

# **Next-generation sequencing Course**

February 23, 2015

Marylyn D. Ritchie and Sarah A. Pendergrass

# Outline

Learning objectives are:

- Technologies for genome/exome sequencing
- Study designs for sequencing experiments
- Learn about new methodological approaches to look for both rare variation, as well as the combination of common and rare variation associated with traits
- Discuss what we can learn from sequence data as well as ethical concerns

# **DNA sequence variation**

## **Lecture 1**



I want to do a genetic study on my phenotype of interest. What do I do next?



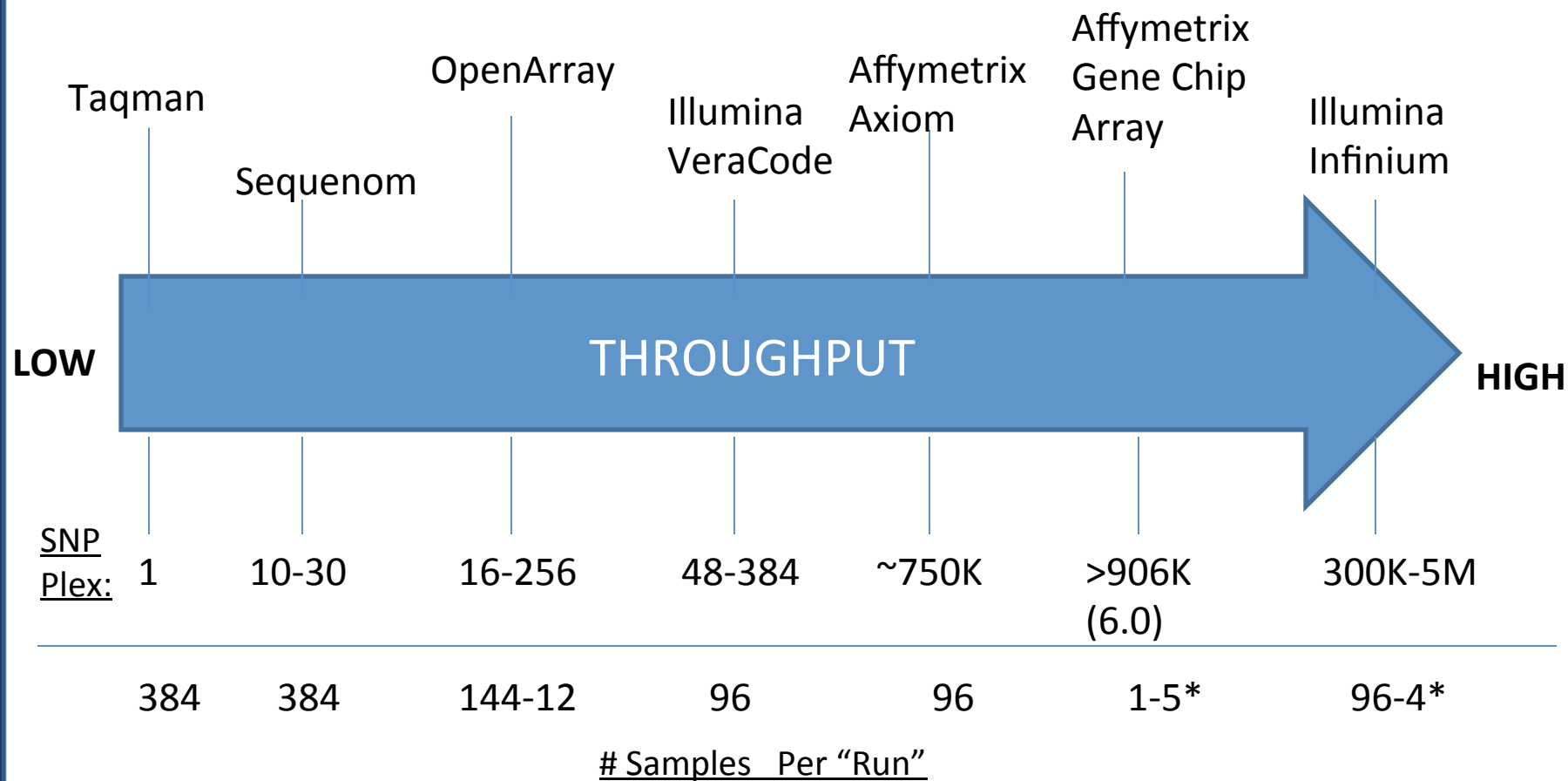


# If only it were that easy.....



- What's your research question?
- What type of genetic variation are you interested in?
  - Rare or common? Both? Either?
- How many samples do you have access to?
  - What type of samples and when/how were they extracted?
  - How much DNA do you have?
- How many variants do you want to (or can you) analyze?
  - Sample all variation in the genome?
  - Tag variants?
- What's your data analysis plan?
- What's your timeline?
- What's your budget?
- What do you wish to accomplish?

