Lecture 20. Disease Genomics

Michael Schatz

April 5, 2021

JHU 600.749: Applied Comparative Genomics

Preliminary Project Report

Assignment Date: March 24, 2021
Due Date: Monday, April 7, 2021 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Wednesday April 7.

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 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. Overleaf is recommended for LaTex submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

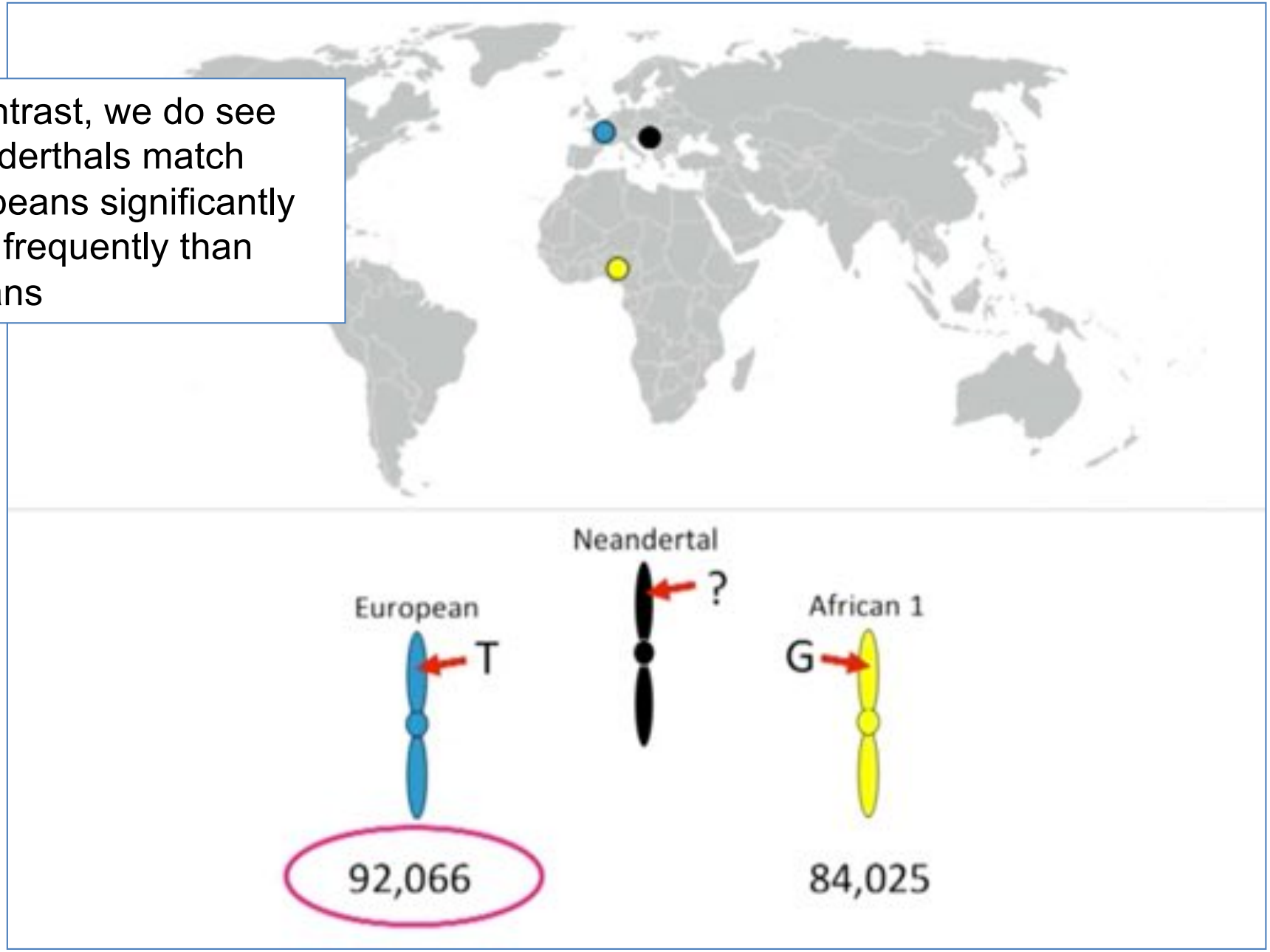
Later, you will present your project in class starting the week of April 21. You will also submit your final written report (5-7 pages) of your project by May 13

Please use Piazza if you have any general questions!

Part 1:
Ancient Hominds

Did we mix?

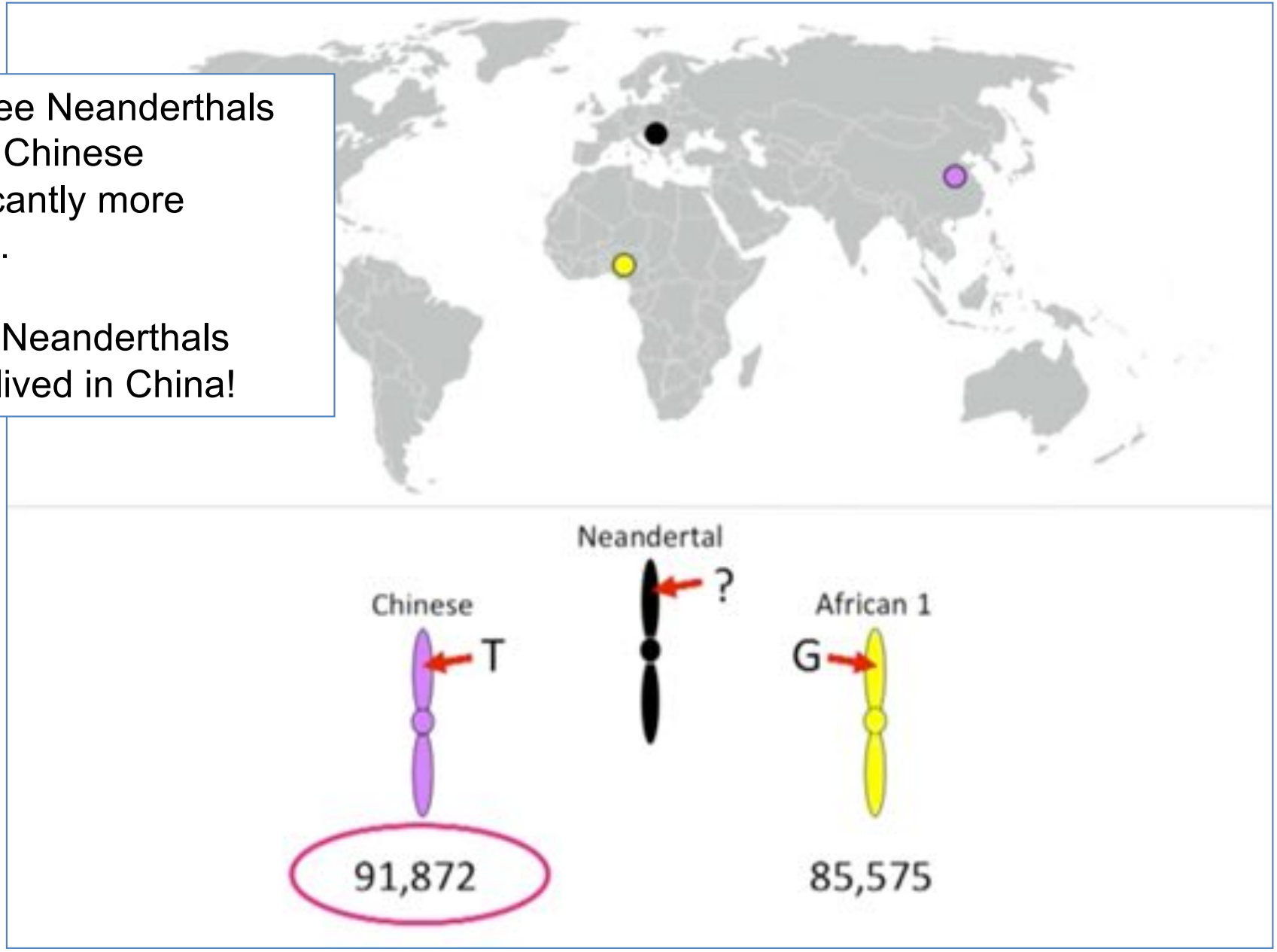
In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans



