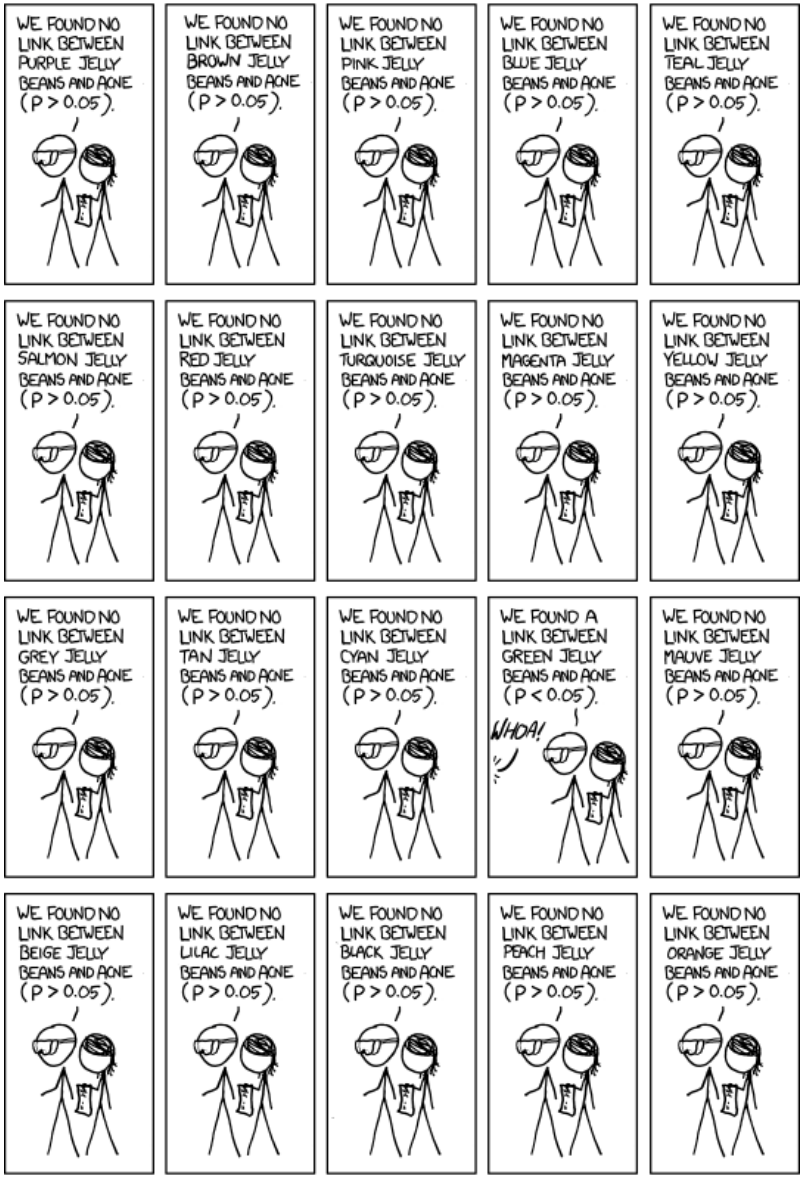
With a (much) larger population, this might be a significant difference in rate: 25320/60000 => p = 5e-7

