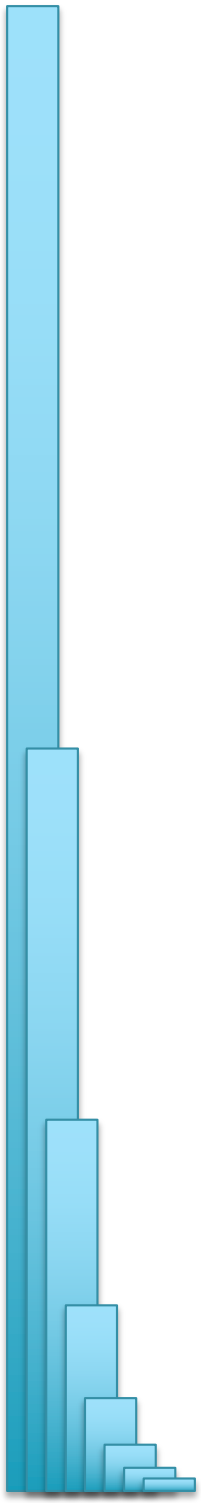
Disease Genetics

Michael Schatz

Sept 30, 2019

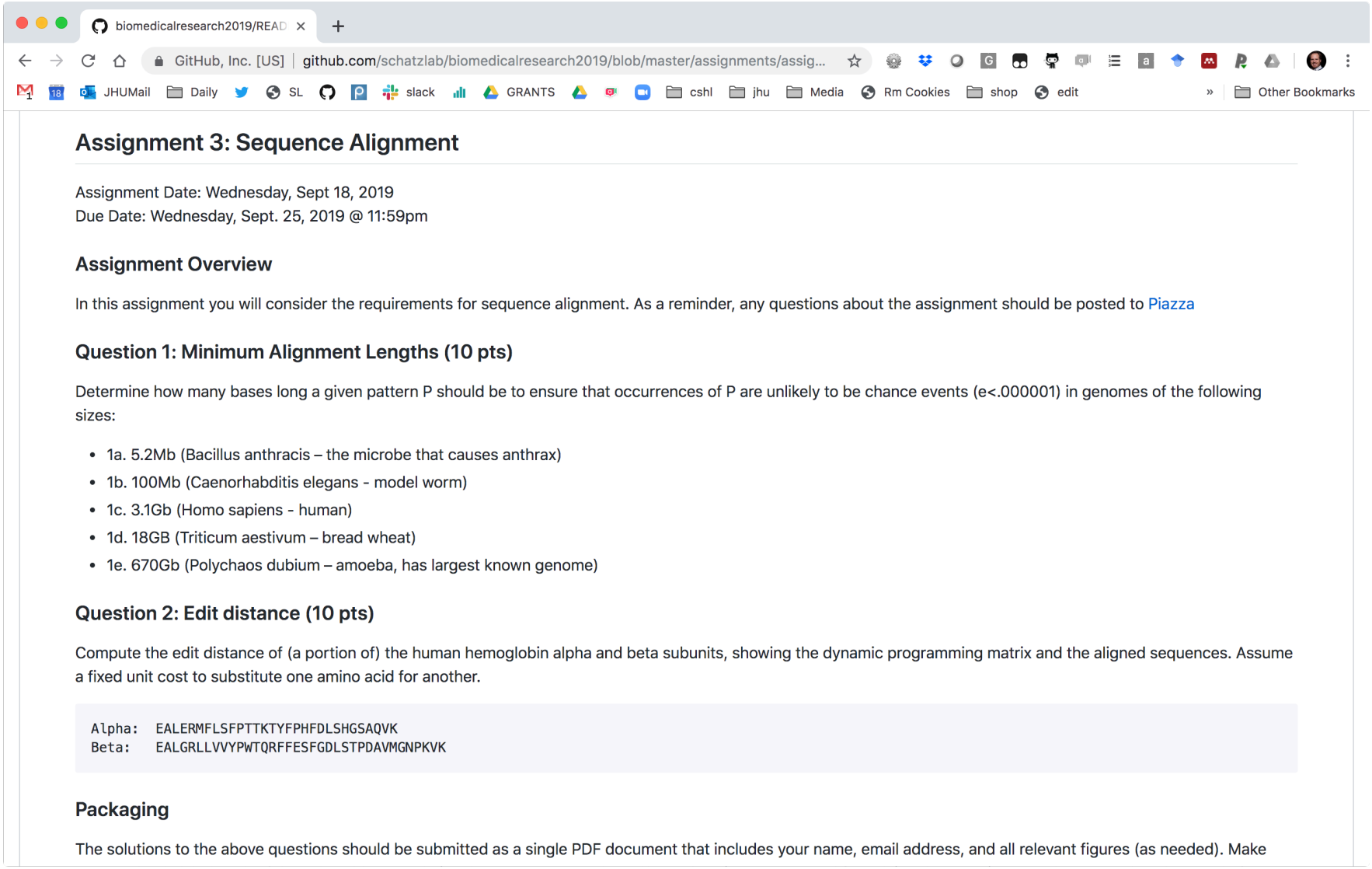
Lecture 9: Computational Biomedical Research



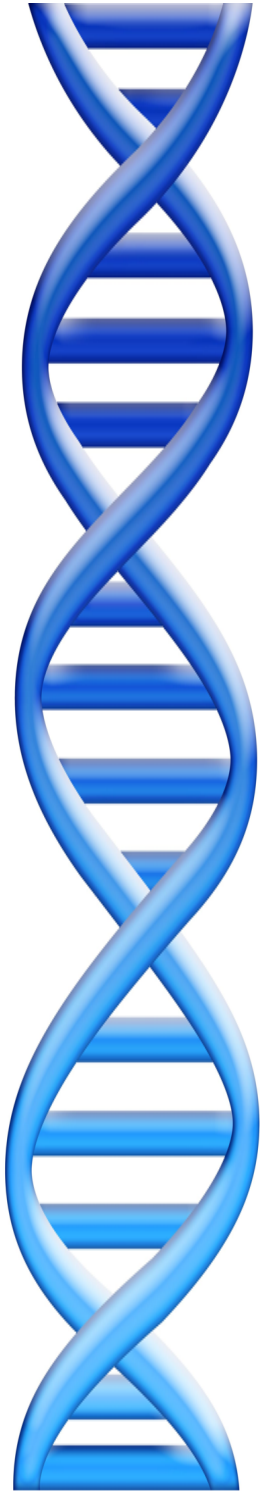
Next class:

Meet in Maryland 217!!!

Assignment 3: Sequence Alignment
Due Monday Sept 30 @ 11:59pm

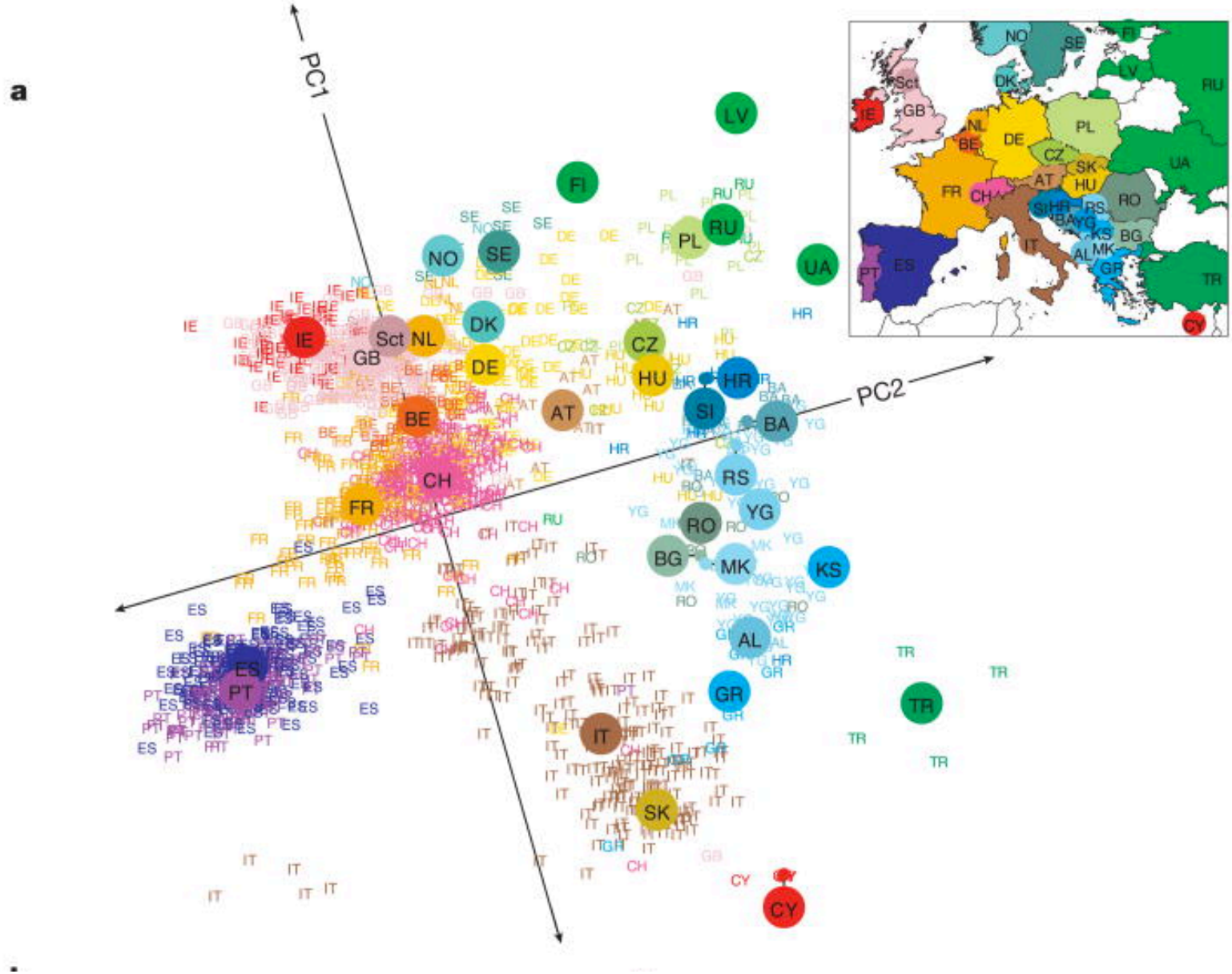


https://github.com/schatzlab/biomedicalresearch2019



Part 0:

Recap



Genes mirror geography within Europe

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

High Dimensional Data

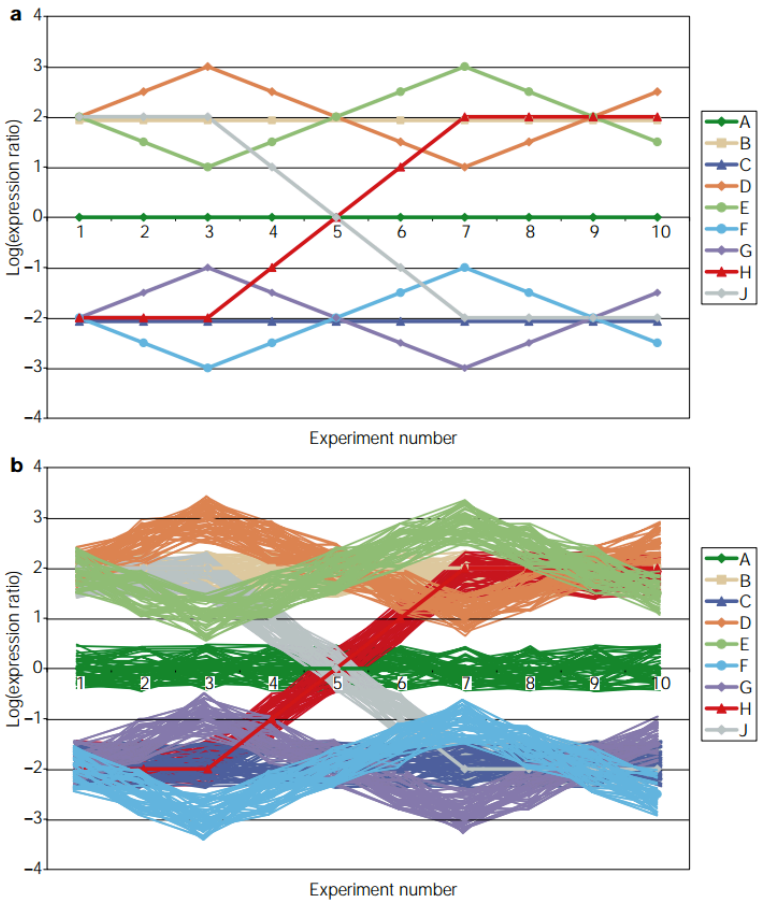
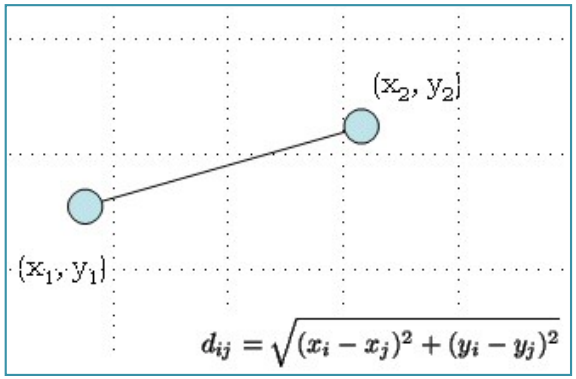
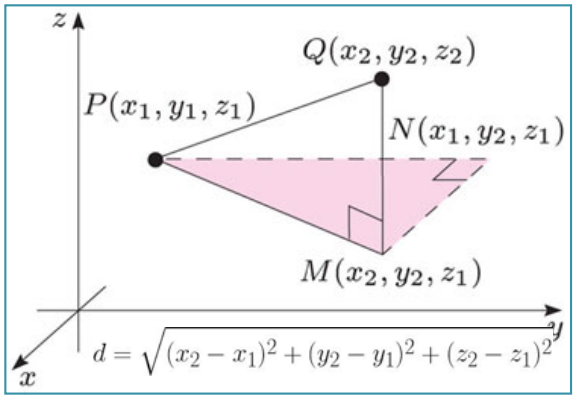


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments. **b** | For each expression pattern, 50 additional genes were generated, representing variations on the basic patterns.