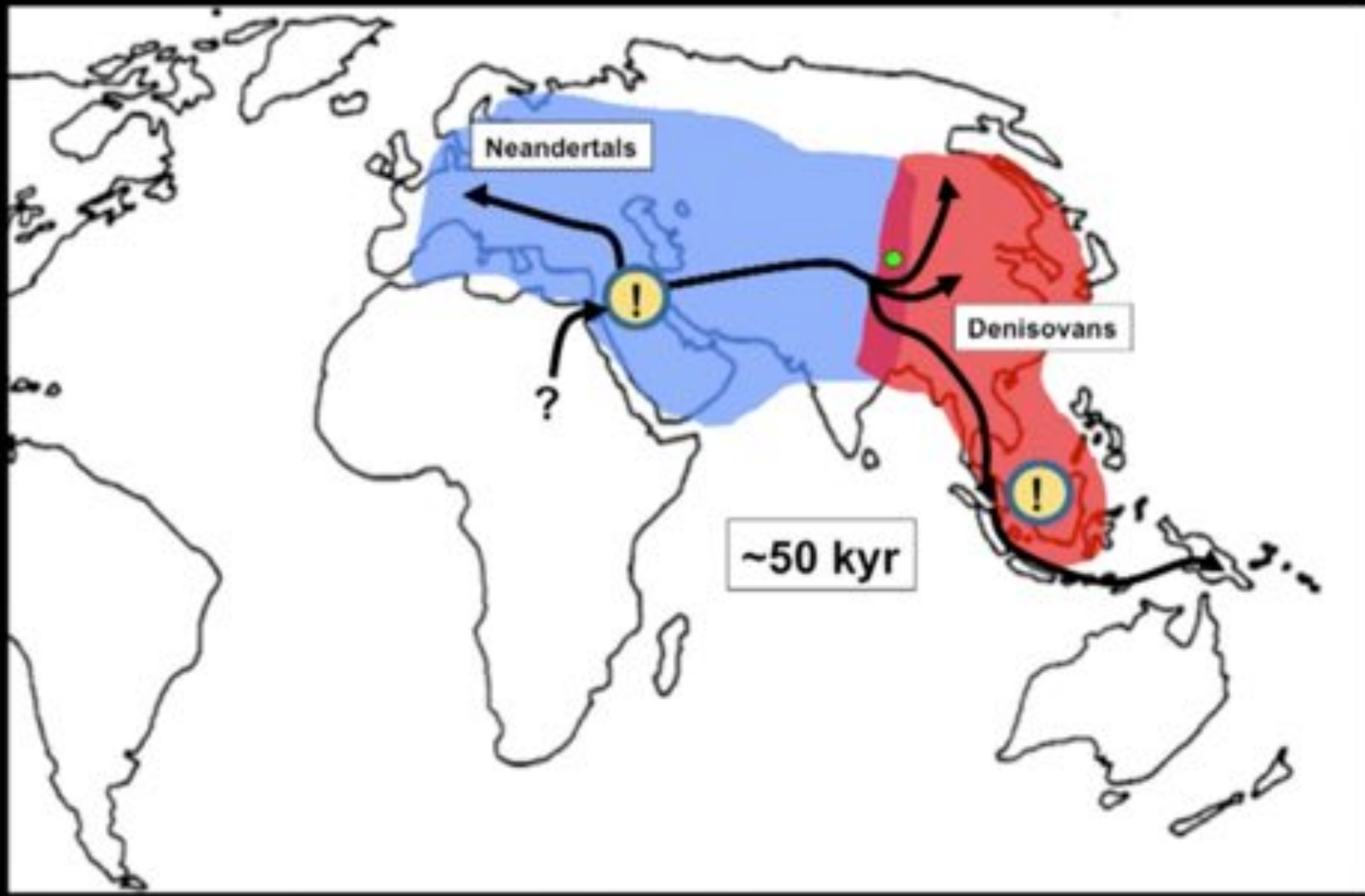
Did we mix?

Also see Neanderthals match Chinese significantly more often…

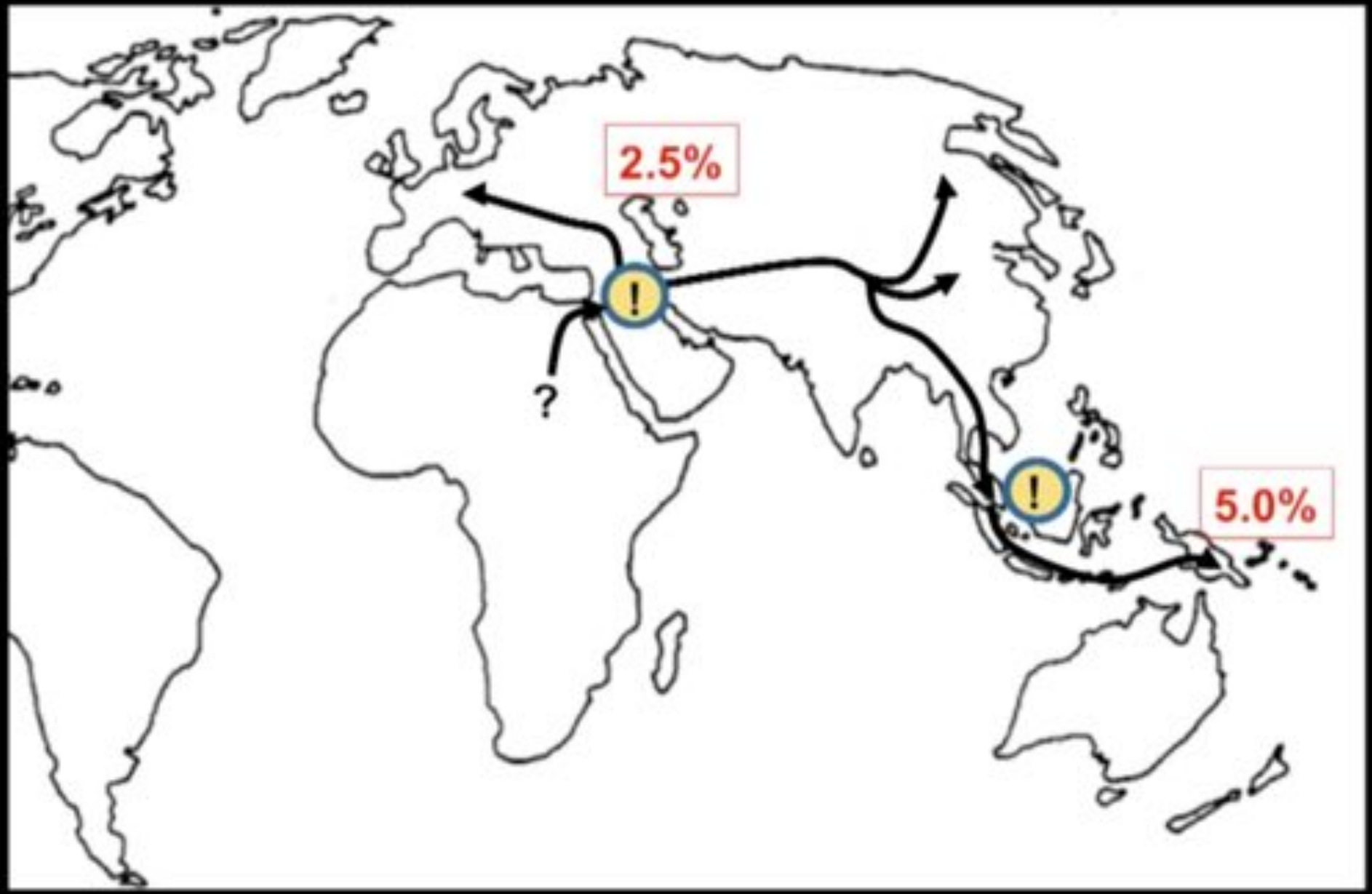
… but Neanderthals never lived in China!



