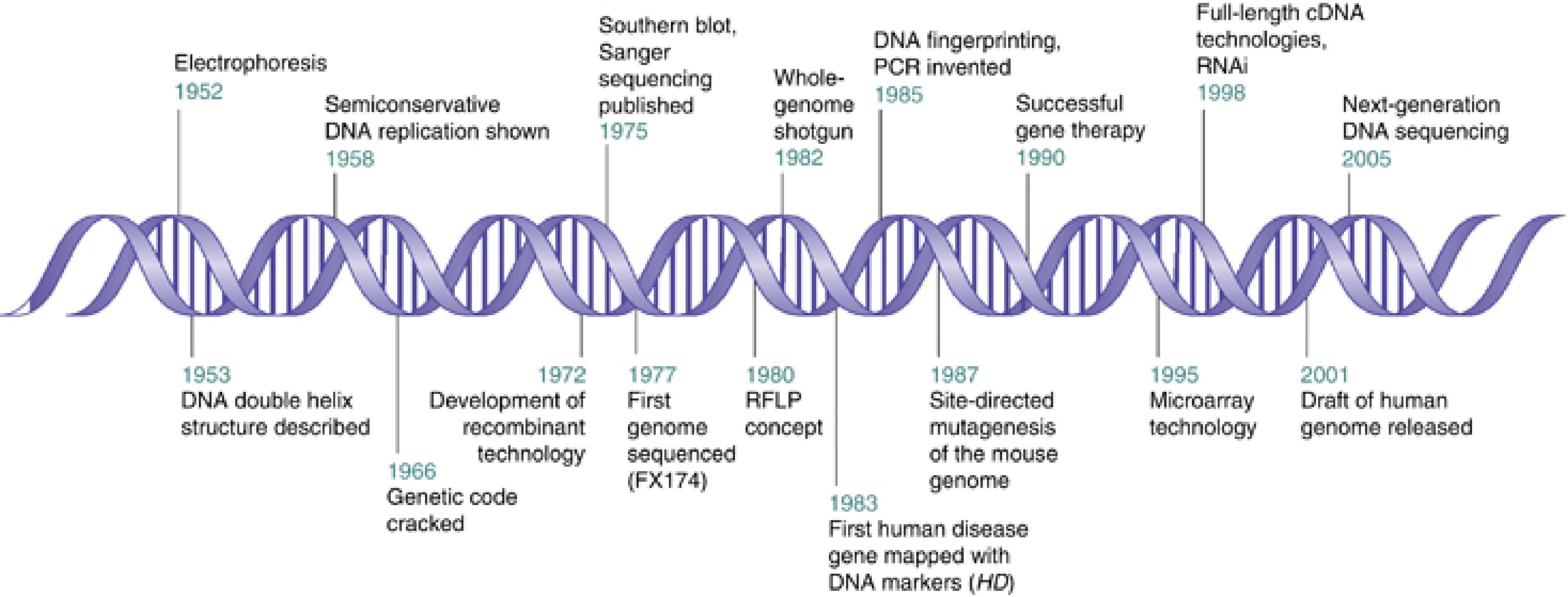
# Genetic and Genomics Laboratory Tools and Approaches

**Meredith Yeager, PhD**
Cancer Genomics Research Laboratory
Division of Cancer Epidemiology and Genetics
yeagerm@mail.nih.gov

**DCEG Radiation Epidemiology and Dosimetry Course 2019**

# (Recent) history of genetics



Electrophoresis
1952

1953
DNA double helix structure described

Semiconservative DNA replication shown
1958

1966
Genetic code cracked

1972
Development of recombinant technology

Southern blot, Sanger sequencing published
1975

1977
First genome sequenced (FX174)

1980
RFLP concept

Whole-genome shotgun
1982

1983
First human disease gene mapped with DNA markers (*HD*)

DNA fingerprinting, PCR invented
1985

1987
Site-directed mutagenesis of the mouse genome

Successful gene therapy
1990

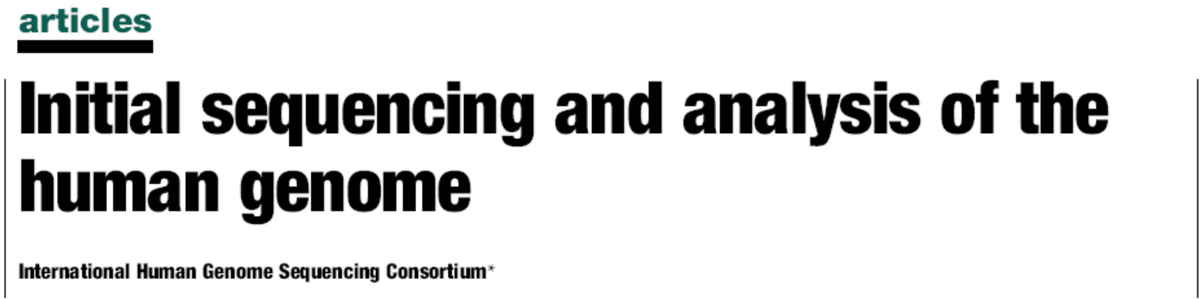
1995
Microarray technology

Full-length cDNA technologies, RNAi
1998

2001
Draft of human genome released

Next-generation DNA sequencing
2005

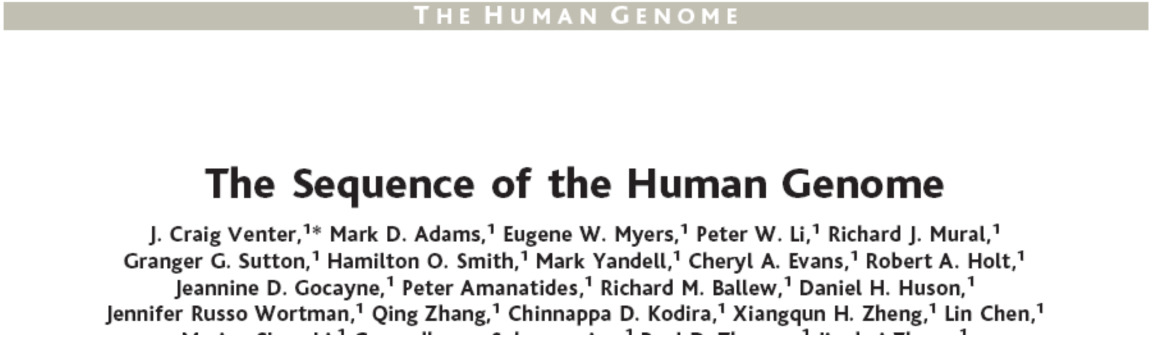
# Sequencing of the Human Genome



articles

**Initial sequencing and analysis of the human genome**

International Human Genome Sequencing Consortium*



THE HUMAN GENOME

**The Sequence of the Human Genome**

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]

*Science* **291**, 1304-1351 (2001)

# The Human Genome – 2019

- ~3.3 billion bases (A, C, G, T)
- ~20,000 protein-coding genes, many non-coding RNAs (~2% of the genome)
- Annotation ongoing – the initial sequencing in 2001 *is still being refined, assembled and annotated, even now – hg38*
- Variation (polymorphism) present within humans
  - Population-specific
  - Cosmopolitan

# Types of polymorphisms

- Single nucleotide polymorphisms (SNPs)
- Common SNPs are defined as > 5% in at least one population
- Abundant in genome (~50 million and counting)

`ATGGAACGA(G/C)AGGATA(T/A)TACGCACTATGAAG(C/A)CGGTGAGAGG`

- Repeats of DNA (long, short, complex, simple), insertions/deletions
- **A small fraction of SNPs and other types of variation are very or slightly deleterious and may contribute by themselves or with other genetic or environmental factors to a phenotype or disease**

# Different mutation rates at the nucleotide level

| Mutation type | Mutation rate (per generation) |
|---|---|
| Transition on a CpG | $1.6X10^{-7}$ |
| Transversion on a CpG | $4.4X10^{-8}$ |
| Transition out of CpG | $1.2X10^{-8}$ |
| Transversion out of CpG | $5.5X10^{-9}$ |
| Substitution (average) | $2.3X10^{-8}$ |
| Insertion/deletion (average) | $2.3X10^{-9}$ |
| Mutation rate (average) | $2.4X10^{-8}$ |

*Transition: purine to purine*
*Transversion: purine to pyrimidine*

*A and G are purines*
*C and T are pyrimidines*

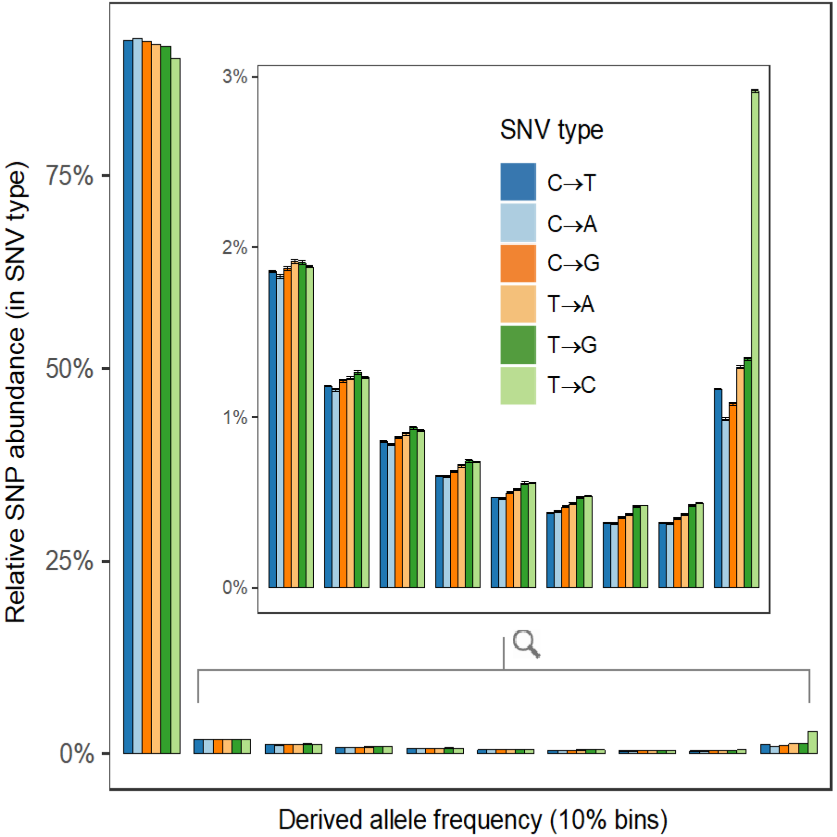
- Size of haploid genome : $3.3X10^9$ nucleotides
- 80 new mutations per haploid genome per generation.
- Assume 2% of genome under natural selection
- About 1.6 new deleterious mutations in each gamete
- Most of these deleterious mutations are recessive

