

# Lecture 20. Disease Genetics

Michael Schatz

April 6, 2020

JHU 600.749: Applied Comparative Genomics



# Preliminary Project Report

---

Assignment Date: March 30, 2019

Due Date: Monday, April 13, 2019 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Monday April 13.

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
- 5+ References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at [https://academic.oup.com/bioinformatics/pages/submission\\_online](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.

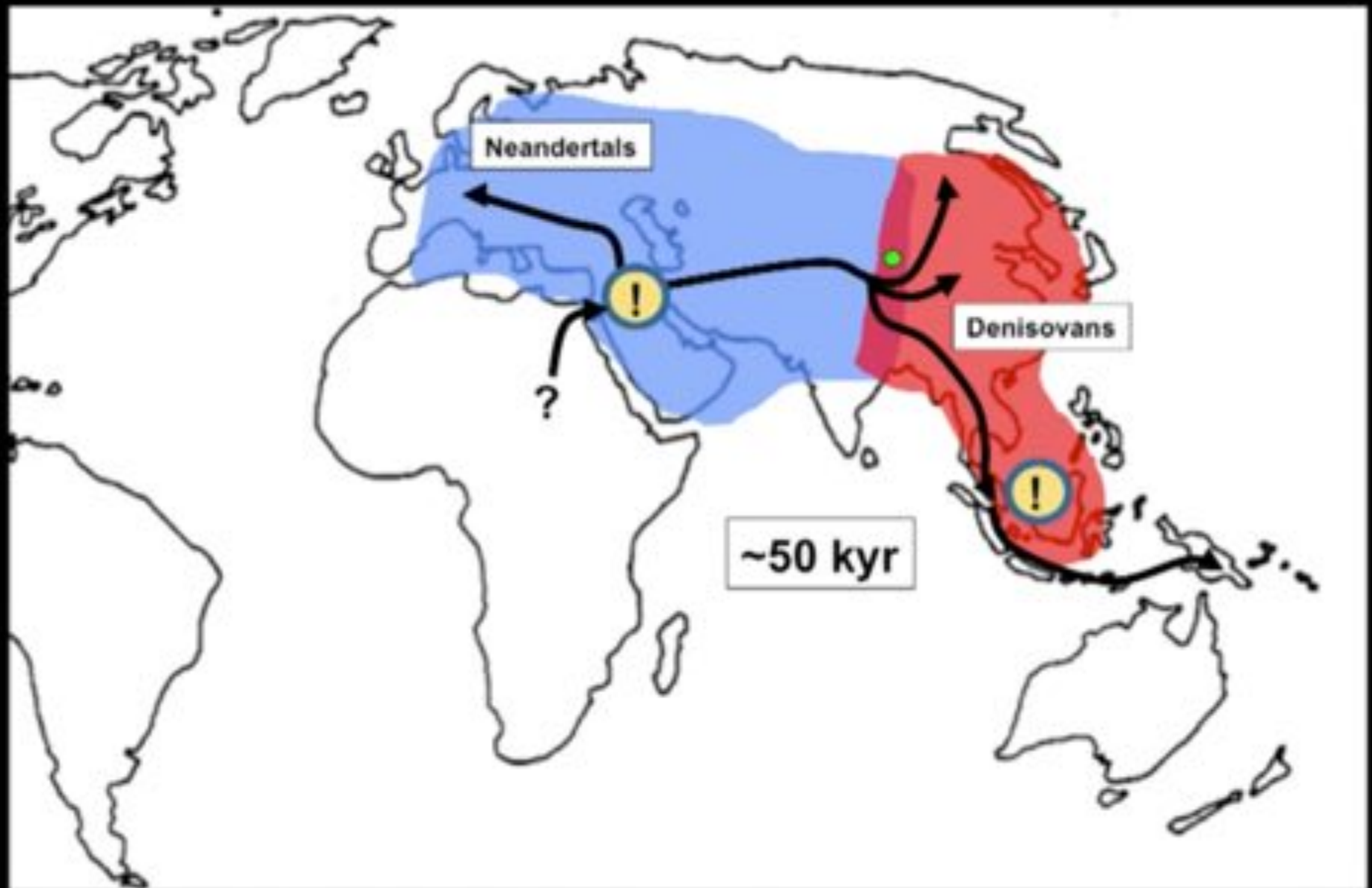
Later, you will present your project in class starting the week of April 22. You will also submit your final written report (5-7 pages) of your project by May 13

Please use Piazza if you have any general questions!



# Part I: Ancient Hominds

# Timeline of ancient hominids





# Part II: Modern Humans

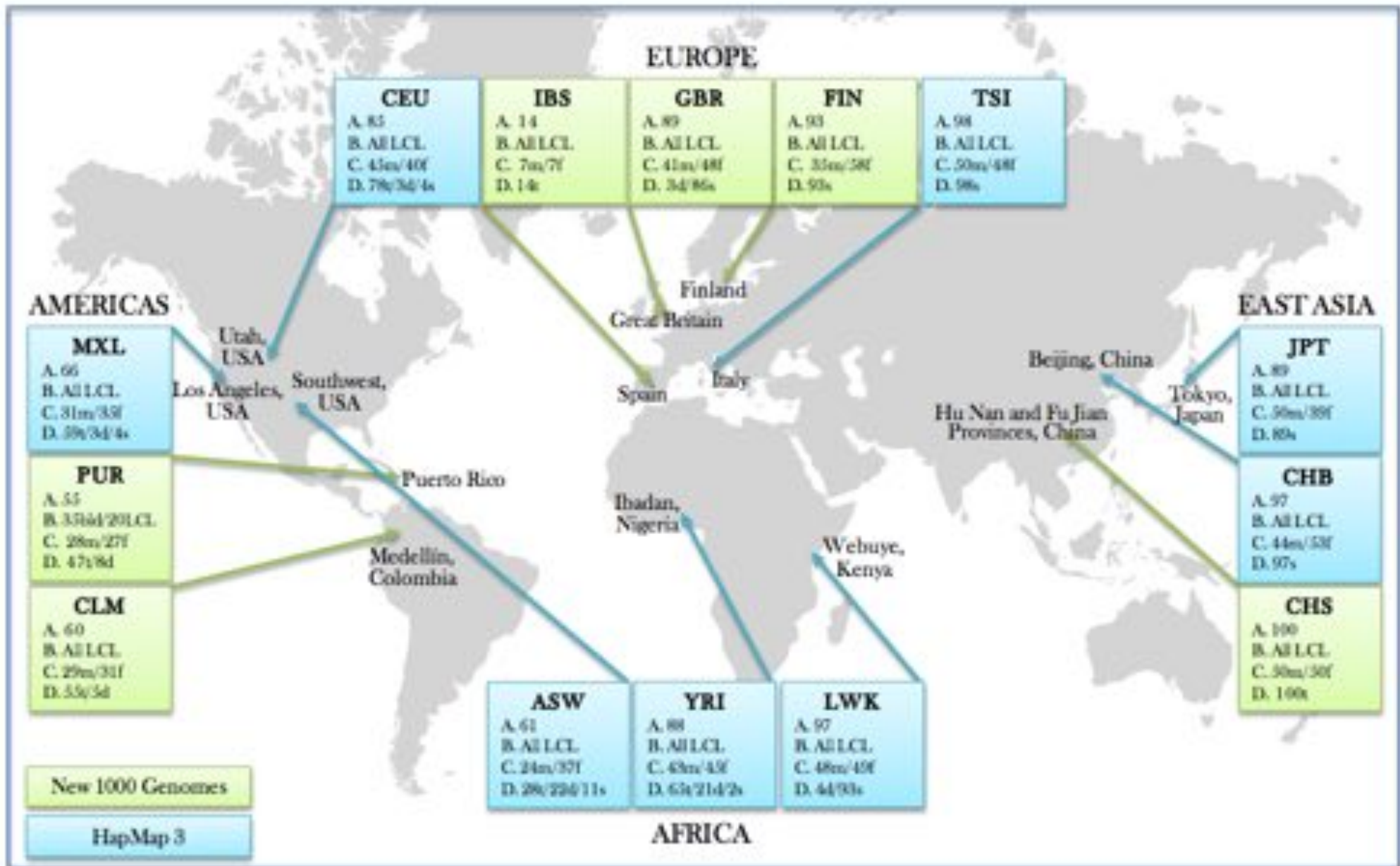


## An integrated map of genetic variation from 1,092 human genomes

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.

# 1000 Genomes Populations



# 1000 Genomes Populations

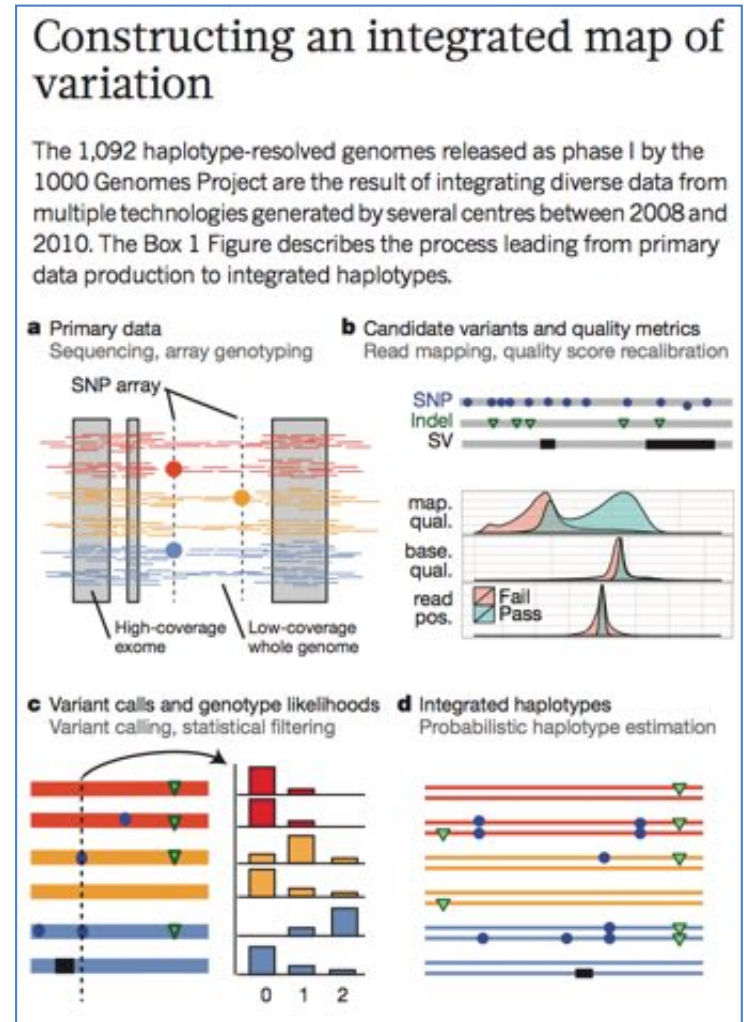
Population	DNA sequenced from blood	Offspring Samples from Trios Available	Pilot Samples	Phase 1 Samples	Final Phase Discovery Sample	Final Release Sample	Total
Chinese Dai in Xishuangbanna, China (CDX)	no	yes	0	0	99	93	99
Han Chinese in Beijing, China (CHB)	no	no	91	97	103	103	306
Japanese in Tokyo, Japan (JPT)	no	no	94	89	104	104	301
Kinh in Ho Chi Minh City, Vietnam (KHV)	yes	yes	0	0	101	99	101
Southern Han Chinese, China (CHS)	no	yes	0	100	108	105	112
<b>Total East Asian Ancestry (EAS)</b>			<b>185</b>	<b>286</b>	<b>515</b>	<b>504</b>	<b>833</b>
Bengali in Bangladesh (BEB)	no	yes	0	0	86	86	86
Gujarati Indian in Houston, TX (GHI)	no	yes	0	0	106	103	106
Indian Telugu in the UK (ITU)	yes	yes	0	0	103	102	103
Punjabi in Lahore, Pakistan (PJL)	yes	yes	0	0	96	96	96
Sri Lankan Tamil in the UK (STU)	yes	yes	0	0	103	102	103
<b>Total South Asian Ancestry (SAS)</b>			<b>0</b>	<b>0</b>	<b>494</b>	<b>489</b>	<b>494</b>
African Ancestry in Southwest US (ASW)	no	yes	0	61	66	62	66
African Caribbean in Barbados (ACB)	yes	yes	0	0	96	96	96
Esan in Nigeria (ESN)	no	yes	0	0	99	99	99
Gambian in Western Division, The Gambia (GWD)	no	yes	0	0	113	113	113
Luhya in Webuye, Kenya (LWK)	no	yes	102	97	101	99	116
Mende in Sierra Leone (MSL)	no	yes	0	0	85	85	85
Yoruba in Ibadan, Nigeria (YRI)	no	yes	106	88	109	108	116
<b>Total African Ancestry (AFR)</b>			<b>208</b>	<b>246</b>	<b>609</b>	<b>601</b>	<b>691</b>
British in England and Scotland (GBR)	no	yes	0	89	92	90	94
Finnish in Finland (FIN)	no	no	0	93	99	99	100
Iberian populations in Spain (IBS)	no	yes	0	14	107	107	107
Toscani in Italy (TSI)	no	no	66	98	108	107	110
Utah residents with Northern and Western European ancestry (CEU)	no	yes	94	85	99	99	103
<b>Total European Ancestry (EUR)</b>			<b>160</b>	<b>379</b>	<b>505</b>	<b>503</b>	<b>834</b>
Colombian in Medellin, Colombia (CLM)	no	yes	0	60	94	94	95
Mexican Ancestry in Los Angeles, California (MXL)	no	yes	0	66	67	64	69
Peruvian in Lima, Peru (PEL)	yes	yes	0	0	86	85	86
Puerto Rican in Puerto Rico (PUR)	yes	yes	0	55	103	104	105
<b>Total Americas Ancestry (AMR)</b>				<b>181</b>	<b>312</b>	<b>347</b>	<b>398</b>
<b>Total</b>			<b>343</b>	<b>1092</b>	<b>2830</b>	<b>2564</b>	<b>2877</b>