# Options for Genotyping SNPs



# Hardware for Genotyping



ay Genotyping System, see the feature article,  
[com/tagmanopenarray](http://com/tagmanopenarray)



Figure 1. TaqMan® OpenArray™ Genotyping System.



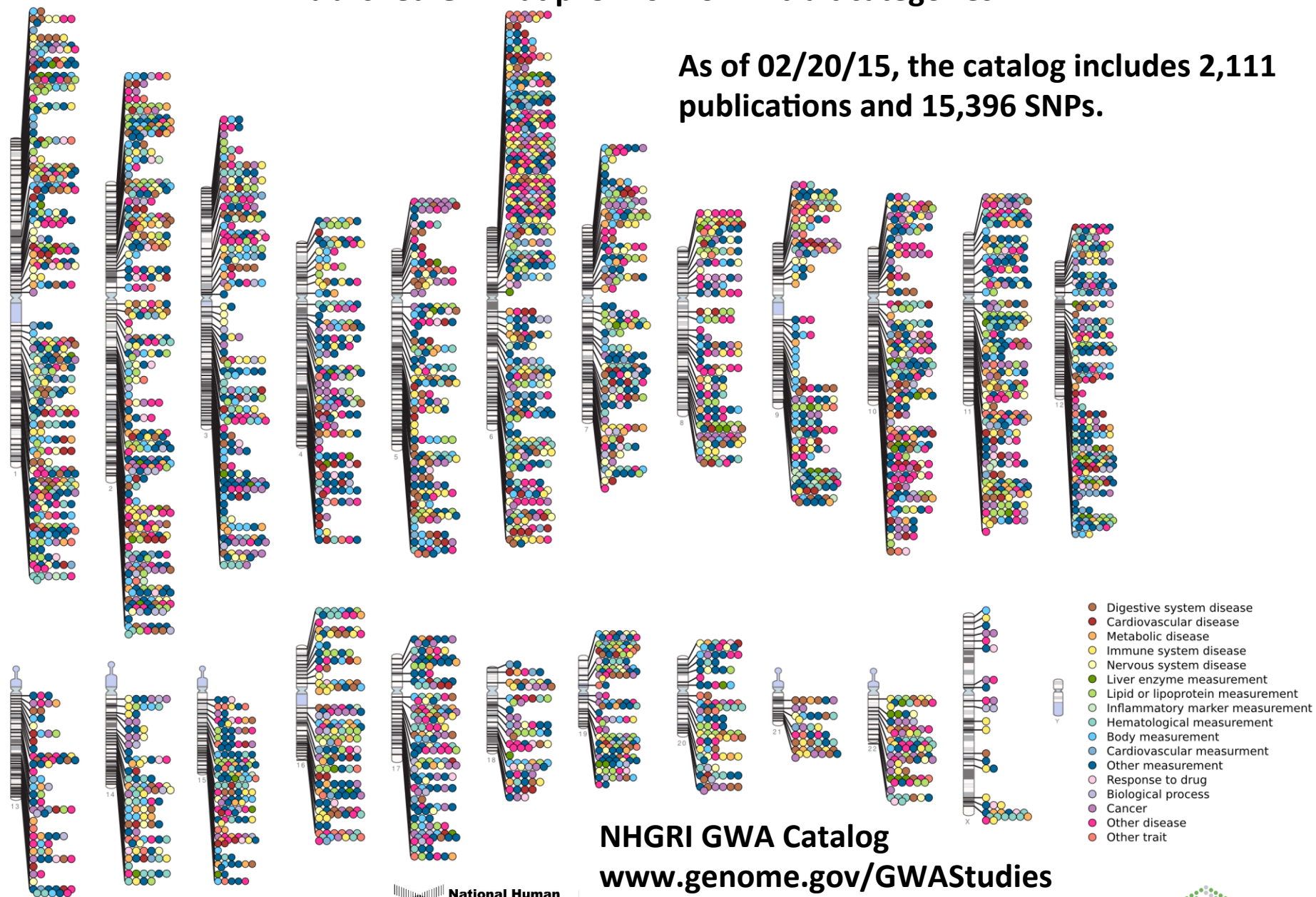
# Genome-wide Common Variant Panels

- 10,000-5 million SNPs
- Affymetrix, Illumina
- Random SNPs – spaced across the genome
- Selected haplotype tag SNPs
- Copy Number Probes
- Considerations for genotyping...
  - Mix up Cases/Controls/ethnic groups
  - Run duplicates
  - Run trios
  - Run HapMap Controls (Positive Control)
  - Blanks/no blanks? (Negative Control)

# Published Genome-Wide Associations through 12/2013

Published GWA at  $p \leq 5 \times 10^{-8}$  for 17 trait categories

As of 02/20/15, the catalog includes 2,111 publications and 15,396 SNPs.