Nachman & Crowell, *Genetics* 156:297-304 (2000)

# Confusing terminology, SNPs versus mutations

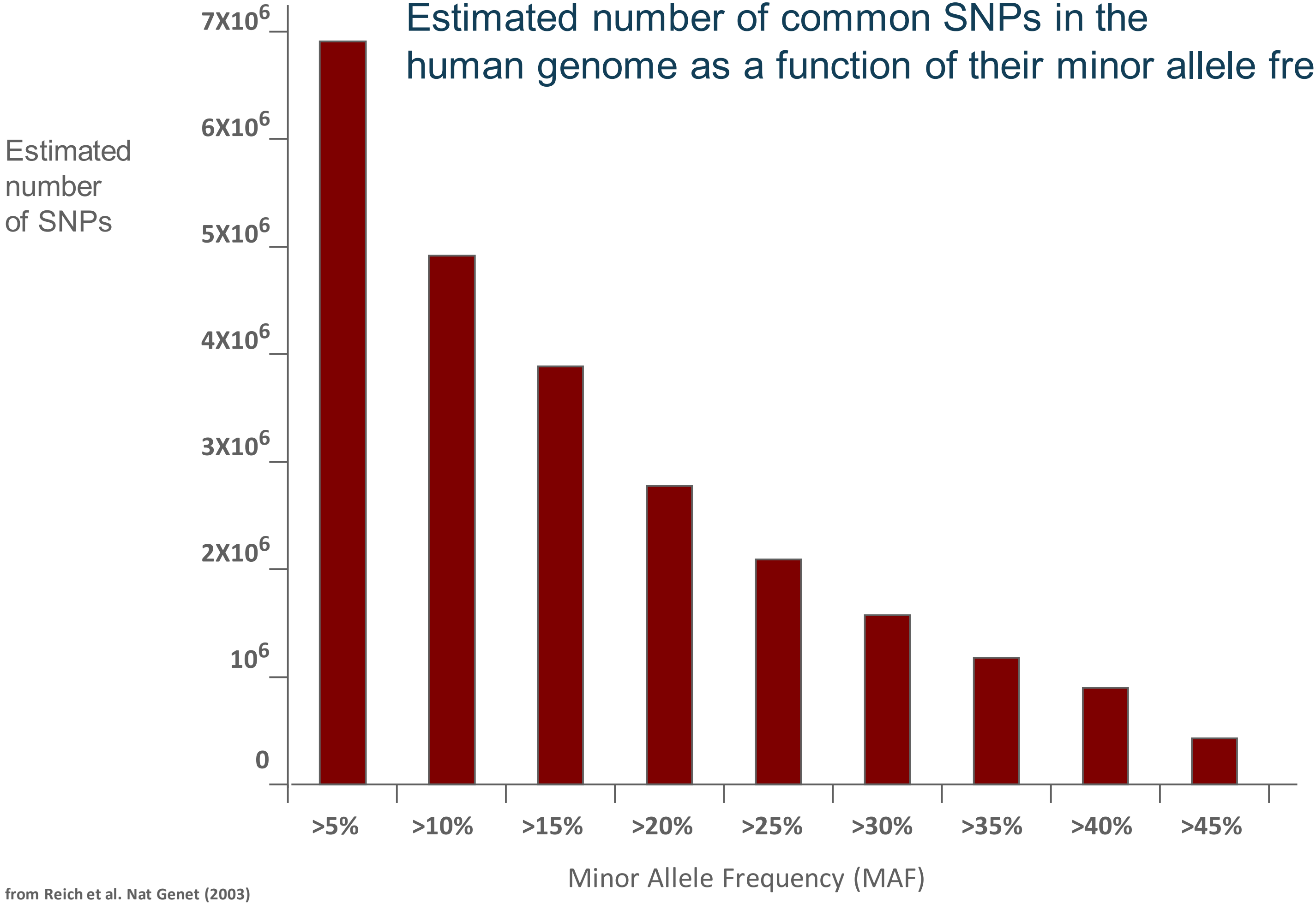
- Germline versus somatic
  - Inherited versus acquired
- Common versus rare germline
  - Common variants are generally called ‘polymorphisms’
  - However, all polymorphisms started as germline mutations (*de novo* mutations)
  - Singletons or exceedingly rare SNPs??

# Most SNPs are rare



(*n* = 67,135,025 SNPs from 1000 Genomes)

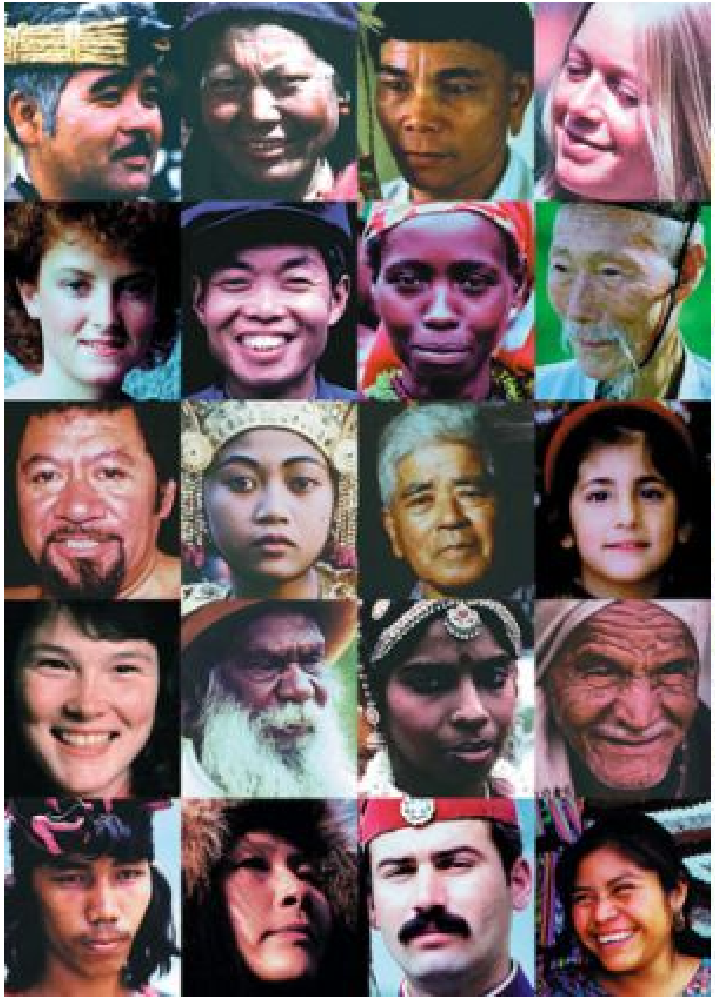
# Estimated number of common SNPs in the human genome as a function of their minor allele frequency



Adapted from Reich et al. Nat Genet (2003)

# SNPs & function

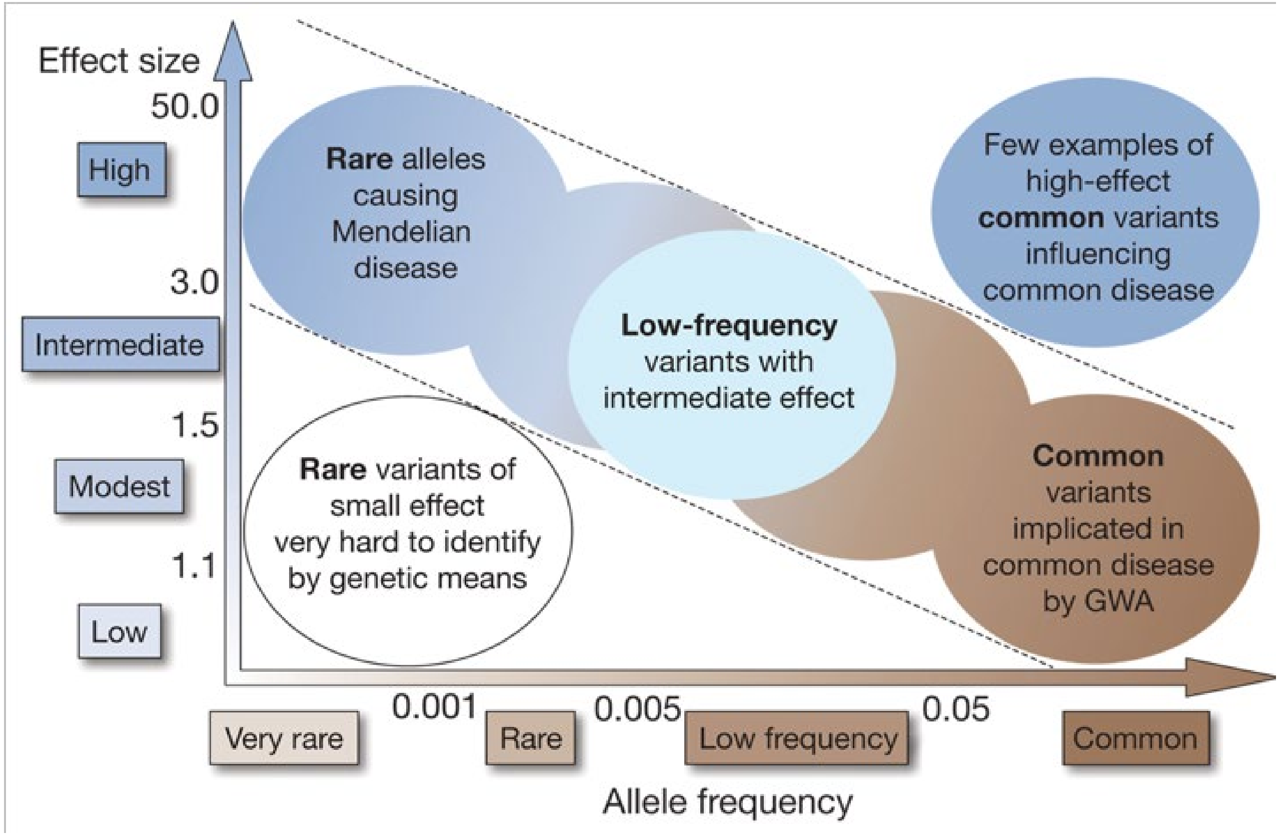
- Vast majority are “silent”
  - No known functional change
- Alter function of gene product
  - Change sequence of protein
- Alter gene expression/regulation
  - Promoter/enhancer
  - mRNA stability
  - Small RNA binding sites
  - Disrupt CpG site