Chi-squared or similar test

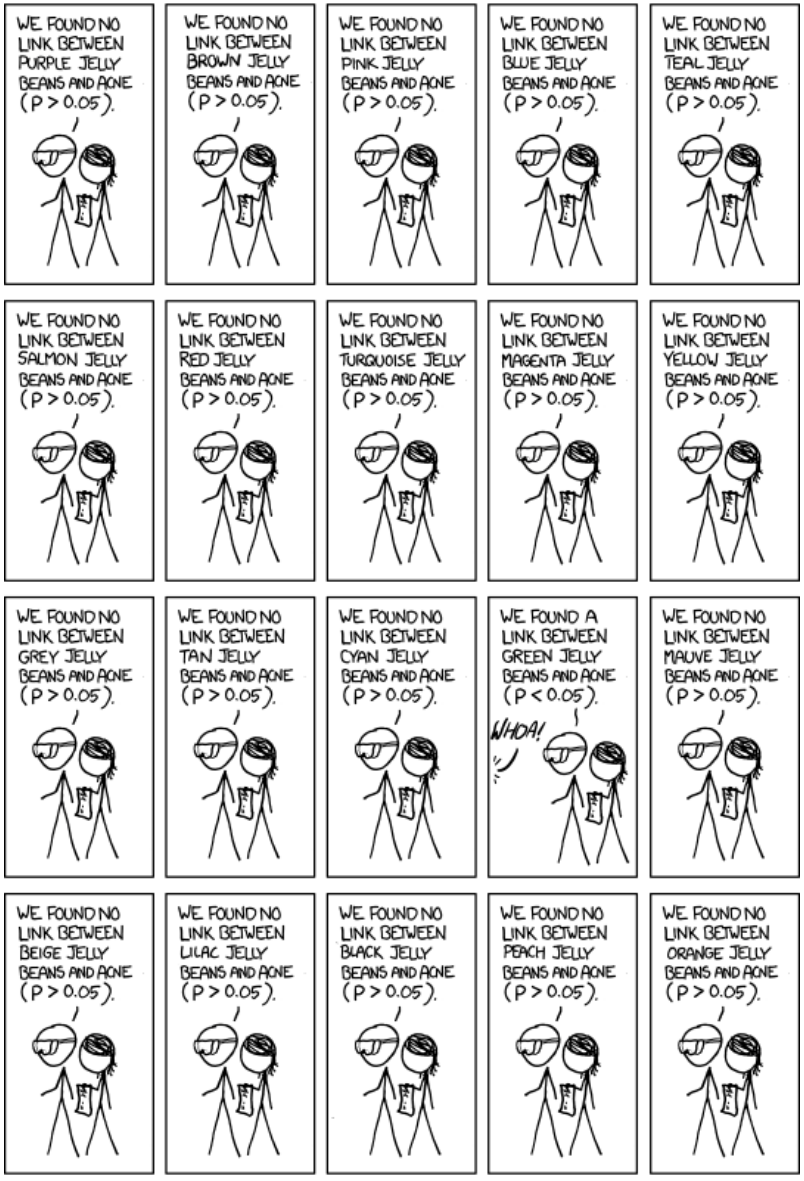
The curse of multiple testing

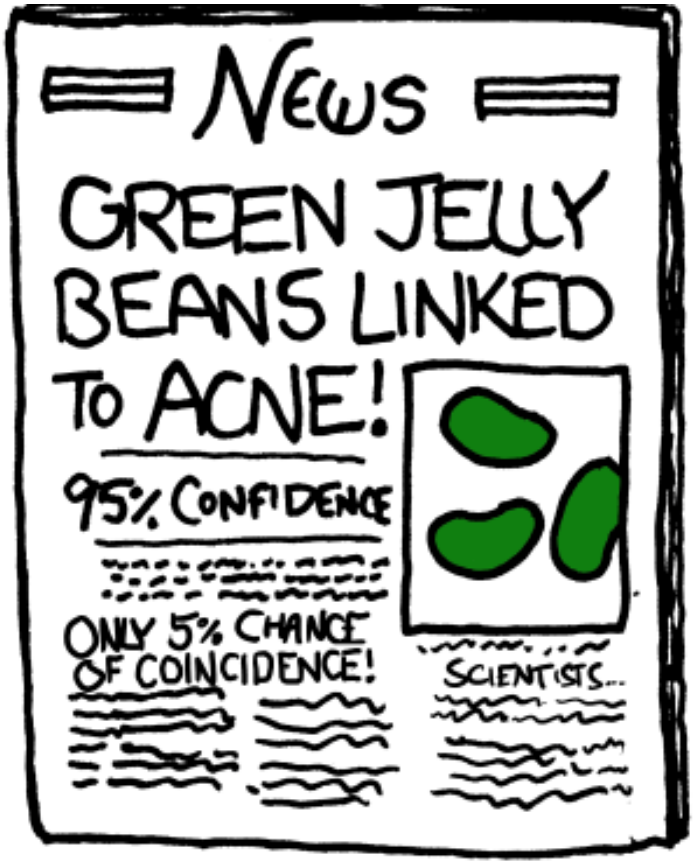


The curse of multiple testing

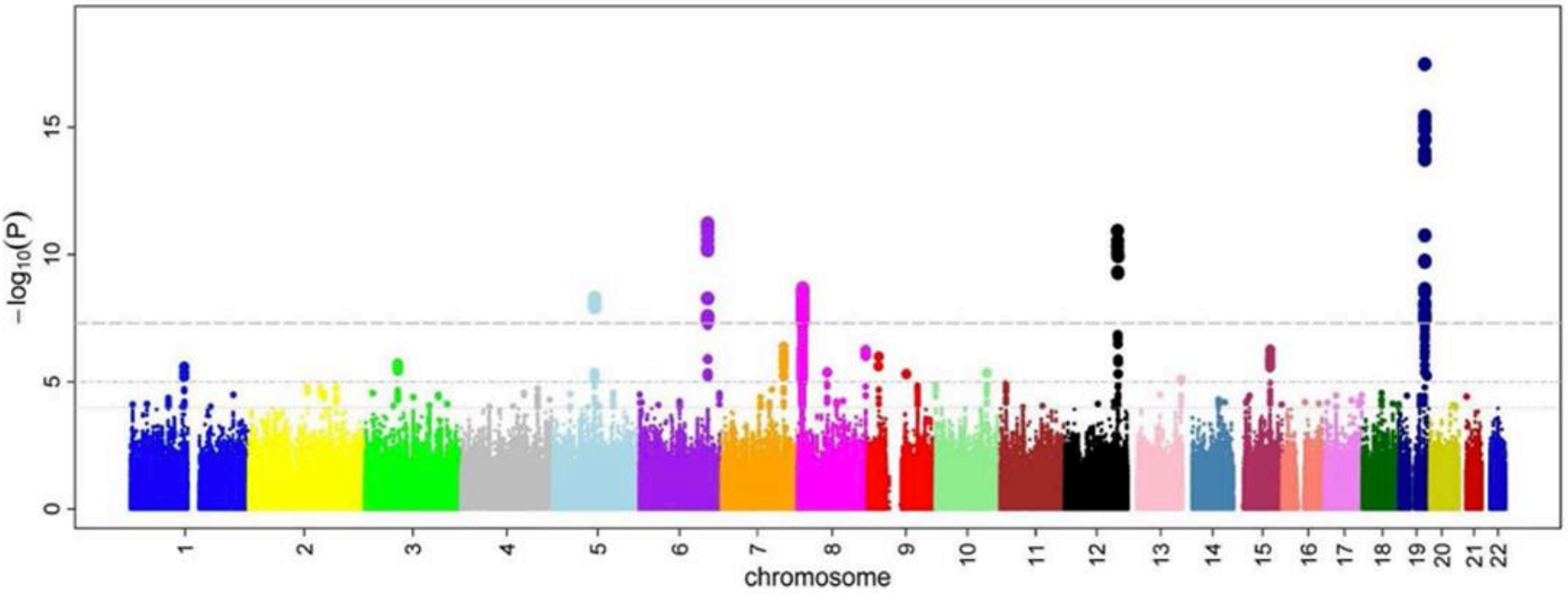


The curse of multiple testing



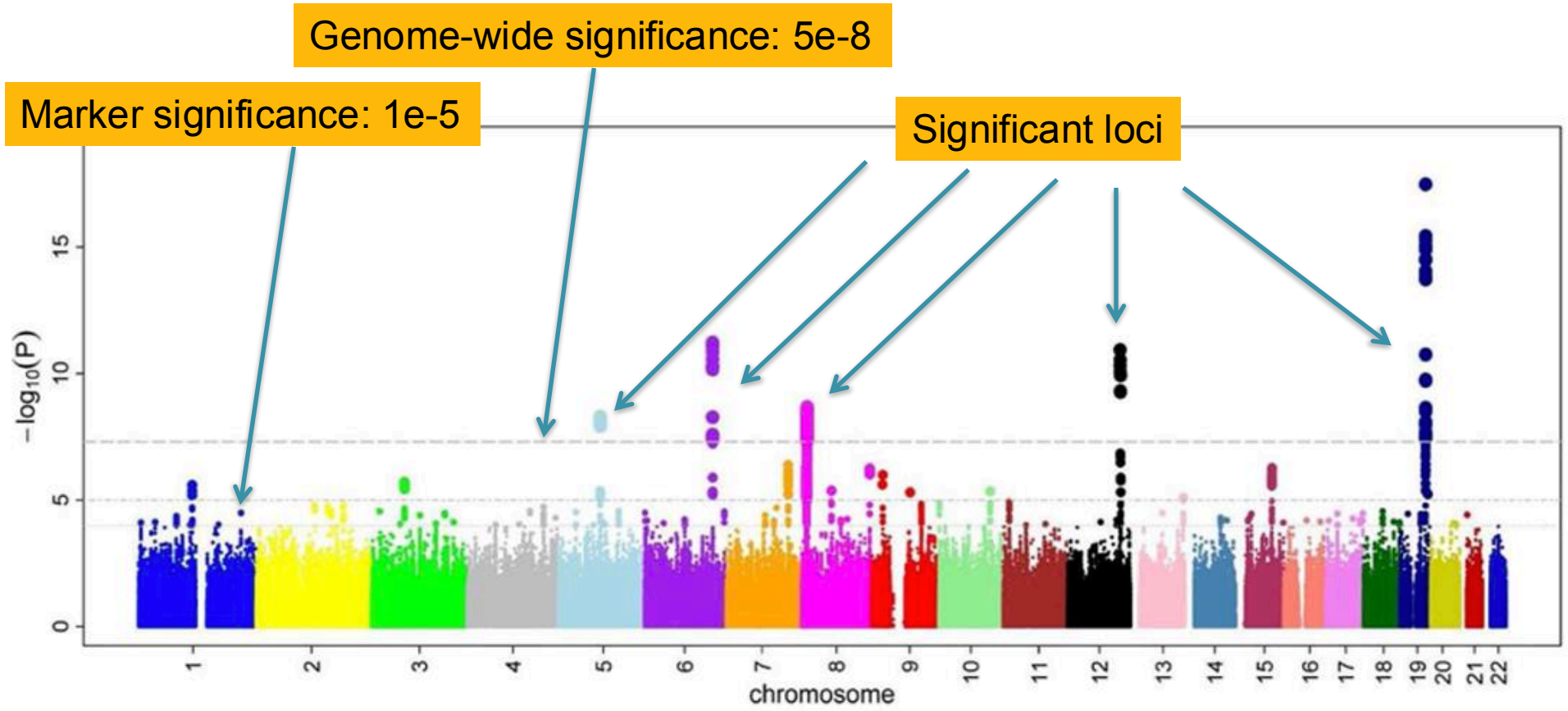


Manhattan Plot



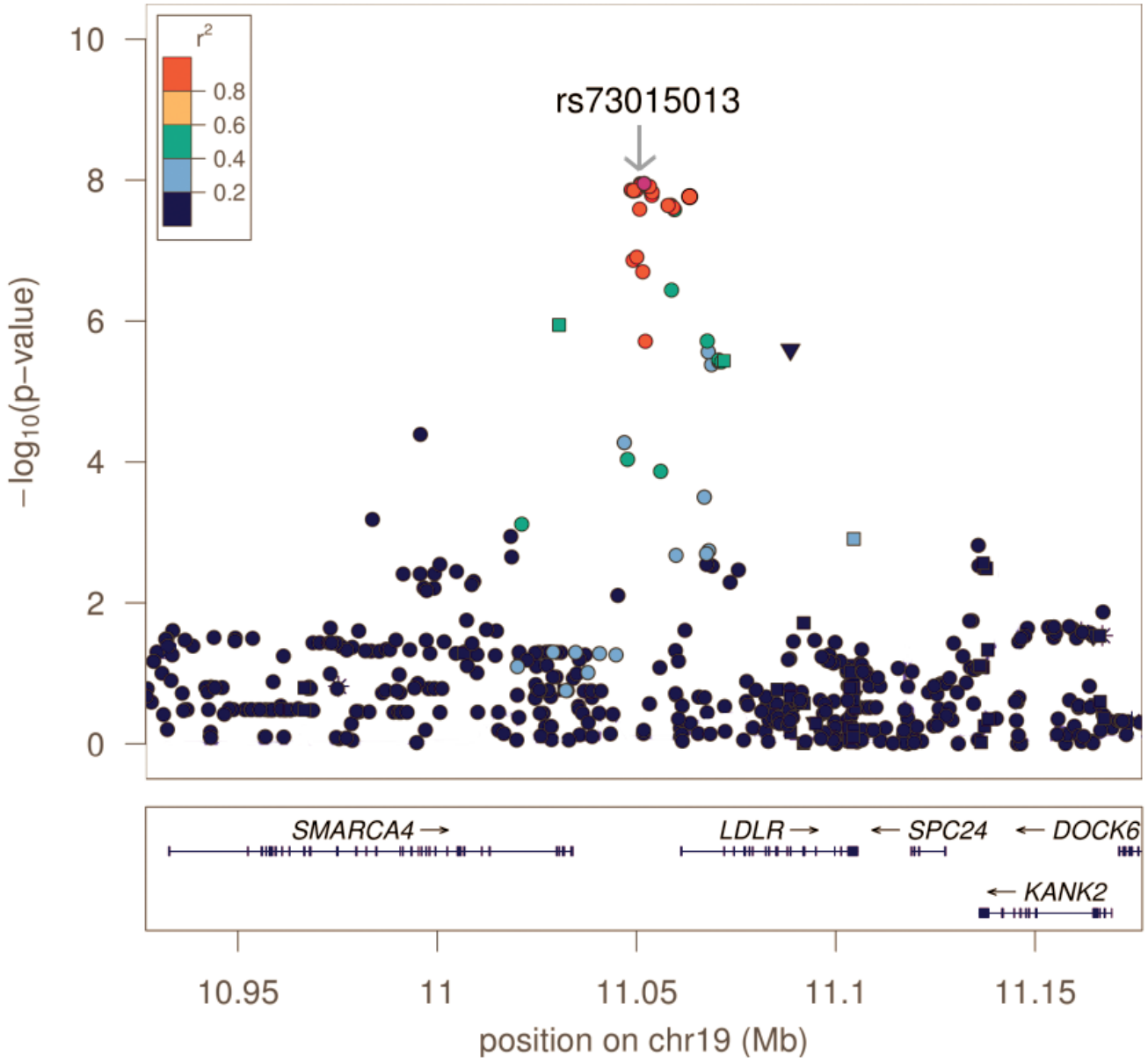
Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