26 populations from 5 major population groups



# 1000 Genomes: Human Mutation Rate

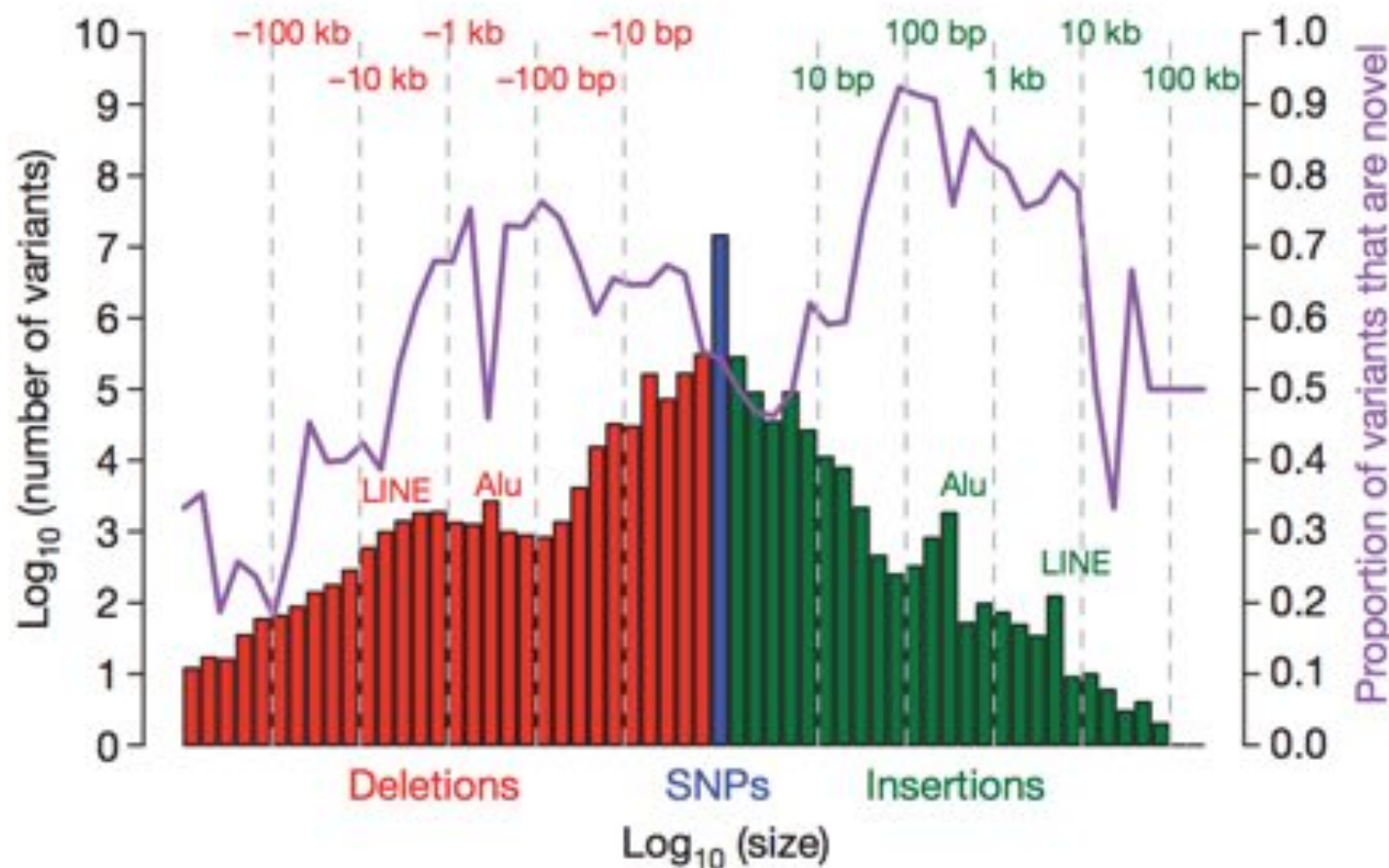
- Phase I Release
  - 1092 individuals from 14 populations
  - Combination of low coverage WGS, deep coverage WES, and SNP genotype data
- Overall SNP rate between any two people is ~1/1200bp to ~1/1300
  - ~3M SNPs between me and you (.1%)
  - ~30M SNPs between human to Chimpanzees (1%)
- De novo mutation rate ~1/100,000,000
  - ~100 de novo mutations from generation to generation
  - ~1-2 de novo mutations within the protein coding genes



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

1000 genomes project (2012) *Nature*. doi:10.1038/nature11632

# 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

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

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

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.



# Homozygous LoF Mutations

## LETTER

doi:10.1038/nature22034

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

Danish Saleheen<sup>1,2\*</sup>, Pradeep Natarajan<sup>1,3\*</sup>, Irina M. Armean<sup>4,5</sup>, Wei Zhao<sup>6</sup>, Asif Rashood<sup>7</sup>, Sarmeet A. Khetarpal<sup>8</sup>, Hong-Hee Won<sup>9</sup>, Konrad I. Karczowski<sup>4,5</sup>, Anne H. O'Donnell-Luria<sup>4,5,6</sup>, Kaitlin E. Samocha<sup>4,5</sup>, Benjamin Weissbourd<sup>4,5</sup>, Namrata Gupta<sup>4</sup>, Moazzam Zaidi<sup>7</sup>, Maria Samuel<sup>7</sup>, Aatif Imran<sup>7</sup>, Shahid Abbas<sup>8</sup>, Faisal Majeed<sup>7</sup>, Madiha Ishaq<sup>7</sup>, Saba Akhtar<sup>8</sup>, Kevin Trindade<sup>9</sup>, Megan Mucksavage<sup>9</sup>, Nadeem Qamar<sup>10</sup>, Khan Shah Zaman<sup>10</sup>, Zia Yaqoob<sup>10</sup>, Tahir Saghir<sup>10</sup>, Syed Nadeem Hasan Rizvi<sup>10</sup>, Anis Merroun<sup>10</sup>, Nadeem Hayyat Malik<sup>11</sup>, Mohammad Ishaq<sup>12</sup>, Syed Zahed Rashood<sup>12</sup>, Fazal-ur-Rehman Memon<sup>13</sup>, Khalid Mahmood<sup>14</sup>, Naveeduddin Ahmed<sup>15</sup>, Ren Do<sup>16,17</sup>, Ronald M. Krauss<sup>18</sup>, Daniel G. MacArthur<sup>4,5</sup>, Stacey Gabriel<sup>4</sup>, Eric S. Lander<sup>4</sup>, Mark I. Daly<sup>4,5</sup>, Philippe Froggatt<sup>19</sup>, John Danesh<sup>19,20</sup>, Daniel I. Rader<sup>4,20</sup> & Sekar Kathiresan<sup>21,22</sup>

A major goal of biomedicine is to understand the function of every gene in the human genome<sup>1</sup>. 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<sup>2</sup>. 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<sup>3</sup>. 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 apolipoprotein-containing lipoprotein subfractions; at either *AIGAL12* 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<sup>4,5</sup>; 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).

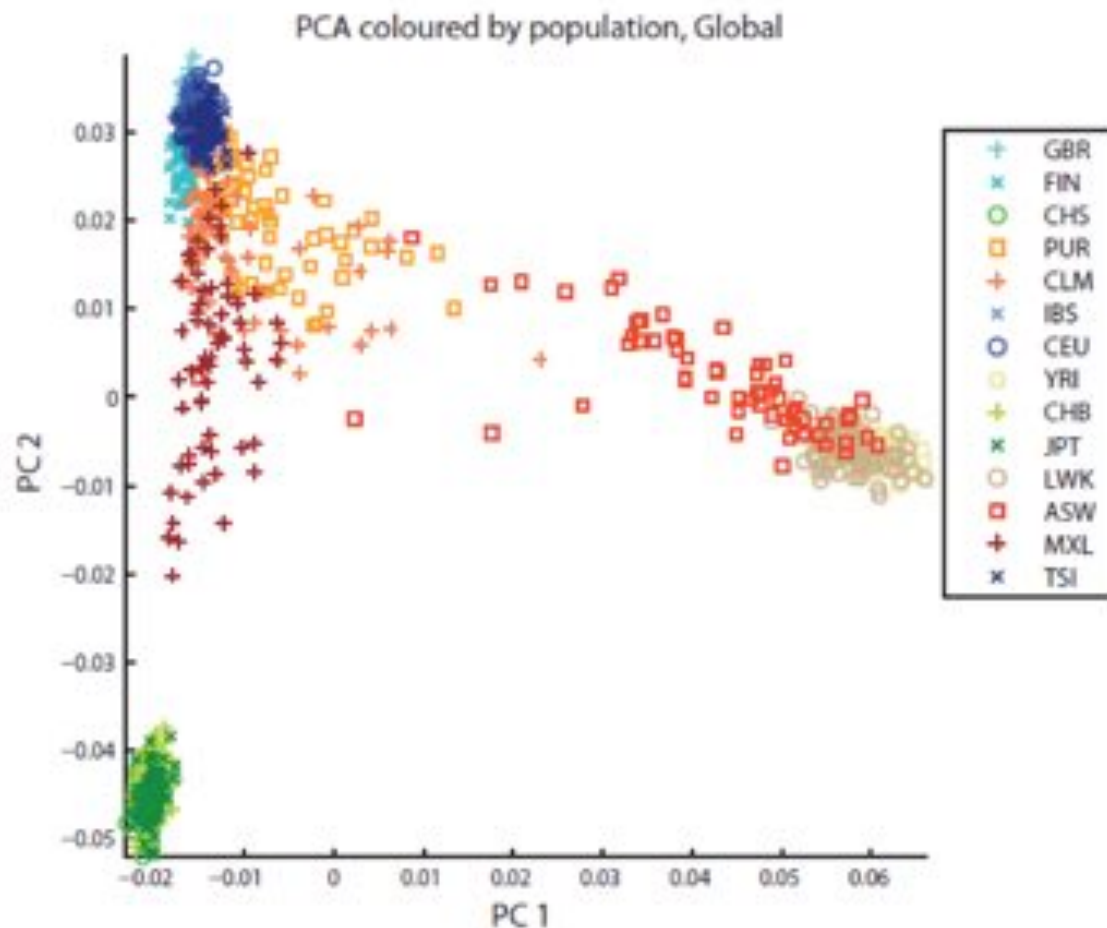
Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated owing to pLoF mutations. A full listing of all 1,317 genes knocked out, the number of knockout participants for each gene, and the specific pLoF mutation(s) are provided in Supplementary Table 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