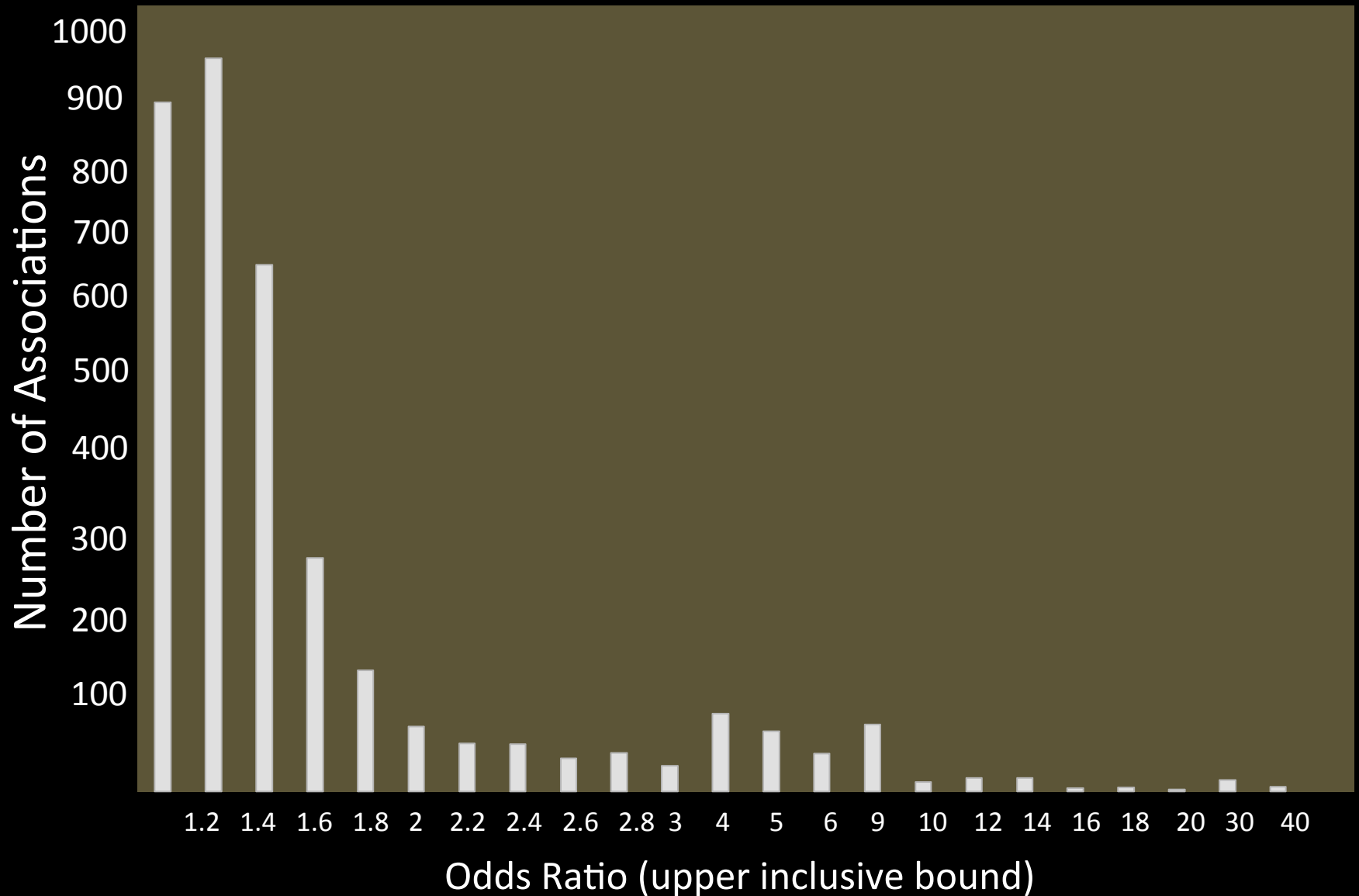


NHGRI GWA Catalog

[www.genome.gov/GWAStudies](http://www.genome.gov/GWAStudies)

[www.ebi.ac.uk/fgpt/gwas/](http://www.ebi.ac.uk/fgpt/gwas/)

# Distribution of Effects







## 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.

PENNSTATE



Maher, B. *Nature* 2008; 456:18-21.