Manhattan Plot

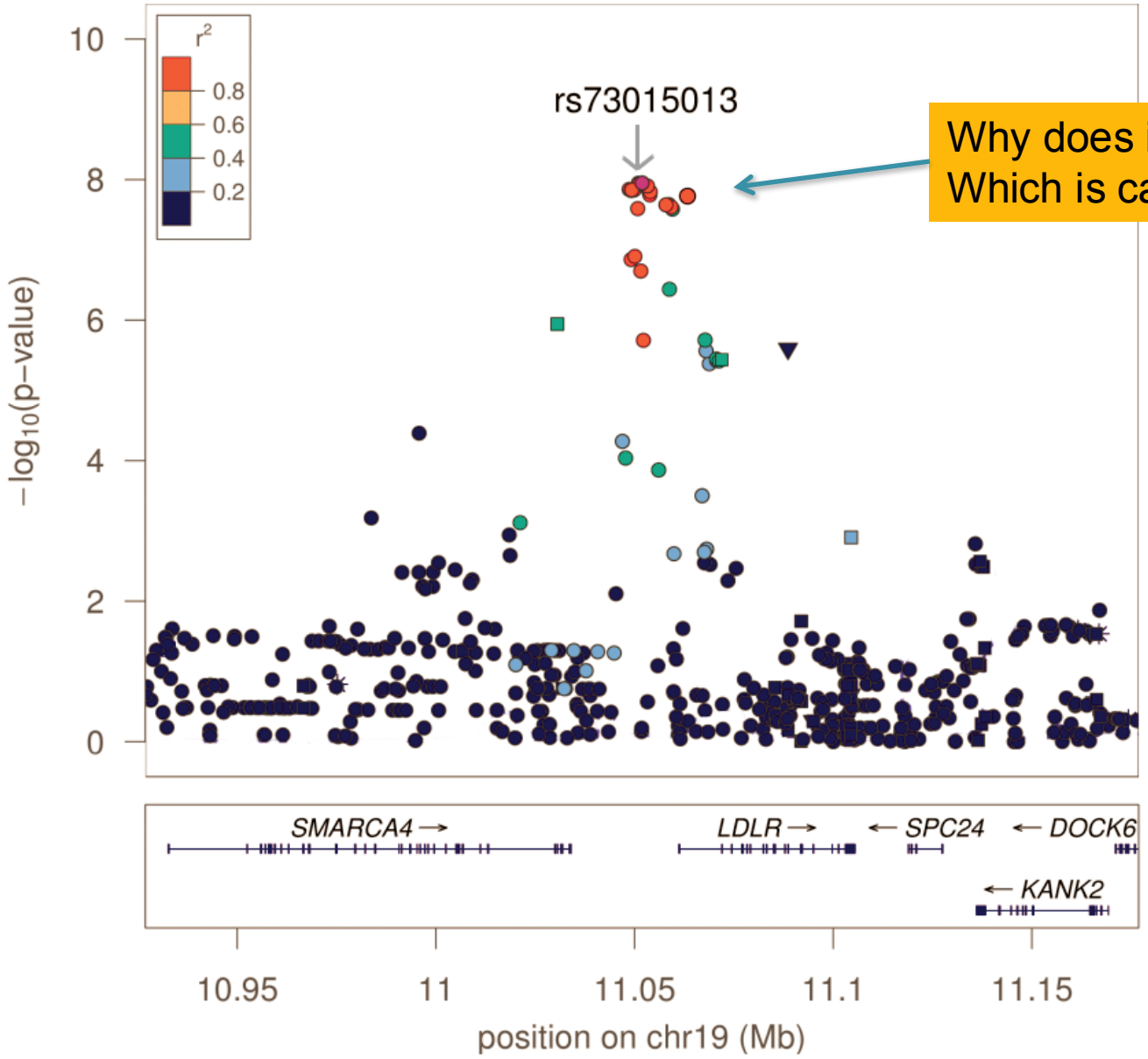


Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

Regional Association Plot



Regional Association Plot



First published GWAS

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world. Its incidence is increasing as the elderly population expands (*1*). AMD is characterized by progressive destruction of the retina's central region (macula), causing central field visual loss (*2*). A key feature of AMD is the formation of extracellular deposits called drusen concentrated in and around the macula behind the retina between the retinal pigment epithelium (RPE) and the choroid. To date, no therapy for this disease has proven to be broadly effective. Several risk factors have been linked to AMD, including age, smoking, and family history (*3*). Candidate-gene studies have not found any genetic differences that can account for a large proportion of the overall prevalence (*2*). Family-based whole-genome linkage scans have identified chromosomal regions that show evidence of linkage to AMD (*4–8*), but the linkage areas have not been resolved to any causative mutations.

Like many other chronic diseases, AMD is caused by a combination of genetic and environmental risk factors. Linkage studies are not as powerful as association studies for the identification of genes contributing to the risk for common, complex diseases (*9*). However, linkage studies have the advantage of searching the whole genome in an unbiased manner without presupposing the involvement of particular genes. Searching the whole genome in an association study requires typing 100,000 or more single-nucleotide polymorphisms (SNPs) (*10*). Because of these technical demands, only one whole-genome association study, on susceptibility to myocardial infarction, has been published to date (*11*).

Study design. We report a whole-genome case-control association study for genes involved in AMD. To maximize the chance of success, we chose clearly defined phenotypes for cases and controls. Case individuals exhibited at least some large drusen in a quantitative photographic assessment combined with evidence of sight-threatening AMD (geographic atrophy or neovascular AMD). Control individuals had either no or only a few small drusen. We analyzed our data using a statistically conservative approach to correct for the large number of SNPs tested, thereby guaranteeing that the probability of a false positive is no greater than our reported *P* values.

We used a subset of individuals who participated in the Age-Related Eye Disease Study (AREDS) (*12*). From the AREDS

[1]Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, NY 10021, USA. [2]Department of Ophthalmology and Visual Science, Yale University School of Medicine, 330 Cedar Street, New Haven, CT 06520, USA. [3]National Eye Institute, Building 10, CRC, 10 Center Drive, Bethesda, MD 20892–1204, USA. [4]Biological Imaging Core, National Eye Institute, 9000 Rockville Pike, Bethesda, MD 20892, USA. [5]The EMMES Corporation, 401 North Washington Street, Suite 700, Rockville MD 20850, USA. [6]W. M. Keck Facility, Yale University, 300 George Street, Suite 201, New Haven, CT 06511, USA. [7]Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: josephine.hoh@yale.edu

First published GWAS

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world. Its incidence is increasing as the elderly population expands (*1*). AMD is characterized by progressive destruction of the retina's central region (macula), causing central field visual loss (*2*). A key feature of AMD is the formation of extracellular deposits called drusen concentrated in and around the macula behind the retina between the retinal pigment epithelium (RPE) and the choroid. To date, no therapy for this disease has proven to be broadly effective. Several risk factors have been linked to AMD, including age, smoking, and family history (*3*). Candidate-gene studies have not found any genetic differences that ca account for a large proportion of the overa prevalence (*2*). Family-based whole-genom linkage scans have identified chromosom regions that show evidence of linkage t AMD (*4–8*), but the linkage areas have n been resolved to any causative mutations.

Like many other chronic diseases, AMD caused by a combination of genetic and env ronmental risk factors. Linkage studies are n as powerful as association studies for th identification of genes contributing to the ris for common, complex diseases (*9*). Howeve linkage studies have the advantage of searc ing the whole genome in an unbiased manne without presupposing the involvement of particular genes. Searching the whole genome in an association study requires typing 100,000 or more single-nucleotide polymorphisms (SNPs) (*10*). Because of these technical demands, only one whole-genome association study, on susceptibility to myocardial infarction, has been published to date (*11*).