Timeline of ancient hominids



Timeline of ancient hominids



Part 11:
Modern Humans

ARTICLE

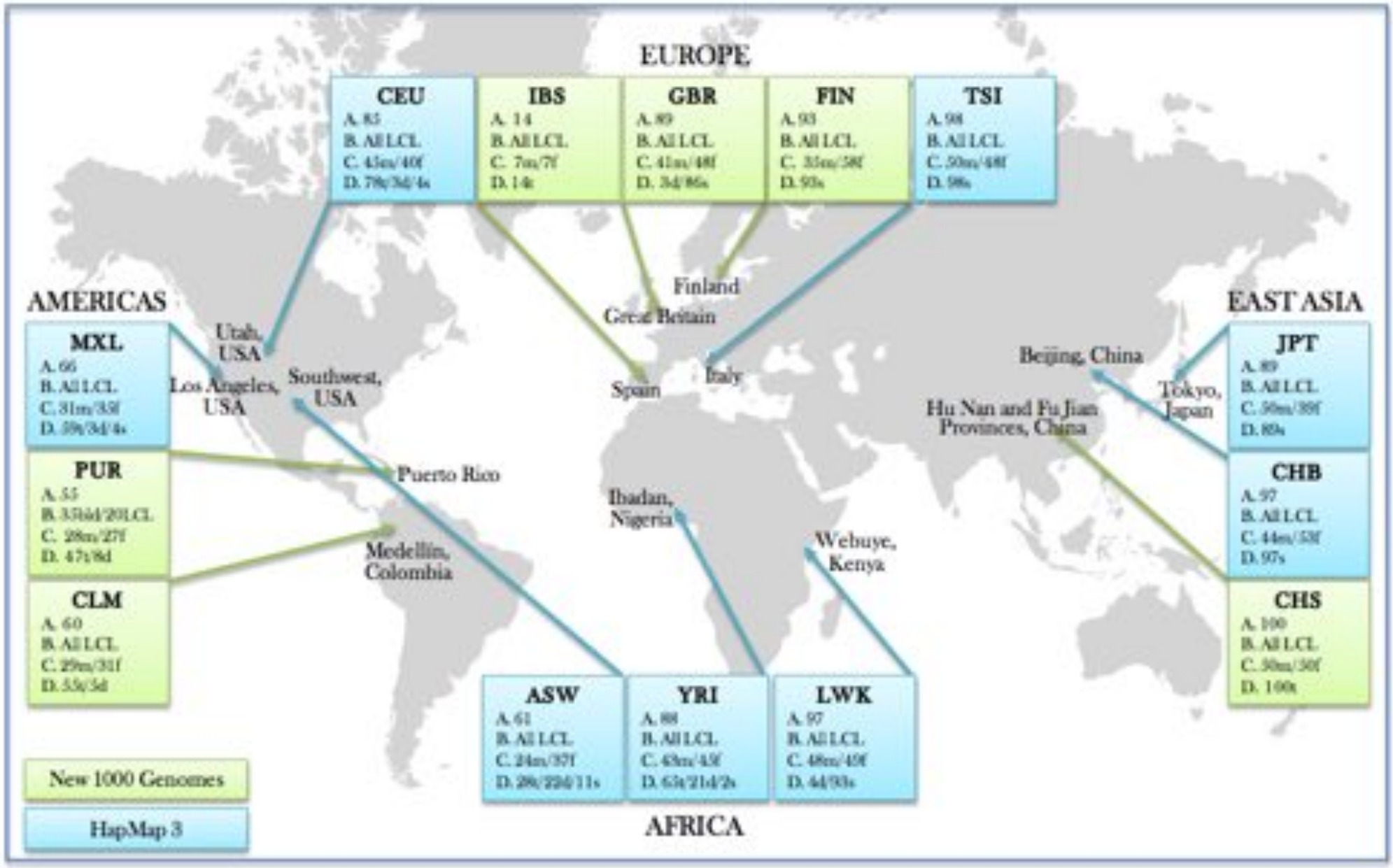
doi:10.1038/nature11632

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

The 1000 Genomes Project Consortium*

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

1000 Genomes Populations



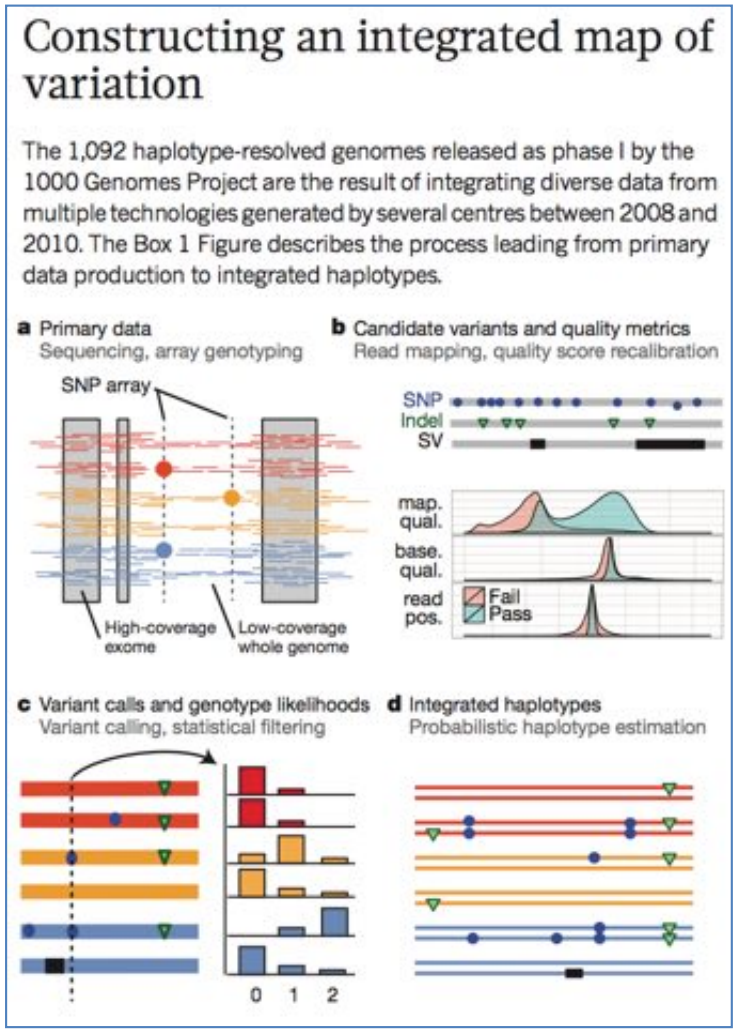
1000 Genomes Populations

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 Bejing, China (**CHB**)	no	no	51	97	103	103	106
Japanese in Tokyo, Japan (**JPT**)	no	no	94	89	104	104	105
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
Total East Asian Ancestry (EAS)			**185**	**286**	**515**	**504**	**523**
Bengali in Bangladesh (**BEB**)	no	yes	0	0	86	86	86
Gujarati Indian in Houston,TX (**GIH**)	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
Total South Asian Ancestry (SAS)			**0**	**0**	**494**	**489**	**494**
African Ancestry in Southwest US (**ASW**)	no	yes	0	61	66	61	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		85	85	85
Yoruba in Ibadan, Nigeria (**YRI**)	no	yes	106	88	109	108	116
Total African Ancestry (AFR)			**208**	**246**	**669**	**661**	**691**
British in England and Scotland (**GBR**)	no	yes	0	89	92	91	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
Total European Ancestry (EUR)			**160**	**379**	**505**	**503**	**514**
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	105	104	105
Total Americas Ancestry (AMR)				**181**	**352**	**347**	**355**
Total			**853**	**1092**	**2030**	**2504**	**2877**