### **The case of the missing heritability**

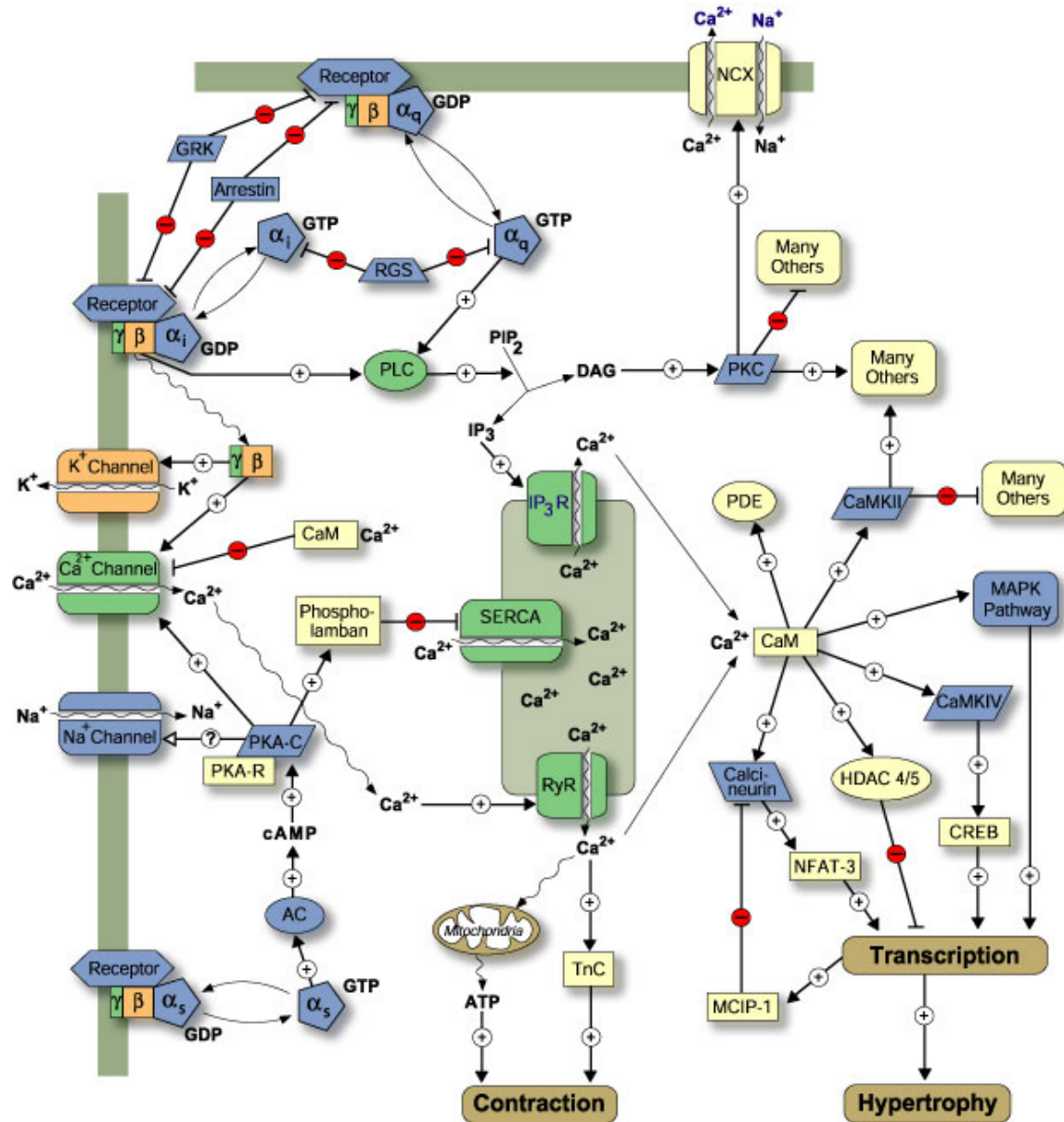
When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