Study design. We report a whole-genome case-control association study for genes involved in AMD. To maximize the chance of

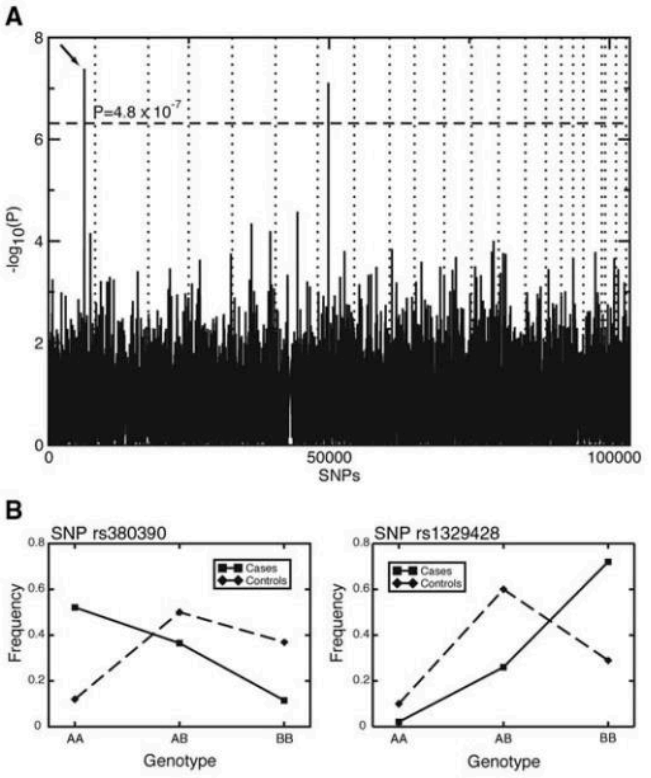


Fig. 1. (**A**) *P* values of genome-wide association scan for genes that affect the risk of developing AMD. $-\log_{10}(p)$ is plotted for each SNP in chromosomal order. The spacing between SNPs on the plot is uniform and does not reflect distances between SNPs on the chromosomes. The dotted horizontal line shows the cutoff for $P = 0.05$ after Bonferroni correction. The vertical dotted lines show chromosomal boundaries. The arrow indicates the peak for SNP rs380390, the most significant association, which was studied further. (**B**) Variation in genotype frequencies between cases and controls.

Polygenetic Risk Scores

nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0183-z

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4*]

A key public health need is to identify individuals at high risk for a given disease to enable enhanced screening or preventive therapies. Because most common diseases have a genetic component, one important approach is to stratify individuals based on inherited DNA variation[1]. Proposed clinical applications have largely focused on finding carriers of rare monogenic mutations at several-fold increased risk. Although most disease risk is polygenic in nature[2–5], it has not yet been possible to use polygenic predictors to identify individuals at risk comparable to monogenic mutations. Here, we develop and validate genome-wide polygenic scores for five common diseases. The approach identifies 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively. For coronary artery disease, this prevalence is 20-fold higher than the carrier frequency of rare monogenic mutations conferring comparable risk[6]. We propose that it is time to contemplate the inclusion of polygenic risk prediction in clinical care, and discuss relevant issues.

For various common diseases, genes have been identified in which rare mutations confer several-fold increased risk in heterozygous carriers. An important example is the presence of a familial hypercholesterolemia mutation in 0.4% of the population, which confers an up to threefold increased risk for coronary artery disease (CAD)[6]. Aggressive treatment to lower circulating cholesterol levels among such carriers can significantly reduce risk[7]. Another example is the p.Glu508Lys missense mutation in *HNF1A*, with a carrier frequency of 0.1% of the general population and 0.7% of Latinos[8], which confers up to fivefold increased risk for type 2 diabetes[9]. Although the ascertainment of monogenic mutations can be highly relevant for carriers and their families, the vast majority of disease occurs in those without such mutations.

For most common diseases, polygenic inheritance, involving many common genetic variants of small effect, plays a greater role than rare monogenic mutations[2–5]. However, it has been unclear whether it is possible to create a genome-wide polygenic score (GPS) to identify individuals at clinically significantly increased risk—for example, comparable to levels conferred by rare monogenic mutations[10,11].

Previous studies to create GPSs had only limited success, providing insufficient risk stratification for clinical utility (for example, identifying 20% of a population at 1.4-fold increased risk relative to the rest of the population)[12]. These initial efforts were hampered by three challenges: (1) the small size of initial genome-wide association studies (GWASs), which affected the precision of the estimated impact of individual variants on disease risk; (2) limited computational methods for creating GPSs; and (3) a lack of large datasets needed to validate and test GPS.

Using much larger studies and improved algorithms, we set out to revisit the question of whether a GPS can identify subgroups of the population with risk approaching or exceeding that of a monogenic mutation. We studied five common diseases with major public health impact: CAD, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer.

For each of the diseases, we created several candidate GPSs based on summary statistics and imputation from recent large GWASs in participants of primarily European ancestry (Table 1). Specifically, we derived 24 predictors based on a pruning and thresholding method, and 7 additional predictors using the recently described LDPred algorithm[13] (Methods, Fig. 1 and Supplementary Tables 1–6). These scores were validated and tested within the UK Biobank, which has aggregated genotype data and extensive phenotypic information on 409,258 participants of British ancestry (average age: 57 years; 55% female)[14,15].

We used an initial validation dataset of the 120,280 participants in the UK Biobank phase 1 genotype data release to select the GPSs with the best performance, defined as the maximum area under the receiver-operator curve (AUC). We then assessed the performance in an independent testing dataset comprised of the 288,978 participants in the UK Biobank phase 2 genotype data release. For each disease, the discriminative capacity within the testing dataset was nearly identical to that observed in the validation dataset.

Taking CAD as an example, our polygenic predictors were derived from a GWAS involving 184,305 participants[16] and evaluated based on their ability to detect the participants in the UK Biobank validation dataset diagnosed with CAD (Table 1). The predictors had AUCs ranging from 0.79–0.81 in the validation set, with the best predictor (GPS_{CAD}) involving 6,630,150 variants (Supplementary Table 1). This predictor performed equivalently well in the testing dataset, with an AUC of 0.81.

[1]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [2]Cardiology Division of the Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Cardiovascular Disease Initiative of the Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]These authors contributed equally: Amit V. Khera, Mark Chaffin. *e-mail: skathiresan1@mgh.harvard.edu

Polygenetic Risk Scores

nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0183-z

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4*]

A key public health need is to identify individuals at high risk for a given disease to enable enhanced screening or preventive therapies. Because most common diseases have a genetic component, one important approach is to stratify individuals based on inherited DNA variation[1]. Proposed clinical applications have largely focused on finding carriers of rare monogenic mutations at several-fold increased risk. Although most disease risk is polygenic in nature[2–5], it has not yet been possible to use polygenic predictors to identify individuals at risk comparable to monogenic mutations. Here, we develop and validate genome-wide polygenic scores for five common diseases. The approach identifies 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively. For coronary artery disease, this prevalence is 20-fold higher than the carrier frequency of rare monogenic mutations conferring comparable risk[4]. We propose that it is time to contemplate the inclusion of polygenic risk prediction in clinical care, and discuss relevant issues.

For various common diseases, genes have been identified in which rare mutations confer several-fold increased risk in heterozygous carriers. An important example is the presence of a familial hypercholesterolemia mutation in 0.4% of the population, which confers an up to threefold increased risk for coronary artery disease (CAD)[6]. Aggressive treatment to lower circulating cholesterol levels among such carriers can significantly reduce risk[7]. Another example is the p.Glu508Lys missense mutation in *HNF1A*, with a carrier frequency of 0.1% of the general population and 0.7% of Latinos[8], which confers up to fivefold increased risk for type 2 diabetes[9]. Although the ascertainment of monogenic mutations can be highly relevant for carriers and their families, the vast majority of disease occurs in those without such mutations.

For most common diseases, polygenic inheritance, involving many common genetic variants of small effect, plays a greater role than rare monogenic mutations[2–5]. However, it has been unclear whether it is possible to create a genome-wide polygenic score (GPS) to identify individuals at clinically significantly increased risk—for example, comparable to levels conferred by rare monogenic mutations[10,11].

Previous studies to create GPSs had only limited success, providing insufficient risk stratification for clinical utility (for example, identifying 20% of a population at 1.4-fold increased risk relative to the rest of the population)[12]. These initial efforts were hampered by three challenges: (1) the small size of initial genome-wide association studies (GWASs), which affected the precision of the estimated impact of individual variants on disease risk; (2) limited computational methods for creating GPSs; and (3) a lack of large datasets needed to validate and test GPS.