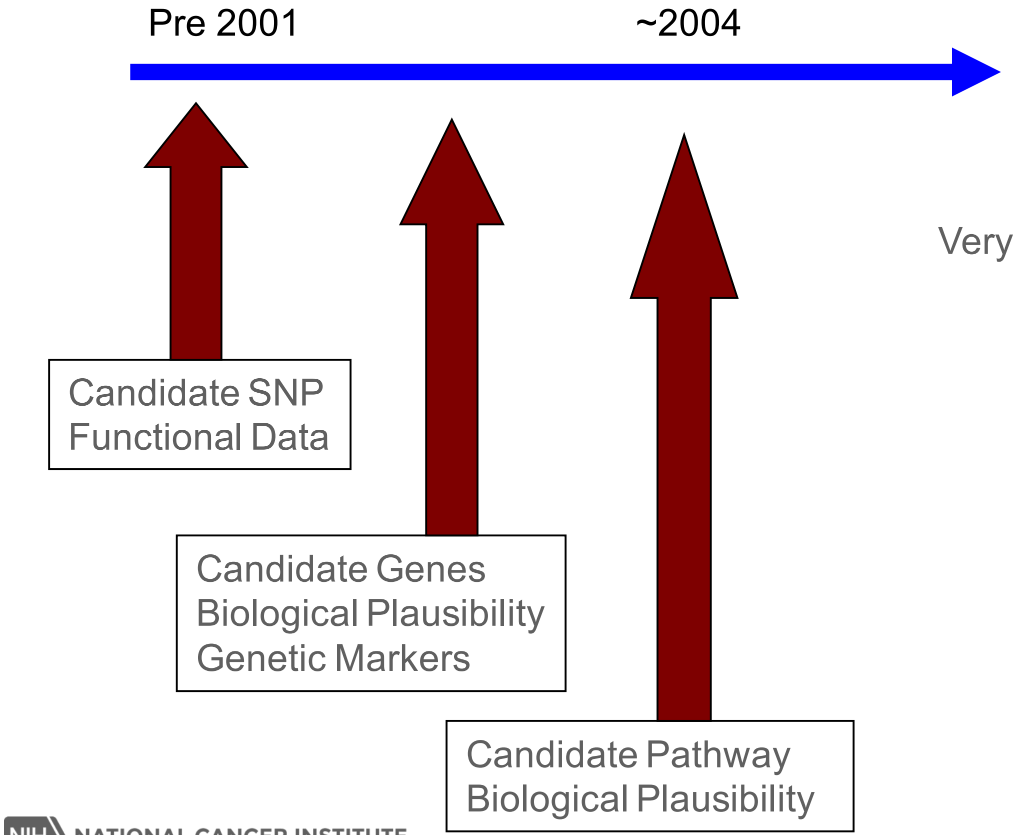
# Which variants contribute to disease?

- Loss of function – premature stops, frame-shifts
- Splice variants
- Amino acid changes
- SNPs in regulatory regions
- Structural variants
- ??
- *There are tools that help predict which variants might be important*

# Mapping Genetic Susceptibility


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

# Trajectory of the Genetics of Cancer Susceptibility



Pre 2001

~2004

Candidate SNP
Functional Data

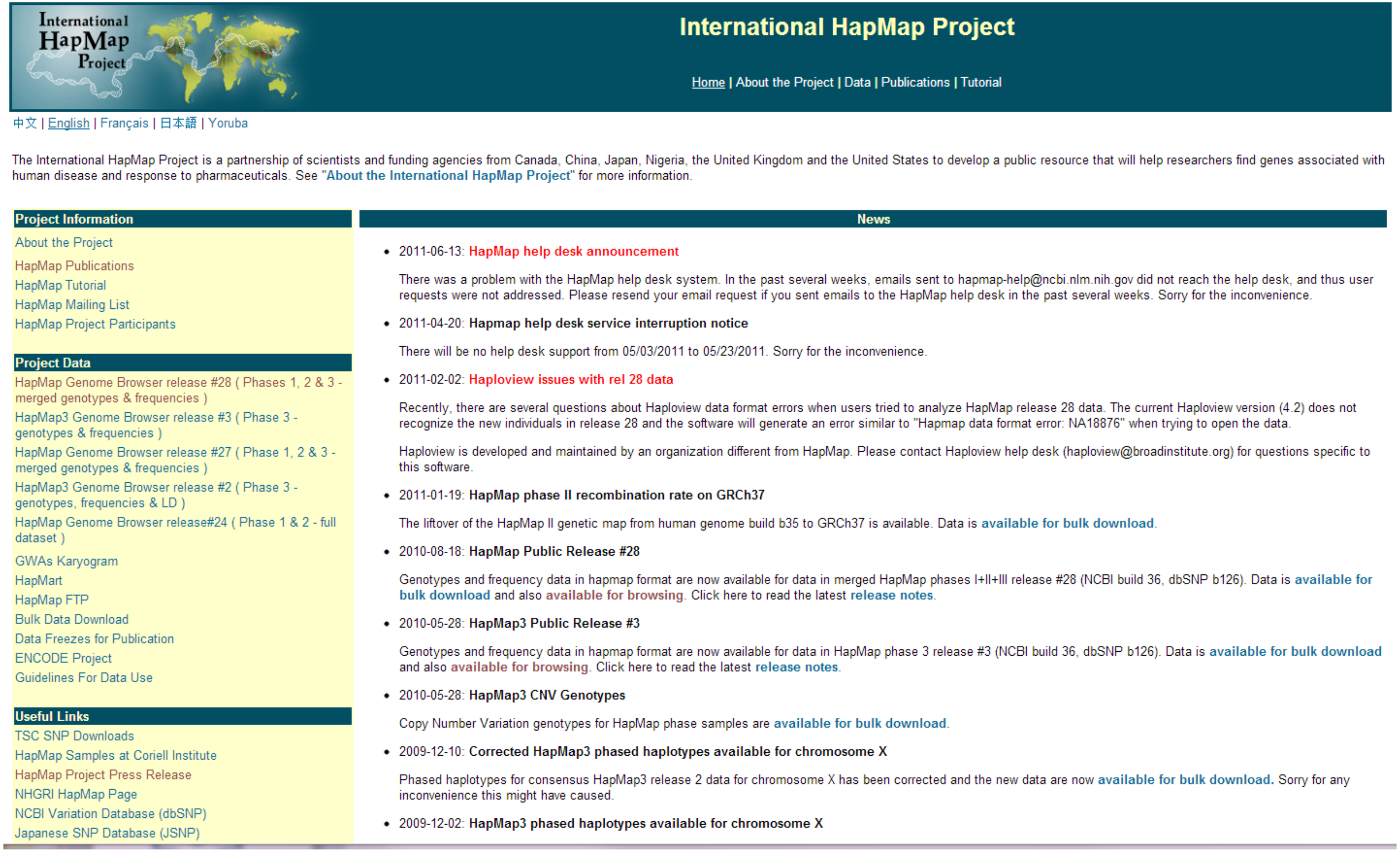
Candidate Genes
Biological Plausibility
Genetic Markers

Candidate Pathway
Biological Plausibility

Very few candidate genes panned out

- Underpowered studies?
- Bad guesses?

*Technological advances + community efforts*



# International HapMap Project

Home | About the Project | Data | Publications | Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap Project" for more information.

**Project Information**

About the Project
HapMap Publications
HapMap Tutorial
HapMap Mailing List
HapMap Project Participants

**Project Data**

HapMap Genome Browser release #28 ( Phases 1, 2 & 3 - merged genotypes & frequencies )
HapMap3 Genome Browser release #3 ( Phase 3 - genotypes & frequencies )
HapMap Genome Browser release #27 ( Phase 1, 2 & 3 - merged genotypes & frequencies )
HapMap3 Genome Browser release #2 ( Phase 3 - genotypes, frequencies & LD )
HapMap Genome Browser release#24 ( Phase 1 & 2 - full dataset )
GWAs Karyogram
HapMart
HapMap FTP
Bulk Data Download
Data Freezes for Publication
ENCODE Project
Guidelines For Data Use

**Useful Links**

TSC SNP Downloads
HapMap Samples at Coriell Institute
HapMap Project Press Release
NHGRI HapMap Page
NCBI Variation Database (dbSNP)
Japanese SNP Database (JSNP)

**News**

- 2011-06-13: **HapMap help desk announcement**

  There was a problem with the HapMap help desk system. In the past several weeks, emails sent to hapmap-help@ncbi.nlm.nih.gov did not reach the help desk, and thus user requests were not addressed. Please resend your email request if you sent emails to the HapMap help desk in the past several weeks. Sorry for the inconvenience.

- 2011-04-20: **Hapmap help desk service interruption notice**

  There will be no help desk support from 05/03/2011 to 05/23/2011. Sorry for the inconvenience.

- 2011-02-02: **Haploview issues with rel 28 data**

  Recently, there are several questions about Haploview data format errors when users tried to analyze HapMap release 28 data. The current Haploview version (4.2) does not recognize the new individuals in release 28 and the software will generate an error similar to "Hapmap data format error: NA18876" when trying to open the data.

  Haploview is developed and maintained by an organization different from HapMap. Please contact Haploview help desk (haploview@broadinstitute.org) for questions specific to this software.

- 2011-01-19: **HapMap phase II recombination rate on GRCh37**

  The liftover of the HapMap II genetic map from human genome build b35 to GRCh37 is available. Data is available for bulk download.

- 2010-08-18: **HapMap Public Release #28**

  Genotypes and frequency data in hapmap format are now available for data in merged HapMap phases I+II+III release #28 (NCBI build 36, dbSNP b126). Data is available for bulk download and also available for browsing. Click here to read the latest release notes.

- 2010-05-28: **HapMap3 Public Release #3**

  Genotypes and frequency data in hapmap format are now available for data in HapMap phase 3 release #3 (NCBI build 36, dbSNP b126). Data is available for bulk download and also available for browsing. Click here to read the latest release notes.