## **Missing Heritability**

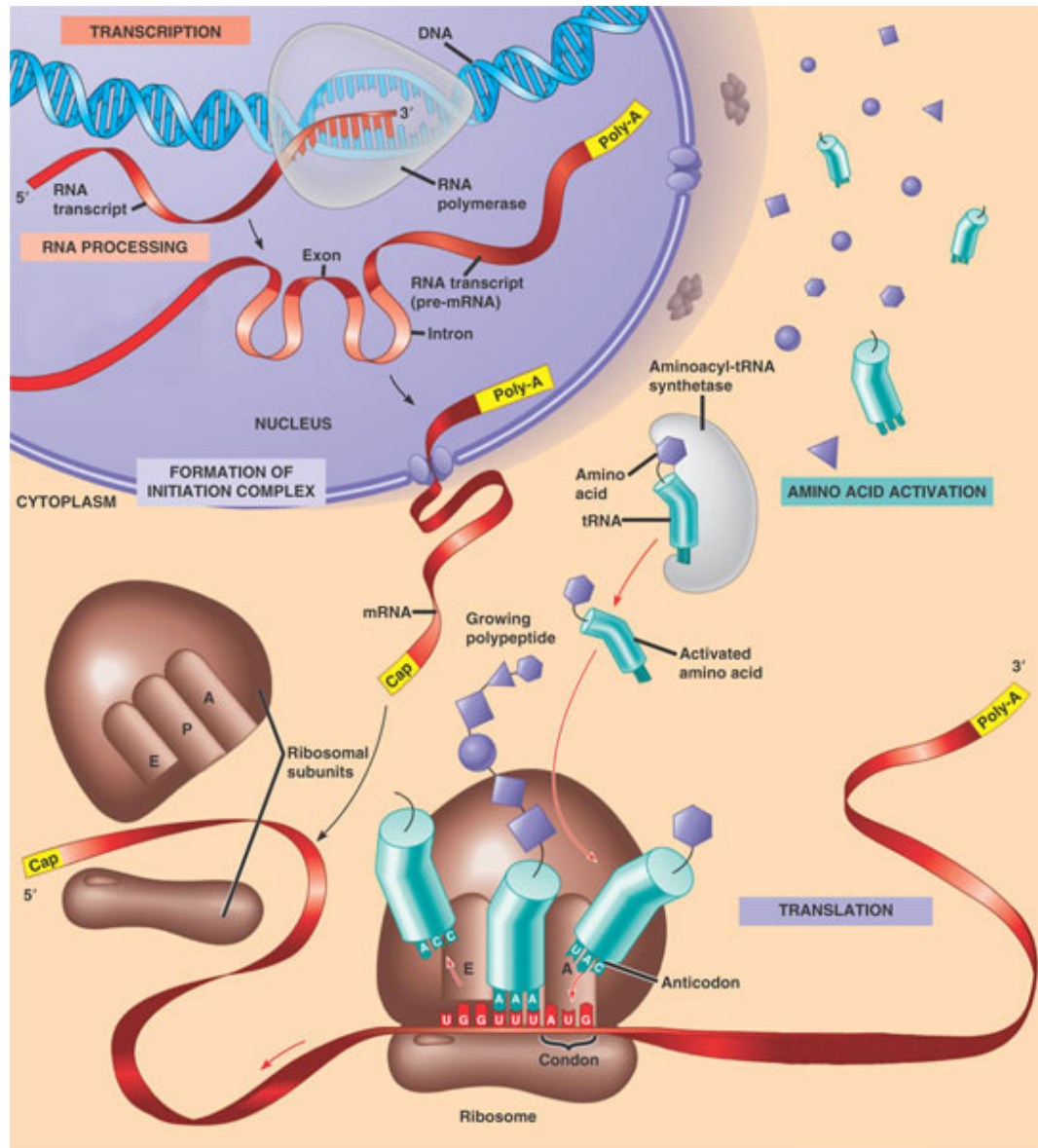
- Under our nose
- Out of sight
- In the architecture
- Underground networks
- Lost in diagnosis
- The great beyond