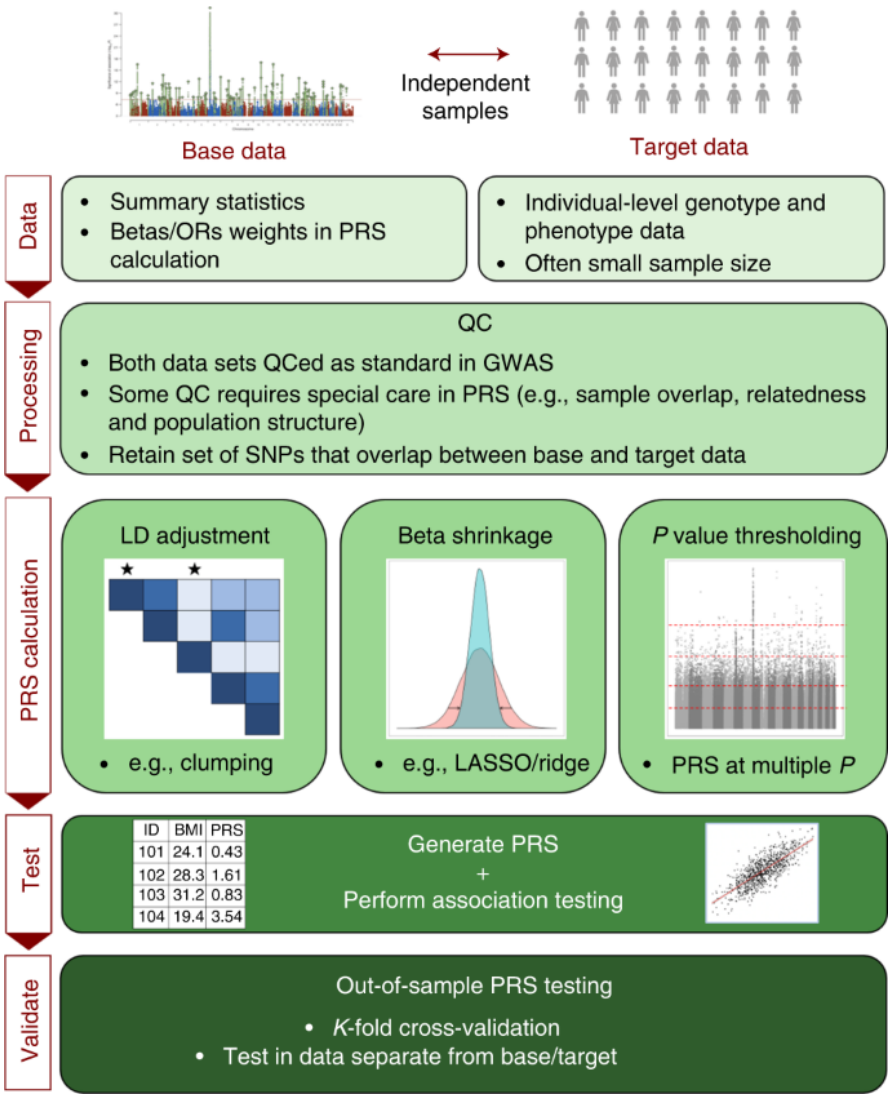
Using much larger studies and improved algorithms, we set out to revisit the question of whether a GPS can identify subgroups of the population with risk approaching or exceeding that of a monogenic mutation. We studied five common diseases with major public health impact: CAD, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer.

For each of the diseases, we created several candidate GPSs based on summary statistics and imputation from recent large GWASs in participants of primarily European ancestry (Table 1). Specifically, we derived 24 predictors based on a pruning and thresholding method, and 7 additional predictors using the recently described LDPred algorithm[13] (Methods, Fig. 1 and Supplementary Tables 1–6). These scores were validated and tested within the UK Biobank, which has aggregated genotype data and extensive phenotypic information on 409,258 participants of British ancestry (average age: 57 years; 55% female)[14,15].

We used an initial validation dataset of the 120,280 participants in the UK Biobank phase 1 genotype data release to select the GPSs with the best performance, defined as the maximum area under the receiver-operator curve (AUC). We then assessed the performance in an independent testing dataset comprised of the 288,978 participants in the UK Biobank phase 2 genotype data release. For each disease, the discriminative capacity within the testing dataset was nearly identical to that observed in the validation dataset.

Taking CAD as an example, our polygenic predictors were derived from a GWAS involving 184,305 participants[16] and evaluated based on their ability to detect the participants in the UK Biobank validation dataset diagnosed with CAD (Table 1). The predictors had AUCs ranging from 0.79–0.81 in the validation set, with the best predictor (GPS_{CAD}) involving 6,630,150 variants (Supplementary Table 1). This predictor performed equivalently well in the testing dataset, with an AUC of 0.81.

[1]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [2]Cardiology Division of the Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Cardiovascular Disease Initiative of the Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]These authors contributed equally: Amit V. Khera, Mark Chaffin. *e-mail: skathiresan1@mgh.harvard.edu



Tutorial: a guide to performing polygenic risk score analyses

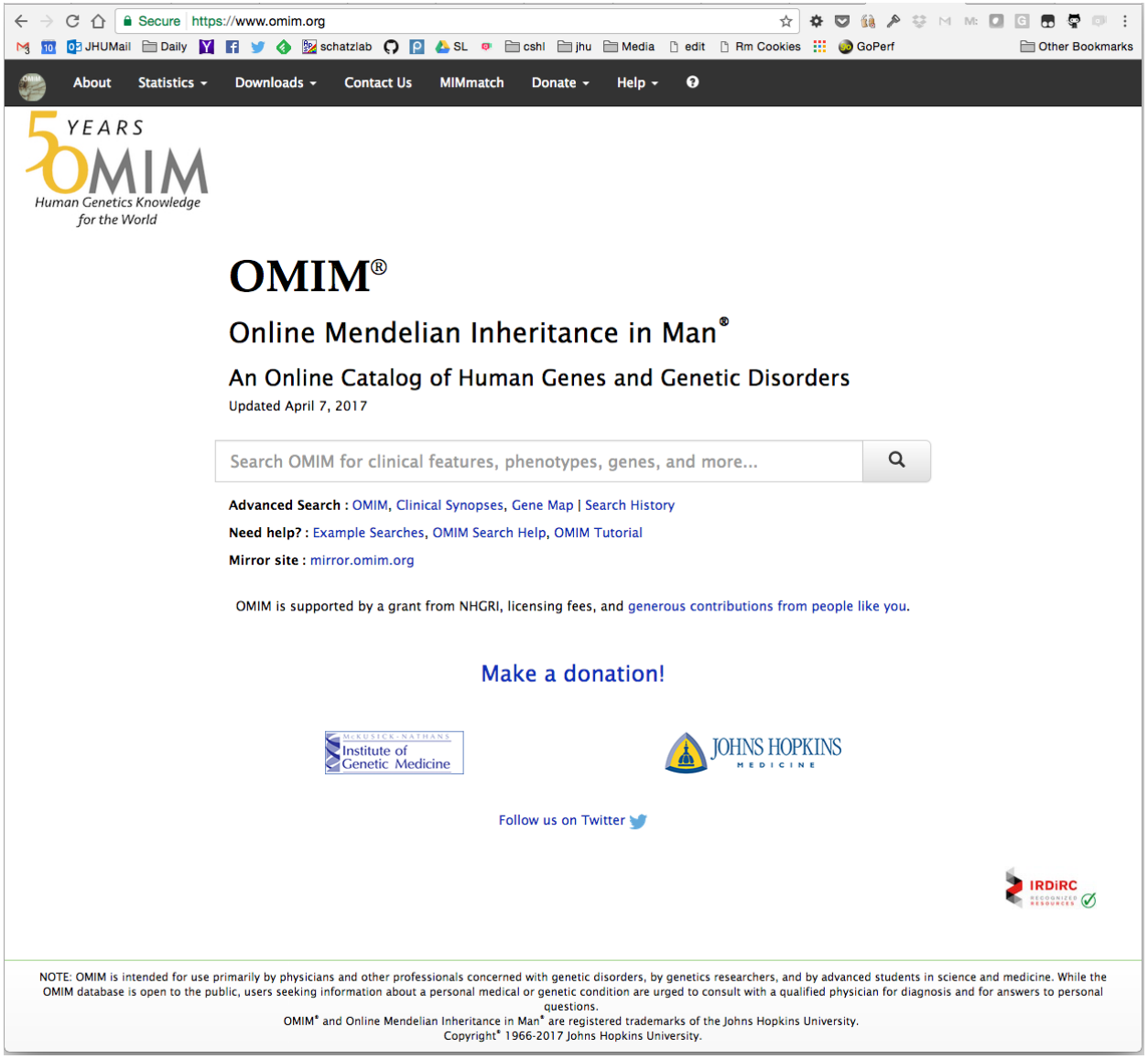
Choi et al (2020) Nature Protocols. https://doi.org/10.1038/s41596-020-0353-1

GWAS Catalog

As of 2024-10-21, the GWAS Catalog contains 7046 publications, 688,820 top associations and 99,586 full summary statistics.

http://www.ebi.ac.uk/gwas/diagram

OMIM



- 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 ☺

ARTICLE

doi:10.1038/nature13595

Biological insights from 108 schizophrenia-associated genetic loci

Schizophrenia Working Group of the Psychiatric Genomics Consortium*

Schizophrenia is a highly heritable disorder. Genetic risk is conferred by a large number of alleles, including common alleles of small effect that might be detected by genome-wide association studies. Here we report a multi-stage schizophrenia genome-wide association study of up to 36,989 cases and 113,075 controls. We identify 128 independent associations spanning 108 conservatively defined loci that meet genome-wide significance, 83 of which have not been previously reported. Associations were enriched among genes expressed in brain, providing biological plausibility for the findings. Many findings have the potential to provide entirely new insights into aetiology, but associations at *DRD2* and several genes involved in glutamatergic neurotransmission highlight molecules of known and potential therapeutic relevance to schizophrenia, and are consistent with leading pathophysiological hypotheses. Independent of genes expressed in brain, associations were enriched among genes expressed in tissues that have important roles in immunity, providing support for the speculated link between the immune system and schizophrenia.