26 populations from 5 major population groups

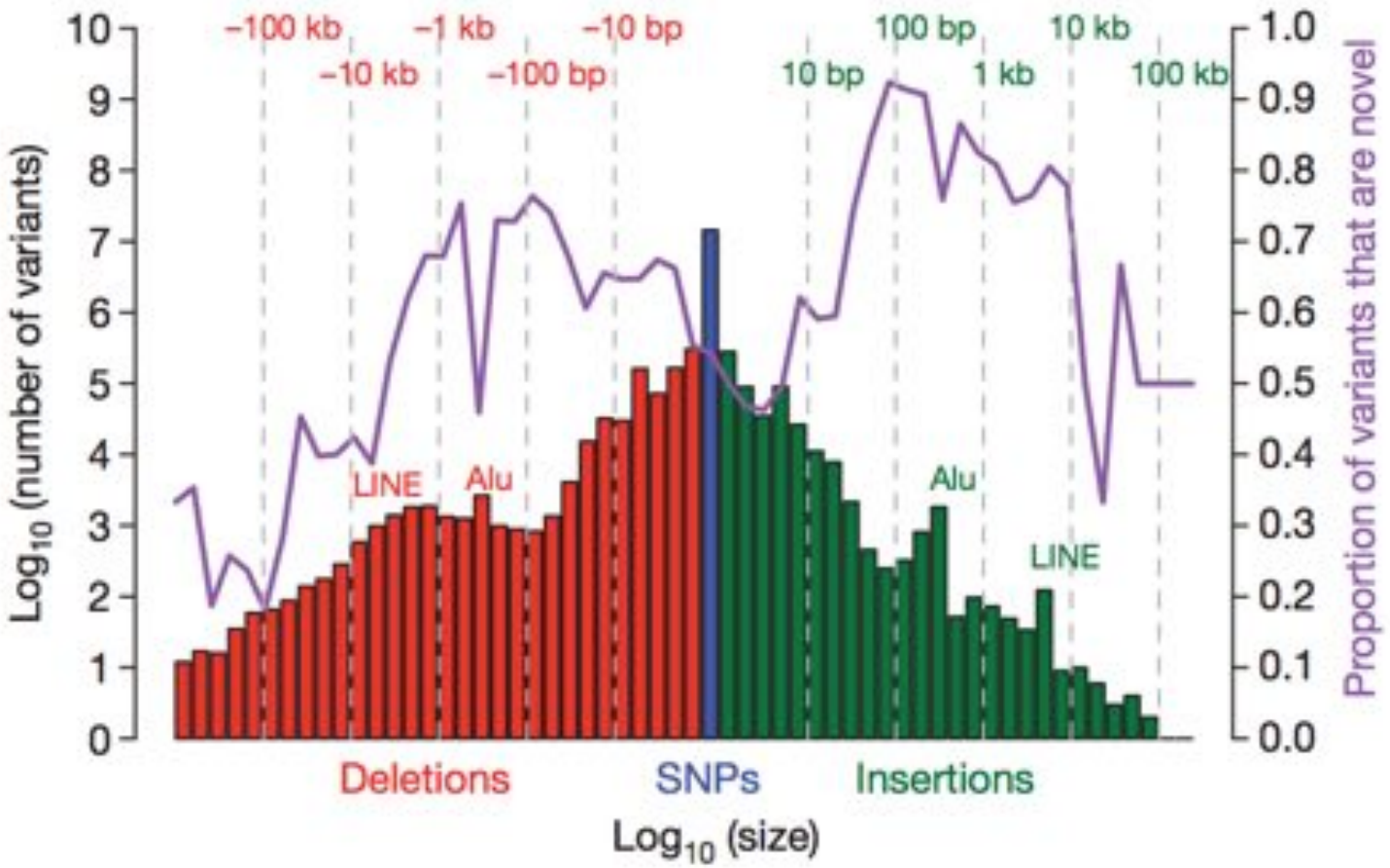
1000 Genomes: Human Mutation Rate

- Phase 1 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

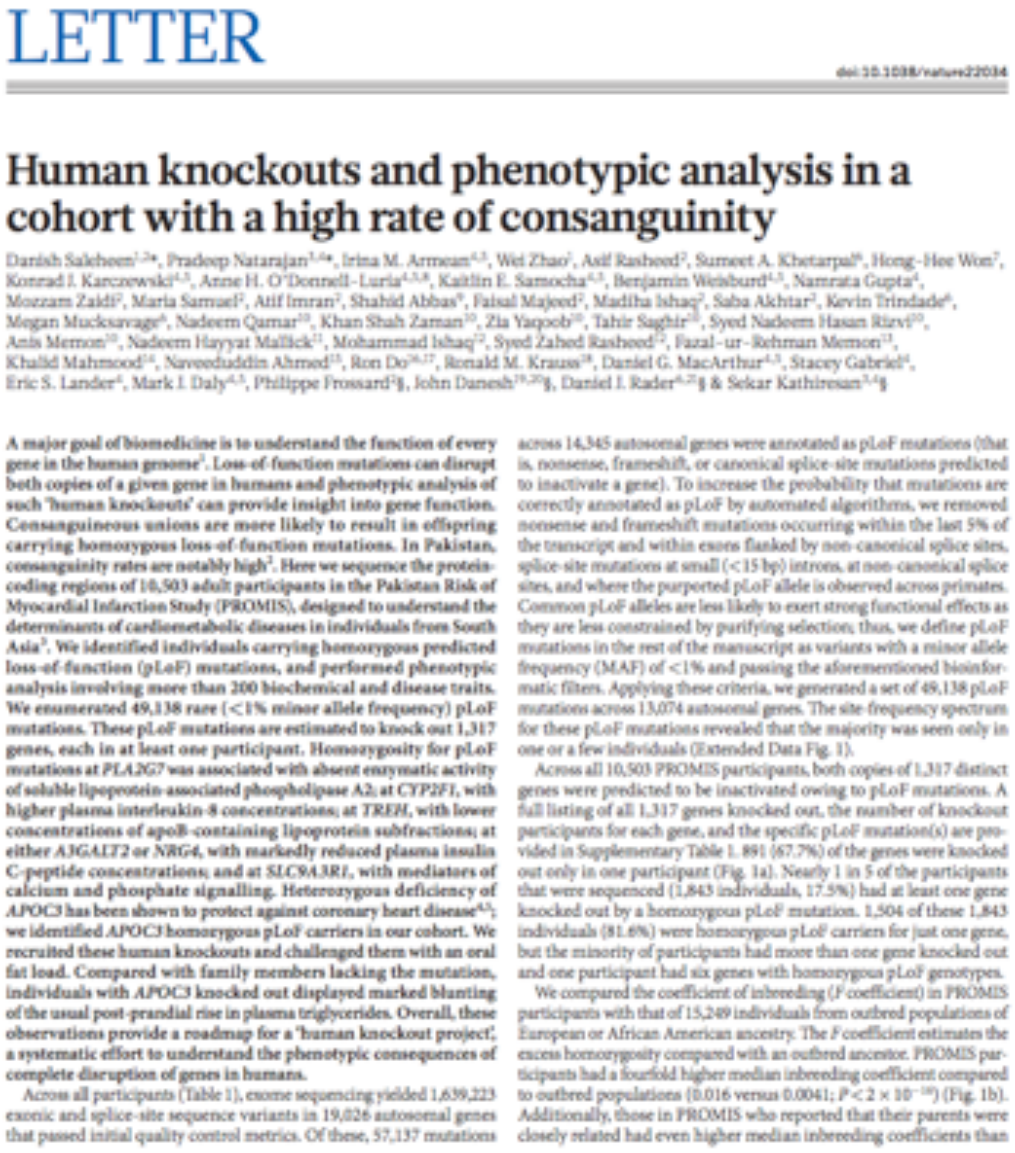
A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes

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

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

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

Homozygous LoF Mutations



LETTER

doi:10.1038/nature22034

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

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

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

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

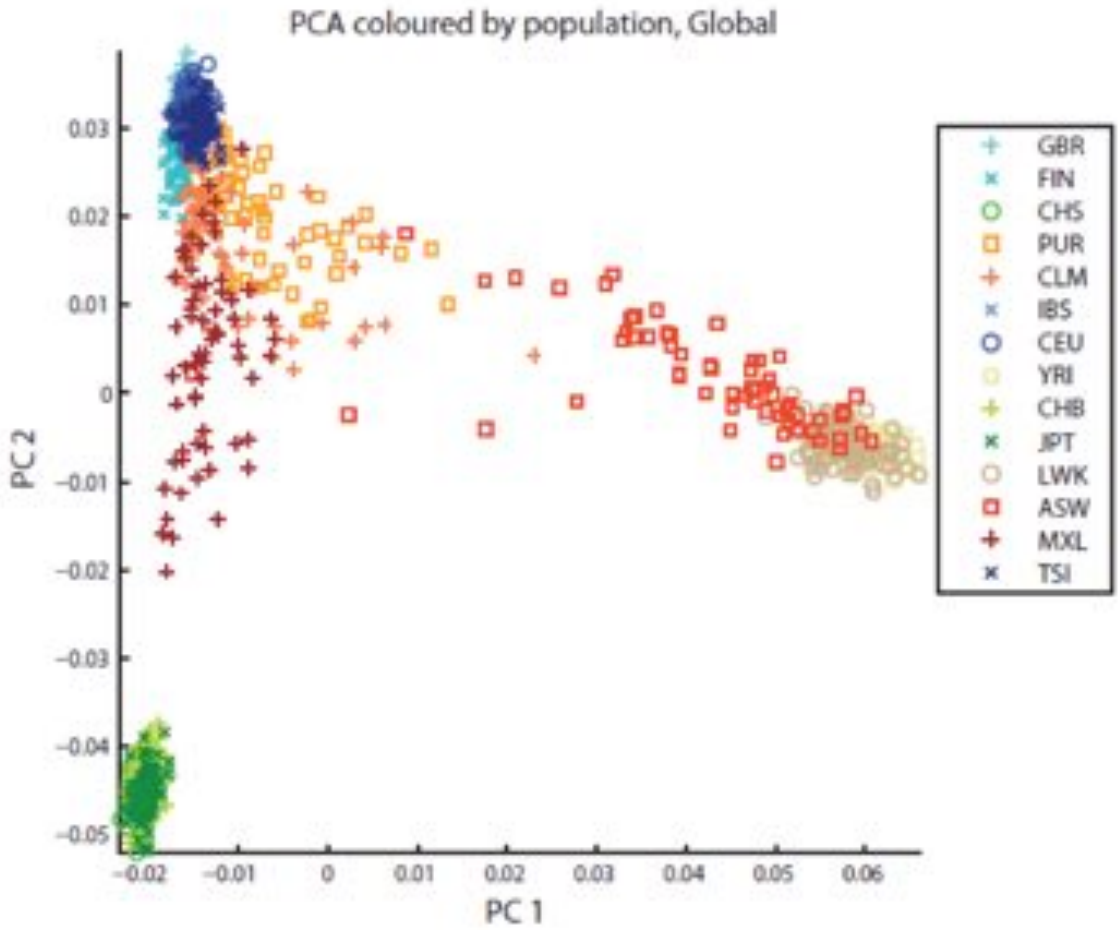
Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated owing to pLoF mutations. A full listing of all 1,317 genes knocked out, the number of knockout participants for each gene, and the specific pLoF mutation(s) are provided in Supplementary Table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