- 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



# Variation across populations



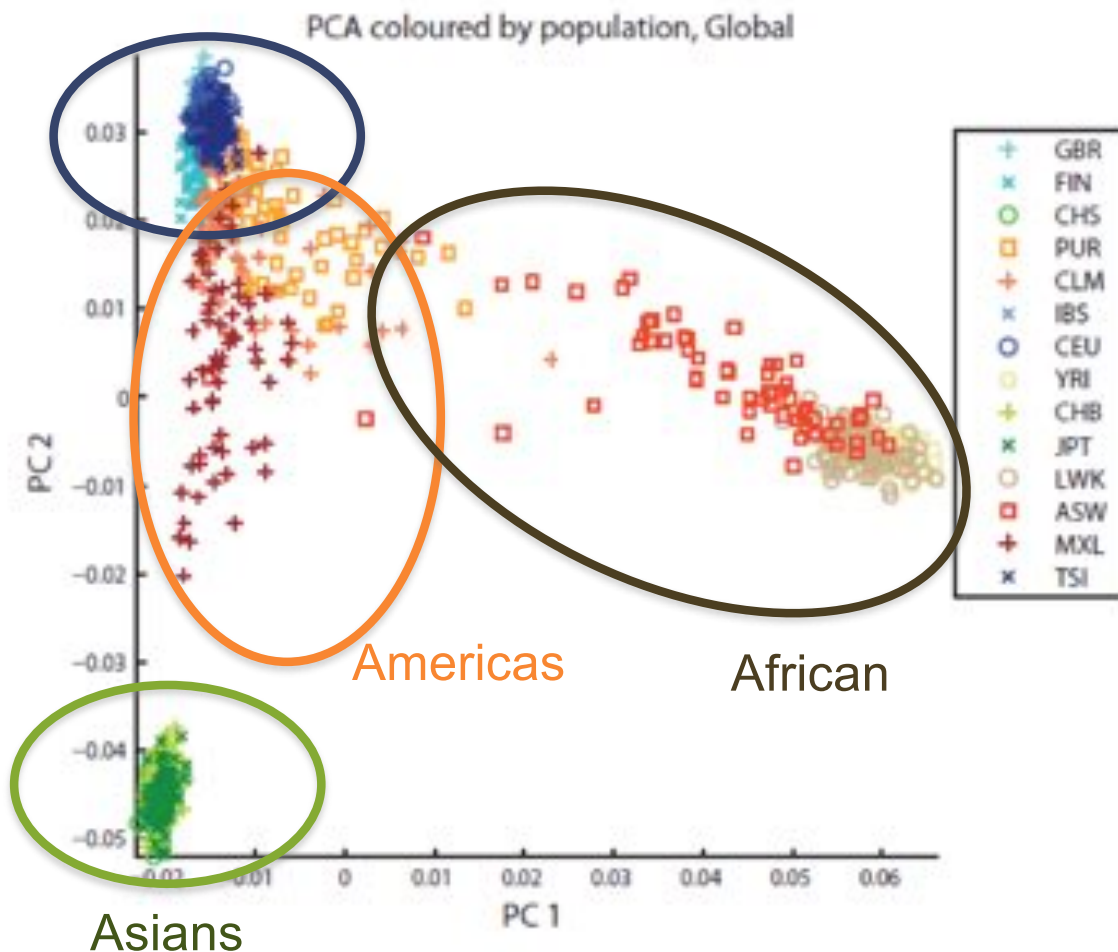
LEVEL	POP_PAIR	# of highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

Europeans

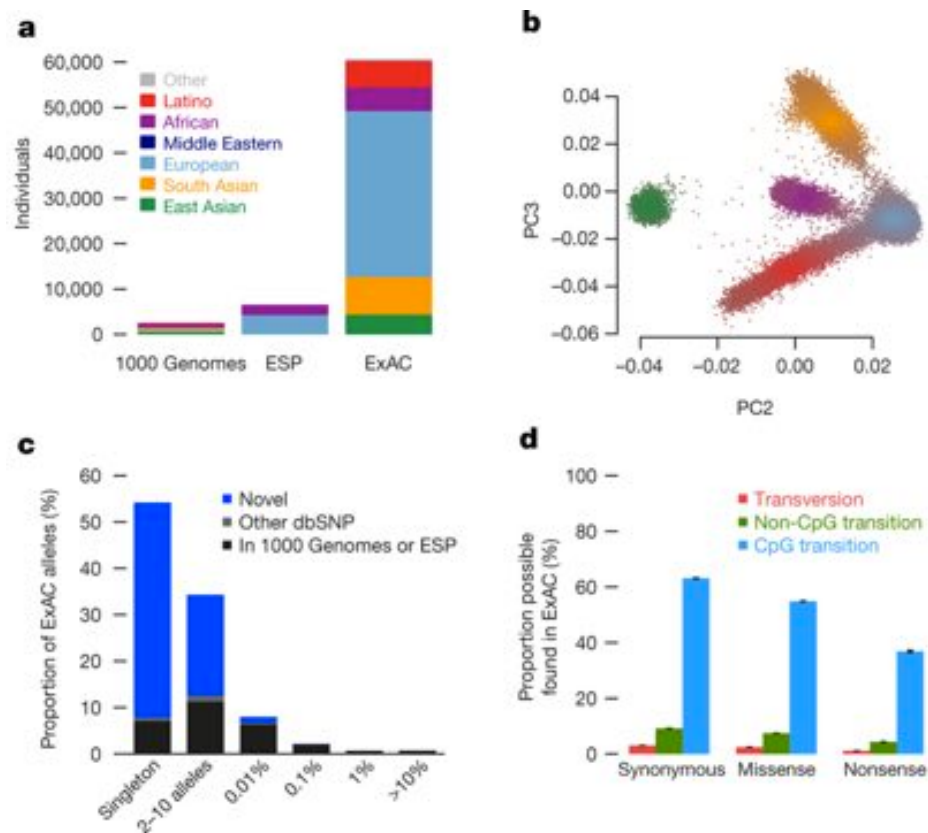


LEVEL	POP_PAIR	# of highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

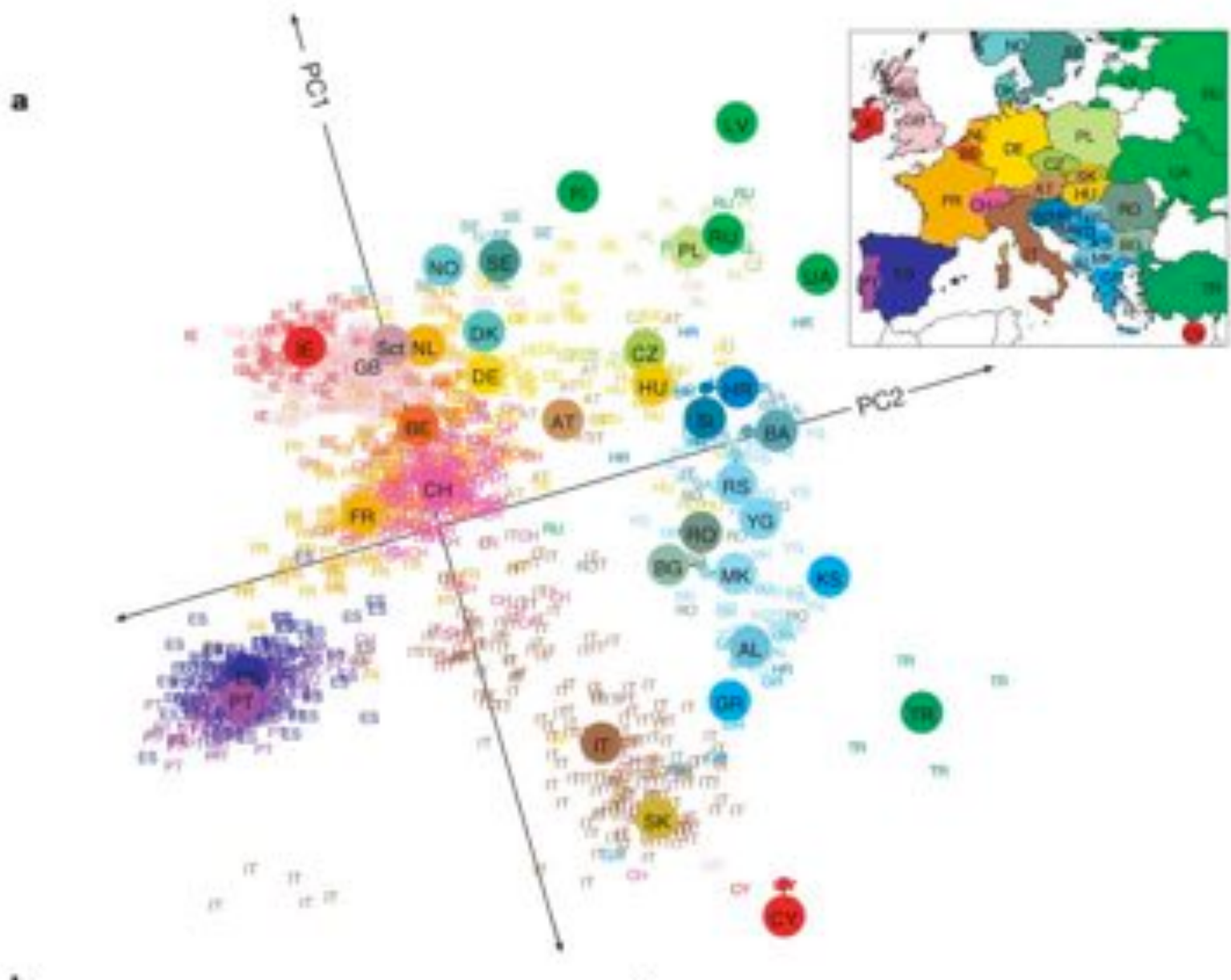
# ExAC: Exome Aggregation Consortium



- The aggregation and analysis of high-quality exome (protein-coding region) DNA sequence data for **60,706 individuals**
- This catalogue of human genetic diversity contains an average of **one variant every eight bases of the exome**
- We have used this catalogue to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation; **identifying 3,230 genes with near-complete depletion of predicted protein-truncating**

**Analysis of protein-coding genetic variation in 60,706 humans**

Lek et al (2016) Nature. doi:10.1038/nature19057



## Genes mirror geography within Europe

Novembre et al (2008) Nature. doi: 10.1038/nature07331



# dbSNP

The screenshot shows the dbSNP website interface. At the top, there's a search bar with the query 'all[sb]' and a 'Search' button. Below the search bar, there's a notification banner about COVID-19. The main content area displays search results for 'all[sb]', showing items 1 to 20 of 686600501. The first result is rs248 (Homo sapiens), which is a SNV with alleles G>A. It is located on chromosome 8:19953315 (GRCh38) and 8:19810826 (GRCh37). The gene is LPL (Vaview). The functional consequence is 'coding\_sequence\_variant,synonymous\_variant'. The clinical significance is 'benign,likely-benign'. The MAF is listed as A=0.038738/194 (1000Genomes), A=0.049866/12534 (GnomAD\_exomes), A=0.051312/6226 (ExAC), A=0.053356/6951 (TOPMED), A=0.065431/851 (GoESP), A=0.066907/2101 (GnomAD), A=0.06876/265 (ALSPAC), A=0.072816/270 (TWINSUK), A=0.081667/49 (NorthernSweden), A=0.121429/544 (Estonian). The HGVS is NC\_000008.11:g.19953315G>A, NC\_000008.10:g.19810826G>A, NG\_008855.2:g.56599G>A, NG\_008855.1:g.19245G>A, NM\_000237.3:c.435G>A, NM\_000237.2:c.435G>A. The second result is rs268 (Homo sapiens), which is also a SNV with alleles A>G. The sidebar on the left contains various filters and annotations, including Variation Class (del, delins, ins, mrv), Clinical Significance (affects, benign, conflicting interpretations of pathogenicity, drug response, likely benign, likely pathogenic, other, pathogenic, pathogenic likely, pathogenic, protective, risk factor), Publication (LIVar Annotated, PubMed Cited, PubMed Linked), and Function Class (frame shift, inframe deletion, inframe indel, inframe insertion, initiator codon variant, intron, missense, non coding transcript variant, synonymous). The right sidebar shows 'Find related data', 'Search details' (all[sb]), 'Recent activity' (all[sb] (686600501)), and a list of related publications.

- Periodic release of databases of known variants and their population frequencies
- Generally assumed to be non-disease related
- However, as catalog grows, almost certainly to contain some medically relevant SNPs.