Nature 511, 421–427 (24 July 2014) doi:10.1038/nature13595

ARTICLE

doi:10.1038/nature13595

Biolog… schiz…

Schizophrenia W…

Schizophrenia… alleles of smal… phrenia genor… ciations span… previously rep… the findings. … and several ge… relevance to sc… in brain, assoc… support for th…

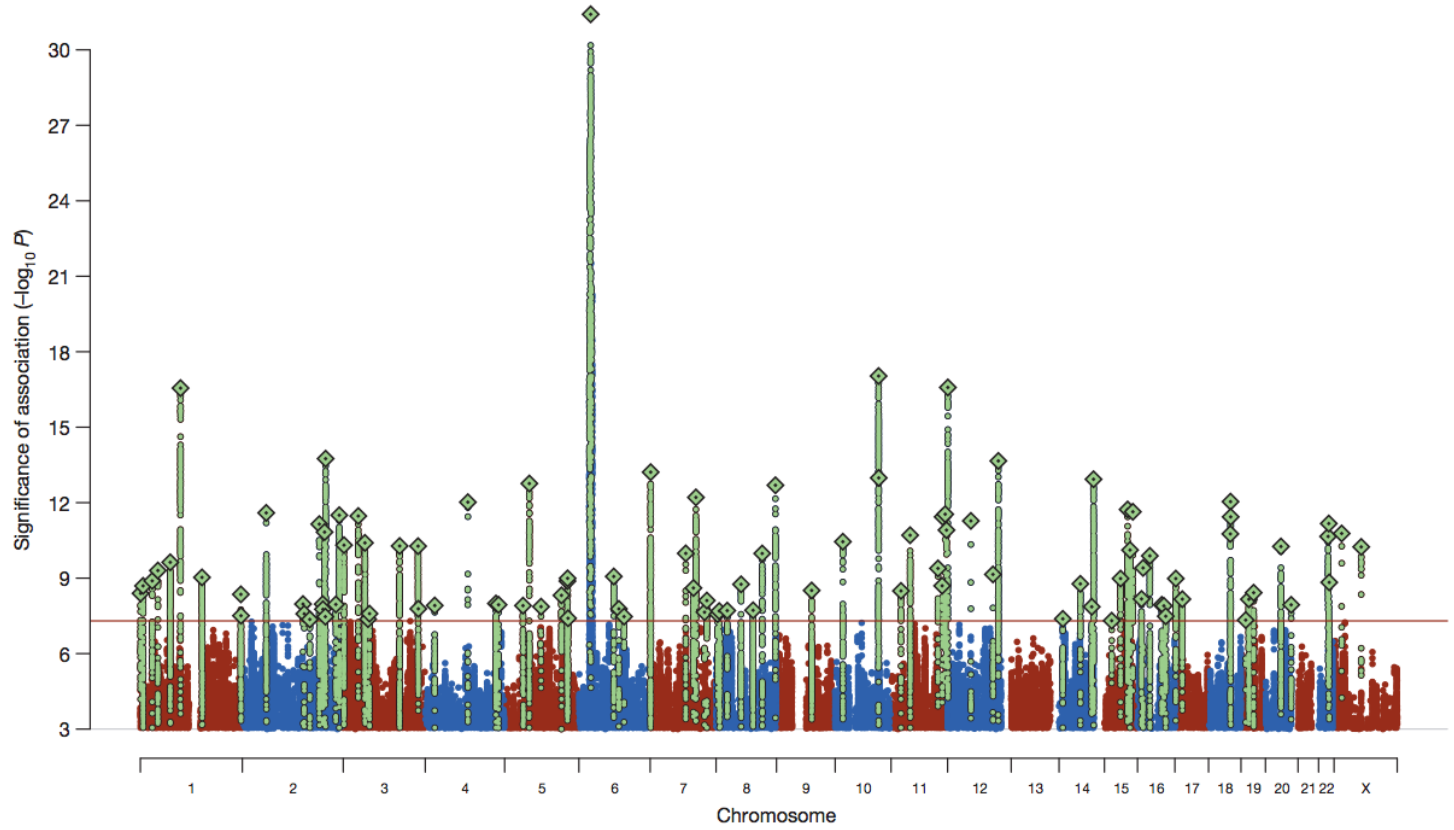


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal position and the *y* axis is the significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

ARTICLE

doi:10.1038/nature13595

Biolog… schiz…

Schizophrenia W…

Schizophrenia… alleles of smal… phrenia genor… ciations span… previously rep… the findings. … and several ge… relevance to sc… in brain, assoc… support for th…

rs115329265: A -> G substitution
86.4% of cases!

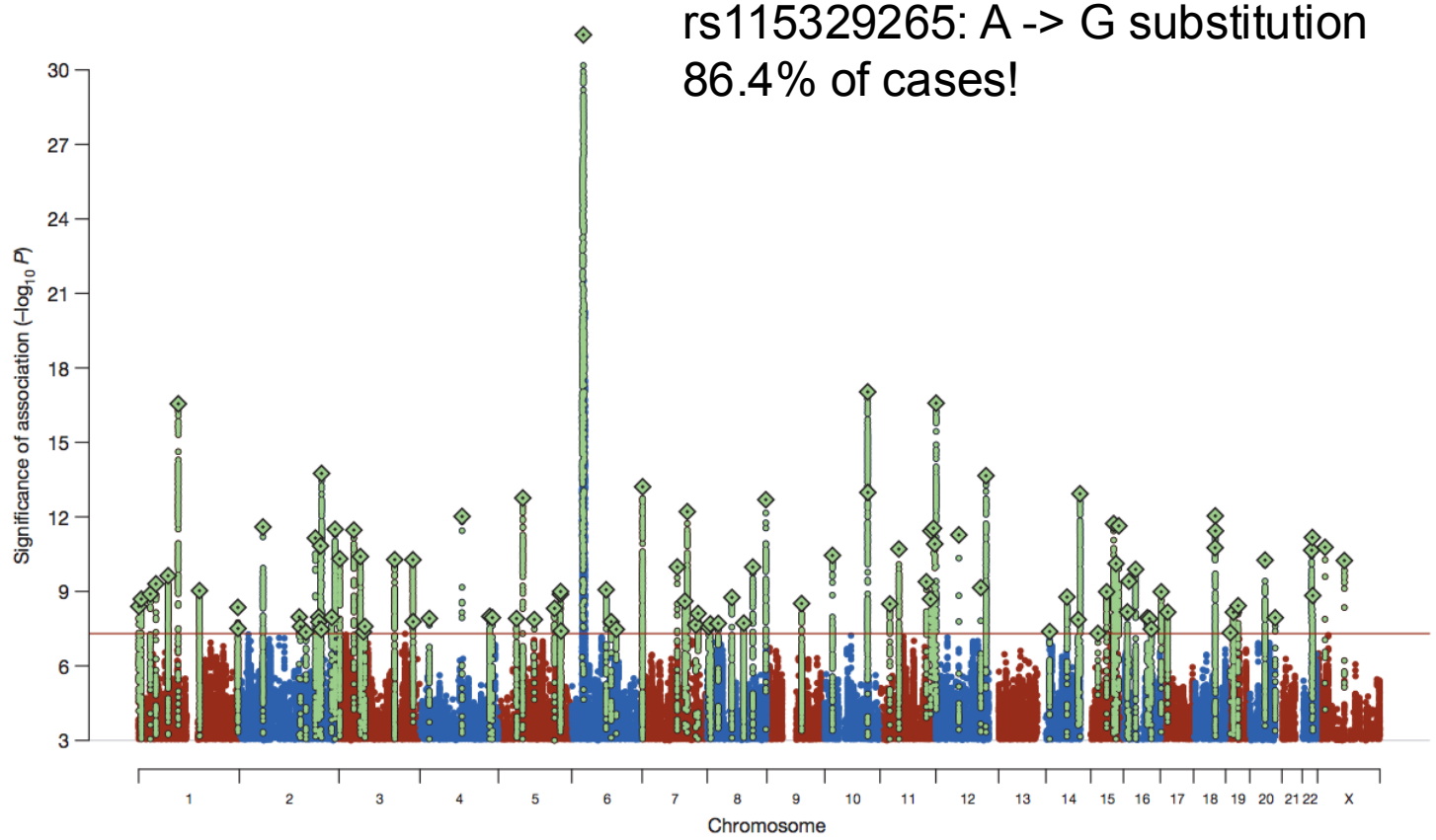


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal position and the *y* axis is the significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