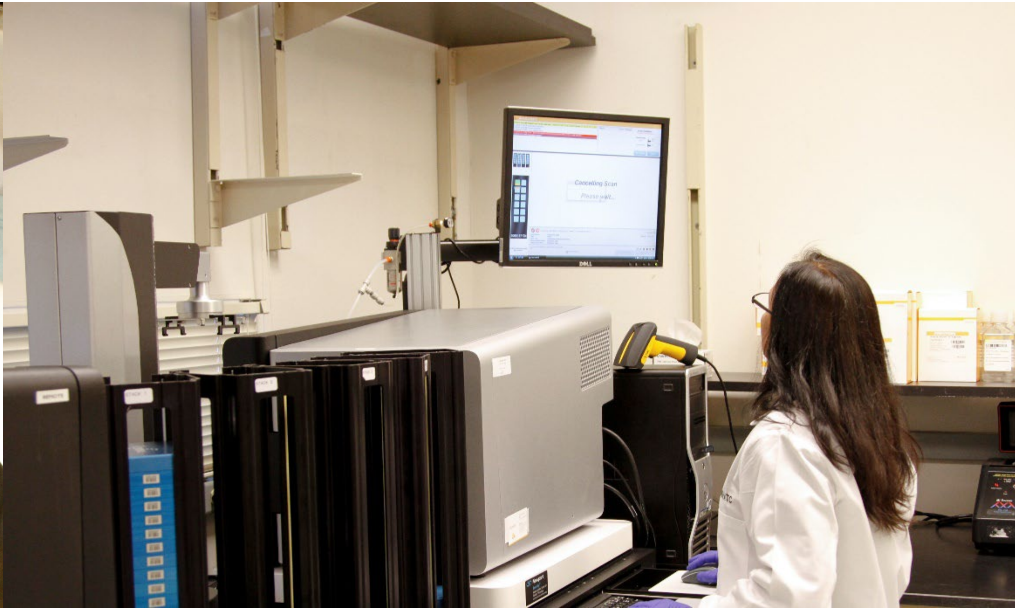
- 2010-05-28: **HapMap3 CNV Genotypes**

  Copy Number Variation genotypes for HapMap phase samples are available for bulk download.

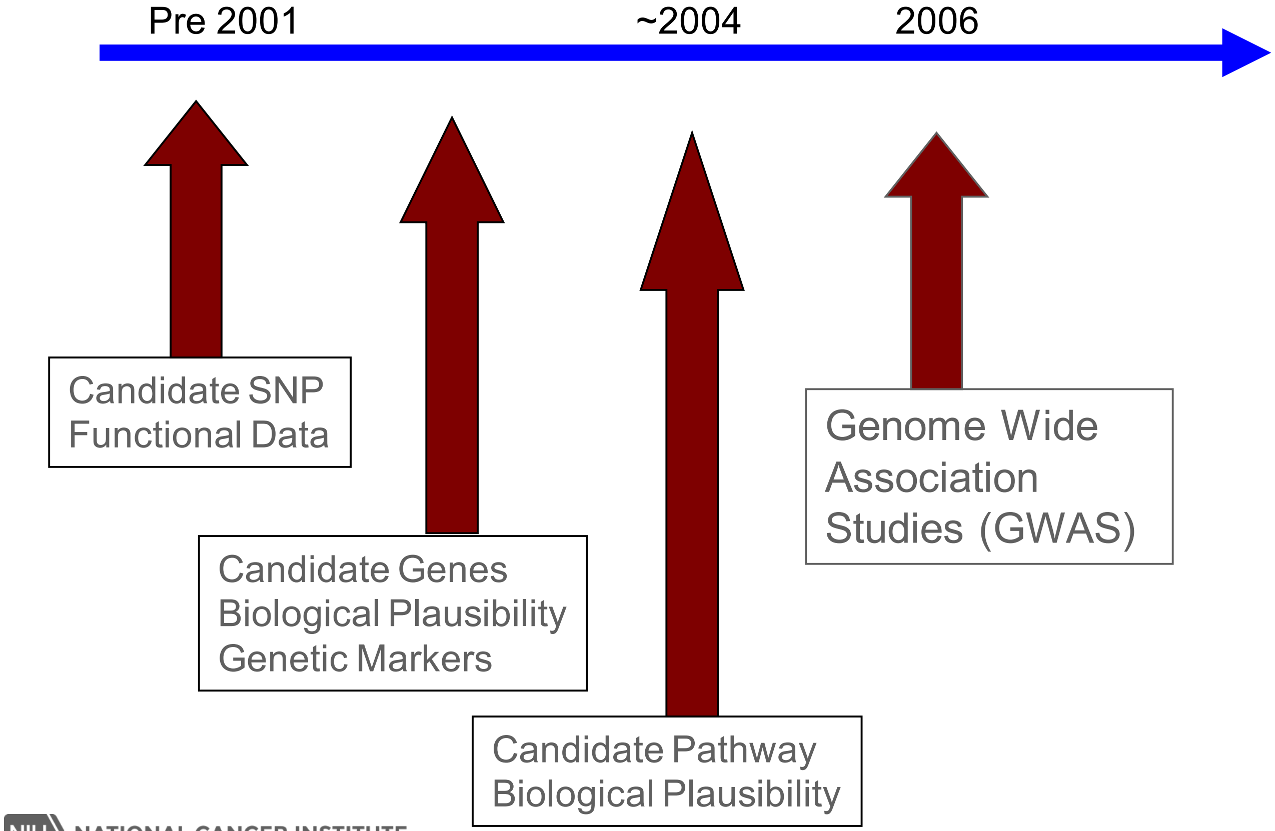
- 2009-12-10: **Corrected HapMap3 phased haplotypes available for chromosome X**

  Phased haplotypes for consensus HapMap3 release 2 data for chromosome X has been corrected and the new data are now available for bulk download. Sorry for any inconvenience this might have caused.

- 2009-12-02: **HapMap3 phased haplotypes available for chromosome X**


Cancelling Scan
Please wait...

# Trajectory of the Genetics of Cancer Susceptibility (2)



Pre 2001

~2004

2006

Candidate SNP
Functional Data

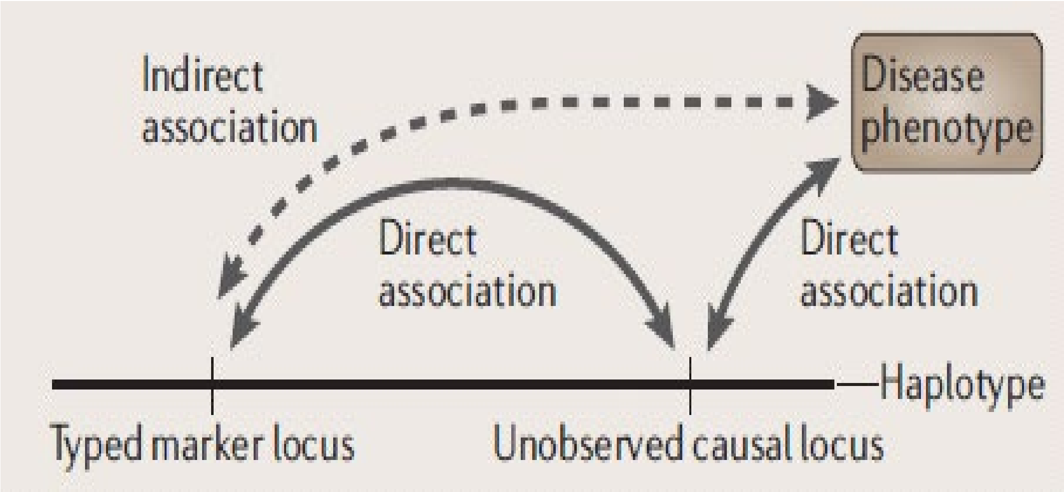
Candidate Genes
Biological Plausibility
Genetic Markers

Candidate Pathway
Biological Plausibility

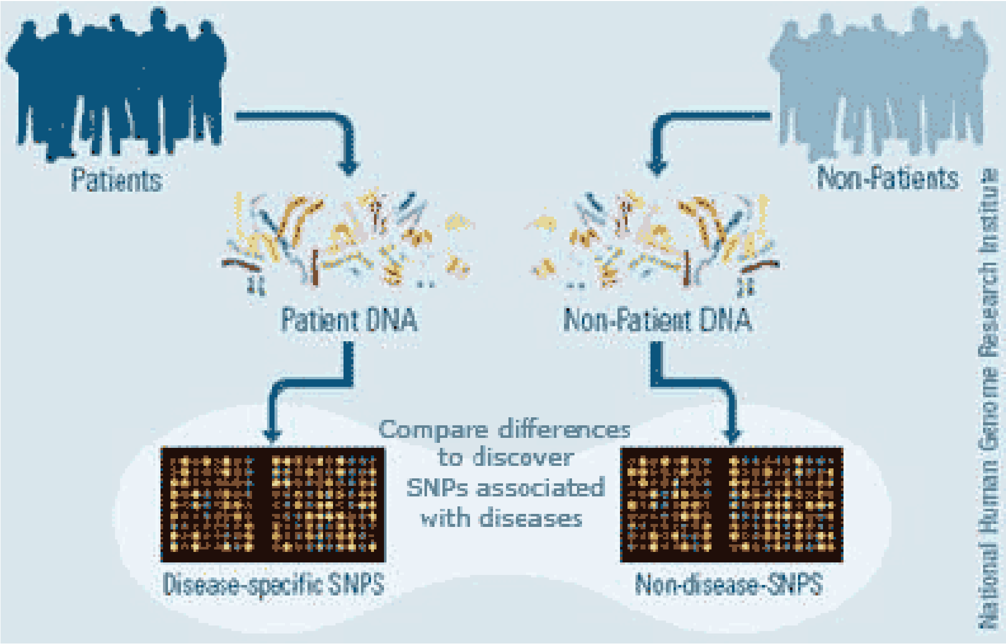
Genome Wide Association Studies (GWAS)

# GWAS

- Somewhat randomly-distributed SNPs are genotyped as “markers”
- Capitalize on phenomenon known as linkage disequilibrium (non-random association of alleles at different loci)
- “Agnostic”