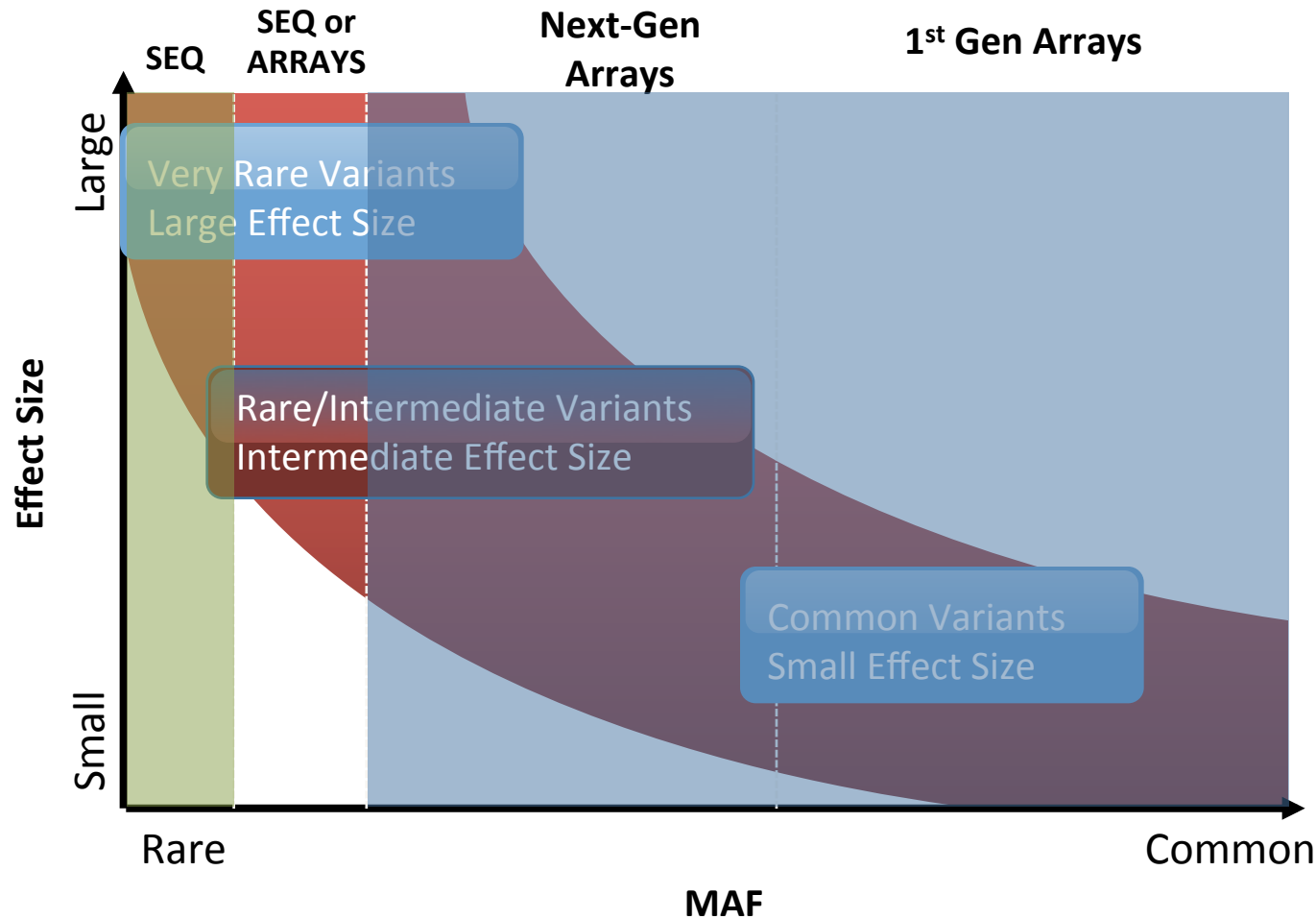
# Biology is complex



# Molecular biology is complex



# Enabling discoveries with Next-Generation GWAS



# Timeline

1. Release of complete human genome
2. HapMap Consortium, collection of common variation
3. Common Variant-Common Disease did not explain much heritability
4. Began to investigate rare variants
5. 1000 Genomes Project

# The International HapMap

- Project started in 2003
- Designed to determine frequencies and patterns of association between common SNPs

## HapMap details

	# of SNPs Genotyped	Targeted SNPs	Populations Studied
Phase I	1 million	Prioritized coding SNPs to attain 1 SNP for each 5-kb region	CEU,YRI,CHB,JPT
Phase II	3 million	Prioritized non-synonymous SNPs in coding regions	CEU,YRI,CHB,JPT
Phase III	1.4 million	Prioritized rare variants	CEU,YRI,CHB,JPT, ASW,CHB,GIH,LWK, MXL,MKK,TSI

# The 'Common Disease-Common Variant' Hypothesis and Familial Risks

Kari Hemminki<sup>1,2\*</sup>, Asta Försti<sup>1,2</sup>, Justo Lorenzo Bermejo<sup>1</sup>

**1** Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, **2** Center for Family and Community Medicine, Karolinska Institute, Huddinge, Sweden

## Abstract

The recent large genotyping studies have identified a new repertoire of disease susceptibility loci of unknown function, characterized by high allele frequencies and low relative risks, lending support to the common disease-common variant (CDCV) hypothesis. The variants explain a much larger proportion of the disease etiology, measured by the population attributable fraction, than of the familial risk. We show here that if the identified polymorphisms were markers of rarer functional alleles they would explain a much larger proportion of the familial risk. For example, in a plausible scenario where the marker is 10 times more common than the causative allele, the excess familial risk of the causative allele is over 10 times higher than that of the marker allele. However, the population attributable fractions of the two alleles are equal. The penetrance mode of the causative locus may be very difficult to deduce from the apparent penetrance mode of the marker locus.