Nature 511, 421–427 (24 July 2014) doi:10.1038/nature13595

ARTICLE

doi:10.1038/nature13595

Biolog… schiz…

Schizophrenia W…

Schizophrenia… alleles of smal… phrenia genor… ciations spann… previously rep… the findings. M… and several ge… relevance to sc… in brain, assoc… support for th…

rs115329265: A -> G substitution
86.4% of cases!
But 85.0% of controls!?!
:-(

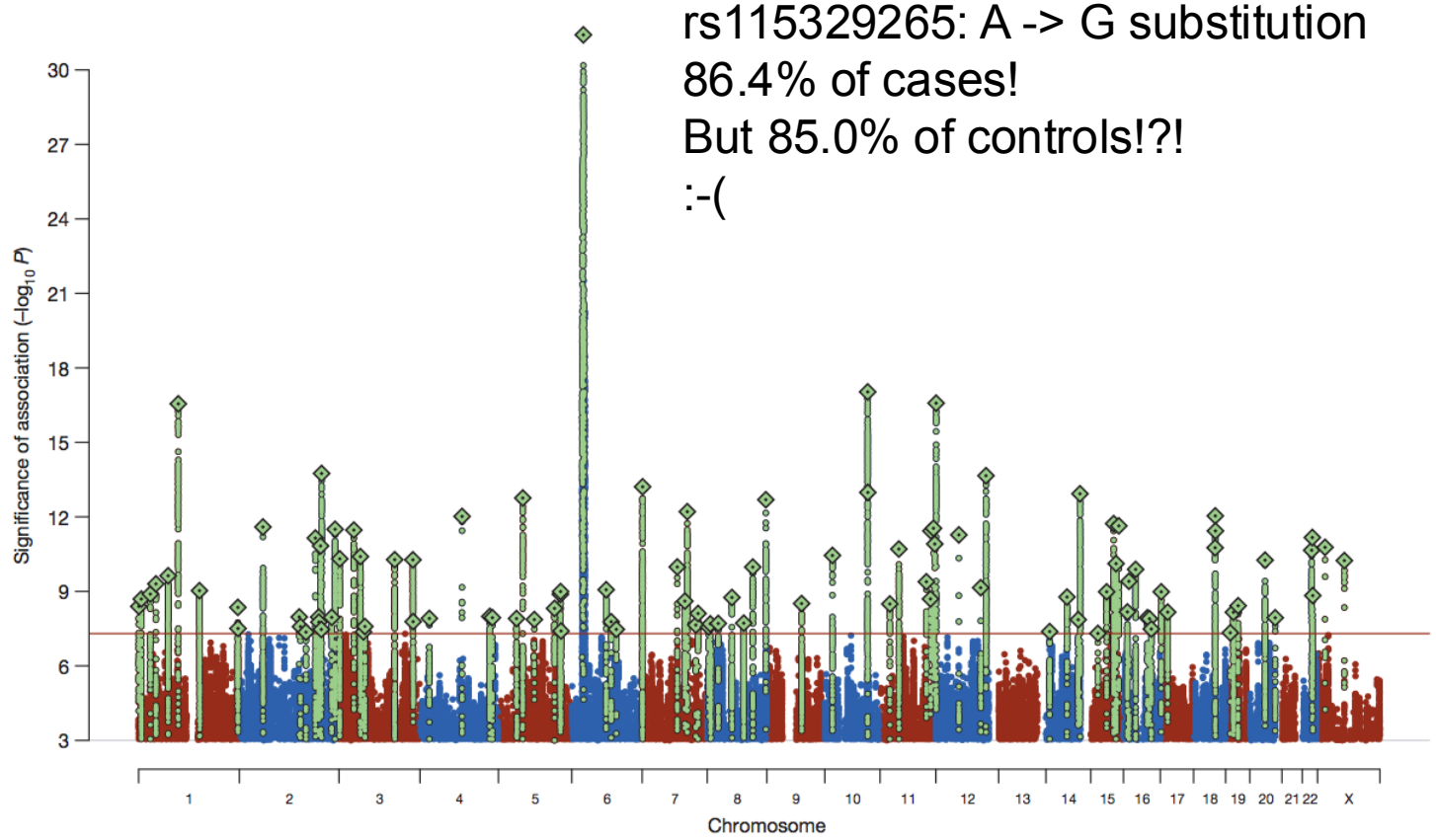


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal position and the *y* axis is the significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

Nature 511, 421–427 (24 July 2014) doi:10.1038/nature13595

ARTICLE

doi:10.1038/nature13595

Compared to the brains of healthy individuals, those of people with schizophrenia have higher expression of a gene called *C4*, according to a paper published in Nature today (January 27). The gene encodes an immune protein that moonlights in the brain as an eradicator of unwanted neural connections (synapses). The findings, which suggest increased synaptic pruning is a feature of the disease, are a direct extension of genome-wide association studies (GWASs) that pointed to the major histocompatibility (MHC) locus as a key region associated with schizophrenia risk.

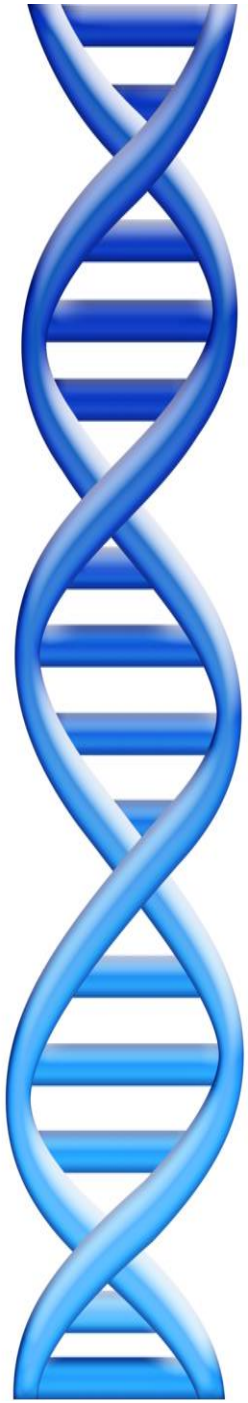
"The MHC [locus] is the first and the strongest genetic association for schizophrenia, but many people have said this finding is not useful," said psychiatric geneticist Patrick Sullivan of the University of North Carolina School of Medicine who was not involved in the study.

-Ruth Williams, The Scientist

plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal

derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

Nature 511, 421–427 (24 July 2014) doi:10.1038/nature13595



Part V:

Post-GWAS Genomic Medicine

Needles in stacks of needles

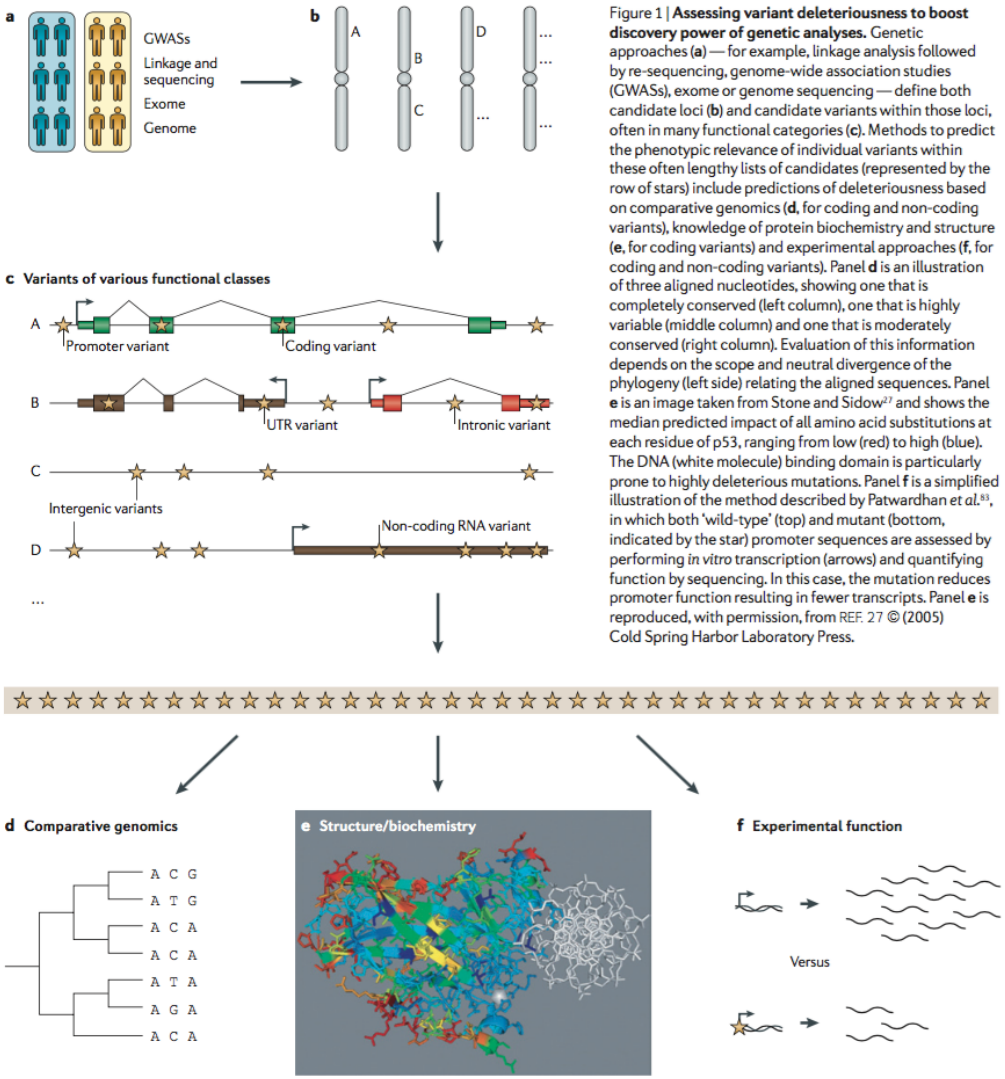


Figure 1 | **Assessing variant deleteriousness to boost discovery power of genetic analyses.** Genetic approaches (**a**) — for example, linkage analysis followed by re-sequencing, genome-wide association studies (GWASs), exome or genome sequencing — define both candidate loci (**b**) and candidate variants 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.