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

(2017) Nature. doi:10.1038/nature22034

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

Variation across populations



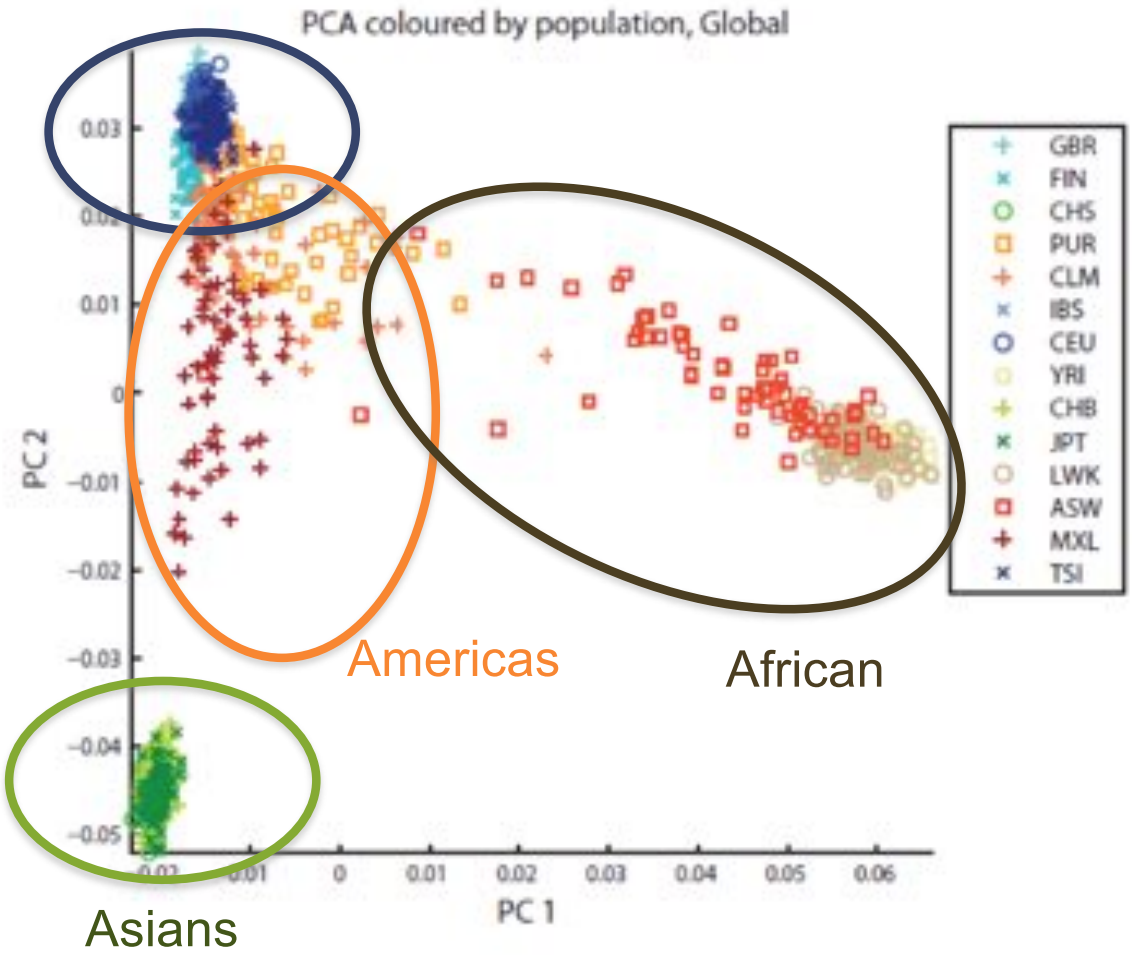
LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