Euclidean Distance

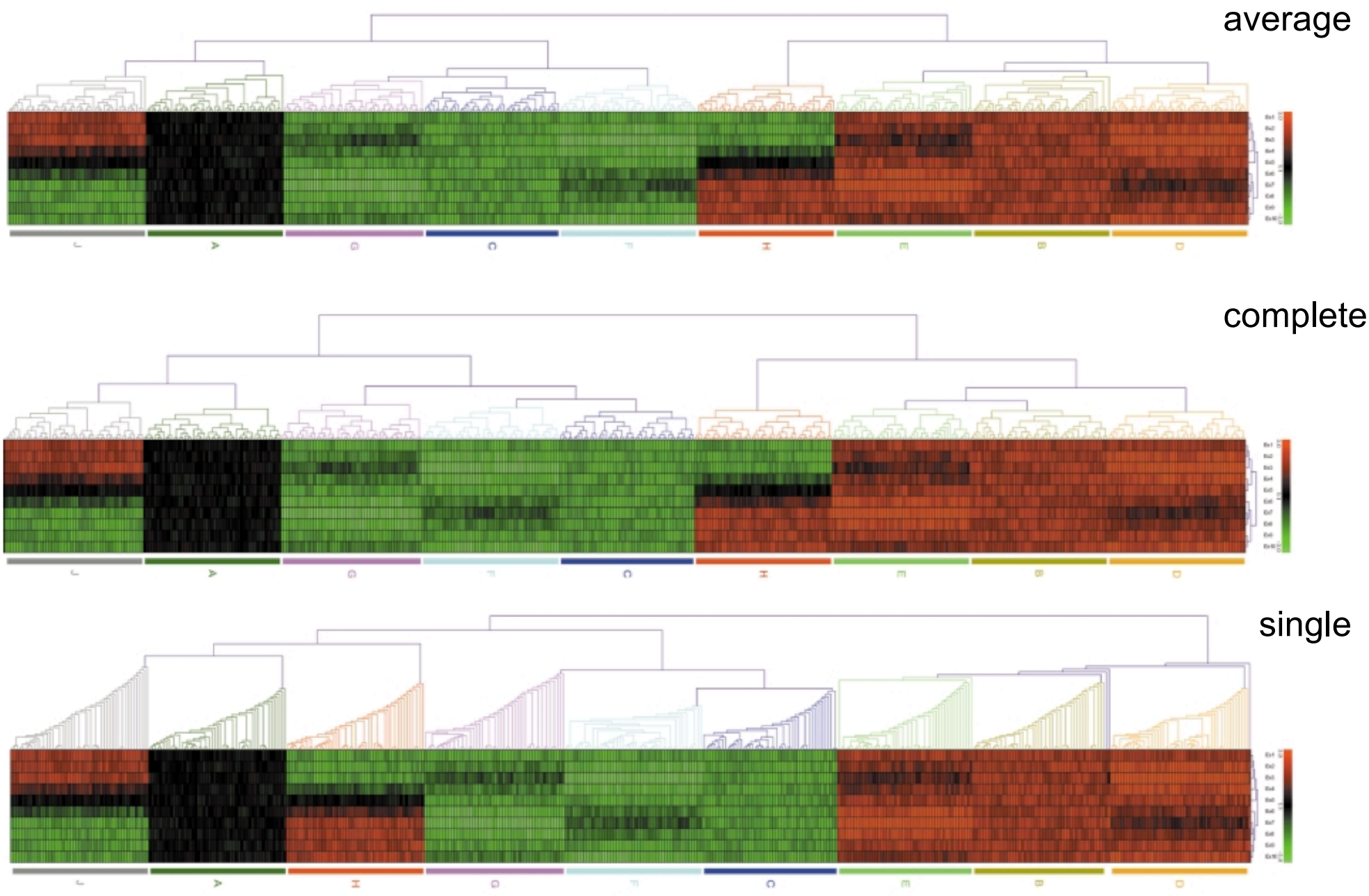




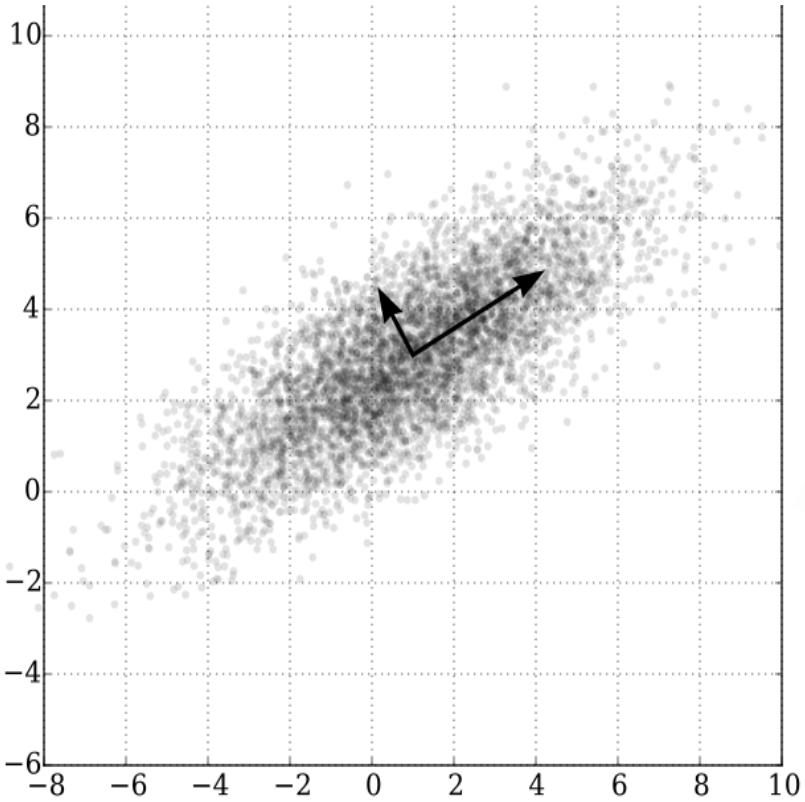
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

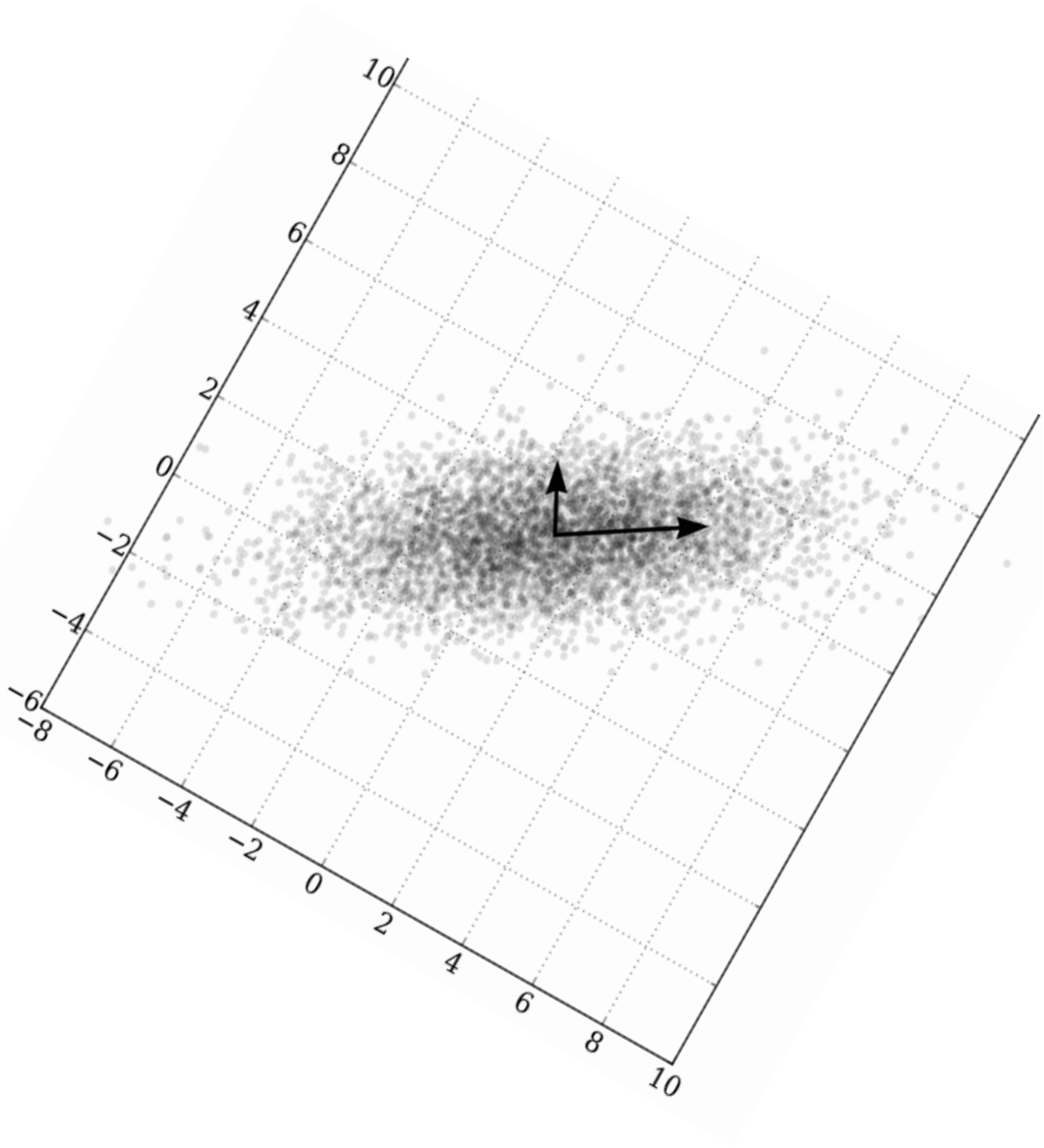
Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

Hierarchical Clustering



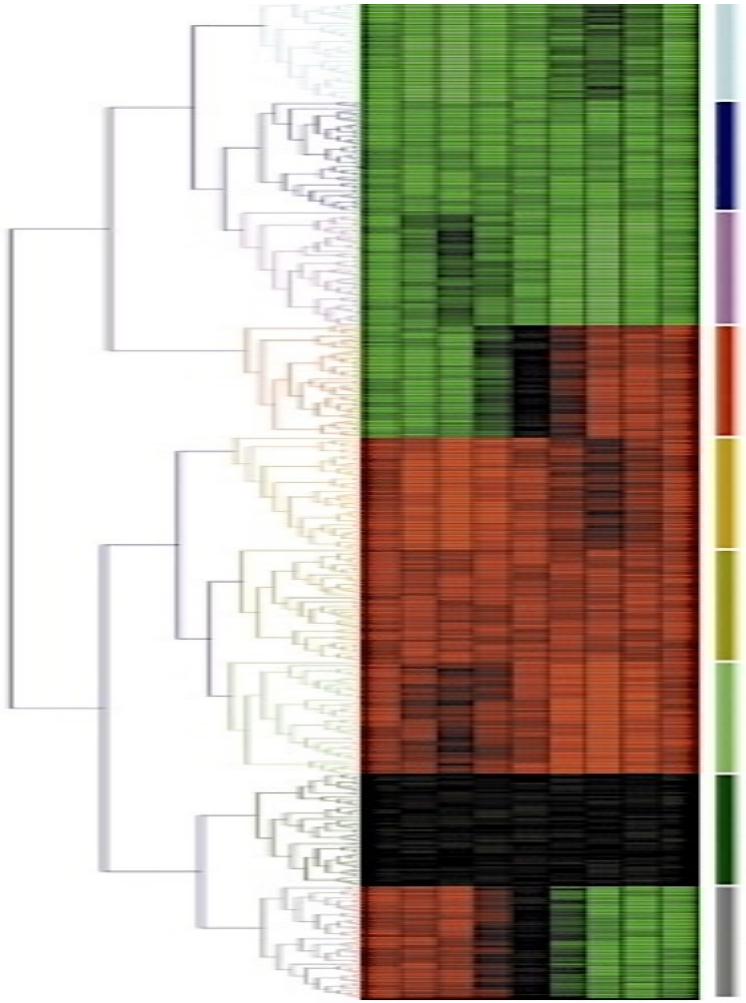
Principle Components Analysis (PCA)

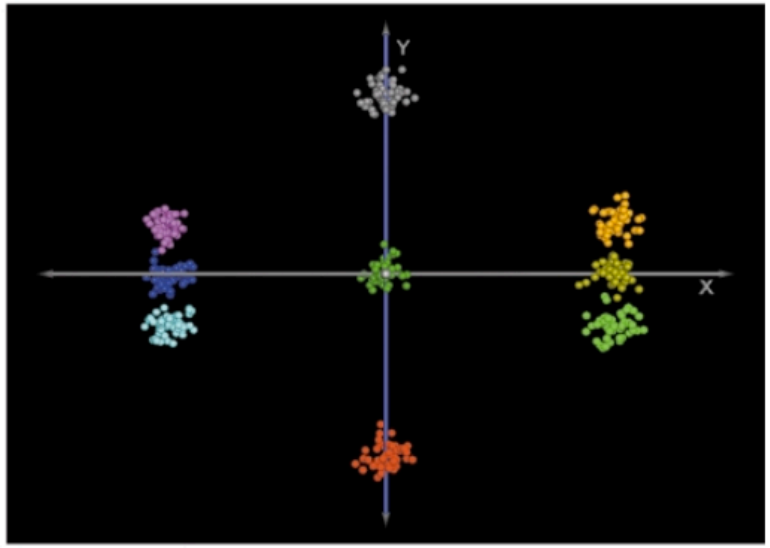




PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)