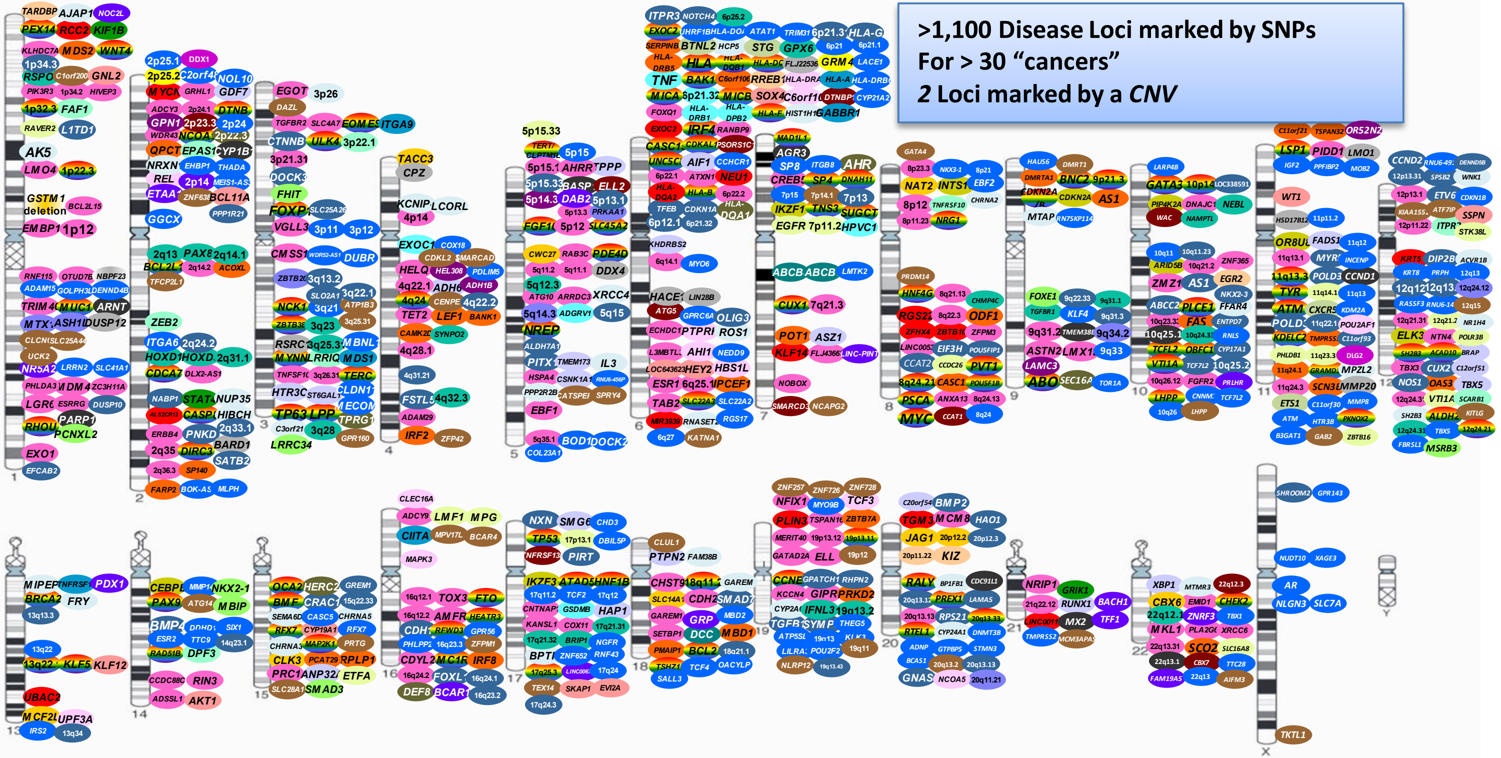


# Basic principle of genetic association studies in unrelated individuals


Patients
Non-Patients
Patient DNA
Non-Patient DNA
Compare differences to discover SNPs associated with diseases
Disease-specific SNPS
Non-disease-SNPS
National Human Genome Research Institute

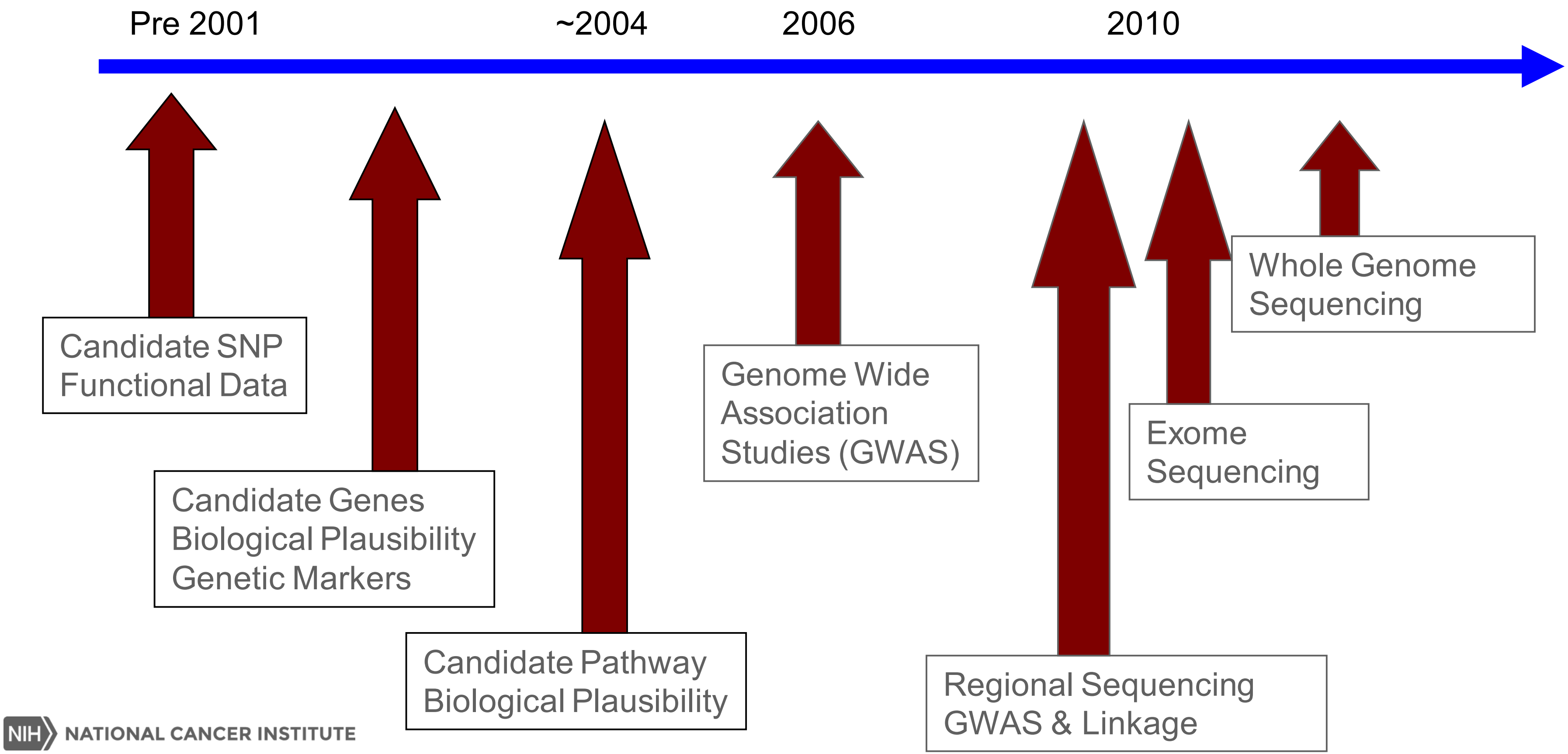
# Published Cancer GWAS Etiology Hits: June 2019

>1,100 Disease Loci marked by SNPs
For > 30 "cancers"
*2* Loci marked by a *CNV*



36 Basal Cell
17 Bladder
184 Breast
15 Cervical
81 CLL
122 Colorectal
15 Cutaneous Sq
22 Endometrial
29 Esophageal
6 Ewing Sarcoma
3 Gallbladder
14 Gastric
38 Glioma
29 Hodgkins
13 Kidney
7 Liver
48 Lung
22 Melanoma
103 Multiple
23 Multiple Myeloma
14 Nasopharyngeal
18 Neuroblastoma
15 Non-Hodgkin
10 Oral
2 Osteosarcoma
39 Ovary
28 Pancreas
16 Pediatric ALL
170 Prostate
60 Testicular
17 Thyroid
3 Wilms

# Trajectory of the Genetics of Cancer Susceptibility (3)



Pre 2001

~2004

2006

2010

Candidate SNP
Functional Data

Candidate Genes
Biological Plausibility
Genetic Markers

Candidate Pathway
Biological Plausibility

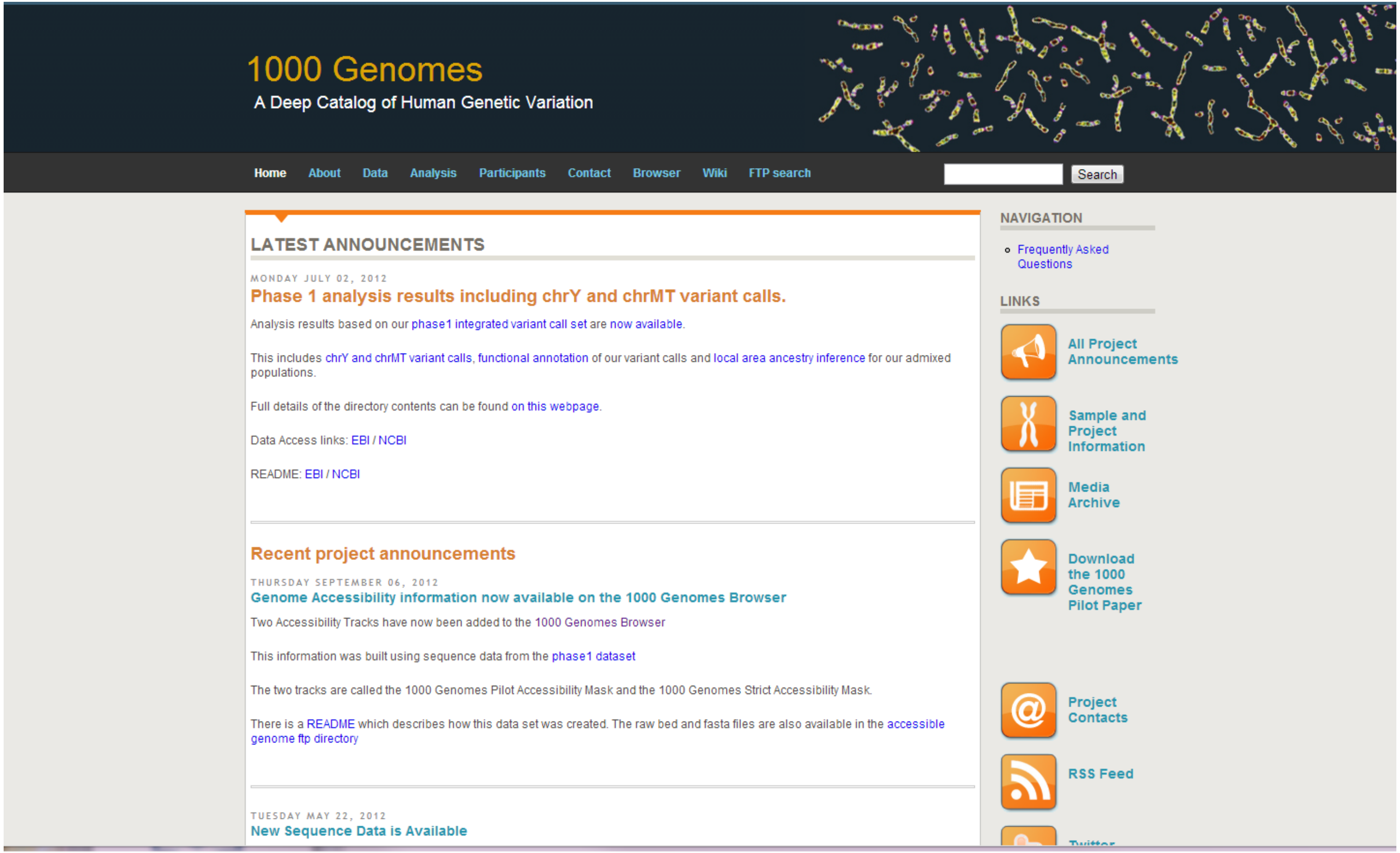
Genome Wide
Association
Studies (GWAS)

Regional Sequencing
GWAS & Linkage

Exome
Sequencing

Whole Genome
Sequencing

NIH NATIONAL CANCER INSTITUTE



# 1000 Genomes

A Deep Catalog of Human Genetic Variation

## LATEST ANNOUNCEMENTS

MONDAY JULY 02, 2012

### Phase 1 analysis results including chrY and chrMT variant calls.

Analysis results based on our phase1 integrated variant call set are now available.