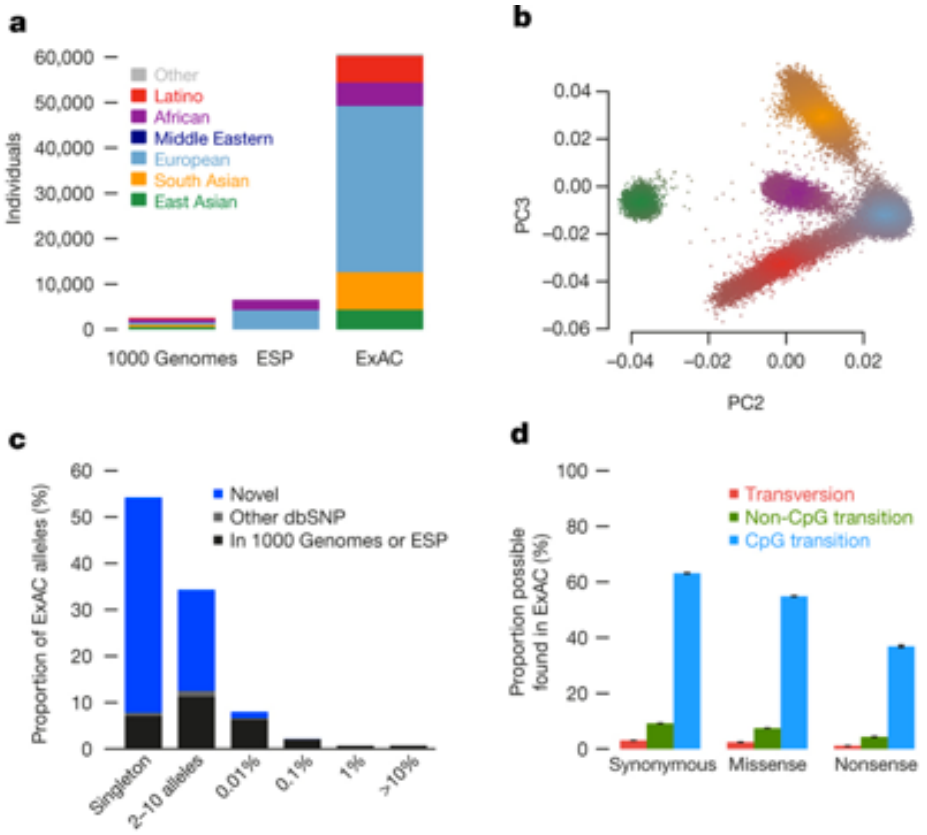


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

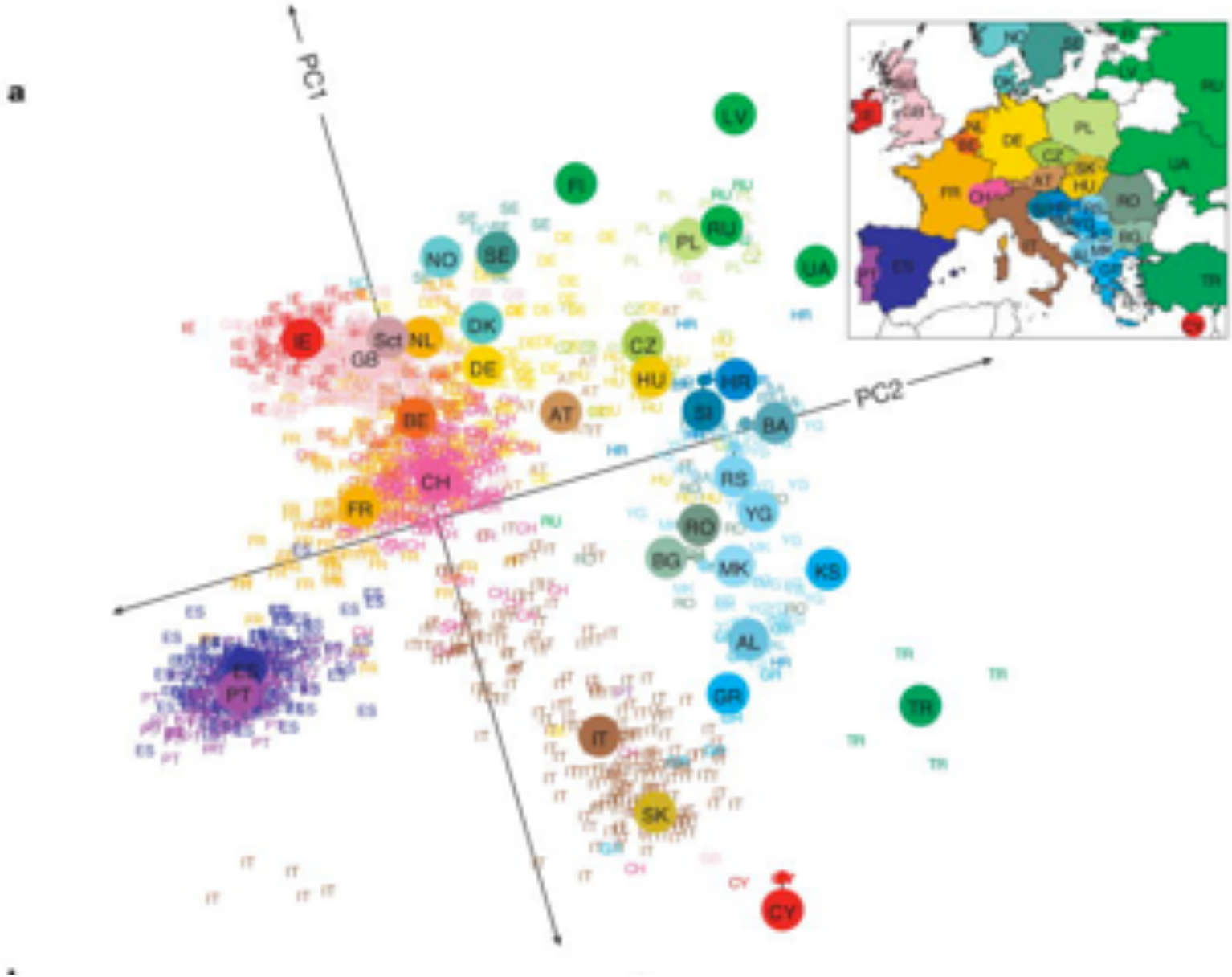
ExAC: Exome Aggregation Consortium



- 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

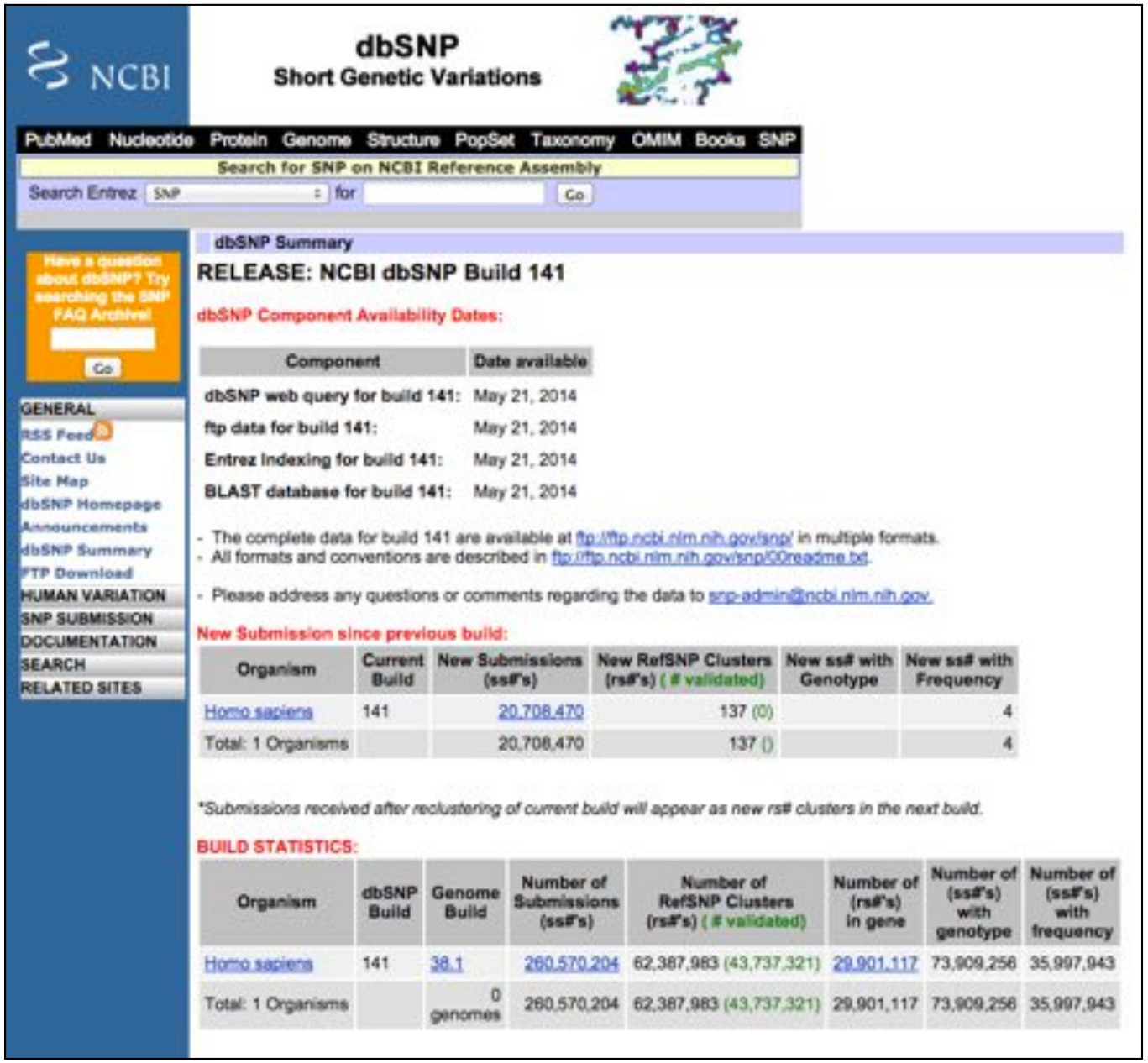
Lek et al (2016) Nature. doi:10.1038/nature19057



Genes mirror geography within Europe

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

dbSNP



NCBI

dbSNP
Short Genetic Variations

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

GENERAL
RSS Feed
Contact Us
Site Map
dbSNP Homepage
Announcements
dbSNP Summary
FTP Download
HUMAN VARIATION
SNP SUBMISSION
DOCUMENTATION
SEARCH
RELATED SITES

dbSNP Summary

RELEASE: NCBI dbSNP Build 141

dbSNP Component Availability Dates:

Component	Date available
dbSNP web query for build 141:	May 21, 2014
ftp data for build 141:	May 21, 2014
Entrez Indexing for build 141:	May 21, 2014
BLAST database for build 141:	May 21, 2014

- The complete data for build 141 are available at ftp://ftp.ncbi.nlm.nih.gov/snp/ in multiple formats.
- All formats and conventions are described in ftp://ftp.ncbi.nlm.nih.gov/snp/00readme.txt.
- Please address any questions or comments regarding the data to snp-admin@ncbi.nlm.nih.gov.

New Submission since previous build:

Organism	Current Build	New Submissions (ss#'s)	New RefSNP Clusters (rs#'s) (# validated)	New ss# with Genotype	New ss# with Frequency
Homo sapiens	141	20,708,470	137 (0)		4
Total: 1 Organisms		20,708,470	137 ()		4

*Submissions received after reclustering of current build will appear as new rs# clusters in the next build.

BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency
Homo sapiens	141	38.1	260,570,204	62,387,983 (43,737,321)	29,901,117	73,909,256	35,997,943
Total: 1 Organisms		0 genomes	260,570,204	62,387,983 (43,737,321)	29,901,117	73,909,256	35,997,943

- 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:

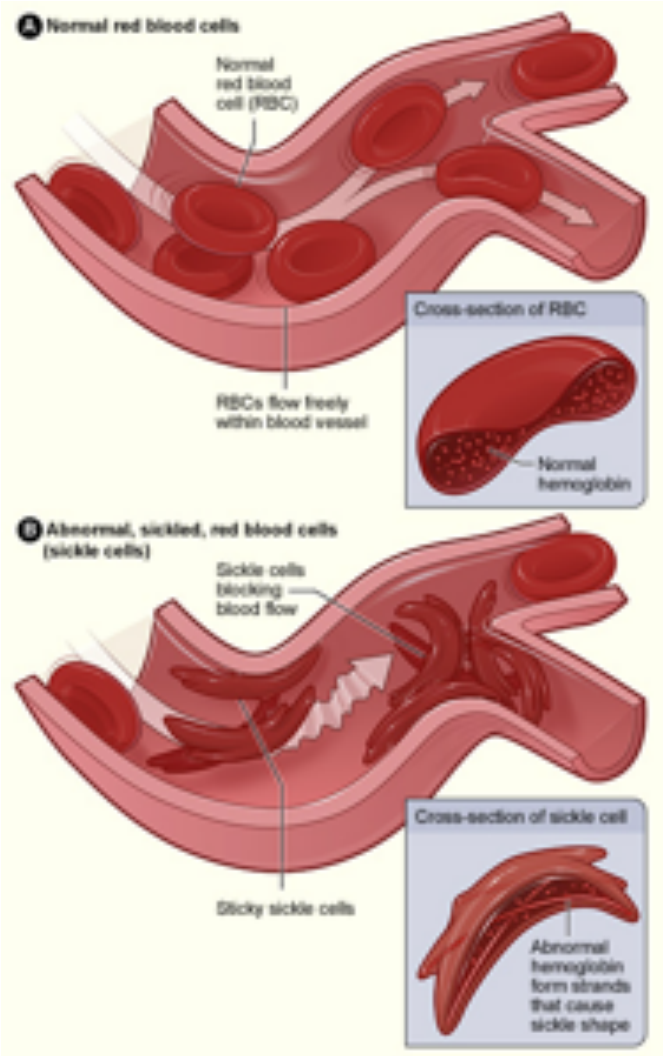
Pre-genome Genetic Medicine

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*, 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

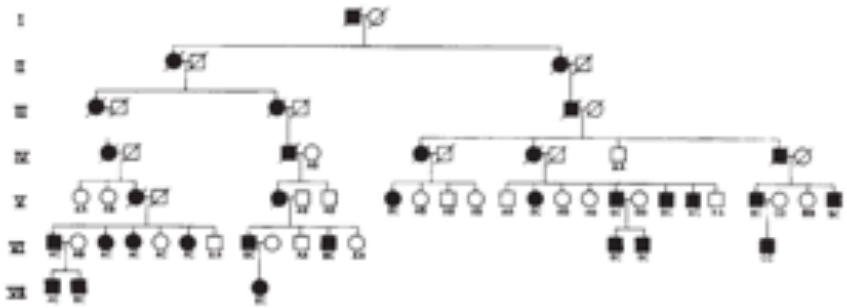


Fig. 2 Pedigree of the Venezuelan Huntington's disease family. This pedigree represents a small part of a much larger pedigree that will be described in detail elsewhere. Permanent EBV-transformed lymphoblastoid cell lines were established from blood samples of these individuals (unpublished data). DNA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

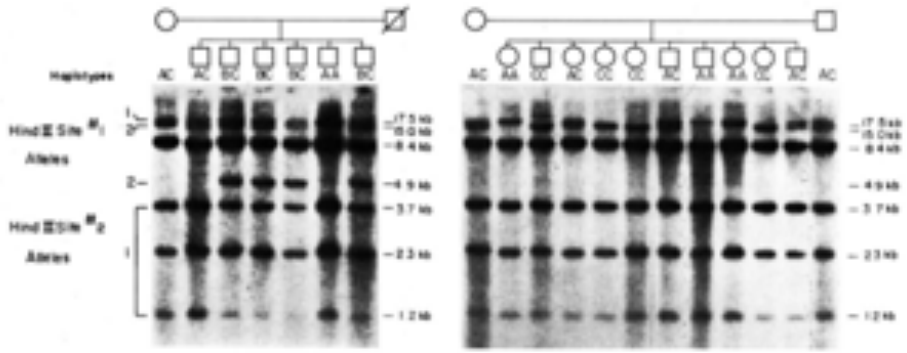


Fig. 3 Hybridization of the G8 Probe to *Hin*dIII-digested human genomic DNA.
Methods: DNA was prepared as described[23] from lymphoblastoid cell lines derived from members of two nuclear families. 5 μg of each DNA was digested to completion with 20 units of *Hin*dIII in a volume of 30 μl using the buffer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

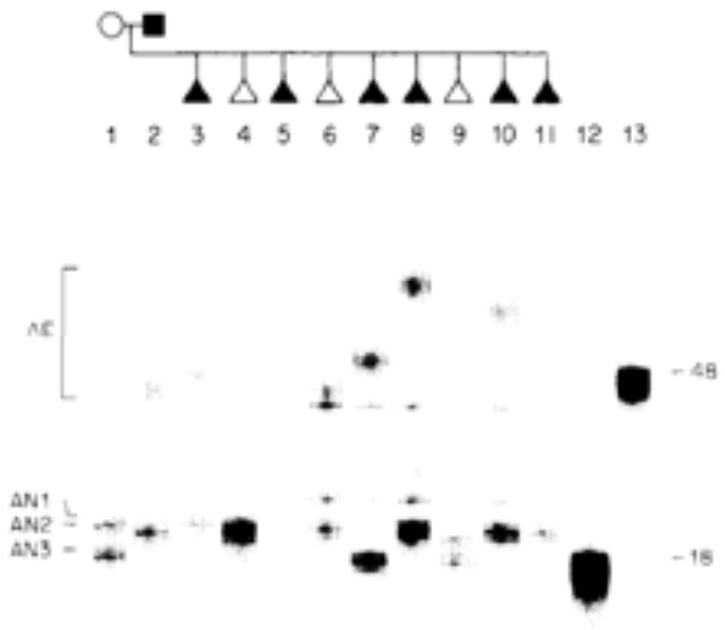


Figure 6. PCR Analysis of the $(CAG)_n$ Repeat in a Venezuelan HD Sibship with Some Offspring Displaying Juvenile Onset

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 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

Old Approaches:
Candidate Genes, Pedigrees, Marker analysis

New Approach

Future

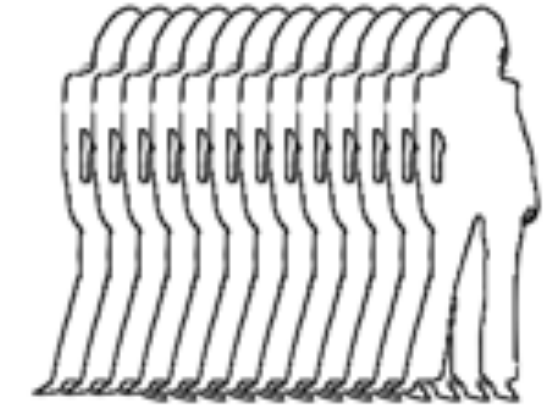
Table 1 | **Approaches to identifying variants underlying complex traits and common diseases**

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	+	+	+	+	+	+/–	–

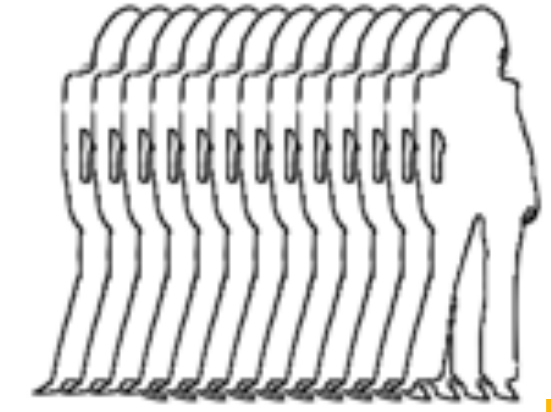
*Candidate-gene studies. ‡Genome-wide studies. §Association and resequencing studies are immune to stratification if they use family-based designs. Symbols indicate whether the potential advantage in the left column applies completely (+), partially (+/–), weakly (–/+) or not at all (–). MAF, minor allele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

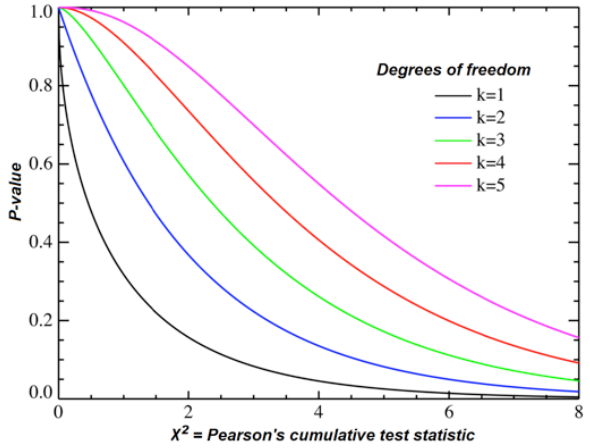
χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.

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.