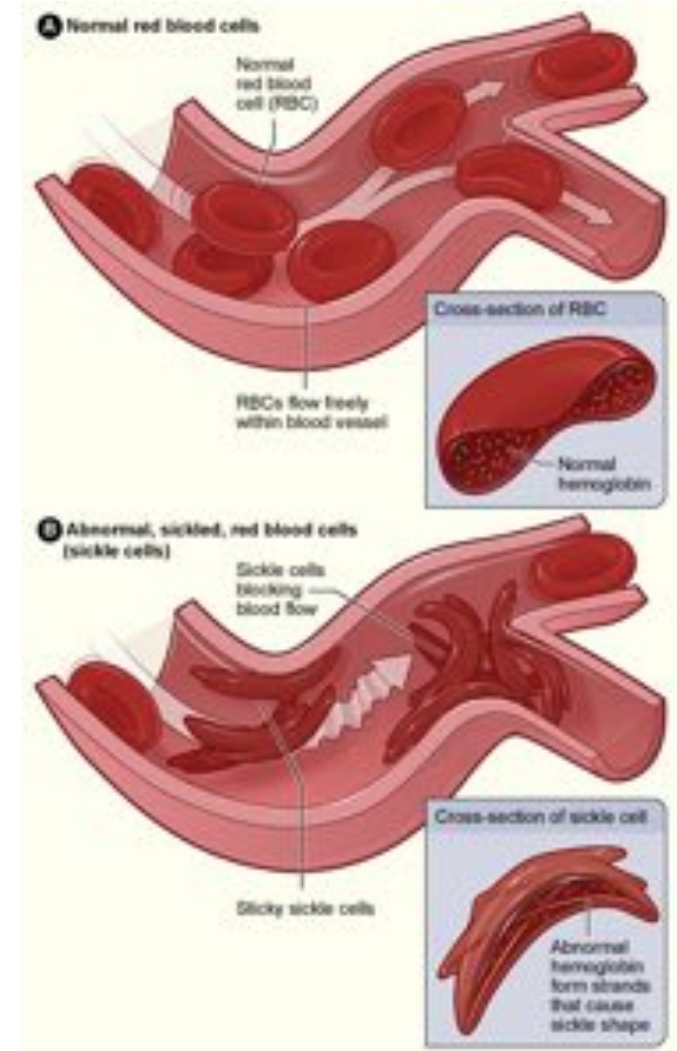


**Part 3:**

# **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<sup>\*</sup>, Nancy S. Wexler<sup>†</sup>, P. Michael Conneally<sup>‡</sup>, Susan L. Naylor<sup>§</sup>,  
Mary Anne Anderson<sup>\*</sup>, Rudolph E. Tanzi<sup>\*</sup>, Paul C. Watkins<sup>\*\*</sup>, Kathleen Ottina<sup>\*</sup>,  
Margaret R. Wallace<sup>‡</sup>, Alan Y. Sakaguchi<sup>§</sup>, Anne B. Young<sup>||</sup>, Ira Shoulson<sup>†</sup>,  
Ernesto Bonilla<sup>†</sup> & Joseph B. Martin<sup>\*</sup>**

<sup>\*</sup> Neurology Department and Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA

<sup>†</sup> Hereditary Disease Foundation, 9701 Wilshire Blvd, Beverly Hills, California 90212, USA

<sup>‡</sup> Department of Medical Genetics, Indiana University Medical Center, Indianapolis, Indiana 46223, USA

<sup>§</sup> Department of Human Genetics, Roswell Park Memorial Institute, Buffalo, New York 14263, USA

<sup>||</sup> Venezuela Collaborative Huntington's Disease Project<sup>\*\*</sup>

---

*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.*

---



# Huntington's Disease

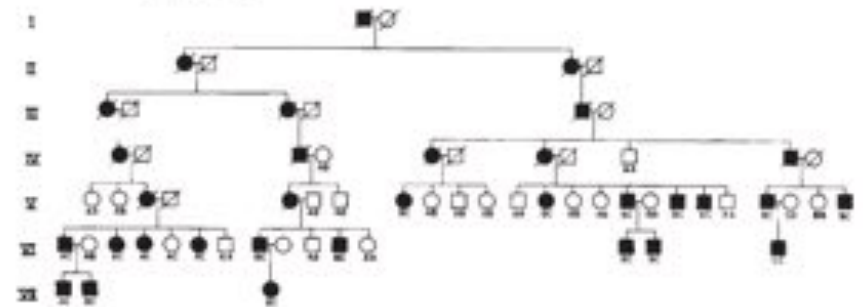
## A polymorphic D to H

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

\* Neurology Department and Genetics Unit, M  
† Hereditary Disease I  
‡ Department of Medical Ge  
§ Department of Human C  
Ive

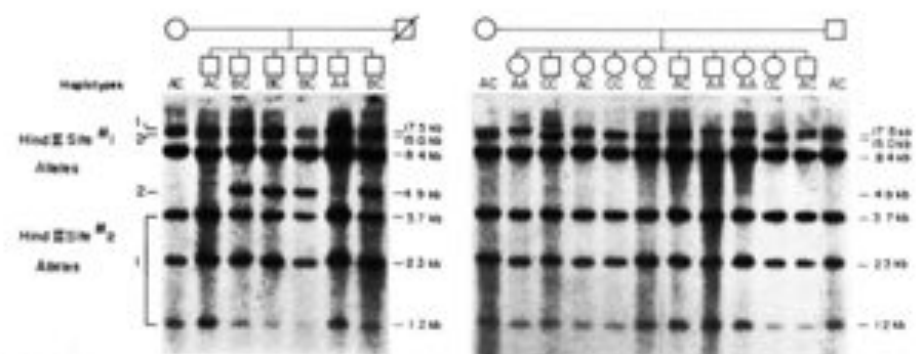
Family studies show that the Huntington  
chromosome 4. The chromosomal loca  
DNA technology to identify the primar

**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 be 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<sup>17</sup> with a correction for the late age of onset<sup>1</sup>. 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 *Hind*III-digested human genomic DNA.

**Methods:** DNA was prepared as described<sup>23</sup> from lymphoblastoid cell lines derived from members of two nuclear families. 5 µg of each DNA was digested to completion with 20 units of *Hind*III 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, pH 8, 89 mM Na borate, 2 mM Na EDTA) for 18 h. *Hind*III-digested λC1857 DNA was loaded in a separate lane as a size marker. The gels were stained with ethidium bromide (0.5 µg ml<sup>-1</sup>) 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, pH 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 pyrrolidone, 0.02% Ficoll), 0.3% SDS and 100 µg ml<sup>-1</sup> 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<sup>-1</sup> denatured salmon sperm DNA) containing approximately 5×10<sup>6</sup> c.p.m. of nick-translated G8 DNA (specific activity ~2×10<sup>8</sup> c.p.m. µg<sup>-1</sup>)<sup>24</sup> 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 *Hind*III RFLP (site 1 and 2) as explained in Fig. 4.



# 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)<sub>n</sub> 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)<sub>n</sub> 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 Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes

The Huntington's Disease Collaborative Research Group\*

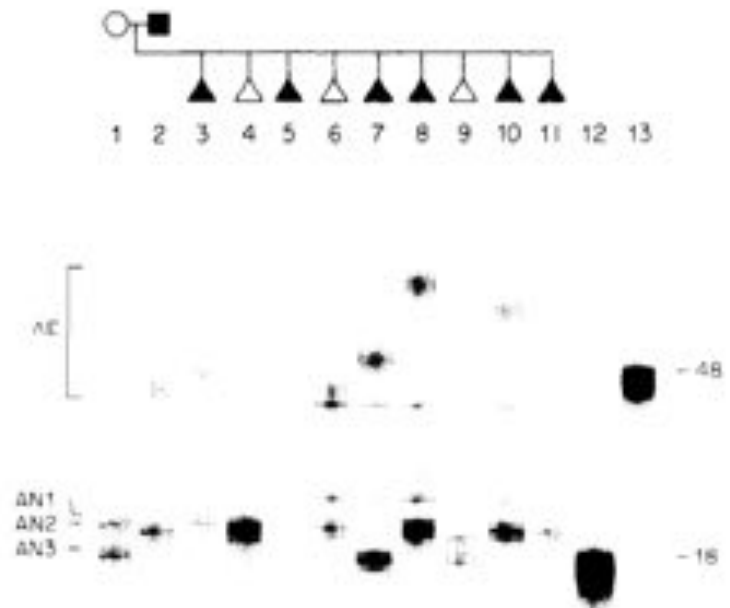
### Summary

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)<sub>n</sub> 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)<sub>n</sub> 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 disorder characterized by motor, cognitive, and psychiatric manifestations (Huntington, 1986). It is inherited in an autosomal dominant fashion and affects ~1 in 10,000 individuals of European origin (Harper et al., 1986). A distinctive choreic movement disorder is a hallmark of HD that typically has a subtle, insidious onset in the fifth decade of life and gradually worsens over a period of 10 to 20 years until death. Occasional juvenile onset, typically manifested by severe symptoms including rigidity and chorea, is associated with a pattern of paternal transmission of the disease. The pathology of HD also displays a distinctive selective loss of neurons that is most prominent in the caudate and putamen. The biochemical basis of HD has not yet been explained, and consequently no treatment effective in delaying the onset and progression of this disease is available.

The genetic defect causing HD was first identified in 1983 in one of the first successful clones using polymorphic DNA markers



**Figure 6. PCR Analysis of the (CAG)<sub>n</sub> 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<sup>1,2</sup>, 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*<sup>2</sup>. We then searched the 'morbid map' and allelic variants listed in the *Online Mendelian Inheritance in Man*<sup>3</sup> (OMIM), an online resource documenting human diseases and their associated genes

([www.ncbi.nlm.nih.gov](http://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 4:**

**Post-genome  
Inherited Diseases**

“Genome-wide linkage analysis has also been carried out for many common diseases and quantitative traits, for which the aforementioned characteristics of Mendelian diseases might not apply. In some cases, genomic regions that show significant linkage to the disease have been identified, leading to the discovery of variants that contribute to susceptibility to diseases such as inflammatory bowel disease (IBD), schizophrenia and type 1 diabetes.

***However, for most common diseases, linkage analysis has achieved only limited success, and the genes discovered usually explain only a small fraction of the overall heritability of the disease.”***

***Genome-wide association studies for common diseases and complex traits***

Hirschhorn and Daly (2005) Nature Review Genetics

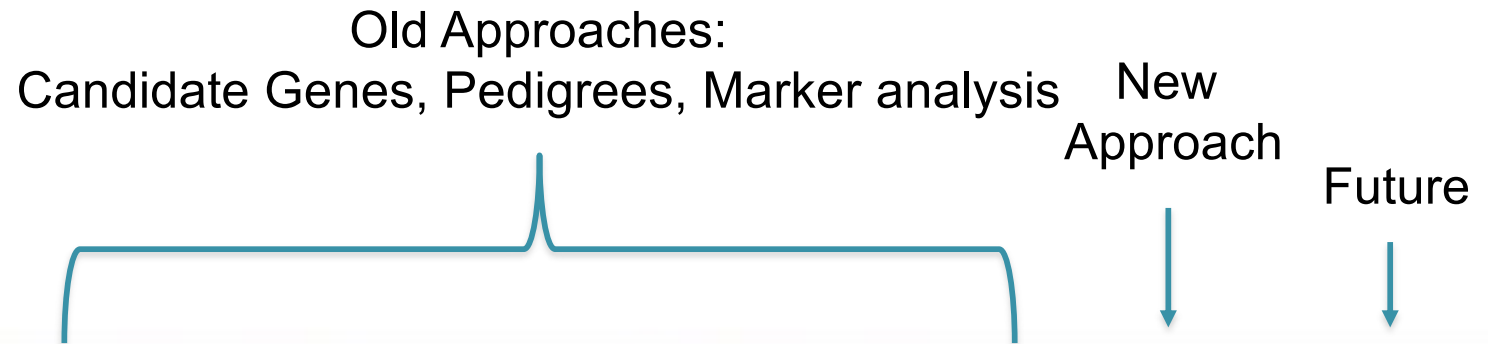


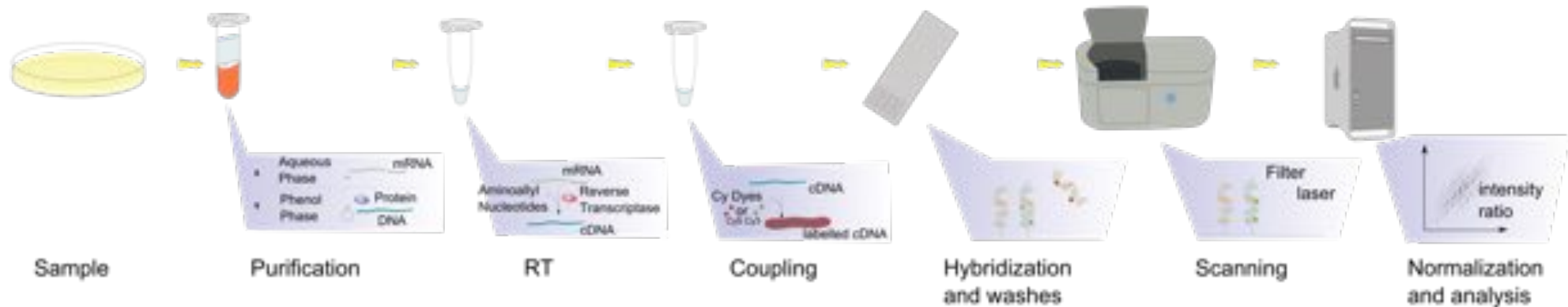
Table 1 | **Approaches to identifying variants underlying complex traits and common diseases**

Potential advantages	Association <sup>*</sup>	Resequencing <sup>*</sup>	Linkage <sup>†</sup>	Admixture <sup>‡</sup>	Missense SNPs <sup>‡</sup>	Association <sup>‡</sup>	Resequencing <sup>‡</sup>
No prior information regarding gene function required	–	–	+	+	+	+	+
Localization to small genomic region	+	+	–	–	+	+	+
Inexpensive	+	–	+	+	+/-	–	Prohibitive
Families not required	+	+	–	+	+	+	+
No assumptions necessary regarding type of variant involved	+	–	+	+	–	+	+
Not susceptible to effects of stratification <sup>§</sup>	-/+	-/+	+	+	-/+	-/+	-/+
No requirement for variation of allele frequency among populations	+	+	+	–	+	+	+
Sufficient power to detect common alleles (MAFs>5%) of modest effect	+	–	-/+	+	+	+	+
Ability to detect rare alleles (MAFs<1%)	–	+	+	–	–	–	+
Reasonable track record for common diseases	+	-/+	+/-	N/A	N/A	N/A	N/A
Tools for analysis available	+	+	+	+	+	+/-	–

<sup>\*</sup>Candidate-gene studies. <sup>†</sup>Genome-wide studies. <sup>‡</sup>Association and resequencing studies are immune to stratification if they use family-based designs. Symbols indicate whether the potential advantage in the left column applies completely (+), partially (+/-), weakly (-/+) or not at all (–). MAF, minor allele frequency; N/A, not yet attempted.