**Citation:** Hemminki K, Försti A, Bermejo JL (2008) The 'Common Disease-Common Variant' Hypothesis and Familial Risks. PLoS ONE 3(6): e2504. doi:10.1371/journal.pone.0002504

**Editor:** A. Cecile J. W. Janssens, Erasmus University Medical Center, Netherlands

**Received:** January 30, 2008; **Accepted:** May 16, 2008; **Published:** June 18, 2008

**Copyright:** © 2008 Hemminki et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits

# Want to learn more about GWAS?

[Methods Mol Biol.](#) 2014;1168:63-81. doi: 10.1007/978-1-4939-0847-9\_5.

## **Bioinformatics challenges in genome-wide association studies (GWAS).**

[De R<sup>1</sup>](#), [Bush WS](#), [Moore JH](#).

### **⊕ Author information**

#### **Abstract**

Genome-wide association studies (GWAS) are a powerful tool for investigators to examine the human genome to detect genetic risk factors, reveal the genetic architecture of diseases and open up new opportunities for treatment and prevention. However, despite its successes, GWAS have not been able to identify genetic loci that are effective classifiers of disease, limiting their value for genetic testing. This chapter highlights the challenges that lie ahead for GWAS in better identifying disease risk predictors, and how we may address them. In this regard, we review basic concepts regarding GWAS, the technologies used for capturing genetic variation, the missing heritability problem, the need for efficient study design especially for replication efforts, reducing the bias introduced into a dataset, and how to utilize new resources available, such as electronic medical records. We also look to what lies ahead for the field, and the approaches that can be taken to realize the full potential of GWAS.

PMID: 24870131 [PubMed - indexed for MEDLINE]

# The 1000 Genomes Project

*“Provide deep characterization of human genome sequence”*

- Expand the investigation of causal variants to include rare variants
- Project started in 2008
- Genotype 95% of 1+% variation

## Details for 1000 Genomes three pilot projects

Pilot Data sets	Populations	Samples	Coverage
Trio	2	6	20-40x
Low coverage	4	179	2-4x
Exon (8,140 exons ~5% of exome)	7	697	20-50x



# The 1,000 Genomes Project

*Sequence 1,000 genomes to complete the picture of genetic variation*

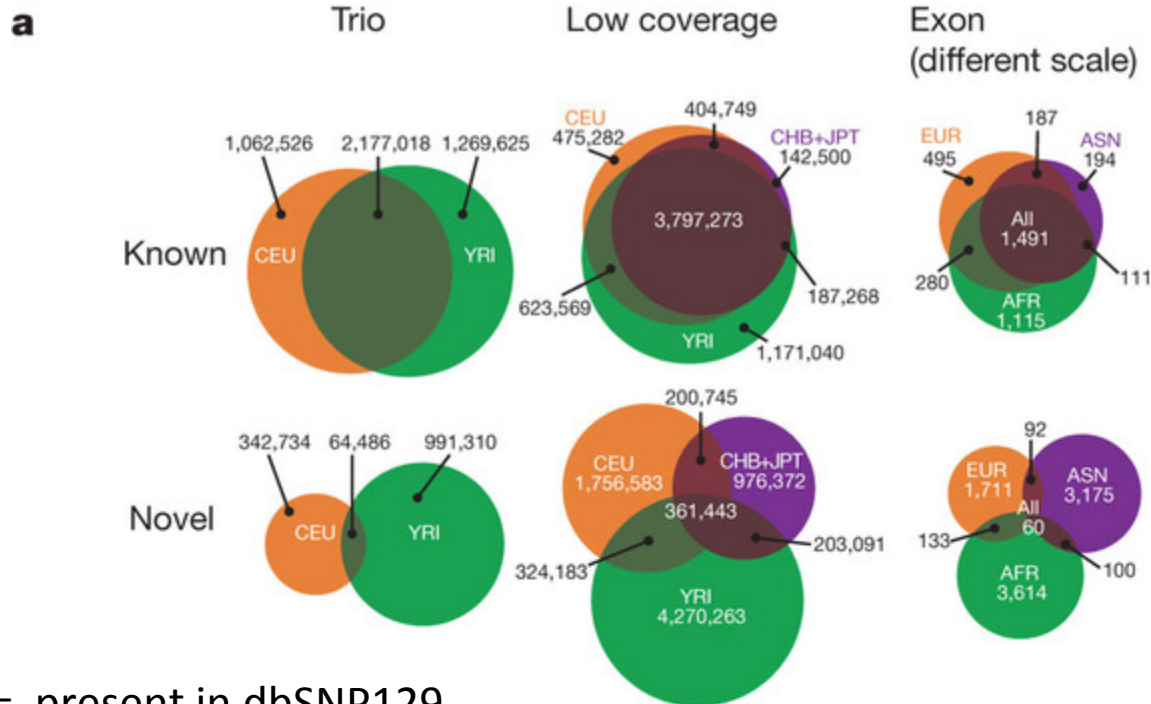
**Achieve a nearly complete catalog of common human genetic variants with frequency 1% or higher.**