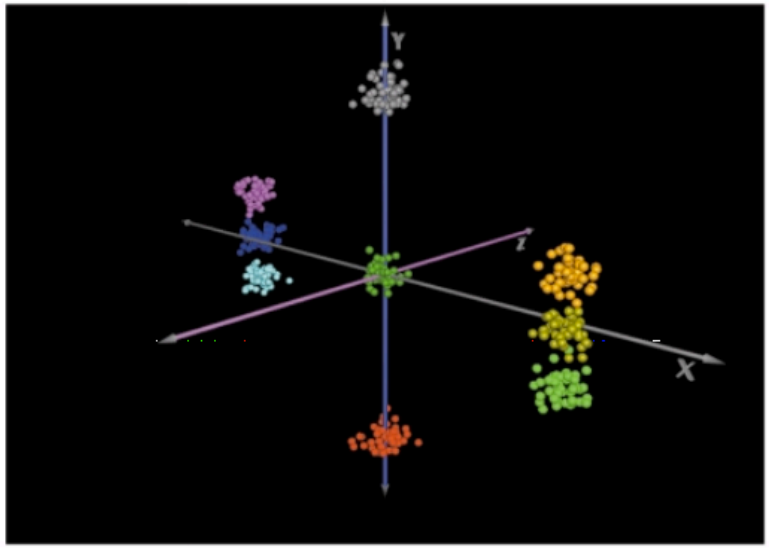
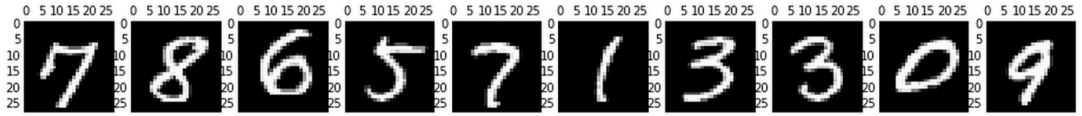
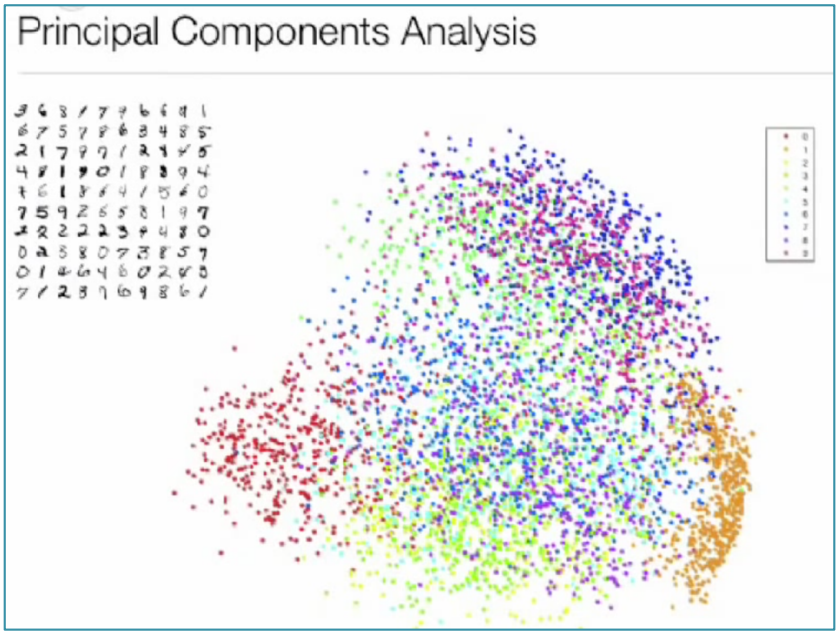
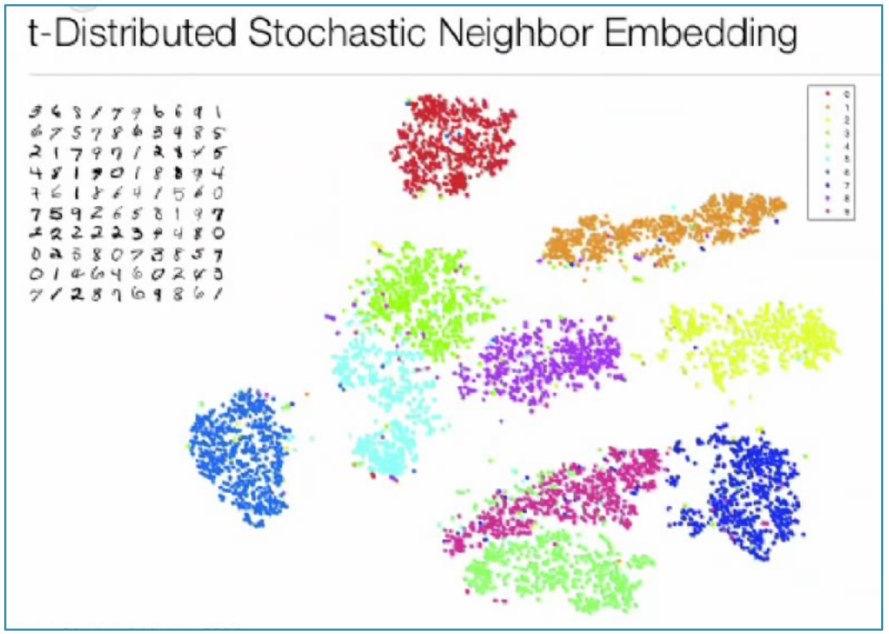


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering results for comparison.

PCA and t-SNE





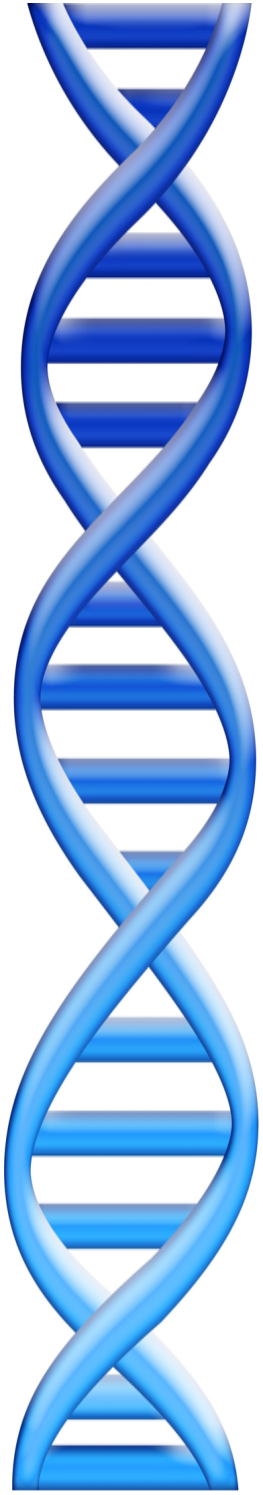


t-distributed Stochastic Neighborhood Embedding

- Non-linear dimensionality reduction technique: distances are only locally meaningful
- Rather than Euclidean distances, for each point fits a Gaussian kernel to fit the nearest N neighbors (perplexity) that define the probabilities that two points should be close together
- Using an iterative spring embedding system to place high probability points nearby

Visualizing Data Using t-SNE

https://www.youtube.com/watch?v=RJVL80Gg3lA



Part 1:

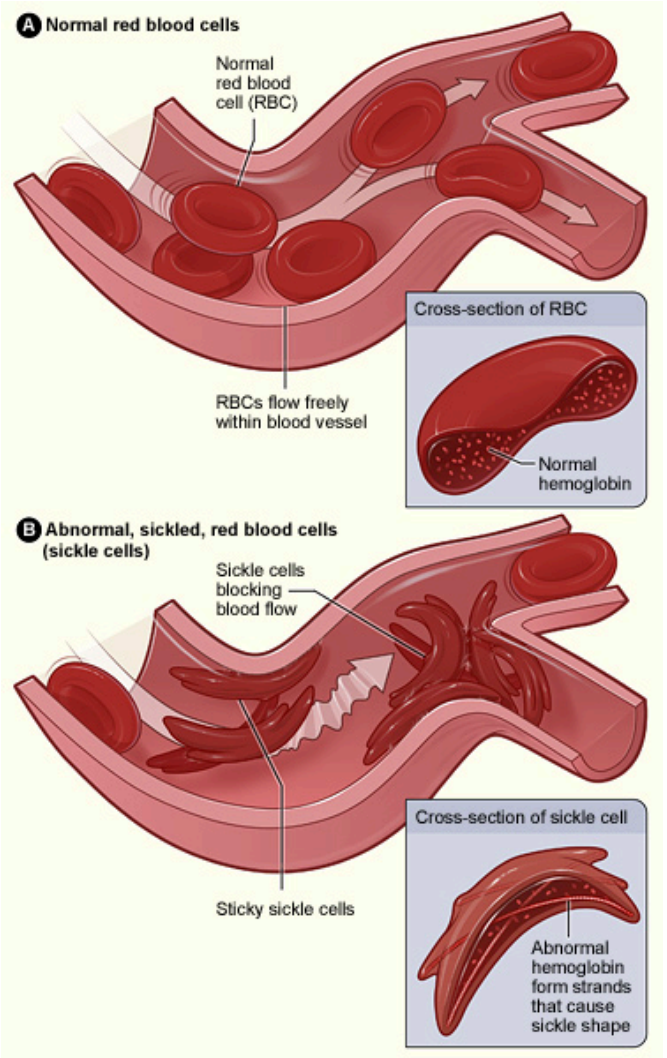
Pre-genome Era

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*

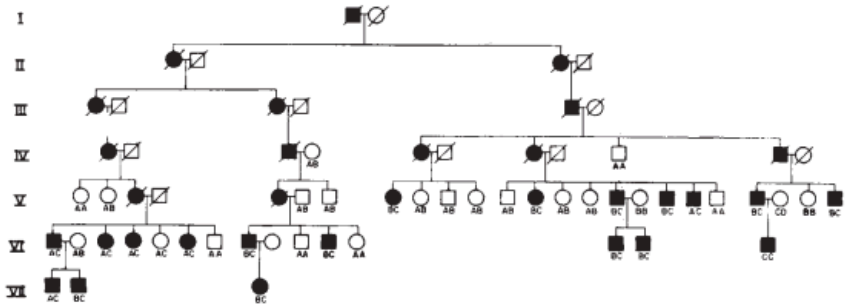


Fig. 2 Pedigree of the Venezuelan Huntington's disease family. This pedigree represents a small part of a much larger pedigree that will be described in detail elsewhere. Permanent EBV-transformed lymphoblastoid cell lines were established from blood samples of these individuals (unpublished data). DNA 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.

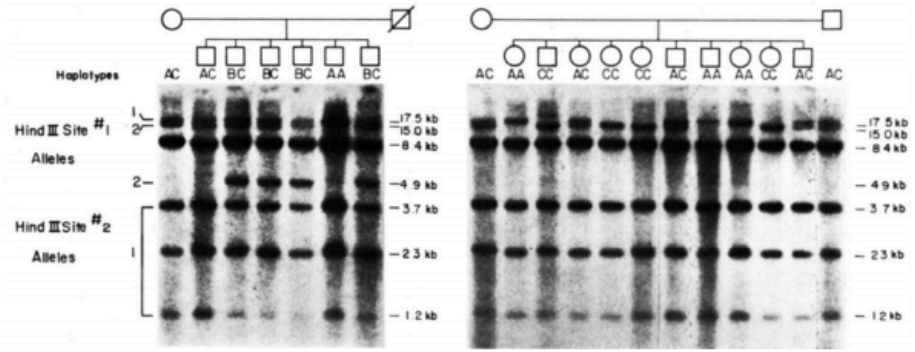


Fig. 3 Hybridization of the G8 Probe to *Hin*dIII-digested human genomic DNA.
Methods: DNA was prepared as described[23] from lymphoblastoid cell lines derived from members of two nuclear families. 5 μg of each DNA was digested to completion with 20 units of *Hin*dIII in a volume of 30 μl using the buffer 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

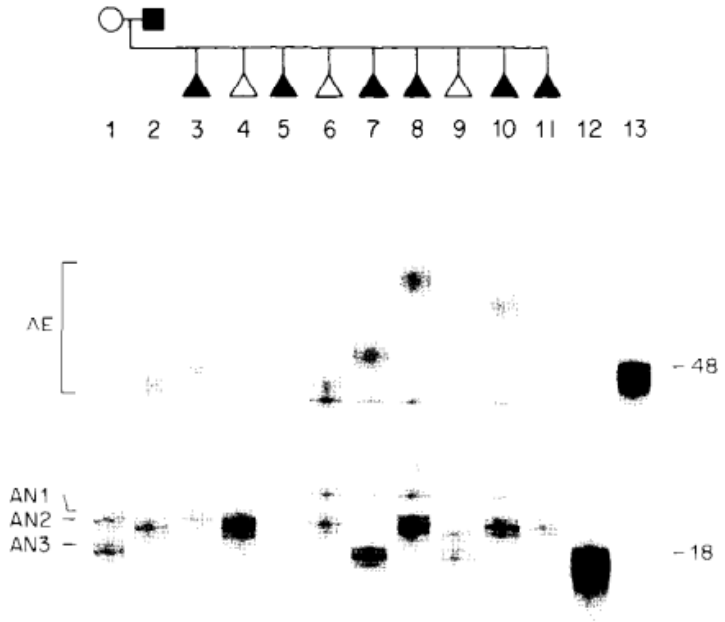


Figure 6. PCR Analysis of the $(CAG)_n$ Repeat in a Venezuelan HD Sibship with Some Offspring Displaying Juvenile Onset

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.

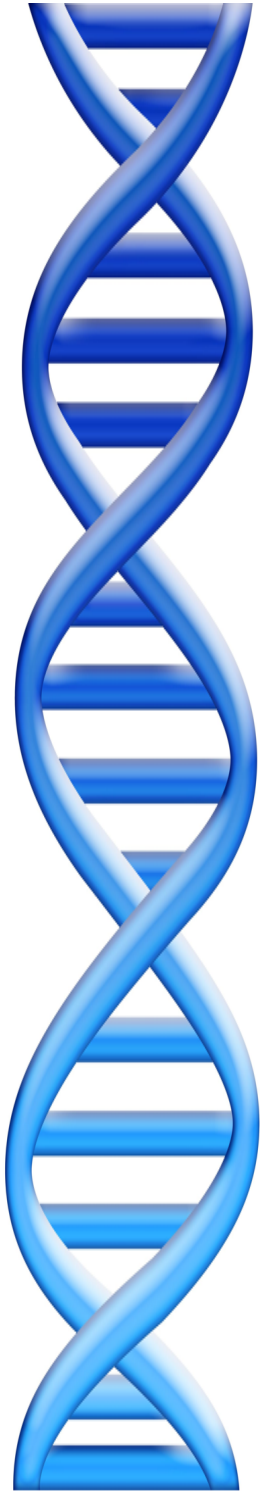
To test the proposal that classifying disease genes and their products according to function will provide general insight into disease processes[1,2], we have compiled and classified a list of disease genes. To assemble the list, we began with 269 genes identified in a survey of the 7th edition of *Metabolic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

Post-genome Inherited Dieases

"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

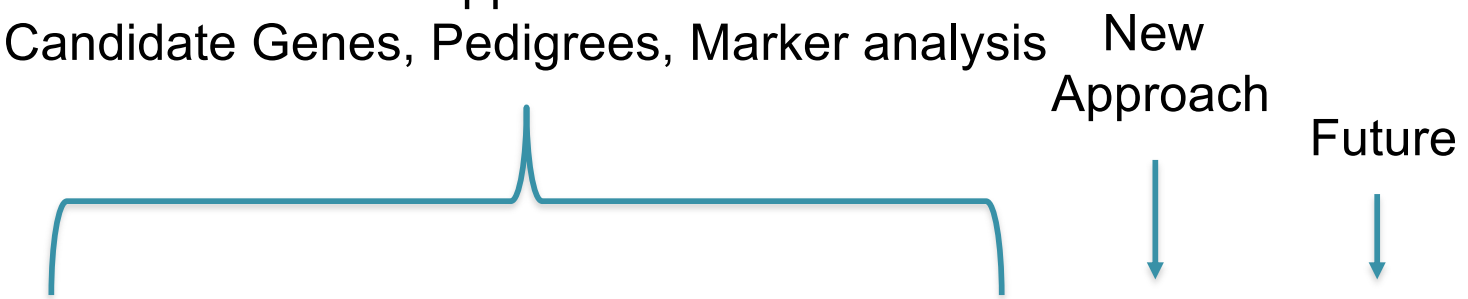


Table 1 | **Approaches to identifying variants underlying complex traits and common diseases**

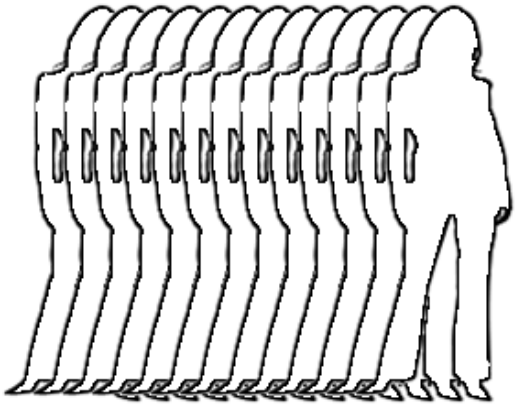
Potential advantages	Association*	Resequencing*	Linkage‡	Admixture‡	Missense SNPs‡	Association‡	Resequencing‡
No prior information regarding gene function required	–	–	+	+	+	+	+
Localization to small genomic region	+	+	–	–	+	+	+
Inexpensive	+	–	+	+	+/–	–	Prohibitive
Families not required	+	+	–	+	+	+	+
No assumptions necessary regarding type of variant involved	+	–	+	+	–	+	+
Not susceptible to effects of stratification§	–/+	–/+	+	+	–/+	–/+	–/+
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	+	+	+	+	+	+/–	–

*Candidate-gene studies. ‡Genome-wide studies. §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.