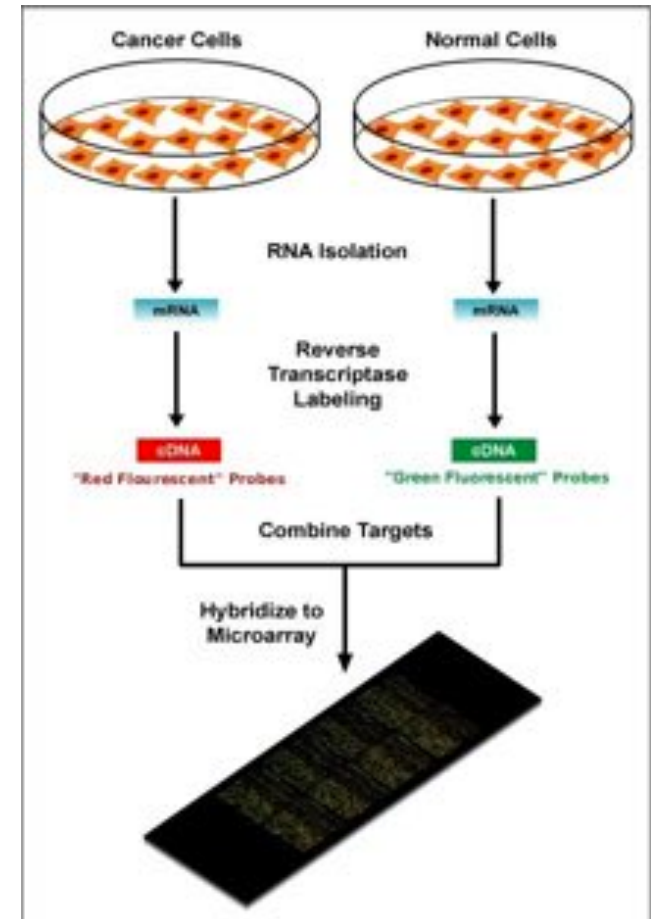
## ***Genome-wide association studies for common diseases and complex traits***

Hirschhorn and Daly (2005) Nature Review Genetics



A DNA microarray is a collection of microscopic DNA “spots” attached to a solid surface.

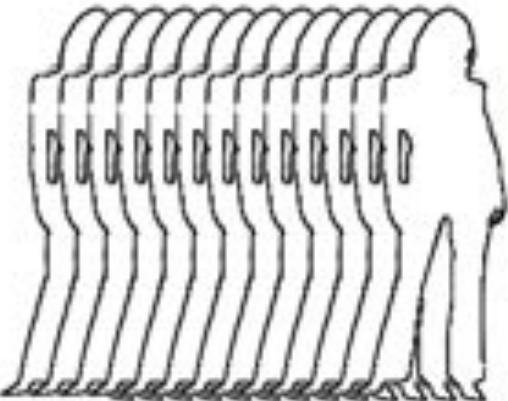
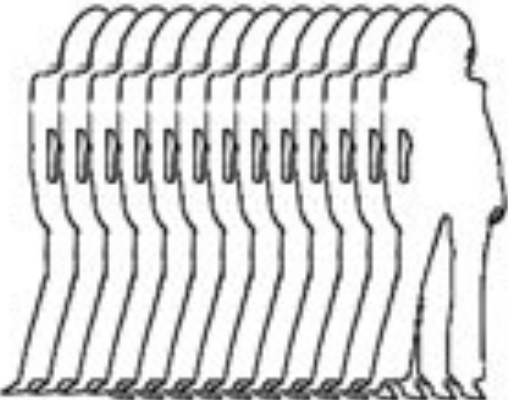
- DNA microarrays can measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome.
- Each DNA spot contains picomoles (10–12 moles) of a specific DNA sequence, known as probes (or reporters or oligos).
- Very cost effective (~\$10) for millions of probes at once







# Genome Wide Association (GWAS)

	SNP1	SNP2	SNP...
	<b>Cases</b> Count of G: 2104 of 4000  Frequency of G: 52.6%	<b>Cases</b> Count of G: 1648 of 4000  Frequency of G: 41.2%	<i>Repeat for all SNPs</i>
GC CC GG GC CC GC GC GG CC GC GG GC GG			
	<b>Controls</b> Count of G: 2676 of 6000  Frequency of G: 44.6%	<b>Controls</b> Count of G: 2532 of 6000  Frequency of G: 42.2%	
GC CC GC GC GG CC CC CC GC GC GG GC GG			

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

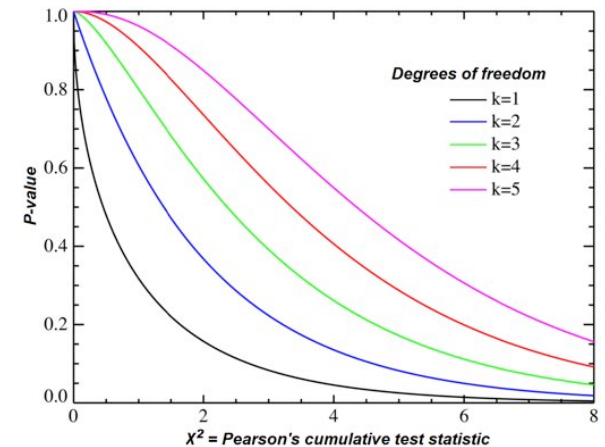
$\chi^2$  = Pearson's cumulative test statistic, which asymptotically approaches a  $\chi^2$  distribution.

$O_i$  = 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.



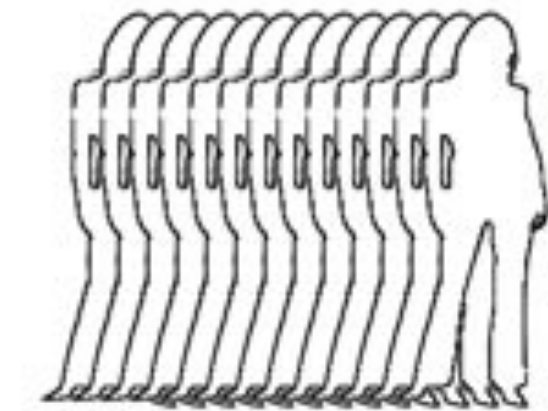
$$P(\chi_P^2(\{p_i\}) > T) \sim C \int_{\sum_{i=1}^{m-1} y_i^2 > T} \left\{ \prod_{i=1}^{m-1} dy_i \right\} \prod_{i=1}^{m-1} \exp \left[ -\frac{1}{2} \left( \sum_{i=1}^{m-1} y_i^2 \right) \right]$$

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

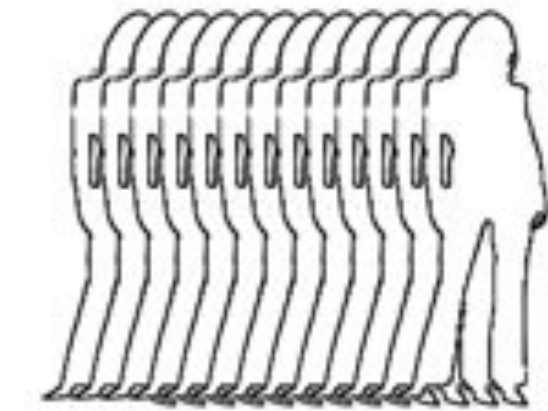
Cases/hasG expected:  $4000 * (4780/10000) = 1912$  expected  
 Cases/hasG squared deviation:  $(2104 - 1912)^2 / 1912 = 19.28$  deviation