This includes chrY and chrMT variant calls, functional annotation of our variant calls and local area ancestry inference for our admixed populations.

Full details of the directory contents can be found on this webpage.

Data Access links: EBI / NCBI

README: EBI / NCBI

### Recent project announcements

THURSDAY SEPTEMBER 06, 2012

**Genome Accessibility information now available on the 1000 Genomes Browser**

Two Accessibility Tracks have now been added to the 1000 Genomes Browser

This information was built using sequence data from the phase1 dataset

The two tracks are called the 1000 Genomes Pilot Accessibility Mask and the 1000 Genomes Strict Accessibility Mask.

There is a README which describes how this data set was created. The raw bed and fasta files are also available in the accessible genome ftp directory

TUESDAY MAY 22, 2012

**New Sequence Data is Available**

NAVIGATION

- Frequently Asked Questions

LINKS

All Project Announcements

Sample and Project Information

Media Archive

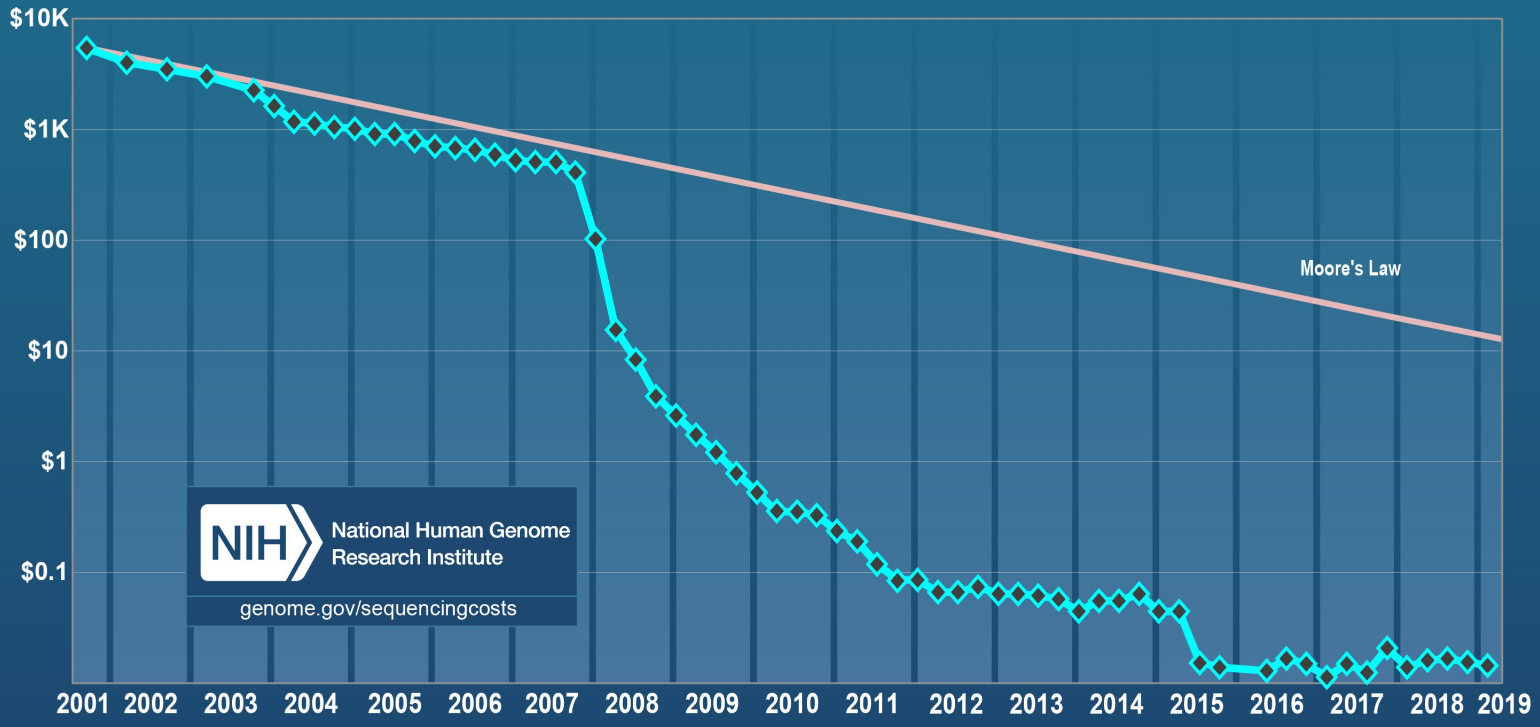
Download the 1000 Genomes Pilot Paper

Project Contacts

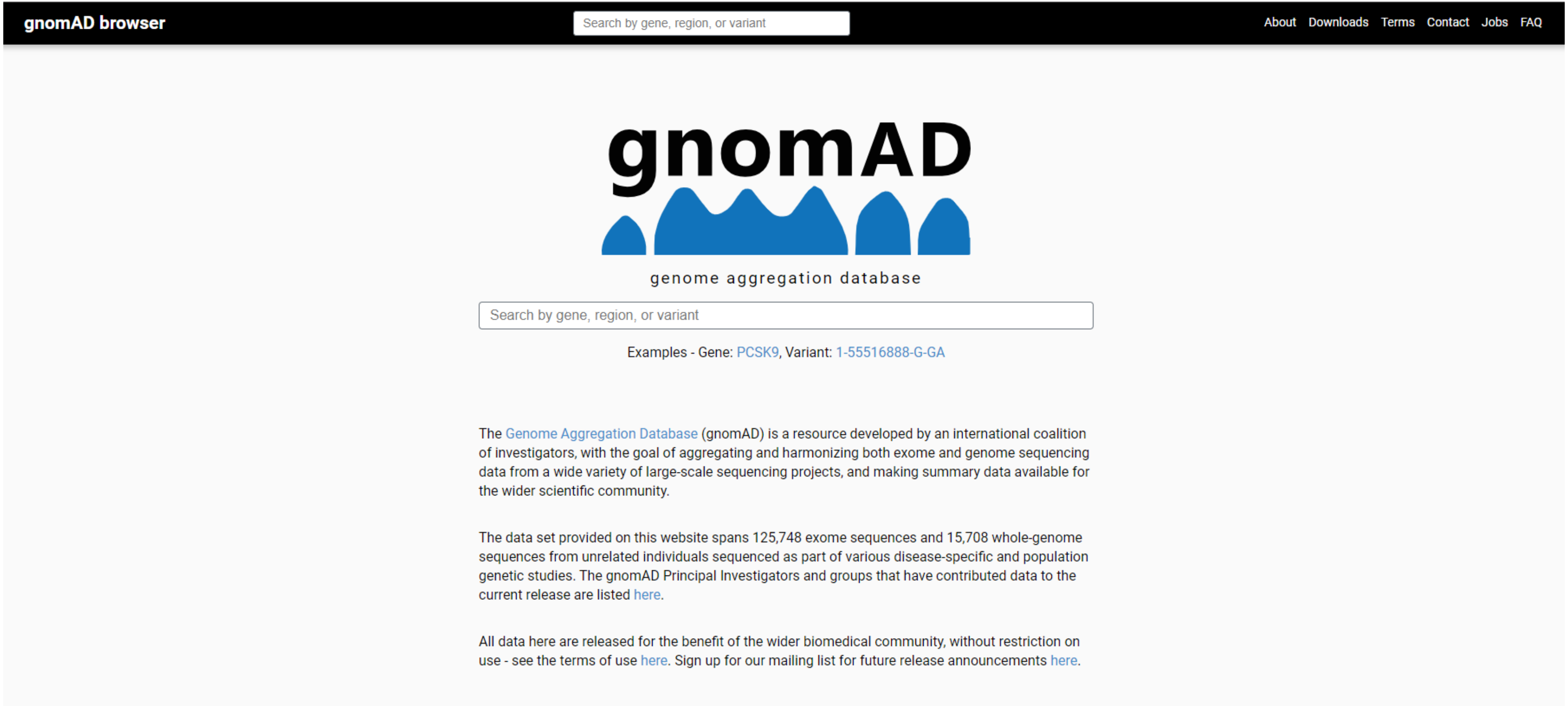
RSS Feed

Twitter

# Cost per Raw Megabase of DNA Sequence



https://gnomad.broadinstitute.org/



gnomAD browser
Search by gene, region, or variant
About
Downloads
Terms
Contact
Jobs
FAQ
gnomAD
genome aggregation database
Search by gene, region, or variant

Examples - Gene: PCSK9, Variant: 1-55516888-G-GA

The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed here.

All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use here. Sign up for our mailing list for future release announcements here.