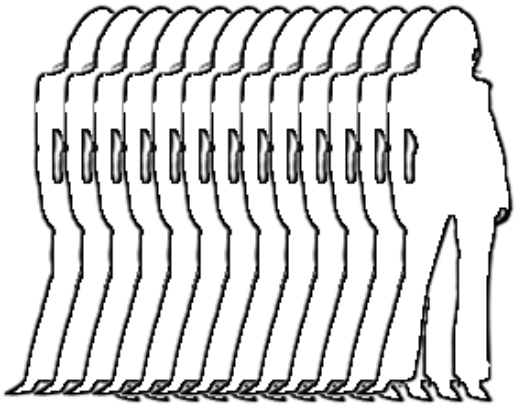
Genome-wide association studies for common diseases and complex traits

Hirschhorn and Daly (2005) Nature Review Genetics

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

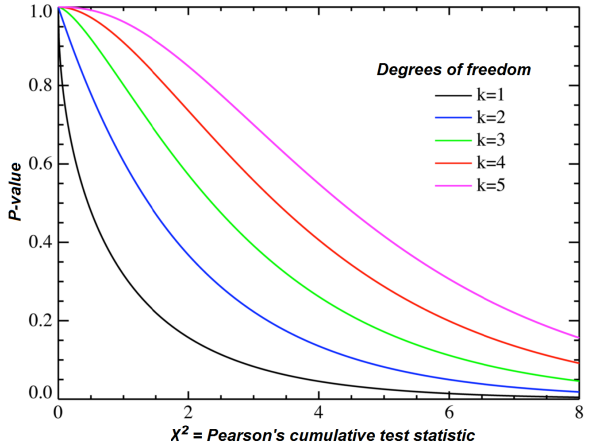
χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.

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.



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

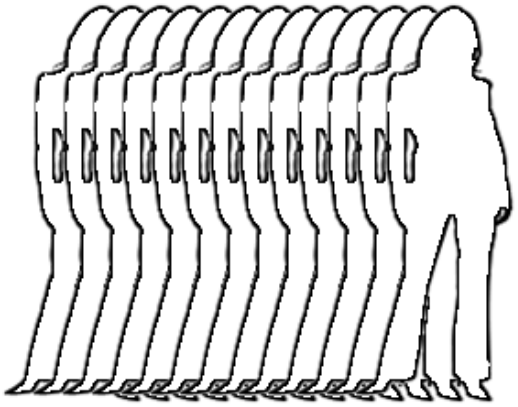
	has G	Not G	*Marginal Row Totals*
Cases	2104 (1912) [19.28]	1896 (2088) [17.66]	4000
Controls	2676 (2868) [12.85]	3324 (3132) [11.77]	6000
Marginal Column Totals	4780	5220	10000 (Grand Total)

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

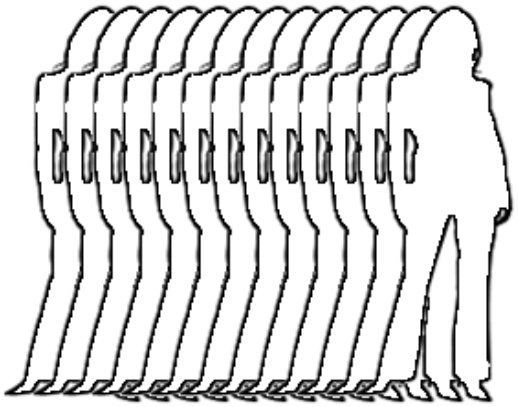
Cases/hasG squared deviation: $(2104 - 1912)^2 / 1912 = 19.28$ deviation

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

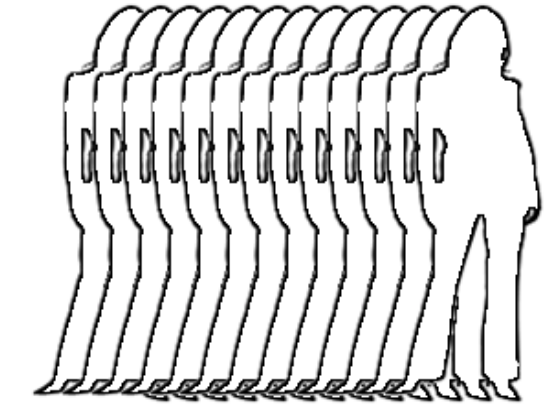


GC CC GC GC GG CC CC
CC GC GC GG GC GG

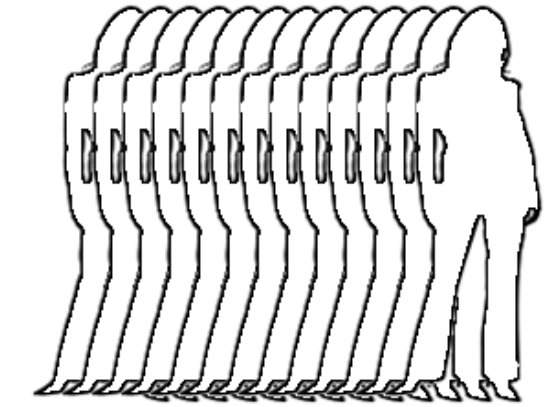
SNP1	*SNP2*	*SNP …*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:**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

Chi-squared or similar test

Genome Wide Association (GWAS)



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



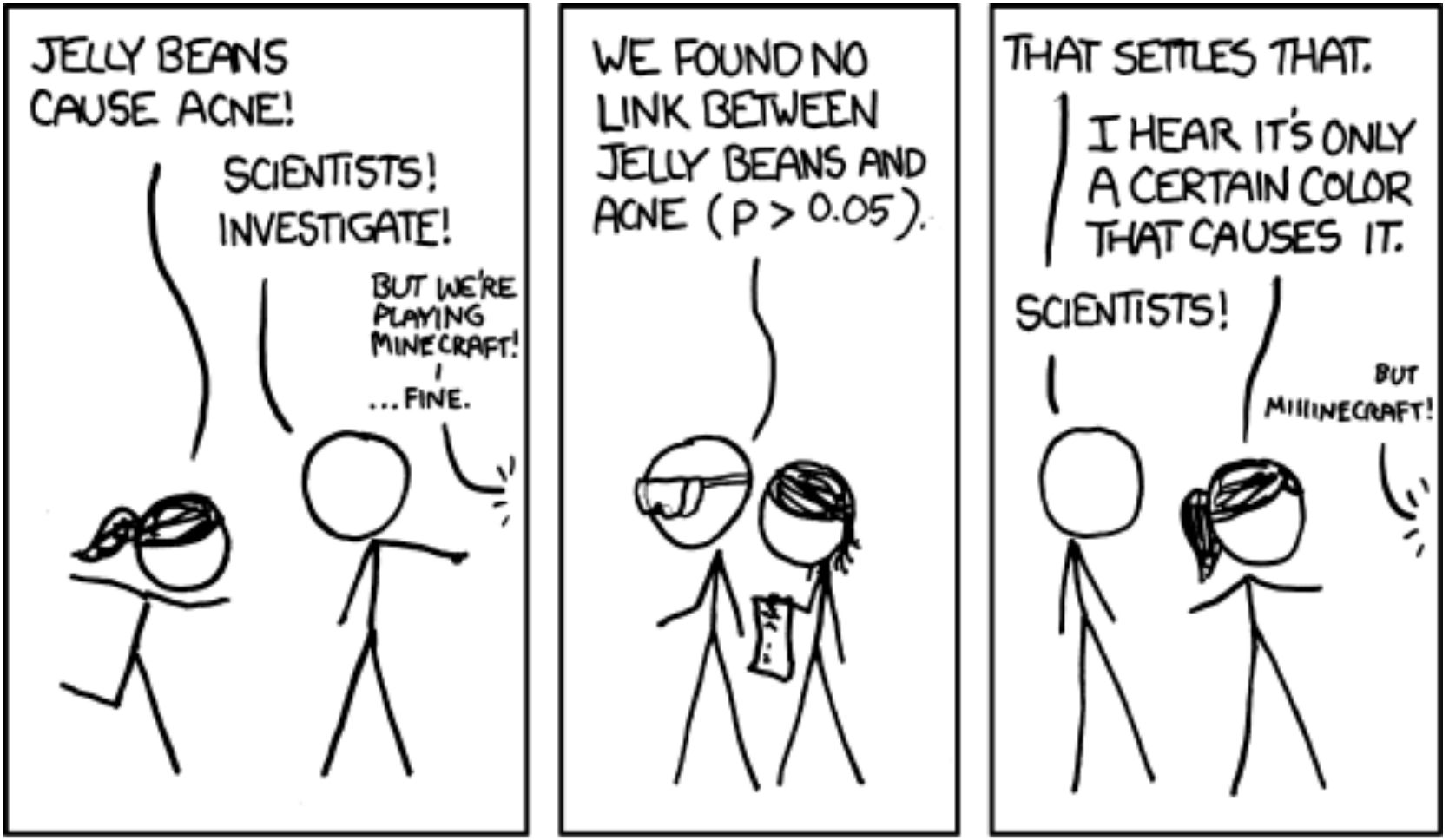
GC CC GC GC GG CC CC
CC GC GC GG GC GG

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

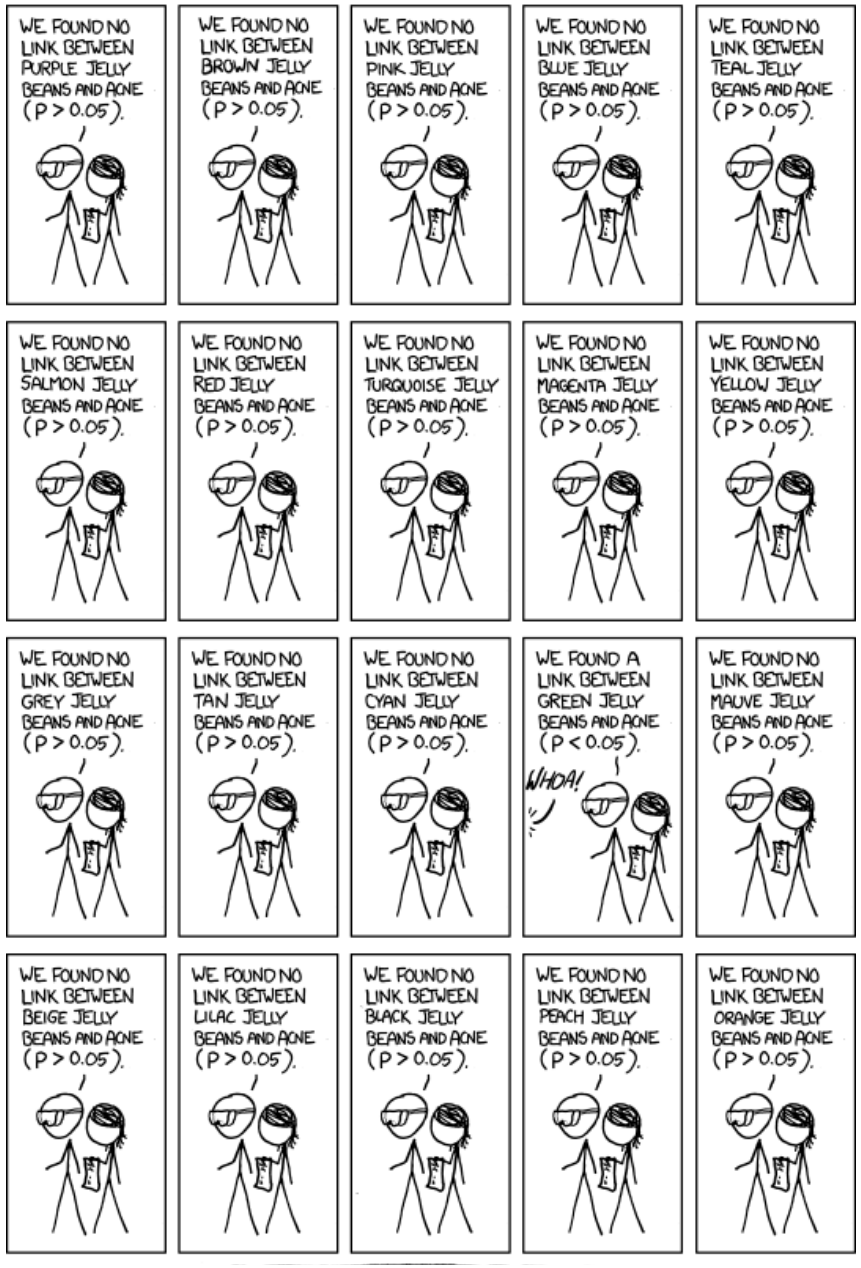
With a (much) larger population, this might be a significant difference in rate:
25320/60000 =>
p = 5e-7