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*

**Cases**

Count of G:  
2104 of 4000

Frequency of G:  
52.6%

**Controls**

Count of G:  
2676 of 6000

Frequency of G:  
44.6%

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

*SNP2*

**Cases**

Count of G:  
1648 of 4000

Frequency of G:  
41.2%

**Controls**

Count of G:  
2532 of 6000

Frequency of G:  
42.2%

**P-value:**  
0.33

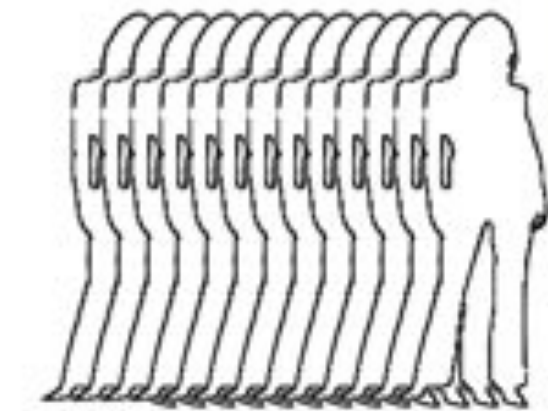
*SNP...*

*Repeat for all  
SNPs*

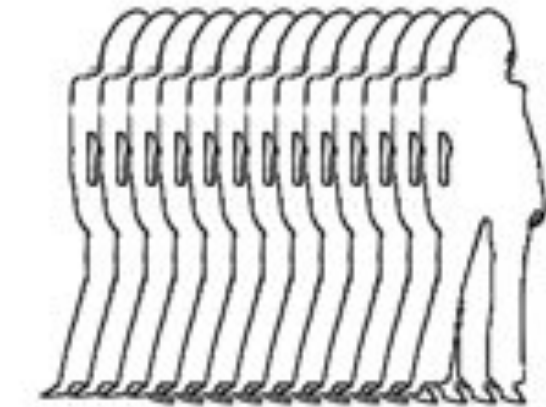
Chi-squared or  
similar test



# Genome Wide Association (GWAS)



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



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

*SNP1*

**Cases**

Count of G:  
2104 of 4000

Frequency of G:  
52.6%

**Controls**

Count of G:  
2676 of 6000

Frequency of G:  
44.6%

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

*SNP2*

**Cases**

Count of G:  
1648 of 4000

Frequency of G:  
41.2%

**Controls**

Count of G:  
2532 of 6000

Frequency of G:  
42.2%

**P-value:**  
0.33

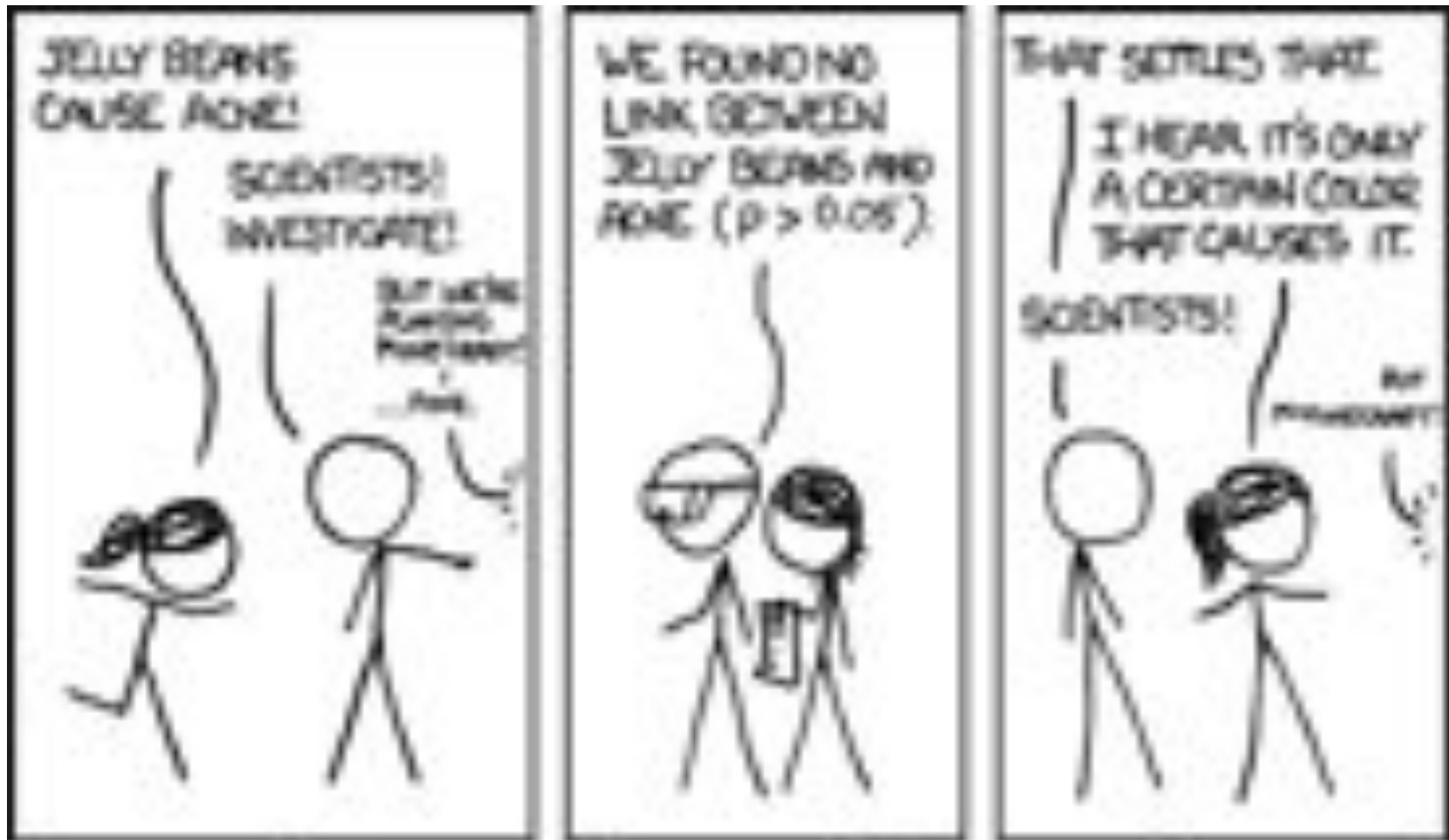
*SNP...*

*Repeat for all  
SNPs*

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

Chi-squared or  
similar test

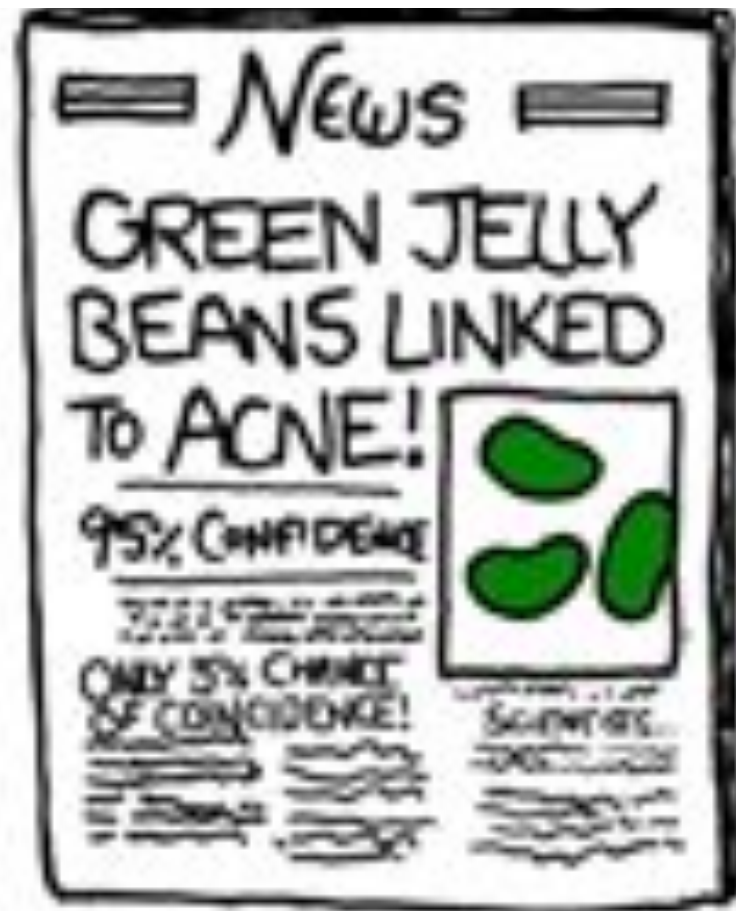
# The curse of multiple testing



# The curse of multiple testing

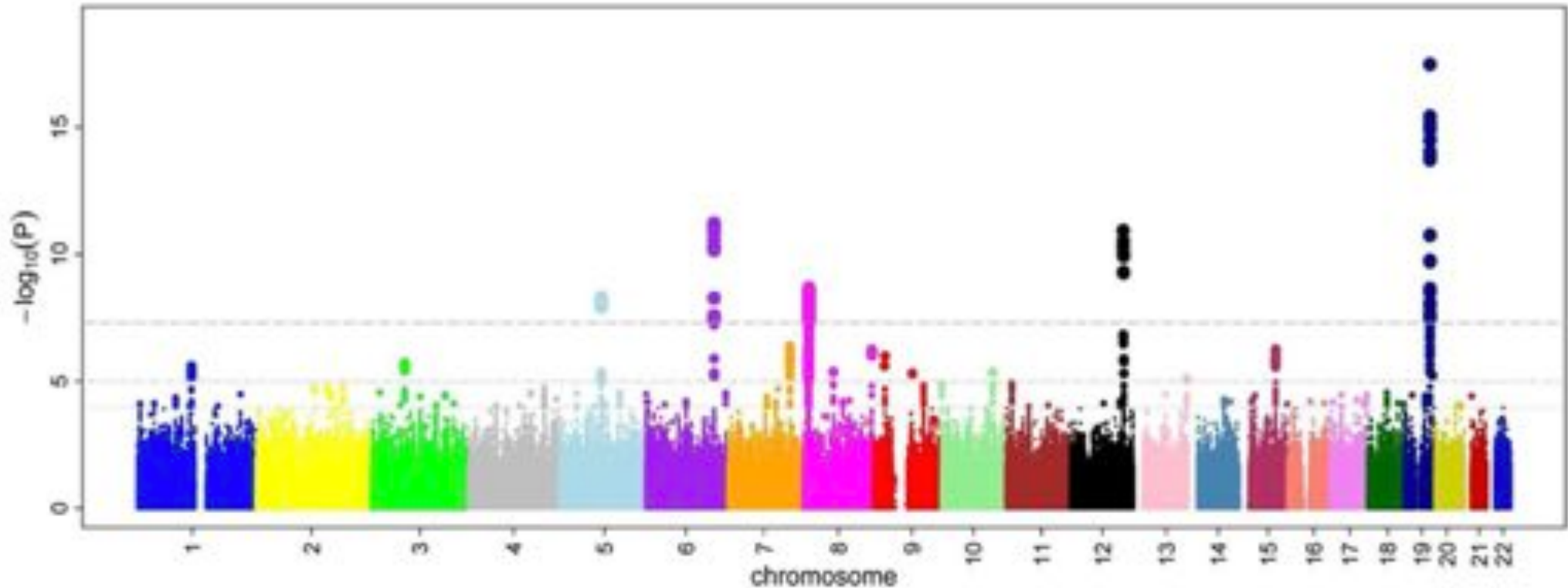


# The curse of multiple testing





# Manhattan Plot



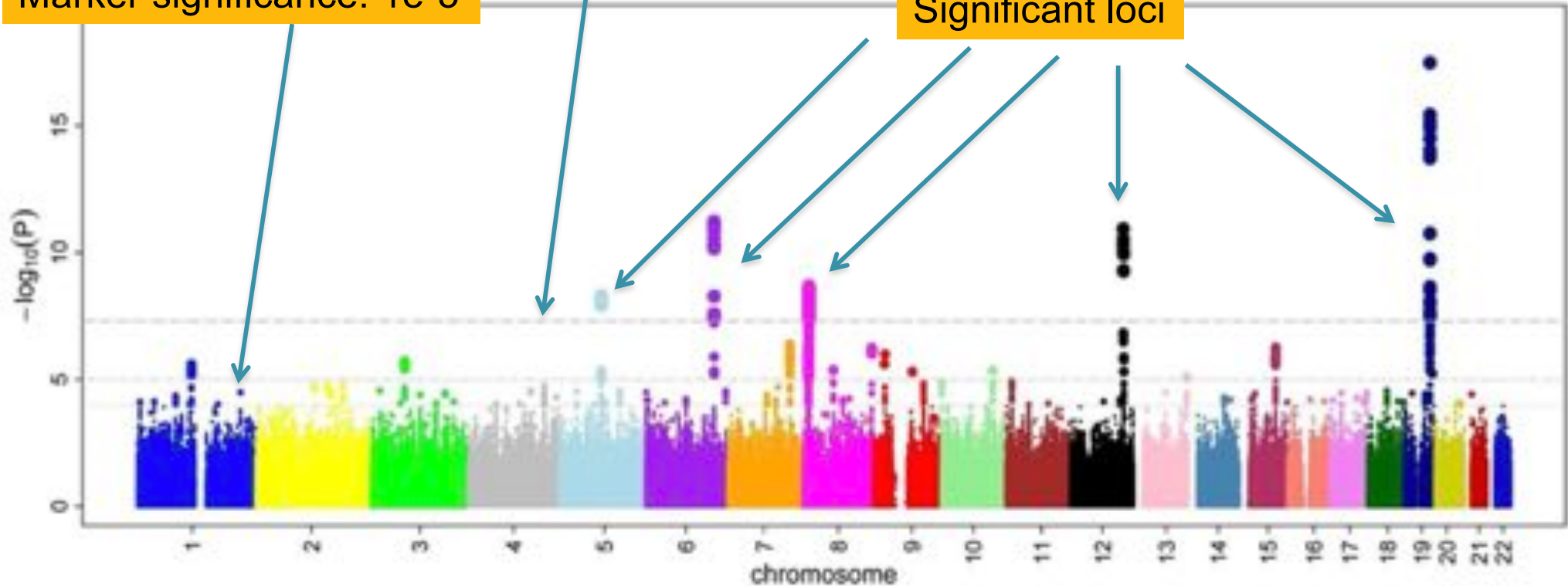
***Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo***  
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