125,748 exomes

15,708 whole genomes

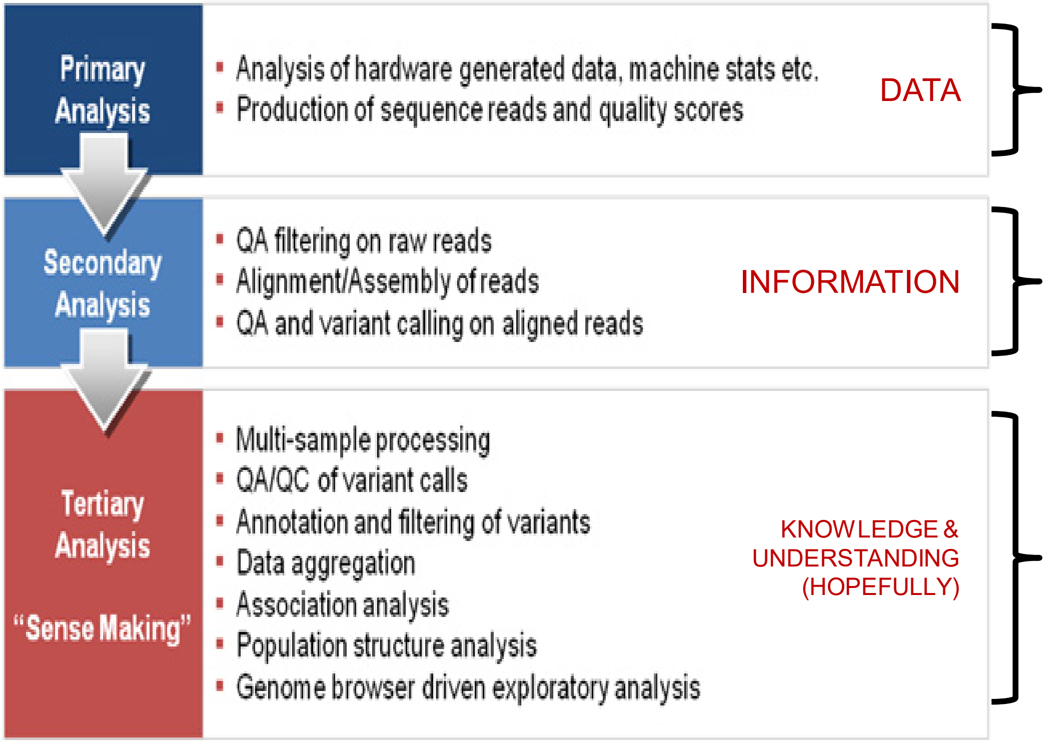
# Examples: Genetic studies

- Genome-wide association studies – use LD to find SNPs associated with a trait or disease.
  - Most GWAS “hits” intergenic
  - Most associations relatively weak (OR ~ 1.2) with common SNPs
- Culprit probably a variant in a regulatory region?
- Exome sequencing studies – sequence the protein-coding regions of the genome
  - Rare variant of strong effect

# Examples: Genetic studies, continued

- Target-gene(s) sequencing to study only a particular gene or set of genes
  - Example: BRCA1/BRCA2
- Regional sequencing of a contiguous part of a chromosome
  - Example: GWAS follow-up
- Whole genome sequencing
- Look everywhere in the genome; structural variation detection

# NGS Data Analysis: Like panning for gold



**Primary Analysis**

- Analysis of hardware generated data, machine stats etc.
- Production of sequence reads and quality scores

DATA

Generate high-quality raw sequences

**Secondary Analysis**

- QA filtering on raw reads
- Alignment/Assembly of reads
- QA and variant calling on aligned reads

INFORMATION

Reassemble reads to so they represent underlying biology more closely

**Tertiary Analysis**

**"Sense Making"**

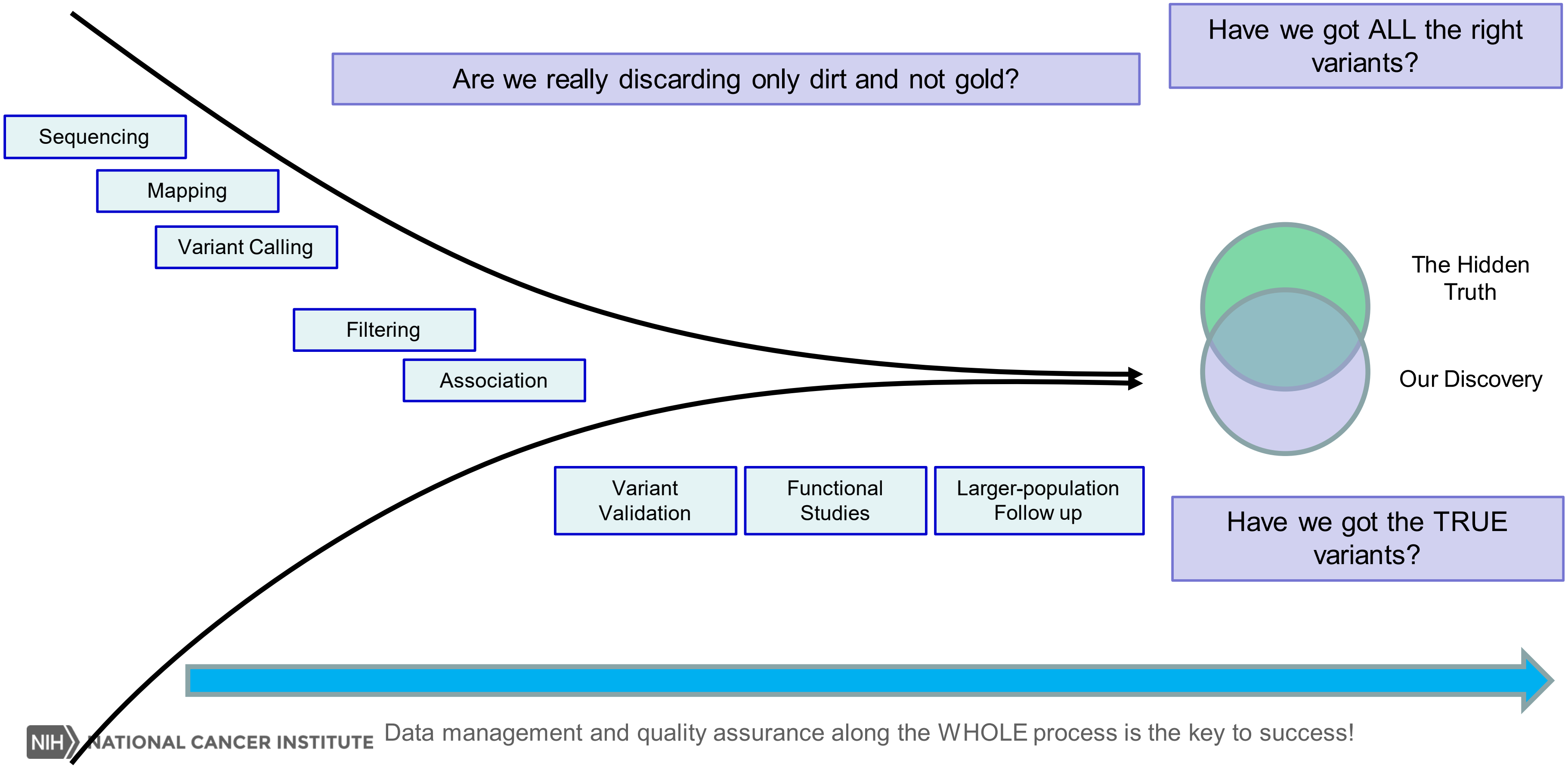
- Multi-sample processing
- QA/QC of variant calls
- Annotation and filtering of variants
- Data aggregation
- Association analysis
- Population structure analysis
- Genome browser driven exploratory analysis

KNOWLEDGE &
UNDERSTANDING
(HOPEFULLY)

Enable hypothesis generation and testing tailored to specific scientific questions

Golden Helix: A Hitchhiker's Guide to Next Generation Sequencing
http://gettinggeneticsdone.blogspot.com/2011/05/golden-helix-hitchhikers-guide-to-next.html

# “Data Cleaning”: A Process of Information Losing



Are we really discarding only dirt and not gold?

Have we got ALL the right variants?

Sequencing

Mapping

Variant Calling

Filtering

Association

Variant Validation

Functional Studies

Larger-population Follow up

Have we got the TRUE variants?

Data management and quality assurance along the WHOLE process is the key to success!

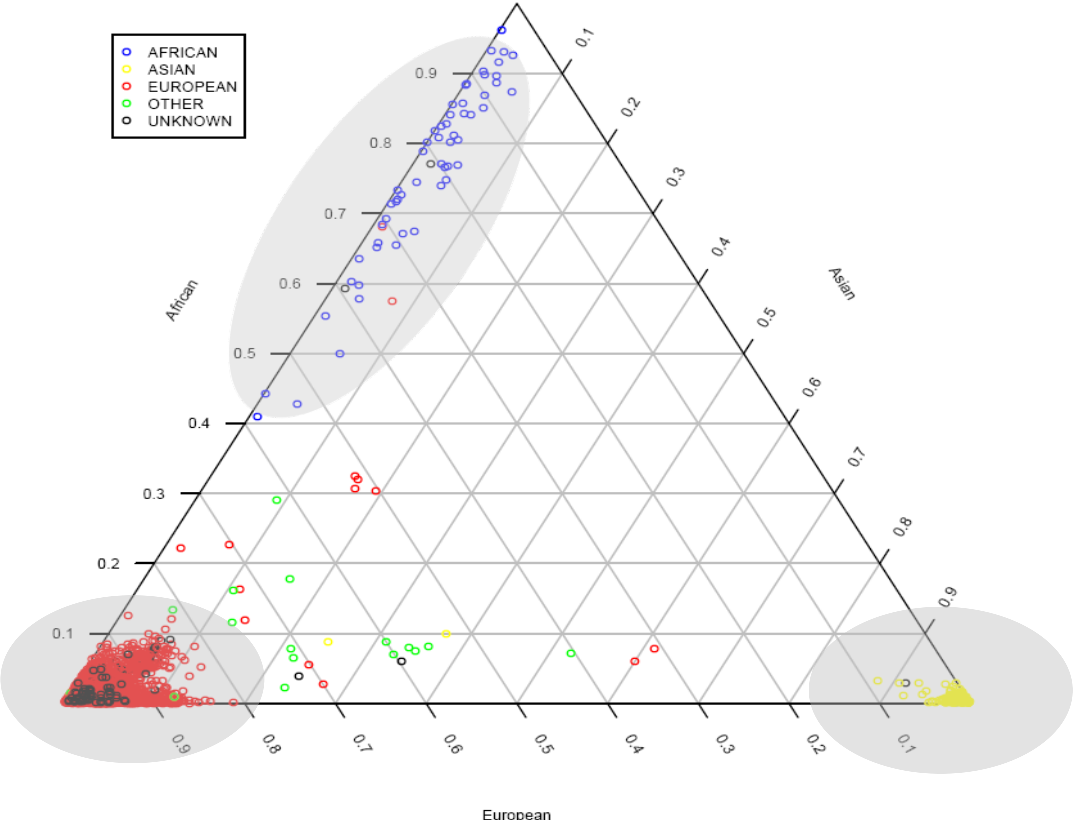
# Other Challenges

- Computational requirements
  - Sequence data storage (many terabytes of data)
  - Raw trace base calling and genomic mapping
  - Variant detection and annotation
  - In silico functional prediction
  - Libraries of in-house and public variation – when to use and how?