Chi-squared or similar test

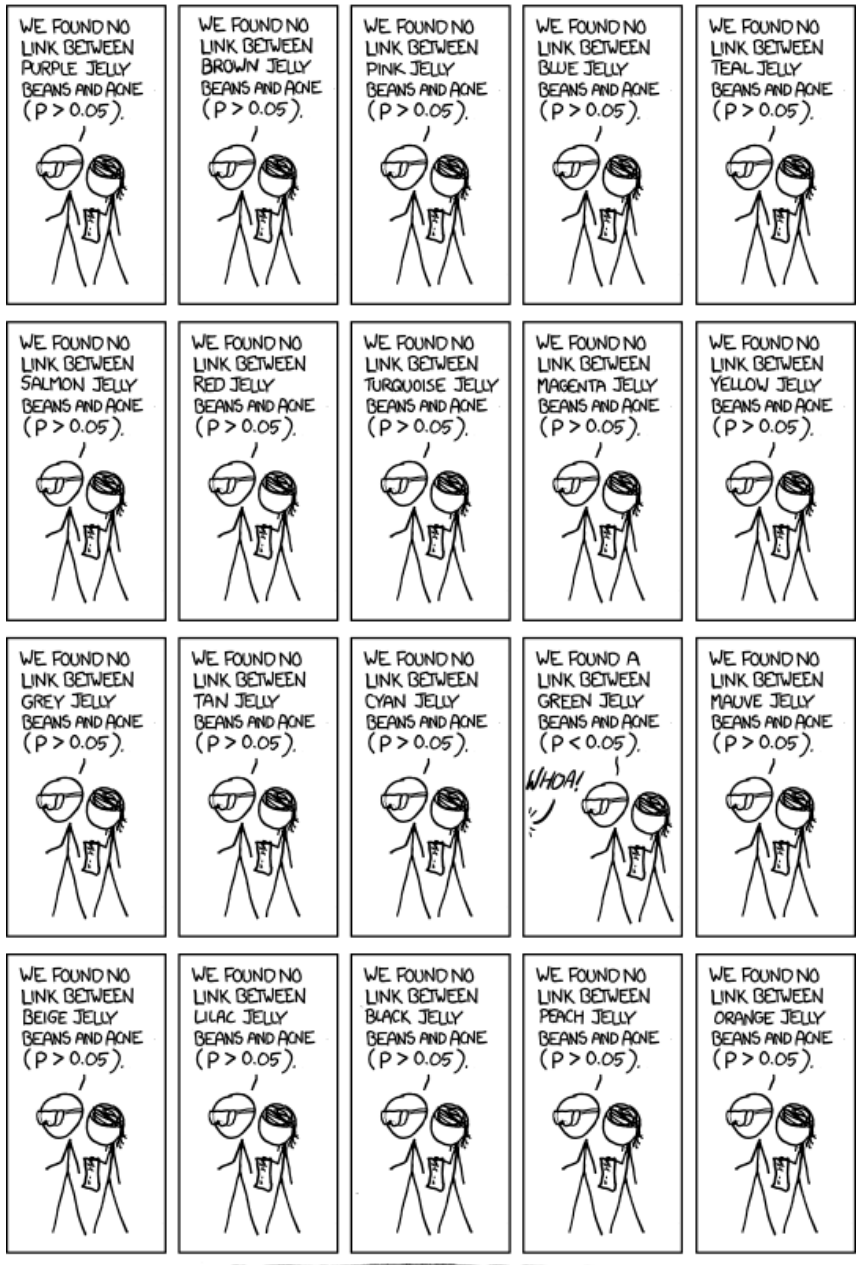
The curse of multiple testing

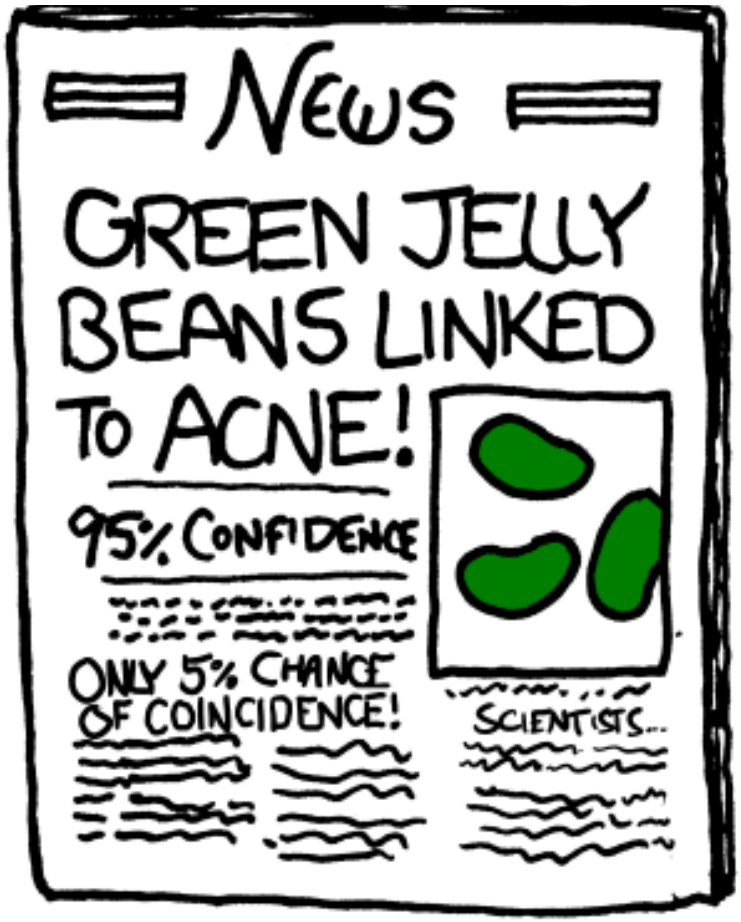


The curse of multiple testing

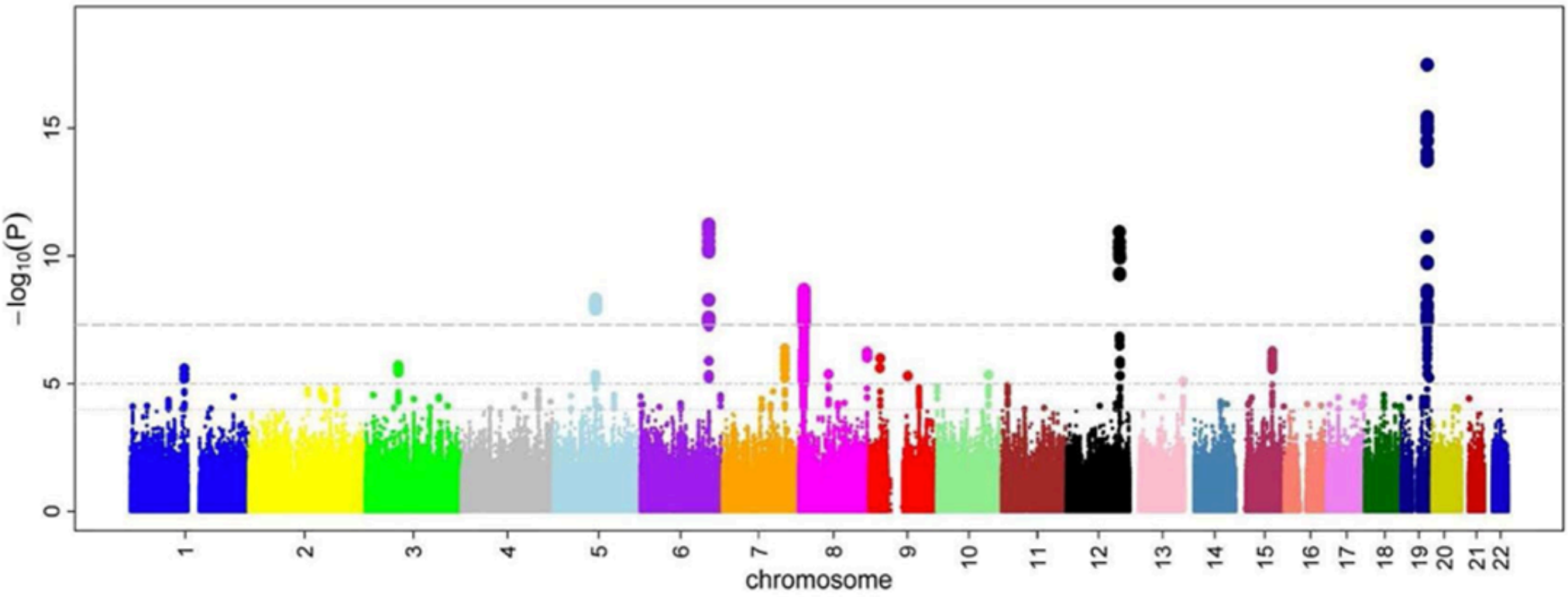


The curse of multiple testing



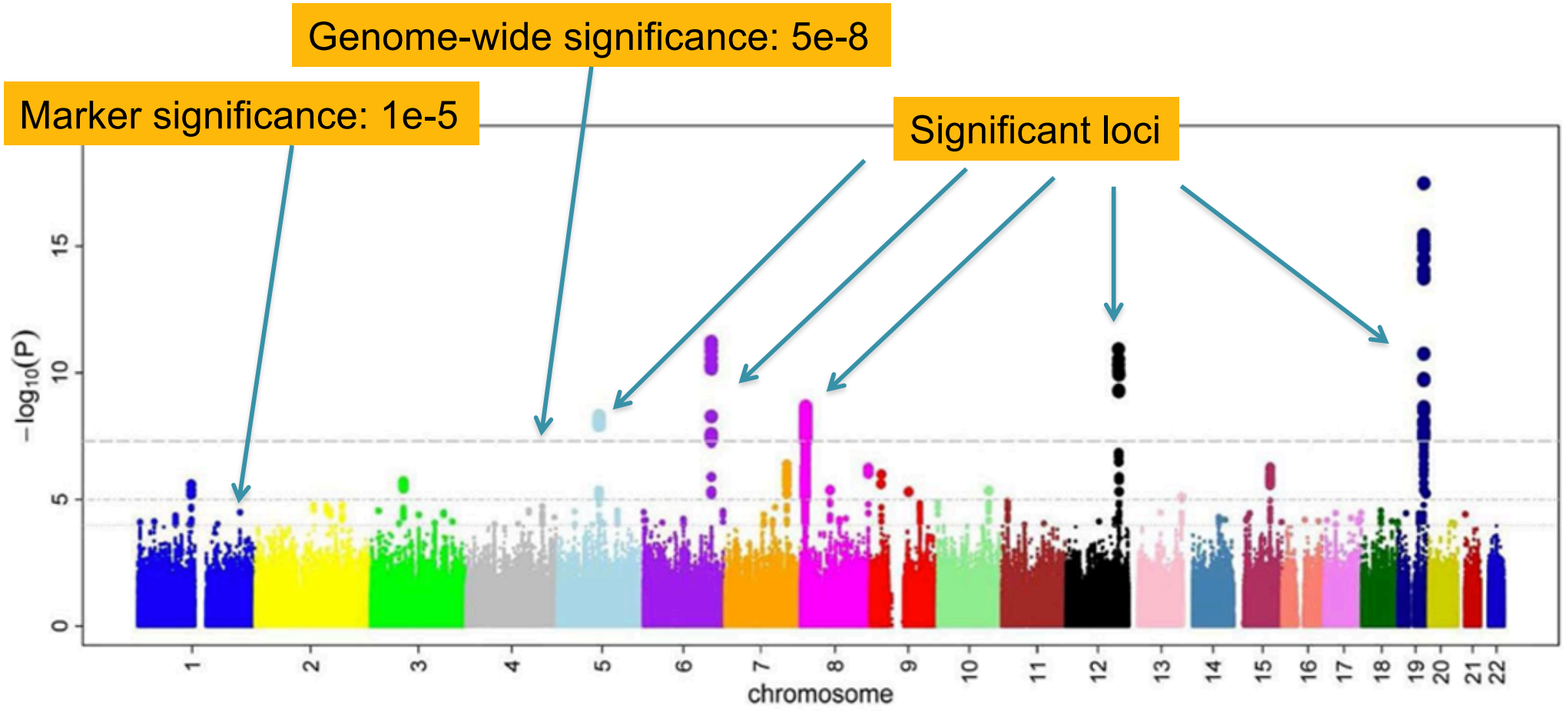


Manhattan Plot



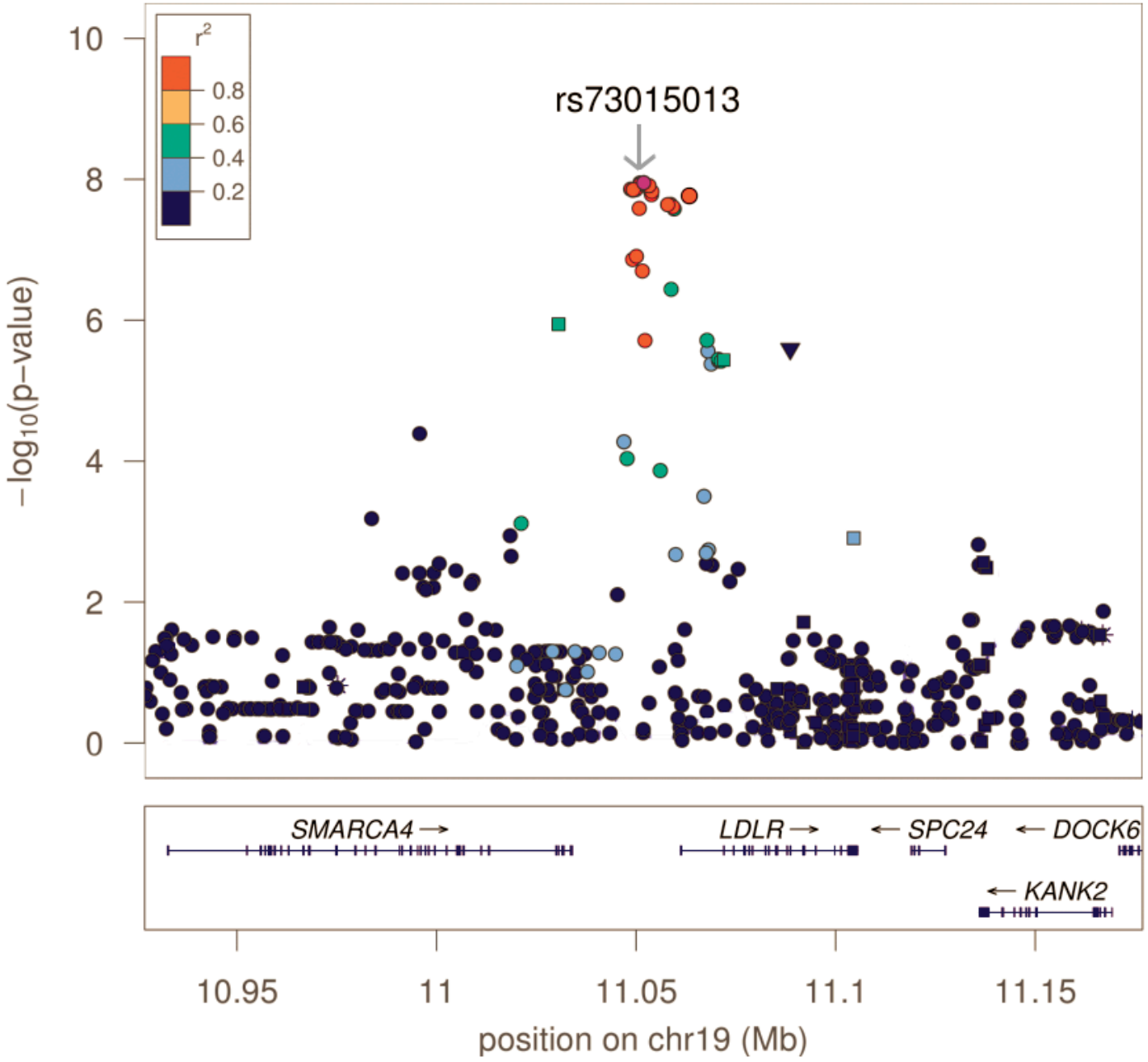
Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