# Manhattan Plot

Genome-wide significance:  $5e-8$

Marker significance:  $1e-5$

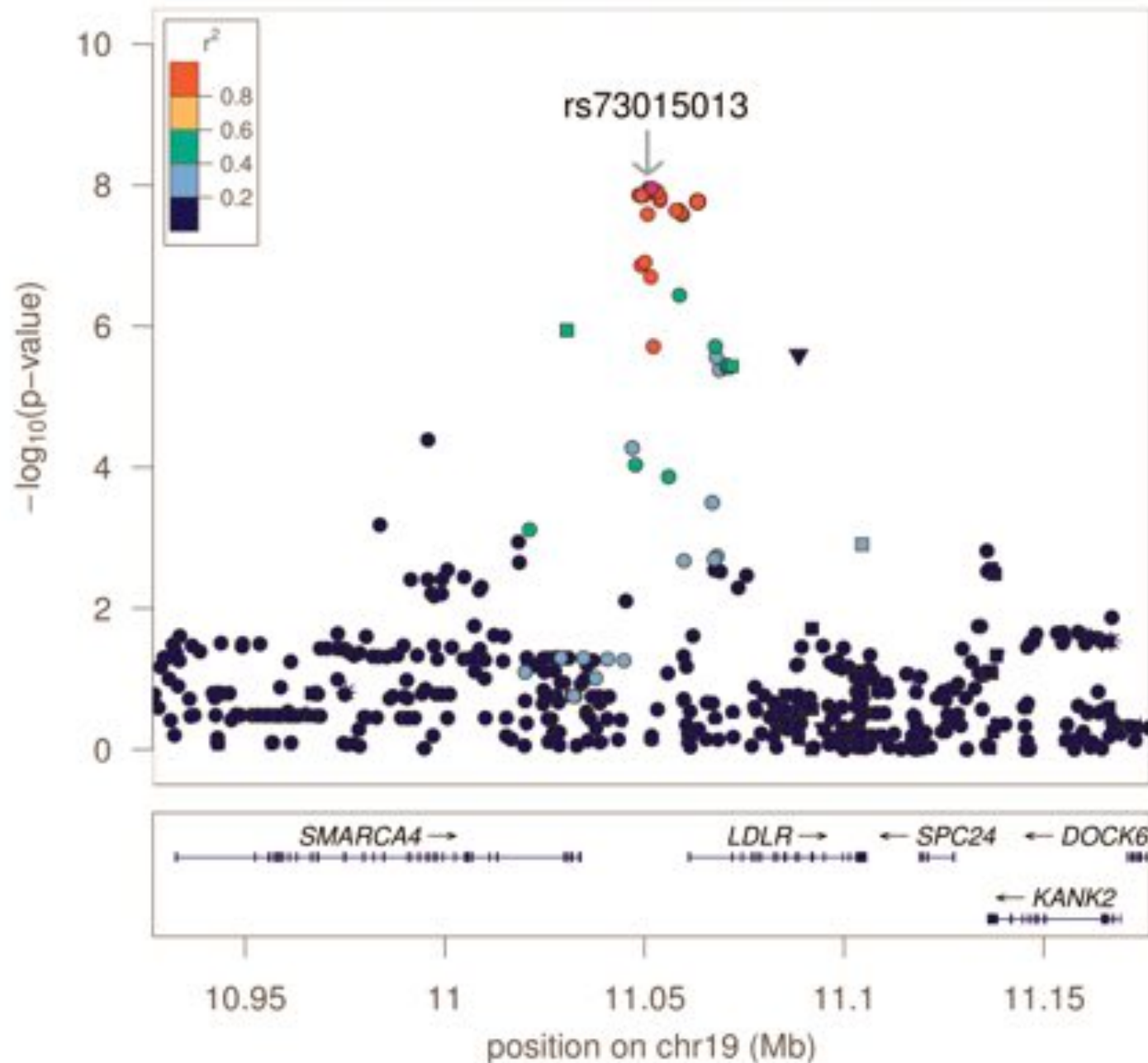
Significant loci



***Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo***

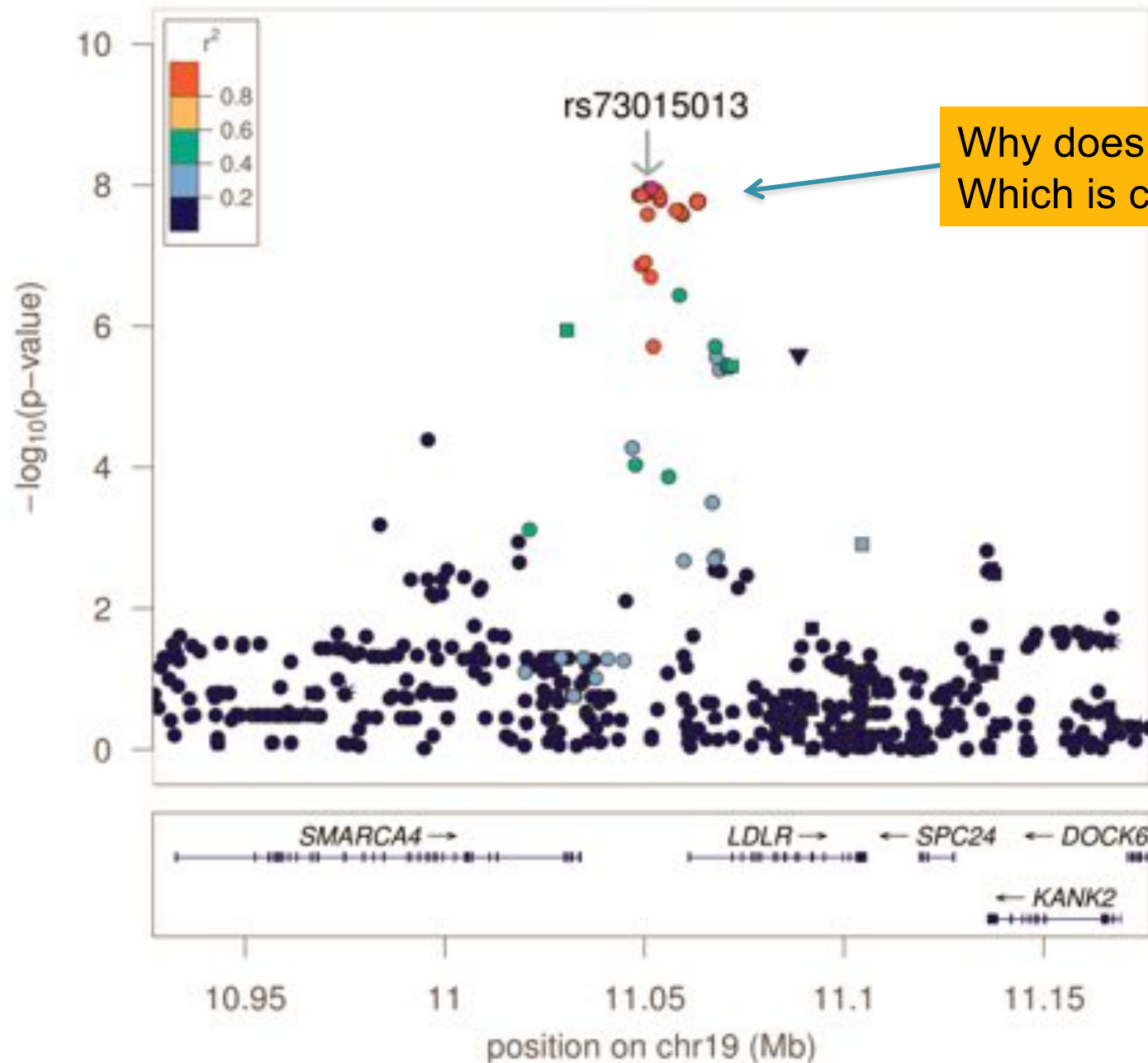
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

# Regional Association Plot





# Regional Association Plot



# First published GWAS

## Complement Factor H Polymorphism in Age-Related Macular Degeneration

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

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

<sup>1</sup>Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, NY 10021, USA. <sup>2</sup>Department of Ophthalmology and Visual Science, Yale University School of Medicine, 330 Cedar Street, New Haven, CT 06520, USA. <sup>3</sup>National Eye Institute, Building 10, CRC, 10 Center Drive, Bethesda, MD 20892–1204, USA. <sup>4</sup>Biological Imaging Core, National Eye Institute, 9000 Rockville Pike, Bethesda, MD 20892, USA. <sup>5</sup>The EMMES Corporation, 401 North Washington Street, Suite 700, Rockville, MD 20850, USA. <sup>6</sup>W. M. Keck Facility, Yale University, 300 George Street, Suite 201, New Haven, CT 06511, USA. <sup>7</sup>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,<sup>1</sup> Caroline Zeiss,<sup>2\*</sup> Emily Y. Chew,<sup>3\*</sup> Jen-Yue Tsai,<sup>4\*</sup> Richard S. Sackler,<sup>1</sup> Chad Haynes,<sup>1</sup> Alice K. Henning,<sup>5</sup> John Paul SanGiovanni,<sup>3</sup> Shrikant M. Mane,<sup>6</sup> Susan T. Mayne,<sup>7</sup> Michael B. Bracken,<sup>7</sup> Frederick L. Ferris,<sup>3</sup> Jurg Ott,<sup>1</sup> Colin Barnstable,<sup>2</sup> Josephine Hoh<sup>7†</sup>

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.

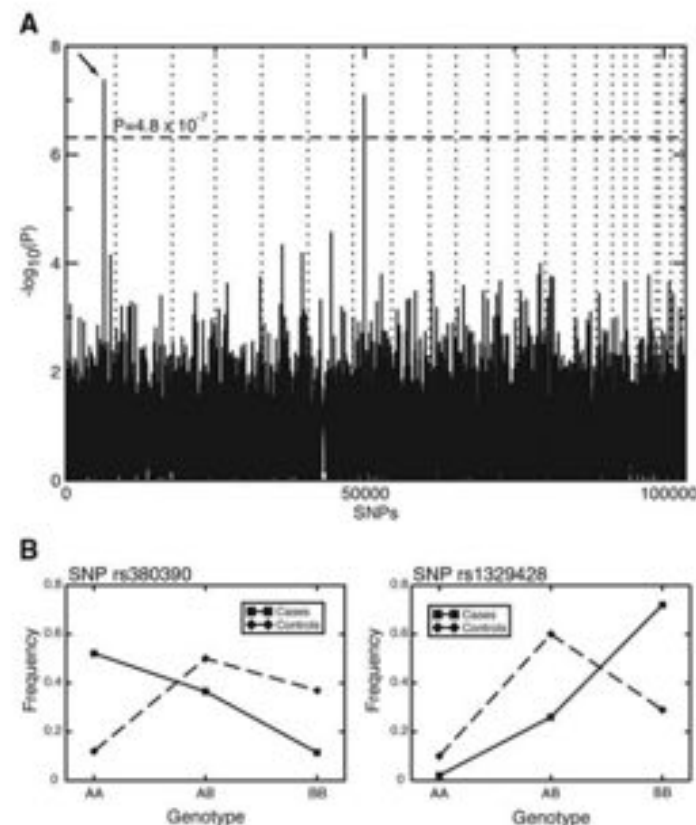
Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world. Its incidence is increasing as the elderly population expands (1). AMD is characterized by progressive destruction of the retina's central region (macula), causing central field visual loss (2). A key feature 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 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 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



**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.

# GWAS Catalog

As of 2020-03-08, the GWAS Catalog contains  
4493 publications and 179364 associations.



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



# ClinVar

The screenshot shows the ClinVar website in a web browser. The address bar displays the URL <https://www.ncbi.nlm.nih.gov/clinvar/>. The page features a navigation bar with links to Home, About, Access, Help, Submit, Statistics, and FTP. A search bar is prominently displayed, allowing users to search for gene symbols, HGVS expressions, or conditions. Below the search bar, there is a section titled 'ClinVar' with a description: 'ClinVar aggregates information about genomic variation and its relationship to human health.' To the left of this description is a DNA sequence: ACTGATGGTATGGGGCCAAGAGATATATCT, CAGGTACGGCTGTCATCACTTAGACCTCAC, CAGGGCTGGGCATAAAAGTCAGGGCAGAGC, CCATGGTGCATCTGACTCCTCAGGAGAAGT, GCAGGTTGGTATCAAGGTTACAAGACAGGT, GGCACCTGACTCTCTCTGCTATTGGTCTAT. The page is organized into three main columns: 'Using ClinVar' (containing links like About ClinVar, Data Dictionary, Downloads/FTP site, FAQ, Contact Us, RSS feed/What's new?, and Factsheet), 'Tools' (containing links like ACMG Recommendations for Reporting of Incidental Findings, ClinVar Submission Portal, Submissions, Variation Viewer, Clinical Remapping - Between assemblies and RefSeqGenes, and RefSeqGene/LRG), and 'Related Sites' (containing links like ClinGen, GeneReviews®, GTR®, MedGen, OMIM®, and Variation). At the bottom, there is a 'Submitter highlights' section and a 'Disclaimer' section.

- ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence
- Currently has 295k mutations
- Most (179k) variants have uncertain affect, only 23 have “4 stars” of significance

# 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 😊

# 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.

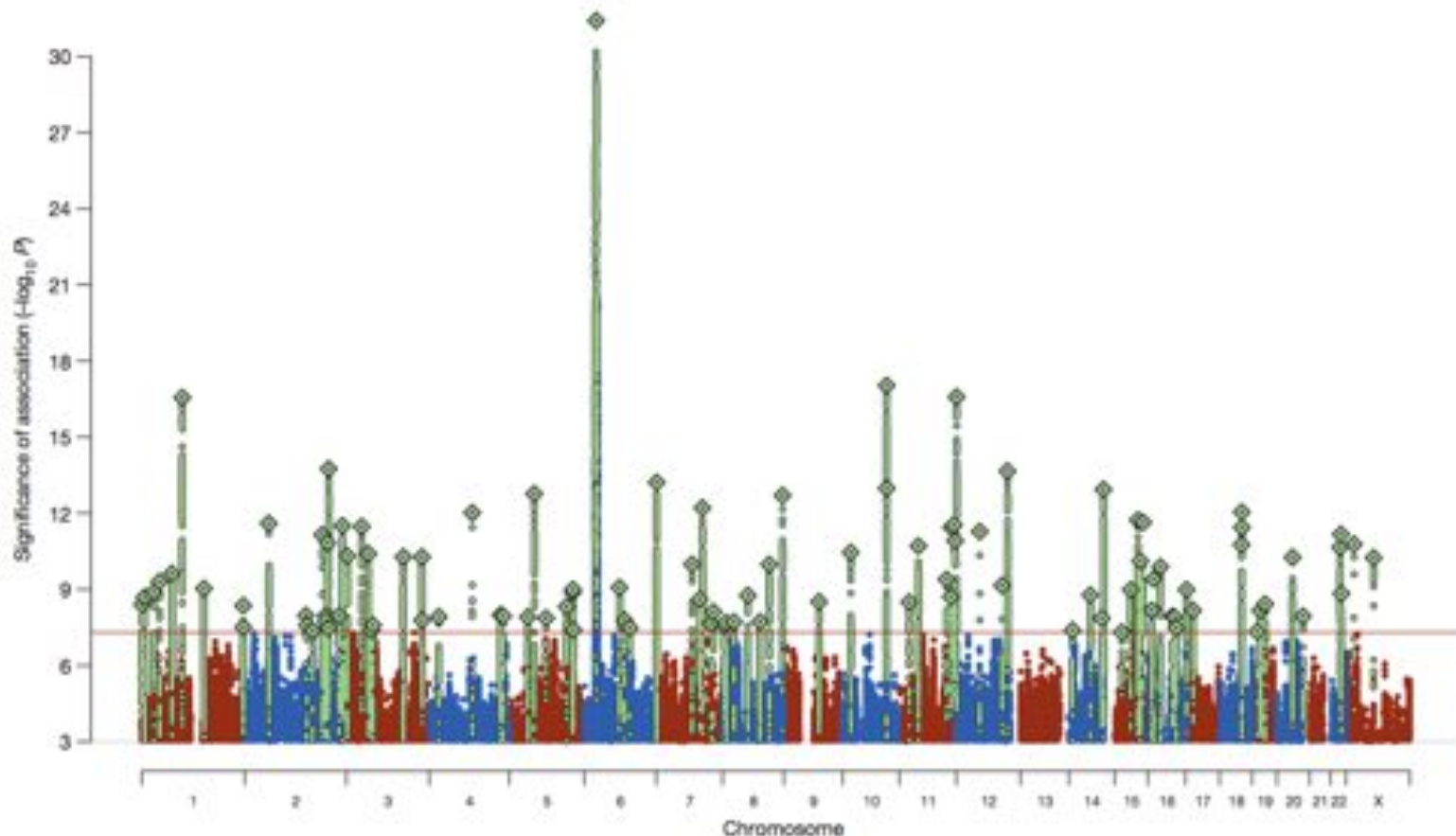


# Biological insights from 108

## schizophrenia

Schizophrenia W

Schizophrenia alleles of small effect sizes. The findings span across multiple studies and several genes, providing support for the