Methods

Predicting Deleterious Amino Acid Substitutions

Pauline C. Ng[1,2] and Steven Henikoff[1,3,4]

[1]Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA; [2]Department of Bioengineering, University of Washington, Seattle, Washington 98105, USA; [3]Howard Hughes Medical Institute, Seattle, Washington 98109, USA

Many missense substitutions are identified in single nucleotide polymorphism (SNP) data and large-scale random mutagenesis projects. Each amino acid substitution potentially affects protein function. We have constructed a tool that uses sequence homology to predict whether a substitution affects protein function. SIFT, which sorts intolerant from tolerant substitutions, classifies substitutions as tolerated or deleterious. A higher proportion of substitutions predicted to be deleterious by SIFT gives an affected phenotype than substitutions predicted to be deleterious by substitution scoring matrices in three test cases. Using SIFT before mutagenesis studies could reduce the number of functional assays required and yield a higher proportion of affected phenotypes. SIFT may be used to identify plausible disease candidates among the SNPs that cause missense substitutions.

SIFT Key Idea: Substituting one amino acid for another with another with very similar biochemical properties is probably less significant that a more dissimilar substitution. Learn those similarities by comparing orthologs across species

Genome Research. 2001 May;11(5):863-74.

Resource

A probabilistic disease-gene finder for personal genomes

Mark Yandell,[1,3,4] Chad Huff,[1,3] Hao Hu,[1,3] Marc Singleton,[1] Barry Moore,[1] Jinchuan Xing,[1] Lynn B. Jorde,[1] and Martin G. Reese[2]

[1]*Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah and School of Medicine, Salt Lake City, Utah 84112, USA;* [2]*Omicia, Inc., Emeryville, California 94608, USA*

VAAST (the Variant Annotation, Analysis & Search Tool) is a probabilistic search tool for identifying damaged genes and their disease-causing variants in personal genome sequences. VAAST builds on existing amino acid substitution (AAS) and aggregative approaches to variant prioritization, combining elements of both into a single unified likelihood framework that allows users to identify damaged genes and deleterious variants with greater accuracy, and in an easy-to-use fashion. VAAST can score both coding and noncoding variants, evaluating the cumulative impact of both types of variants simultaneously. VAAST can identify rare variants causing rare genetic diseases, and it can also use both rare and common variants to identify genes responsible for common diseases. VAAST thus has a much greater scope of use than any existing methodology. Here we demonstrate its ability to identify damaged genes using small cohorts ($n = 3$) of unrelated individuals, wherein no two share the same deleterious variants, and for common, multigenic diseases using as few as 150 cases.

[Supplemental material is available for this article.]

VAAST Key Idea: Evaluate amino acid substitutions in evolution AND allele frequencies in 1000 genomes project

Genome Research 2011. doi:10.1101/gr.123158.111

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

A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko[1], Melissa J Hubisz[2], Ilan Gronau[2,3] & Adam Siepel[1–3]

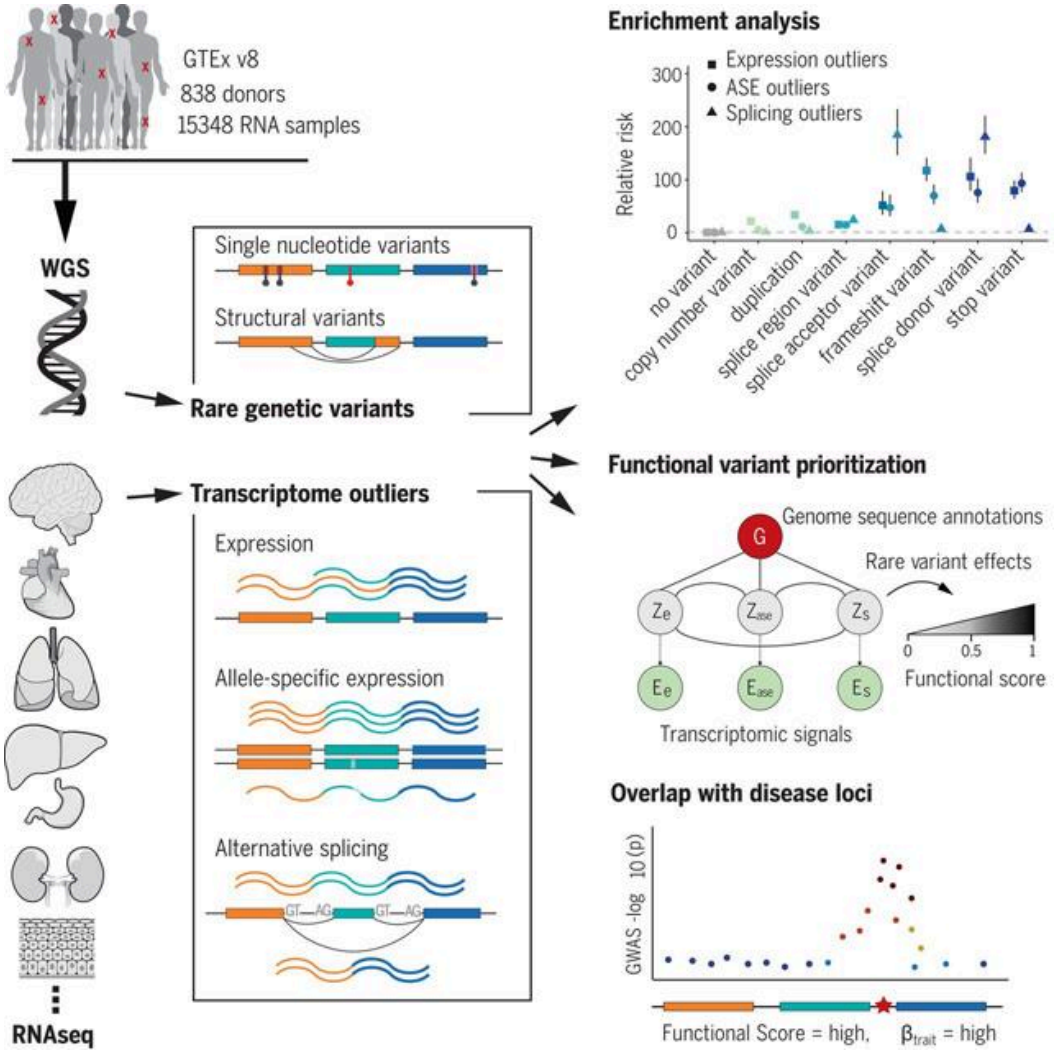
We describe a new computational method for estimating the probability that a point mutation at each position in a genome will influence fitness. These 'fitness consequence' (fitCons) scores serve as evolution-based measures of potential genomic function. Our approach is to cluster genomic positions into groups exhibiting distinct 'fingerprints' on the basis of high-throughput functional genomic data, then to estimate a probability of fitness consequences for each group from associated patterns of genetic polymorphism and divergence. We have generated fitCons scores for three human cell types on the basis of public data from ENCODE. In comparison with conventional conservation scores, fitCons scores show considerably improved prediction power for *cis* regulatory elements. In addition, fitCons scores indicate that 4.2–7.5% of nucleotides in the human genome have influenced fitness since the human-chimpanzee divergence, and they suggest that recent evolutionary turnover has had limited impact on the functional content of the genome.

roles[16–19] by getting at fitness directly through observations of evolutionary change. In essence, the 'experiment' considered by these methods is the one conducted directly on genomes by nature over millennia, and the outcomes of interest are the presence or absence of fixed mutations.

These conservation-based methods, however, depend critically on the assumption that genomic elements are present at orthologous locations and maintain similar functional roles over relatively long evolutionary time periods. Evolutionary turnover may cause inconsistencies between sequence orthology and functional homology that substantially limit this type of analysis. Consequently, investigators have developed two major alternative strategies for the identification and characterization of functional elements. The first strategy is to augment information about interspecies conservation with information about genetic polymorphism[20–28]. The shorter evolutionary time scales associated with intraspecies variation make this approach more robust to evolutionary turnover and less sensitive to errors in alignment and orthology detection. Polymorphic sites tend to be sparse along the genome, however, so this approach requires some type

fitCons Key Idea: Evaluate amino acid substitutions AND allele frequencies in 1000 genomes project AND aggregate by ENCODE regions

Nature Genetics 47, 276–283 (2015) doi:10.1038/ng.3196



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