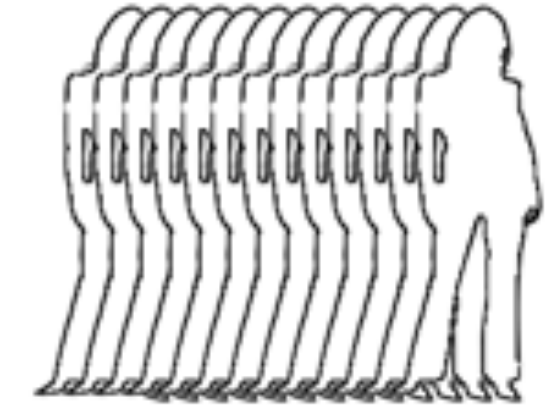
$$P(\chi^2_P(\{p_i\}) > T) \sim C \int_{\sum_{i=1}^{m-1} y_i^2 > T} \left\{ \prod_{i=1}^{m-1} dy_i \right\} \prod_{i=1}^{m-1} \exp\left[-\frac{1}{2} \left(\sum_{i=1}^{m-1} y_i^2 \right) \right]$$

	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)

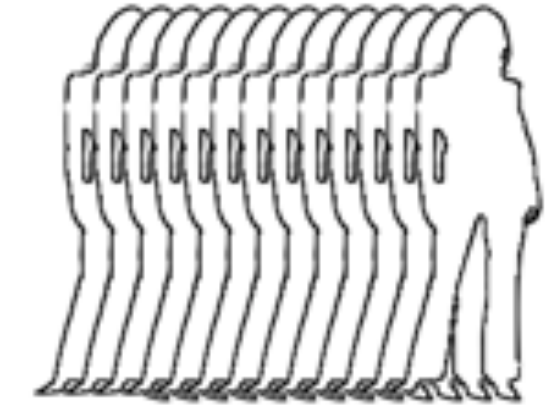
Cases/hasG expected: 4000 * (4780/10000) = 1912 expected
Cases/hasG squared deviation: $(2104 - 1912)^2 / 1912 = 19.28$ deviation

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

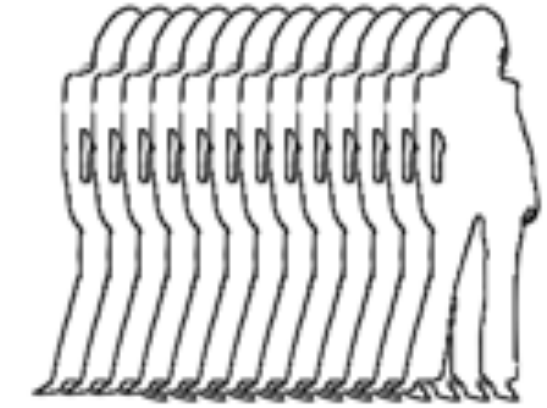


GC CC GC GC GG CC CC
CC GC GC GG GC GG

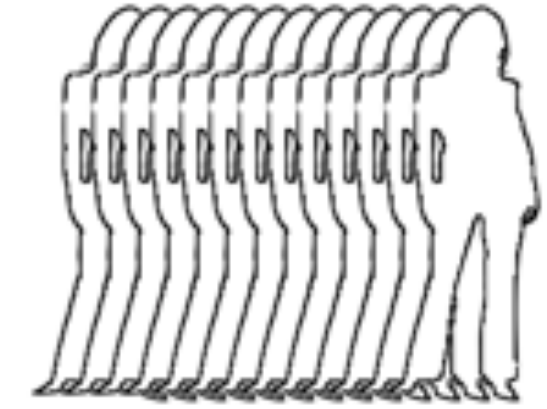
SNP1	*SNP2*	*SNP…*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:**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

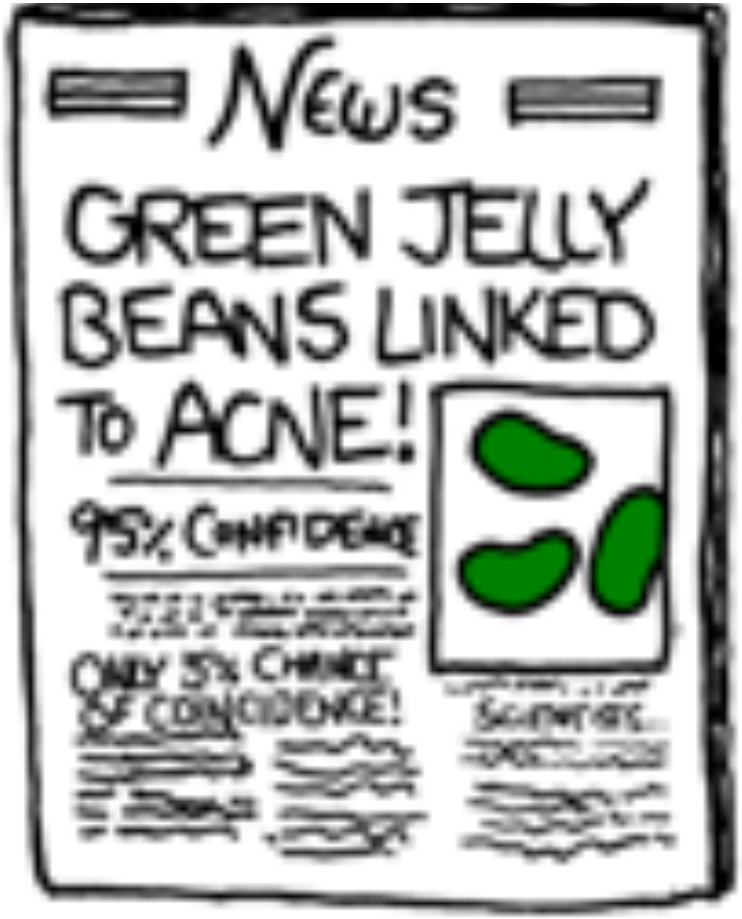
With a (much) larger population, this might be a significant difference in rate:
25320/60000 =>
p = 5e-7

Chi-squared or similar test

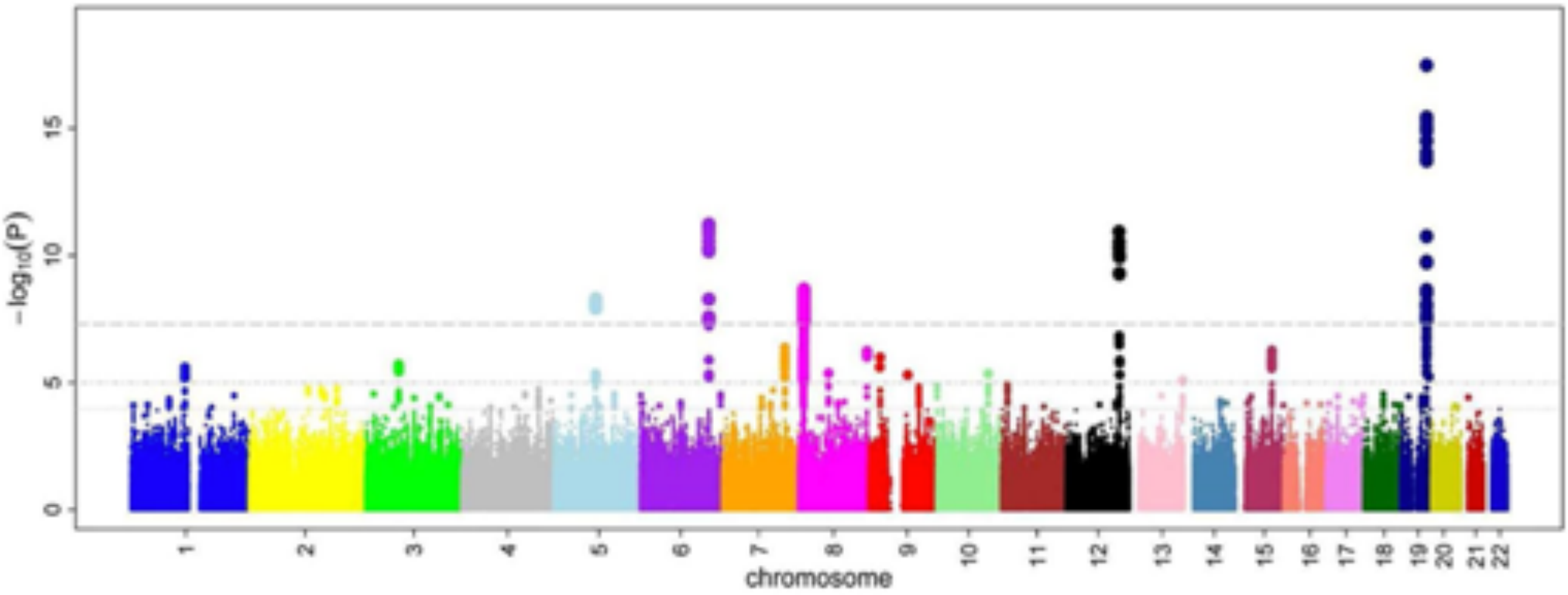
The curse of multiple testing

The curse of multiple testing

The curse of multiple testing



Manhattan Plot



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