**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 ( $5 \times 10^{-8}$ ). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

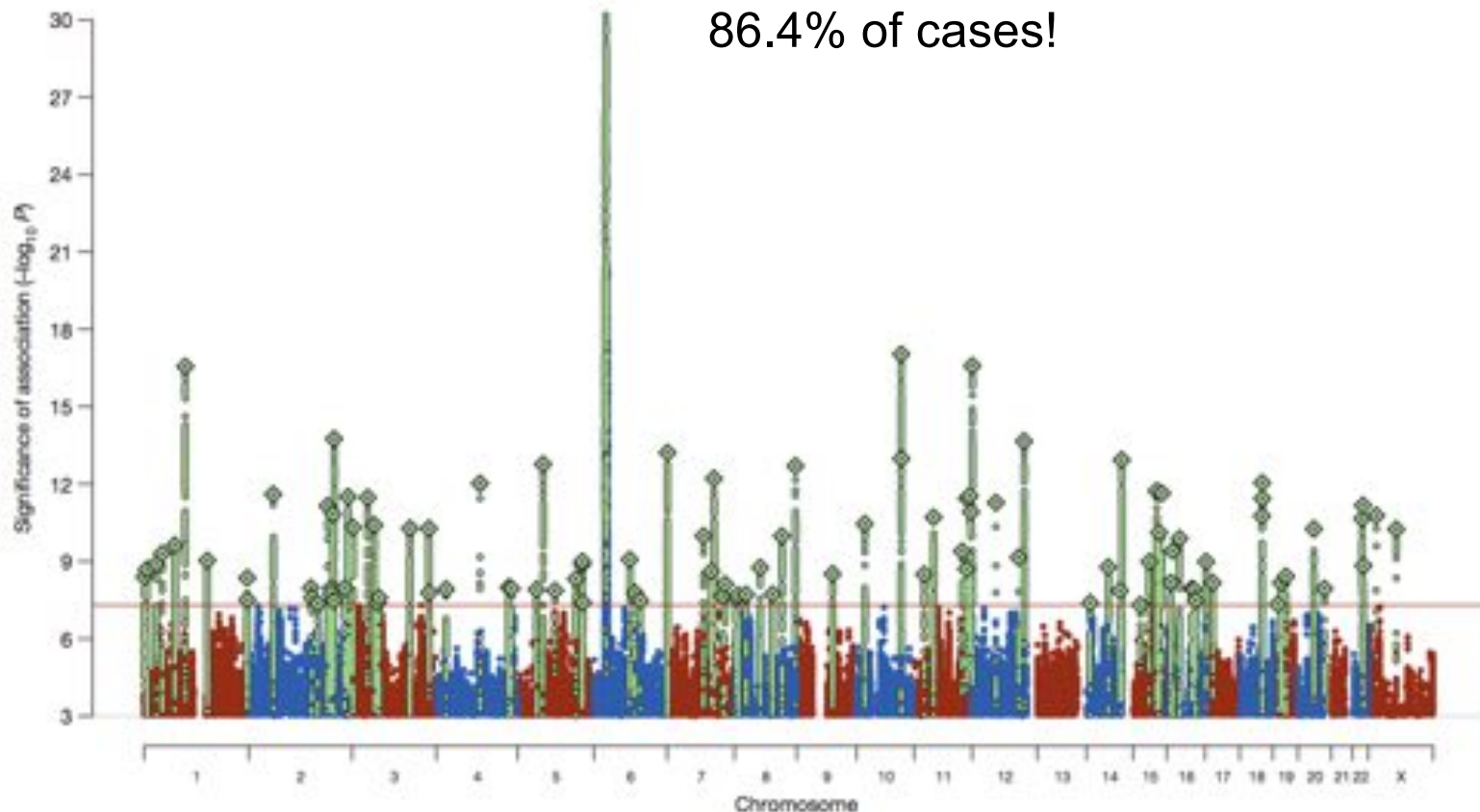


# Biological insights from 108

## schizophrenia

Schizophrenia W

Schizophrenia alleles of small effect sizes. Schizophrenia genetic associations span the genome, and several genes have been found to be relevant to schizophrenia. In brain, associations support for the



**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

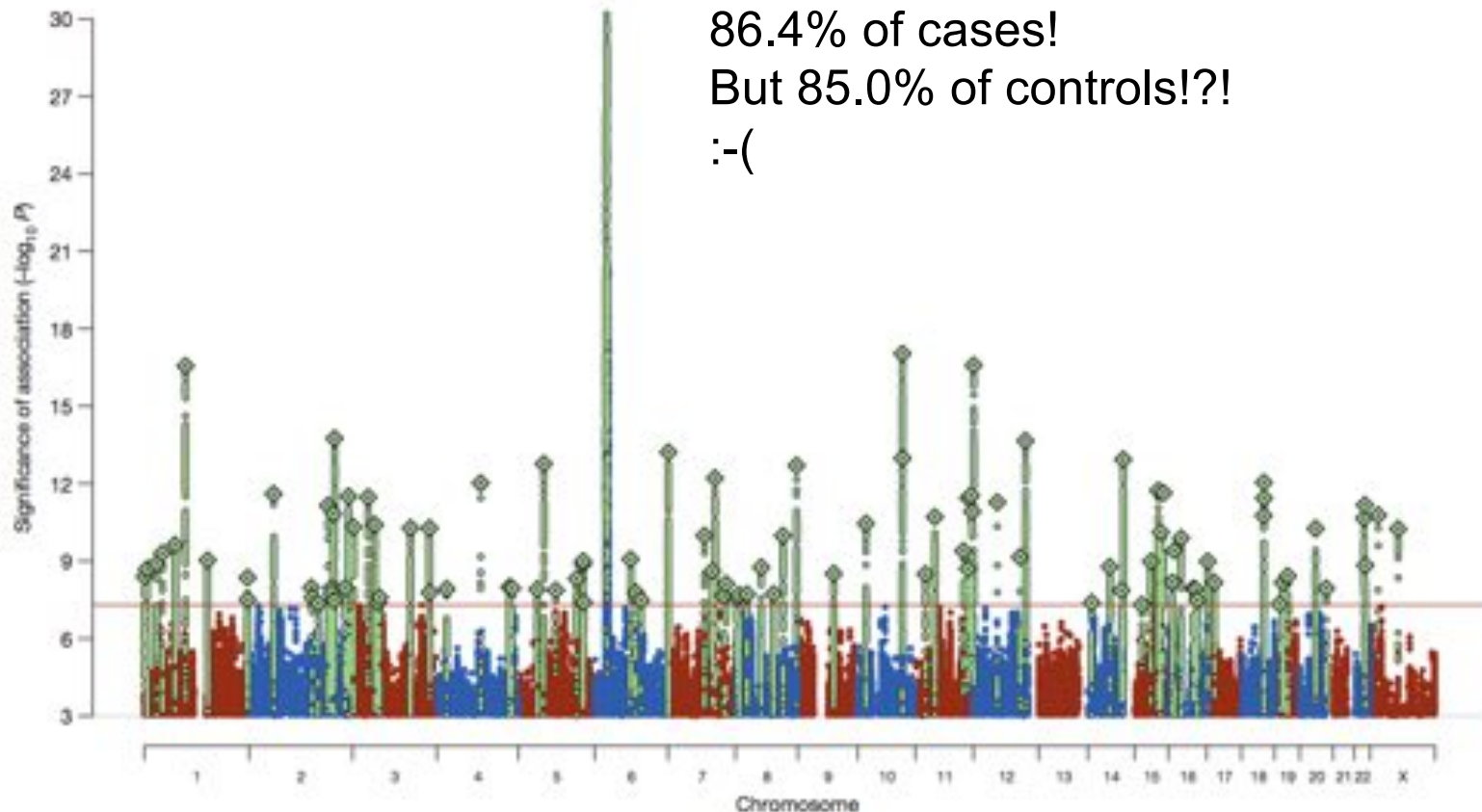
position and the y axis is the significance ( $-\log_{10} P$ ; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level ( $5 \times 10^{-8}$ ). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

## Biological insights from 108

### schizophrenia

Schizophrenia W

Schizophrenia alleles of small phrenia genotiations span previously reported findings. and several genes relevance to schizophrenia in brain, associated support for the



**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 ( $5 \times 10^{-8}$ ). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

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

Bi  
SC

Schizo

Schiz  
allel  
phre  
ciati  
prev  
the  
and  
relev  
in br  
supp

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 ( $5 \times 10^{-8}$ ). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.



# GWAS In Crisis

**Table 1.** Replication and non-replication in associations found by GWA studies of complex diseases published until the end of 2006

Phenotype	Genome-wide association study characteristics				Identified gene/SNPs	Replication status (January 2007)
	platform (SNPs/analyzed)	design	stratification control	n		
Age-related macular degeneration	Affymetrix 100k (116204/103611)	UCC; then sequencing of region	Genomic control, F-ratio	146	<i>CFH</i> /Intronic rs380390; then sequencing showing exonic rs106170 (Y420H) 2kb upstream of 41-kb haplotype block	Meta-analysis of 11 studies (n = 8,991): OR 2.49 and 6.15 (heterozygotes and homozygotes respectively), <b>no large between-study inconsistency in effect sizes; also replicated in large Dutch cohort</b> (n = 5,681); several studies on Asian populations claim no association
Obesity	Affymetrix 100k (116204/86604)	Family-based, 2-stage, followed by mapping 100 neighboring SNPs	Family-based design	694, then up to 923	<i>INSIG2</i> /rs7566605 10kb upstream of the transcription start site	Replication in the same publication in 3 of 4 independent populations of n = 9,881 subjects with modest between-study heterogeneity; 7 more independent populations with over 21,000 subjects total <b>failed to replicate the association</b> ; no effect and no heterogeneity across the independent replication teams
Parkinson disease	Perlegen (248535/198345)	Family-based, second stage with matched case-controls	Family-based design; matching at second stage; also genomic control	443 sib-pairs, then 664	Thirteen genes/ 13 different SNPs identified from analysis of both stages; none with genome-wide significance	Several small replication studies and a large collaborative consortium (n = 12,208) <b>failed to replicate any of the 13 proposed SNPs</b> ; null results were consistent across the teams participating in the consortium
Myocardial infarction	Random gene-based (92788/67671)	UCC	None (just Japanese nationality)	752 (only 94 cases)	<i>LTA</i> /Haplotype of 5 SNPs (2 in <i>LTA</i> and 3 in adjacent genes); the two <i>LTA</i> SNPs had association in larger sample and then Thr26Asn had also functional assay support	Replication in the same publication in additional 1,133 cases and two control groups (n = 1,006 and 872); association not replicated in subsequent ISIS-4 case-control study and meta-analysis (n = 18,325) shows <b>no association (non-significant OR 1.07)</b> without significant between-study heterogeneity vs. 1.77 in originally proposed association for recessive model)
Age-related macular degeneration	Affymetrix 100k (116204/97824)	UCC; then sequencing of region	Genomic control, F-ratio	226	<i>HTRA1</i> /Intragenic rs10490924; then sequencing showing promoter rs11200638 6kb downstream	Independent study (n = 890) published in the same issue starting from dense mapping of locus showing consistent effects with OR 1.90 and 7.51 for heterozygotes and homozygotes, respectively

## Non-Replication and Inconsistency in the Genome-Wide Association Setting

Ioannidis (2007) Hum Hered 2007;64:203–213 <https://doi.org/10.1159/000103512>