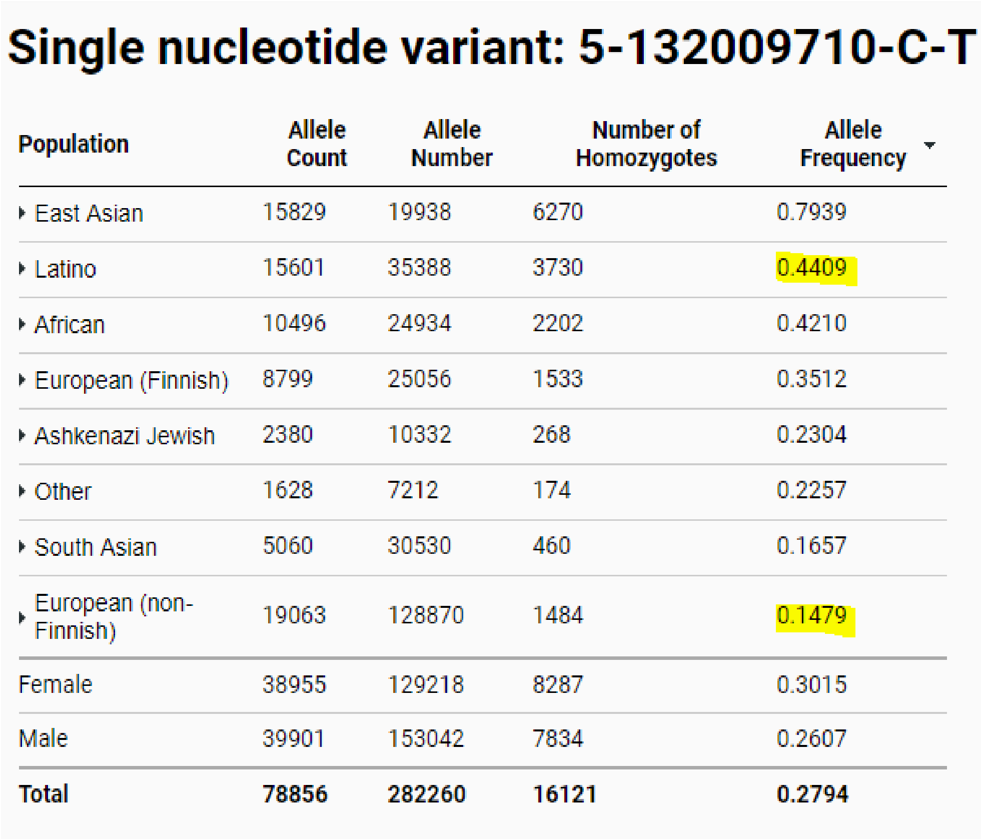
# More things to consider

- Study design – laboratory point of view is also important
  - Case / control
    - Distribute cases and controls evenly on each plate
    - Perform same assay on cases and controls
- Genetic ancestry is important

# Population Structure (Admixture)



# Many SNPs have different allele frequencies in different populations



## Single nucleotide variant: 5-132009710-C-T

| Population | Allele Count | Allele Number | Number of Homozygotes | Allele Frequency ▾ |
|---|---|---|---|---|
| ▸ East Asian | 15829 | 19938 | 6270 | 0.7939 |
| ▸ Latino | 15601 | 35388 | 3730 | 0.4409 |
| ▸ African | 10496 | 24934 | 2202 | 0.4210 |
| ▸ European (Finnish) | 8799 | 25056 | 1533 | 0.3512 |
| ▸ Ashkenazi Jewish | 2380 | 10332 | 268 | 0.2304 |
| ▸ Other | 1628 | 7212 | 174 | 0.2257 |
| ▸ South Asian | 5060 | 30530 | 460 | 0.1657 |
| ▸ European (non-Finnish) | 19063 | 128870 | 1484 | 0.1479 |
| Female | 38955 | 129218 | 8287 | 0.3015 |
| Male | 39901 | 153042 | 7834 | 0.2607 |
| **Total** | **78856** | **282260** | **16121** | **0.2794** |

HEALTH INC.

# How Should Scientists' Access To Health Databanks Be Managed?

September 6, 2019 · 5:04 AM ET
Heard on Morning Edition

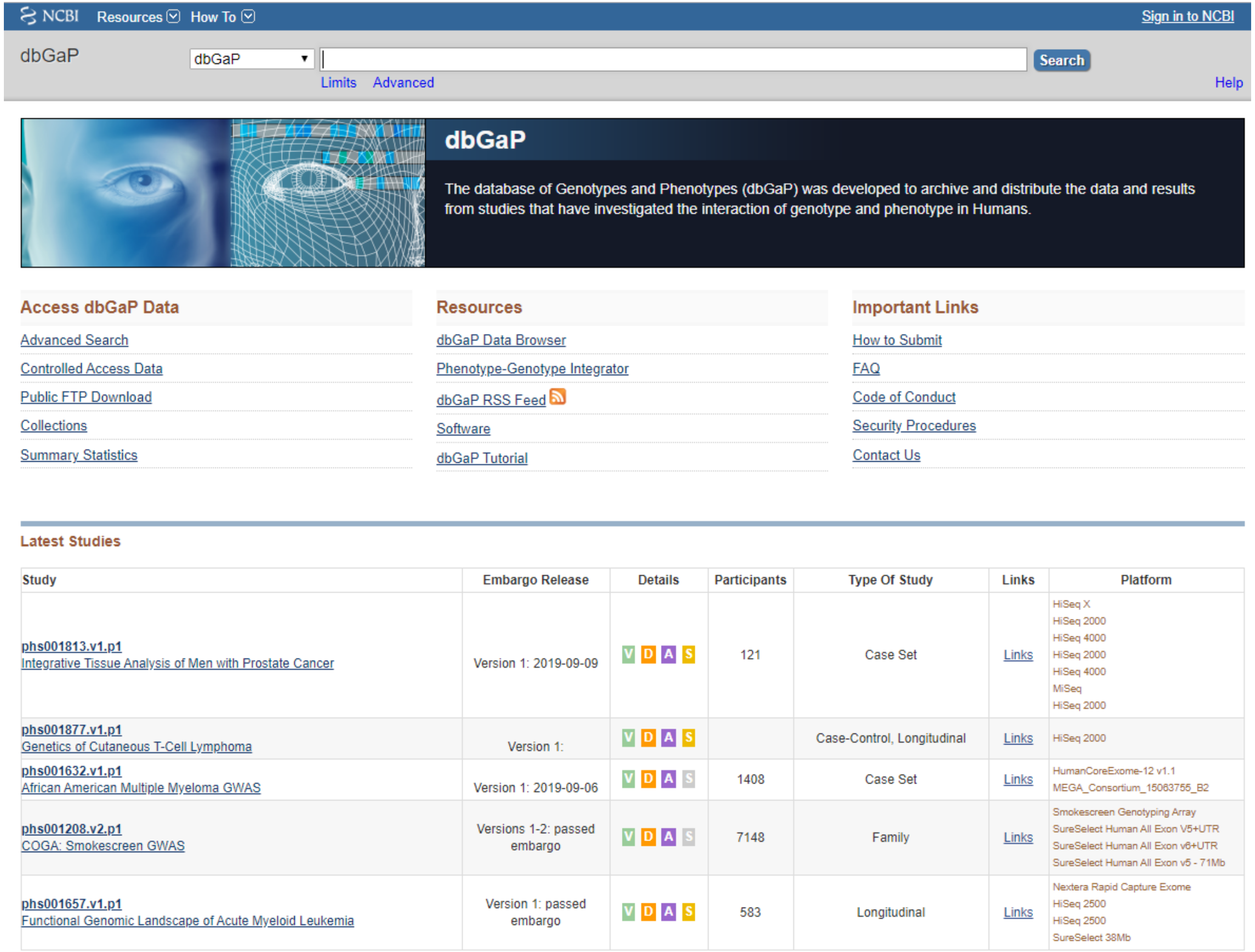
RICHARD HARRIS

UK Biobank has granted 10,000 qualified scientists access to its large database of genetic sequences and other medical data, but other organizations with databases have been far more restrictive in giving access.

*KTSDESIGN/Getty Images/Science Photo Library*

*"The philosophy is straightforward: The more easily smart people can see the data, the more likely they are to make discoveries that can benefit us all."*

https://www.npr.org/sections/health-shots/2019/09/06/755402750/how-should-scientists-access-to-health-databanks-be-managed



## dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

### Access dbGaP Data

- Advanced Search
- Controlled Access Data
- Public FTP Download
- Collections
- Summary Statistics

### Resources

- dbGaP Data Browser
- Phenotype-Genotype Integrator
- dbGaP RSS Feed
- Software
- dbGaP Tutorial

### Important Links

- How to Submit
- FAQ
- Code of Conduct
- Security Procedures
- Contact Us

### Latest Studies

| Study | Embargo Release | Details | Participants | Type Of Study | Links | Platform |
|---|---|---|---|---|---|---|
| **phs001813.v1.p1**<br>Integrative Tissue Analysis of Men with Prostate Cancer | Version 1: 2019-09-09 | V D A S | 121 | Case Set | Links | HiSeq X<br>HiSeq 2000<br>HiSeq 4000<br>HiSeq 2000<br>HiSeq 4000<br>MiSeq<br>HiSeq 2000 |
| **phs001877.v1.p1**<br>Genetics of Cutaneous T-Cell Lymphoma | Version 1: | V D A S | | Case-Control, Longitudinal | Links | HiSeq 2000 |
| **phs001632.v1.p1**<br>African American Multiple Myeloma GWAS | Version 1: 2019-09-06 | V D A S | 1408 | Case Set | Links | HumanCoreExome-12 v1.1<br>MEGA_Consortium_15063755_B2 |
| **phs001208.v2.p1**<br>COGA: Smokescreen GWAS | Versions 1-2: passed embargo | V D A S | 7148 | Family | Links | Smokescreen Genotyping Array<br>SureSelect Human All Exon V5+UTR<br>SureSelect Human All Exon v6+UTR<br>SureSelect Human All Exon v5 - 71Mb |
| **phs001657.v1.p1**<br>Functional Genomic Landscape of Acute Myeloid Leukemia | Version 1: passed embargo | V D A S | 583 | Longitudinal | Links | Nextera Rapid Capture Exome<br>HiSeq 2500<br>HiSeq 2500<br>SureSelect 38Mb |

# Conclusions

- Technological advances and community efforts have allowed for many ways to affordably approach our questions of disease
- Multiple technologies and assays are available in order to best explore biological questions
- There are ever-growing issues in analyzing and storing large datasets
- Study design from a laboratory point-of-view is (and always will be) important
- Data sharing is important for the advancement of science

U.S. Department of Health & Human Services

National Institutes of Health | National Cancer Institute

dceg.cancer.gov/

1-800-4-CANCER

Produced September 2019