Manhattan Plot

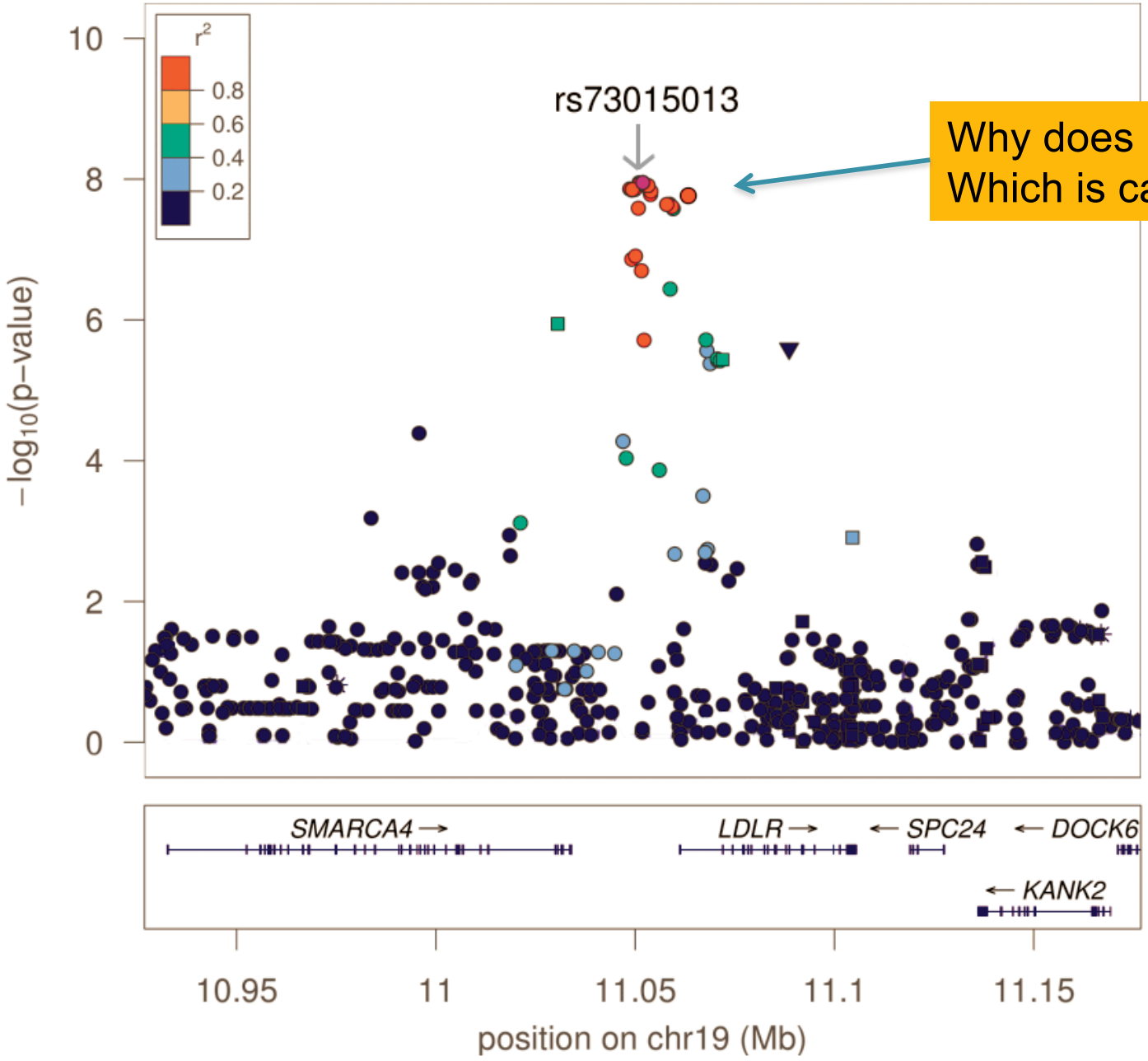


Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo
Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

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

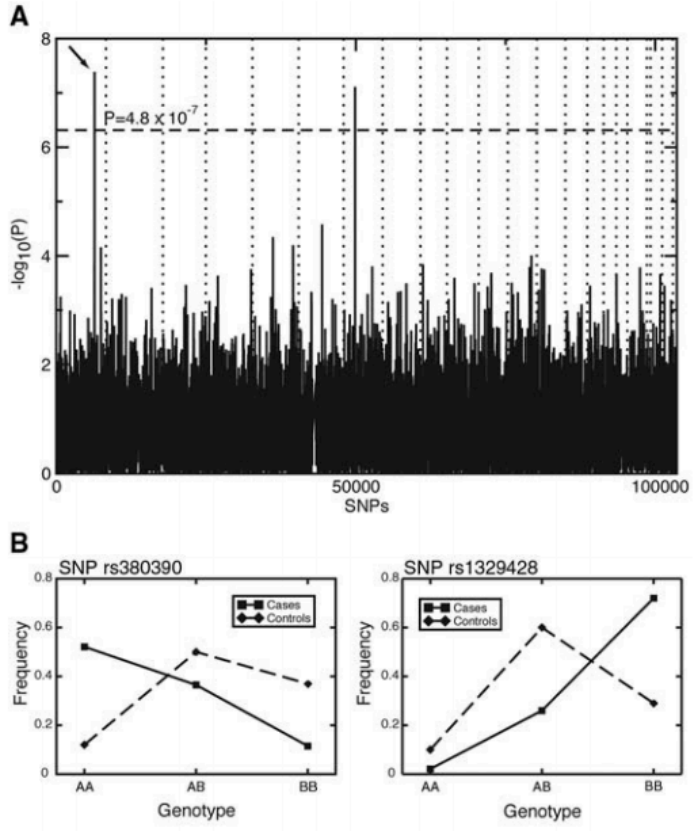


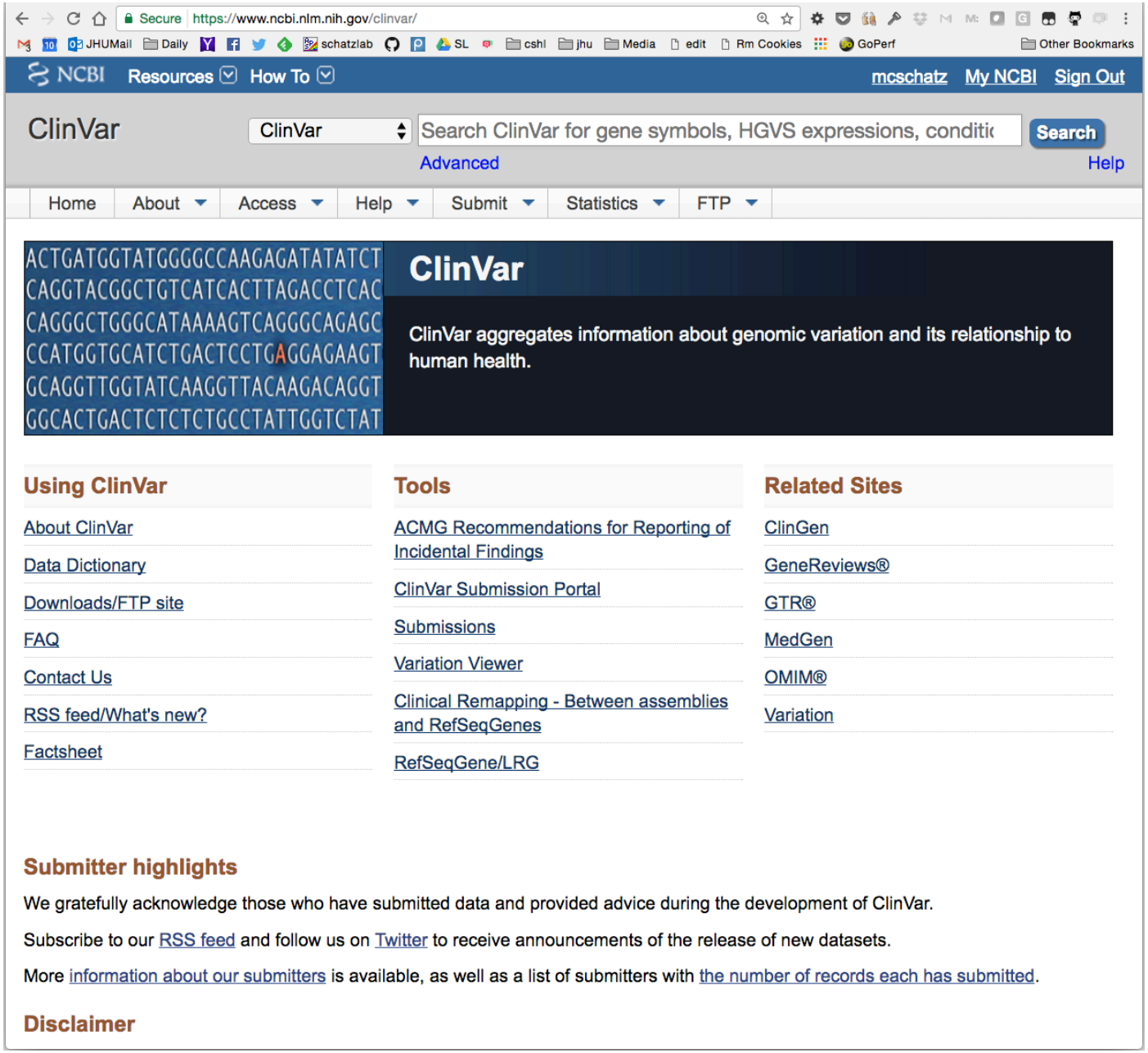
Fig. 1. (**A**) *P* values of genome-wide association scan for genes that affect the risk of developing AMD. $-\log_{10}(p)$ is plotted for each SNP in chromosomal order. The spacing between SNPs on the plot is uniform and does not reflect distances between SNPs on the chromosomes. The dotted horizontal 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.

GWAS Catalog

As of 2019-09-24, the GWAS Catalog contains 4220 publications and 157,336 associations.

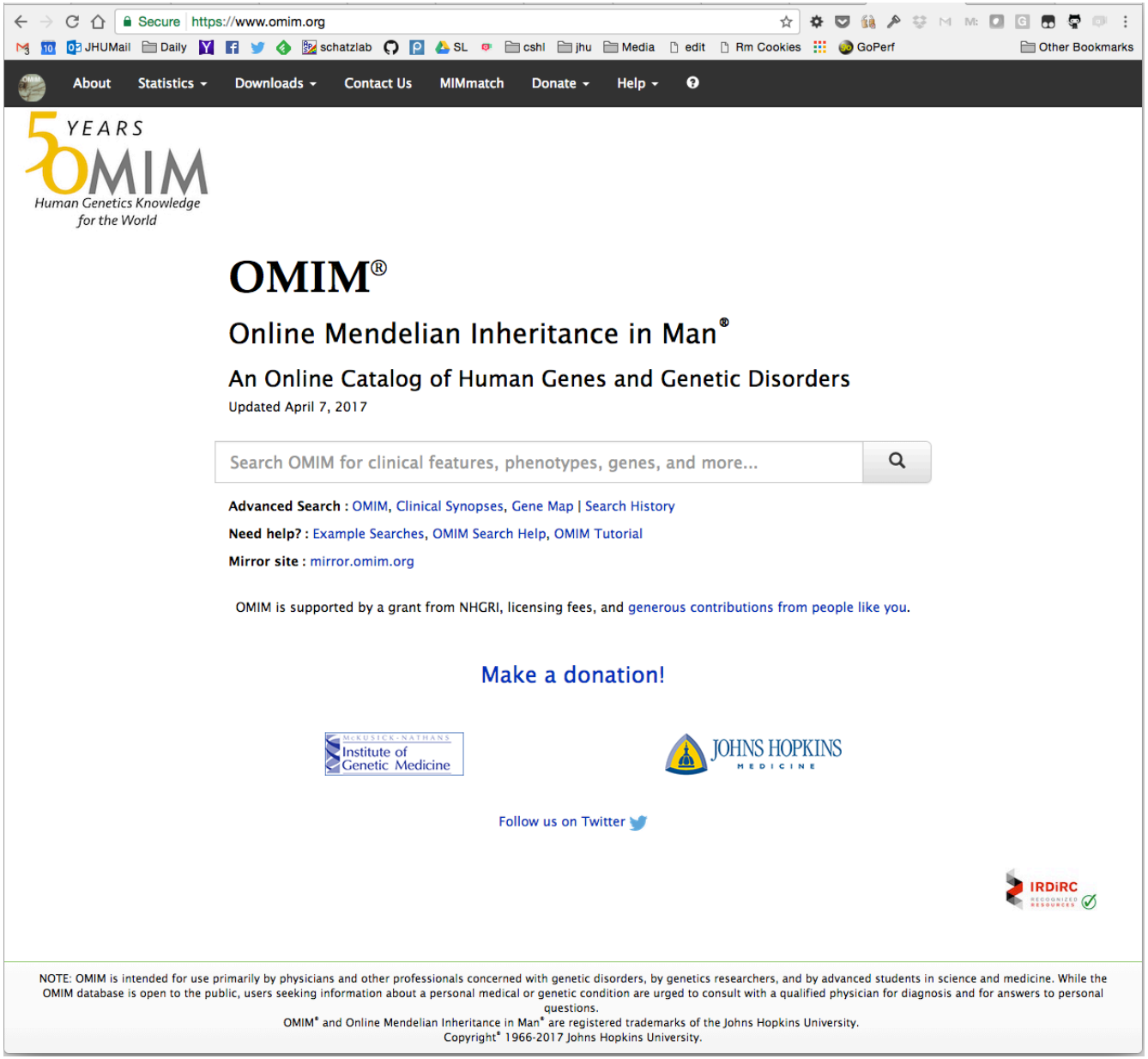
http://www.ebi.ac.uk/gwas/diagram

ClinVar



- 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 signifance

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 spanı
previously rep
the findings. M
and several ge
relevance to sc
in brain, assoc
support for th

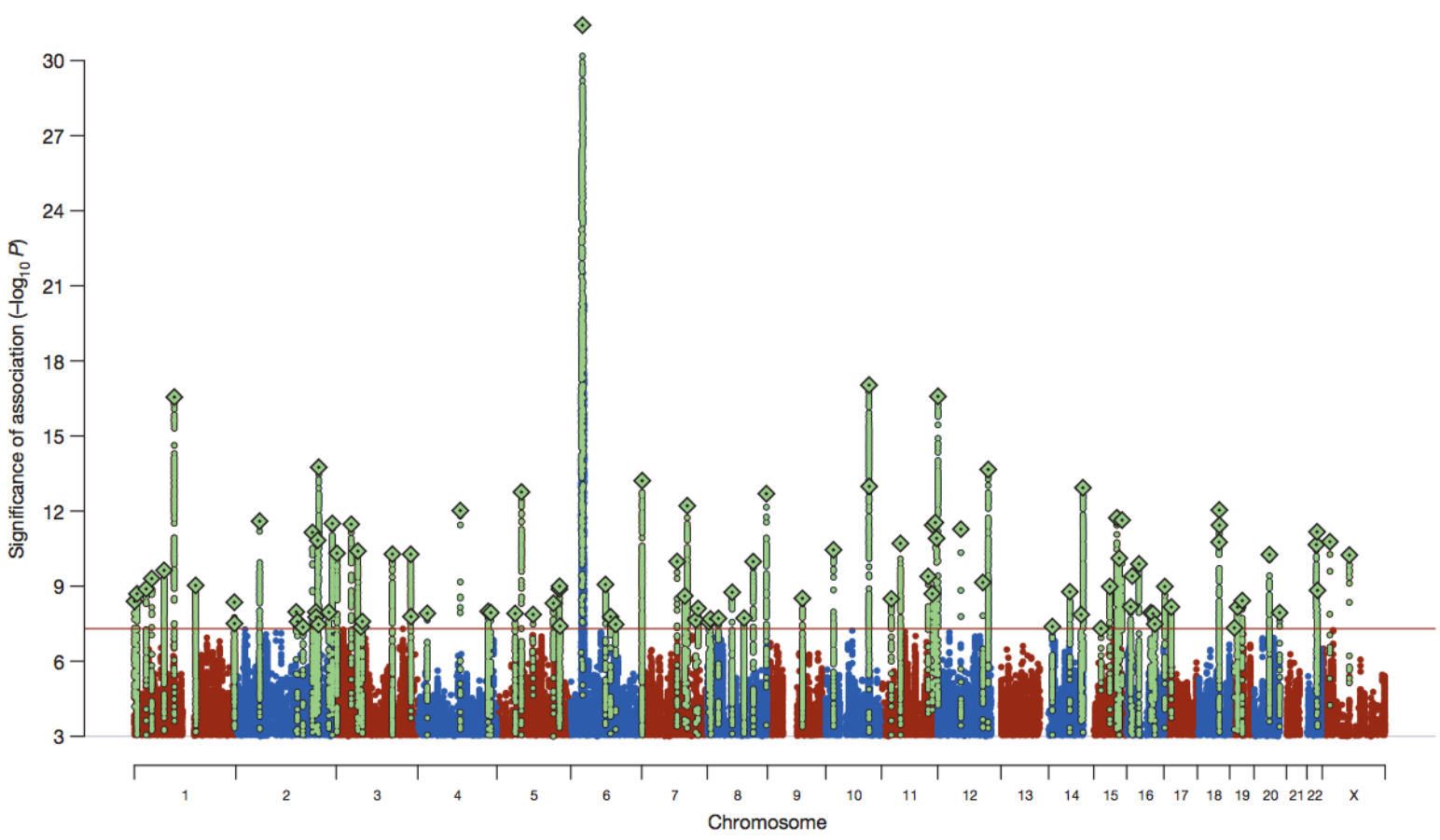


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal 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. M… and several ge… relevance to sc… in brain, assoc… support for th…

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

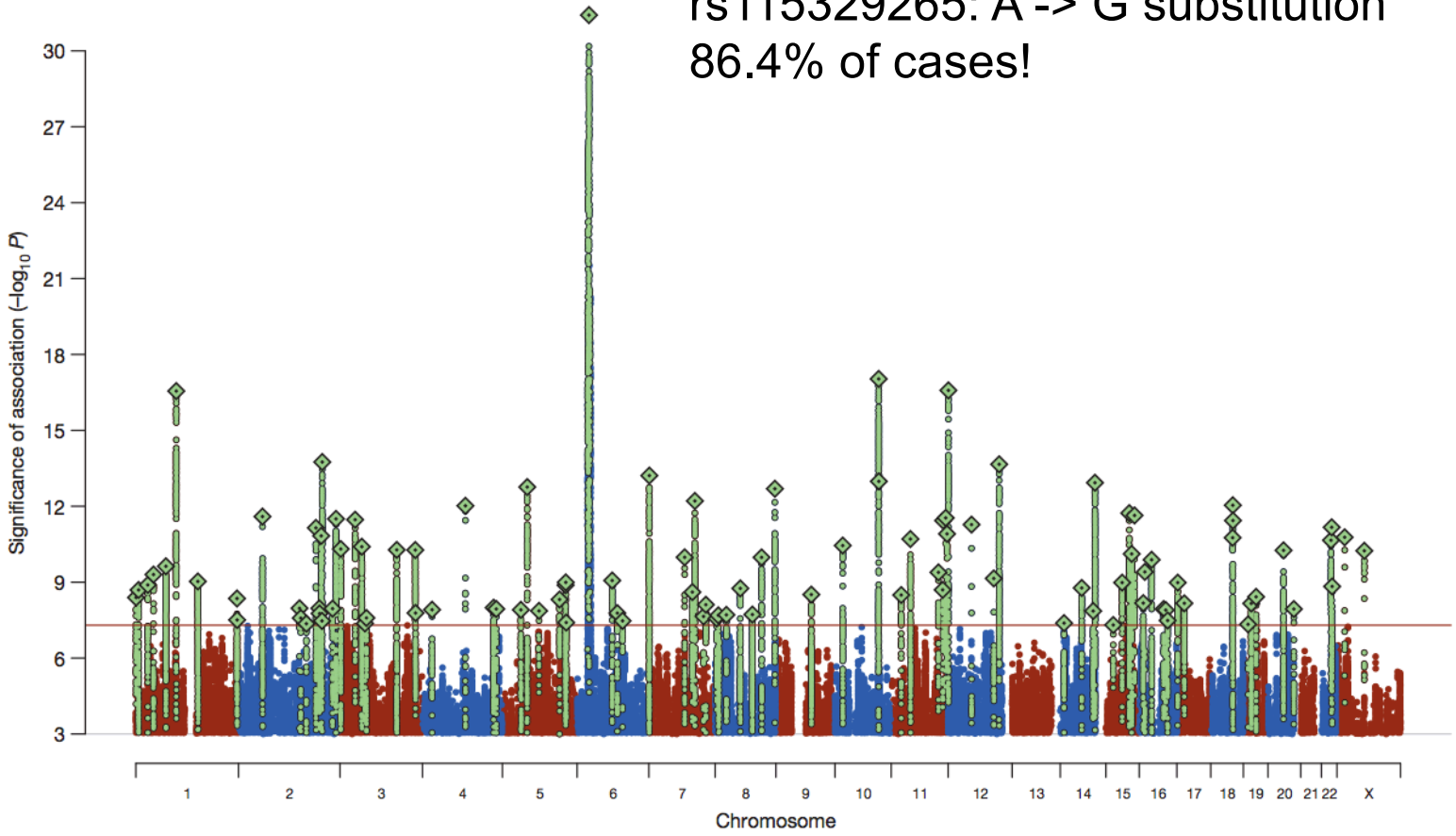


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal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

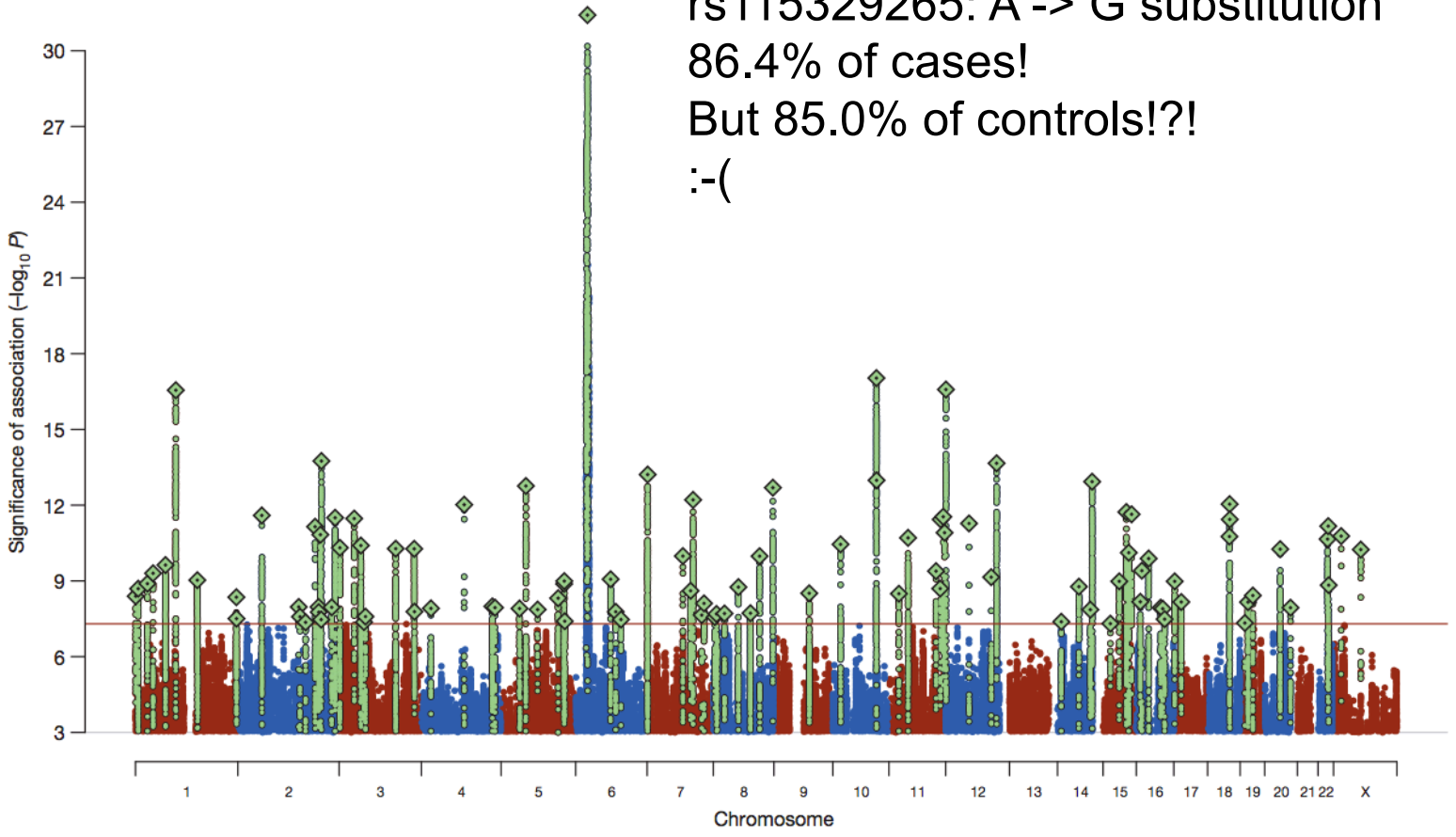


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal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

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	*CFH*/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), no large between-study inconsistency in effect sizes; also replicated in large Dutch cohort (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	*INSIG2*/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 failed to replicate the association: 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) failed to replicate any of the 13 proposed SNPs; 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)	*LTA*/Haplotype of 5 SNPs (2 in *LTA* and 3 in adjucent genes); the two *LTA* 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 no association (non-significant OR 1.07 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	*HTRA1*/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

Missing Heritability

The case of the missing heritability

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

If you want to predict how tall your children might one day be, a good bet would be to look in the mirror, and at your mate. Studies going back almost a century have estimated that height is 80–90% heritable. So if 29 centimetres separate the tallest 5% of a population from the shortest, then genetics would account for as many as 27 of them[1].

This year, three groups of researchers[2–4] scoured the genomes of huge populations (the largest study[4] looked at more than 30,000 people) for genetic variants associated with the height differences. More than 40 turned up.

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

Even though these genome-wide association studies (GWAS) turned up dozens of variants, they did "very little of the prediction that you would do just by asking people how tall their parents are", says Joel Hirschhorn at the Broad Institute in Cambridge, Massachusetts, who led one of the studies[3].

Height isn't the only trait in which genes have gone missing, nor is it the most important. Studies looking at similarities between identical and fraternal twins estimate heritability at more than 90% for autism[5] and more than 80% for schizophrenia[6]. And genetics makes a major contribution to disorders such as obesity, diabetes and heart disease. GWAS, one of the most celebrated techniques of the past five years, promised to deliver many of the genes involved (see 'Where's the reward?', page 20). And to some extent they have, identifying more than 400 genetic variants that contribute to a variety of traits and common diseases. But even when dozens of genes have been linked to a trait, both the individual and cumulative effects are disappointingly small and nowhere near enough to explain earlier estimates of heritability. "It is the big topic in the genetics of common disease right now," says Francis Collins, former head of the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. The unexpected results left researchers at a point "where we all had to scratch our heads and say, 'Huh?'", he says.

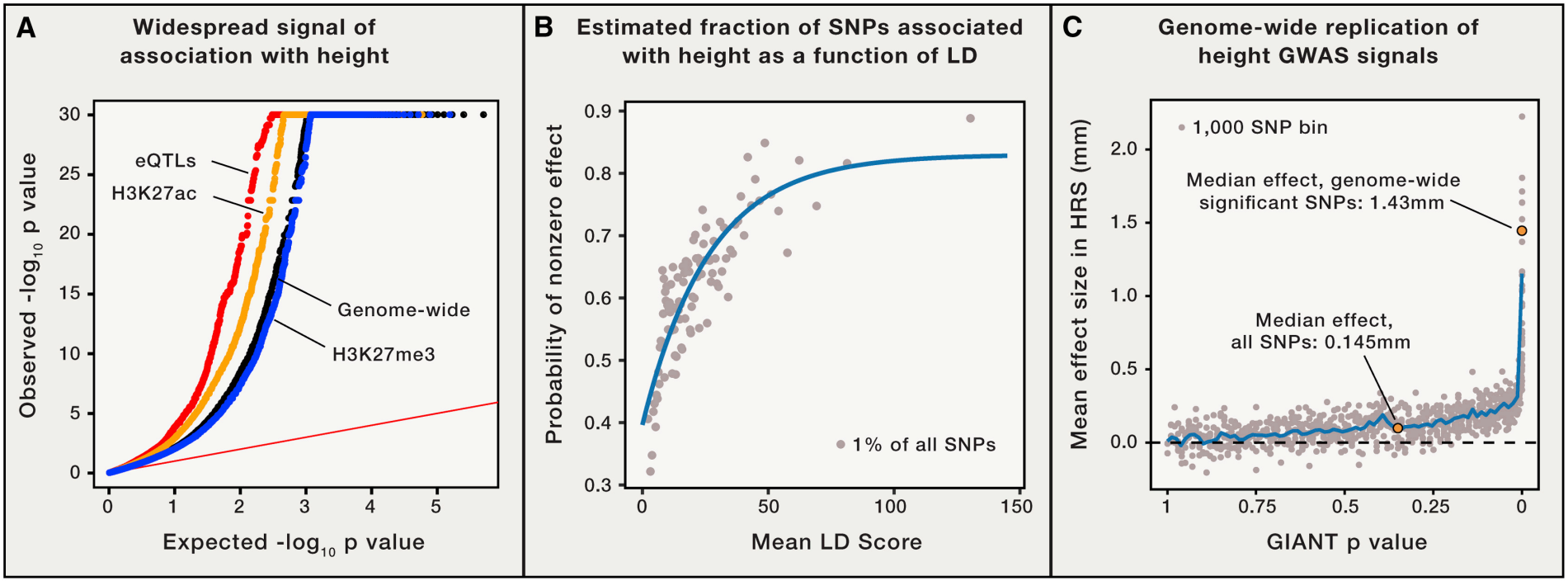
Although flummoxed by this missing heritability, geneticists remain optimistic that they can find more of it. "These are very early days, and there are things that are doable in the next year or two that may well explain another sizeable chunk of heritability," says Hirschhorn. So where might it be hiding?

ILLUSTRATIONS BY D. PARKINS

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

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

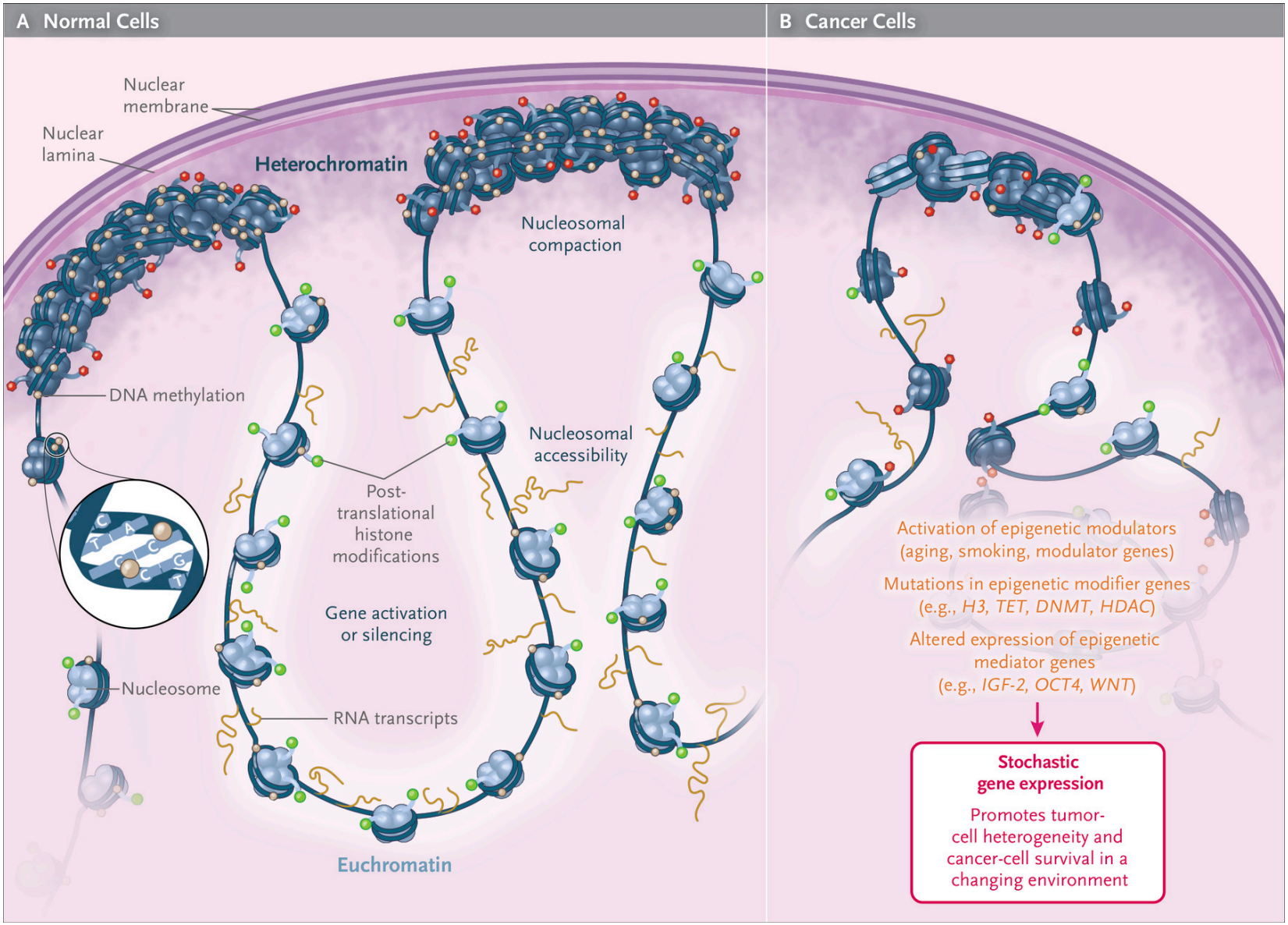
Omnigenics



A central goal of genetics is to understand the links between genetic variation and disease. Intuitively, one might expect disease-causing variants to cluster into key pathways that drive disease etiology. But for complex traits, association signals tend to be spread across most of the genome—including near many genes without an obvious connection to disease. **We propose that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways. We refer to this hypothesis as an "omnigenic" model.**

An Expanded View of Complex Traits: From Polygenic to Omnigenic
Boyle, Li, Pritchard (2017) Cell. https://doi.org/10.1016/j.cell.2017.05.038

Epigenetic Factors



The Key Role of Epigenetics in Human Disease Prevention and Mitigation
Feinberg (2018) NEJM. doi: 10.1056/NEJMra1402513

Needles in stacks of needles

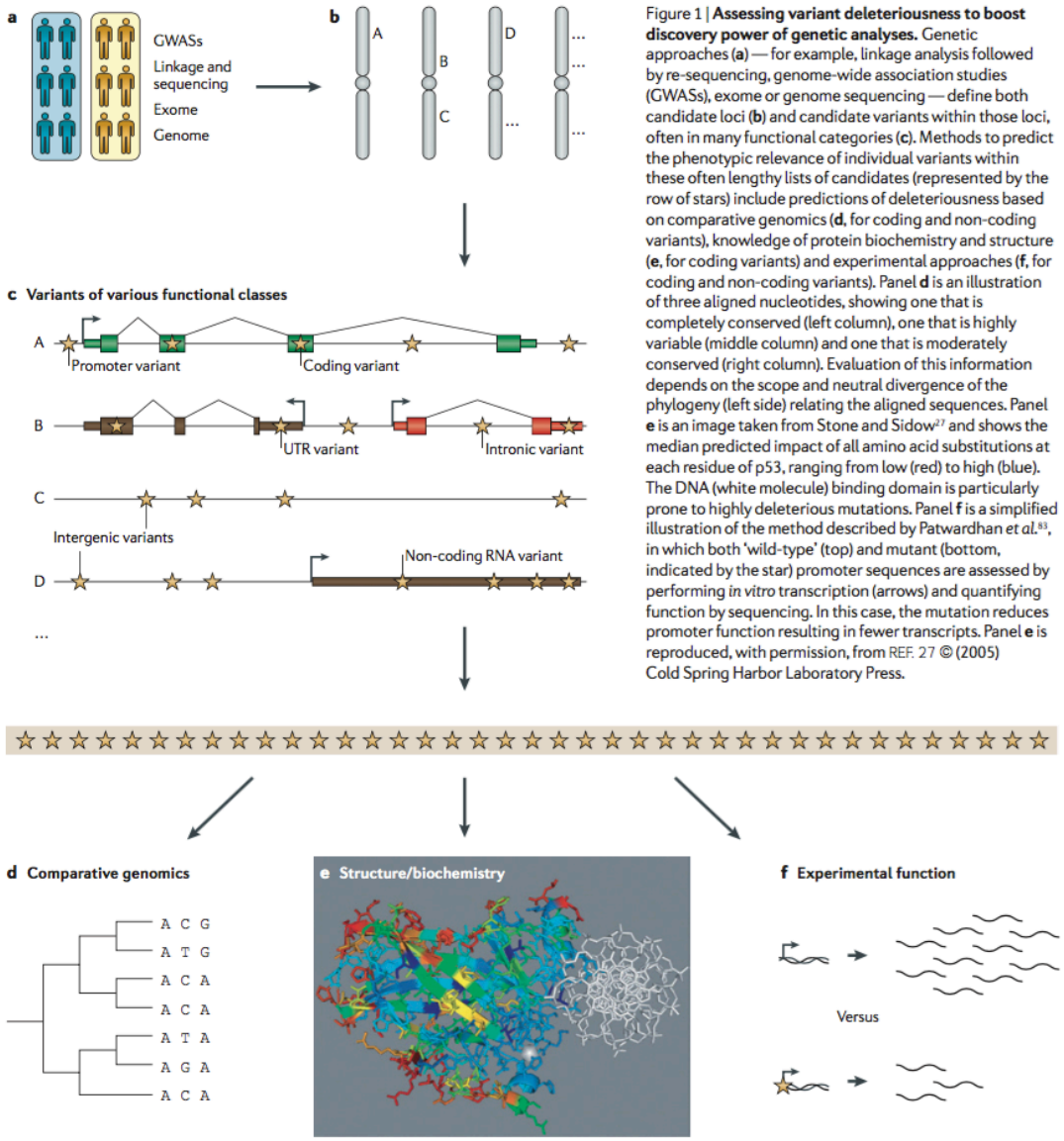


Figure 1 | **Assessing variant deleteriousness to boost discovery power of genetic analyses.** Genetic approaches (**a**) — for example, linkage analysis followed by re-sequencing, genome-wide association studies (GWASs), exome or genome sequencing — define both candidate loci (**b**) and candidate variants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

ARTICLE

Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

Alan F. Rope,[1] Kai Wang,[2,19] Rune Evjenth,[3] Jinchuan Xing,[4] Jennifer J. Johnston,[5] Jeffrey J. Swensen,[6,7] W. Evan Johnson,[8] Barry Moore,[4] Chad D. Huff,[4] Lynne M. Bird,[9] John C. Carey,[1] John M. Opitz,[1,4,6,10,11] Cathy A. Stevens,[12] Tao Jiang,[13,14] Christa Schank,[8] Heidi Deborah Fain,[15] Reid Robison,[15] Brian Dalley,[16] Steven Chin,[6] Sarah T. South,[1,7] Theodore J. Pysher,[6] Lynn B. Jorde,[4] Hakon Hakonarson,[2] Johan R. Lillehaug,[3] Leslie G. Biesecker,[5] Mark Yandell,[4] Thomas Arnesen,[3,17] and Gholson J. Lyon[15,18,20,*]

We have identified two families with a previously undescribed lethal X-linked disorder of infancy; the disorder comprises a distinct combination of an aged appearance, craniofacial anomalies, hypotonia, global developmental delays, cryptorchidism, and cardiac arrhythmias. Using X chromosome exon sequencing and a recently developed probabilistic algorithm aimed at discovering disease-causing variants, we identified in one family a c.109T>C (p.Ser37Pro) variant in *NAA10*, a gene encoding the catalytic subunit of the major human N-terminal acetyltransferase (NAT). A parallel effort on a second unrelated family converged on the same variant. The absence of this variant in controls, the amino acid conservation of this region of the protein, the predicted disruptive change, and the co-occurrence in two unrelated families with the same rare disorder suggest that this is the pathogenic mutation. We confirmed this by demonstrating a significantly impaired biochemical activity of the mutant hNaa10p, and from this we conclude that a reduction in acetylation by hNaa10p causes this disease. Here we provide evidence of a human genetic disorder resulting from direct impairment of N-terminal acetylation, one of the most common protein modifications in humans.

Am J Hum Genet. 2011 Aug 12;89(2):345. doi: 10.1016/j.ajhg.2011.05.017

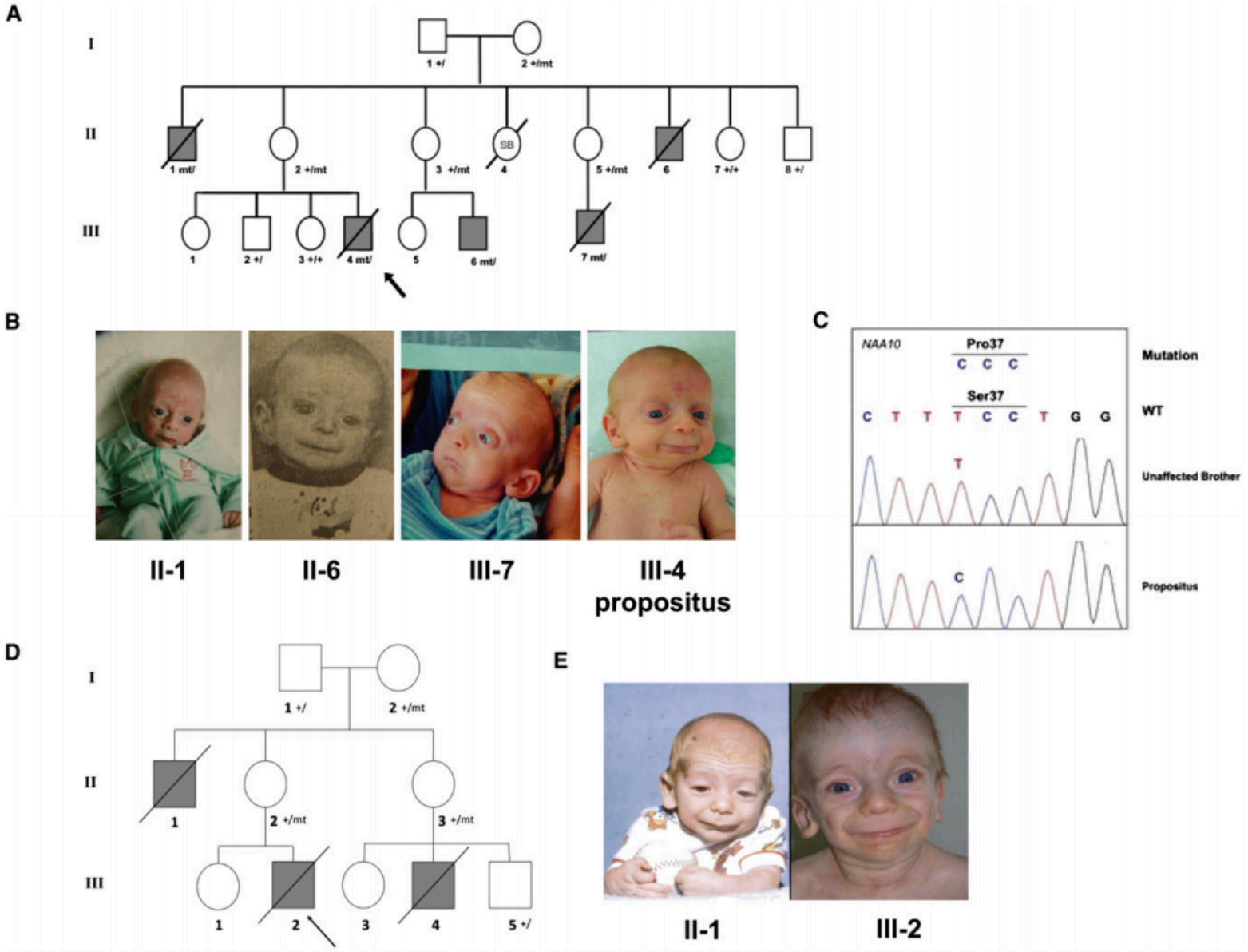


Figure 2. Pedigree Drawing and Pictures of Families 1 and 2

(A) Pedigree drawing for family 1. The most recent deceased individual, III-4, is the most well-studied subject in the family and is indicated by an arrow. Genotypes are marked for those in which DNA was available and tested. The following abbreviations are used: SB, stillborn; +, normal variant; mt, rare mutant variant.

(B) Pictures of four affected and deceased boys in this family, showing the aged appearance.

(C) Sanger sequencing results of *NAA10* in individual III-4 from family 1.

(D) Pedigree for family 2. Individual III-2 is the most well-studied subject in the family and is indicated by an arrow.

(E) Picture of individuals II-I and III-2 in family 2 at ~1 year of age.

Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Work on HW3