## Project Goals

1. Accelerate fine-mapping efforts in gene regions indentified through genome-wide association studies or candidate gene studies
2. Improve the power of future genetic association studies by enabling design of next-generation genotyping microarrays that more fully represent human genetic variation
3. Enhance the analysis of ongoing and already completed association studies by improving our ability to “impute” or “predict” untyped genetic variants



# Properties of Variants Found



Known = present in dbSNP129

91% SNPs found in coding regions were already present in dbSNP

The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." Nature 2010

## 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.

# Better Reference Catalog?

Which database should be used as a reference?

- Which database was most recently released?
- Which contains the most variants?
- Which ethnicities were included?
- What is the allele frequency distribution captured?

# New Content for Next-gen GWAS Arrays

*Rich content to explore new hypotheses and enable new discoveries*

Sequence to discover SNPs >1% MAF (1000-Genomes project)

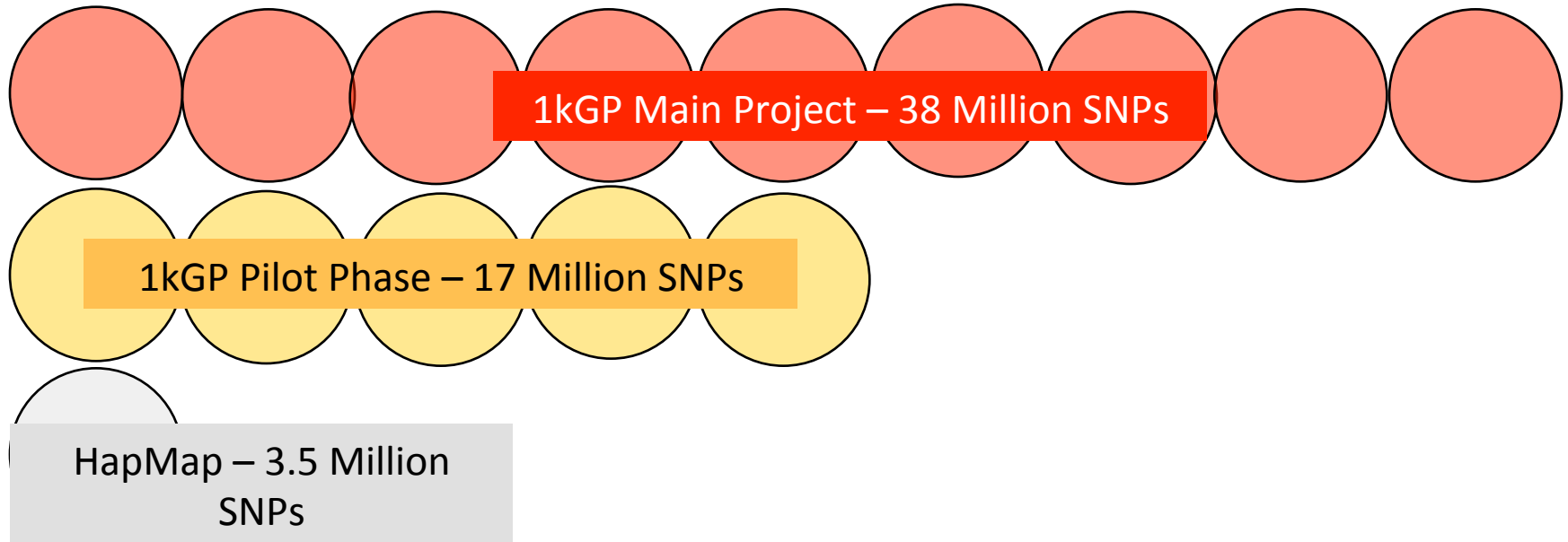
Leverage the power of LD to select tagSNPs and remove redundancy

Include progressively more SNPs at lower allele frequencies (5%, 2.5%, 1%)

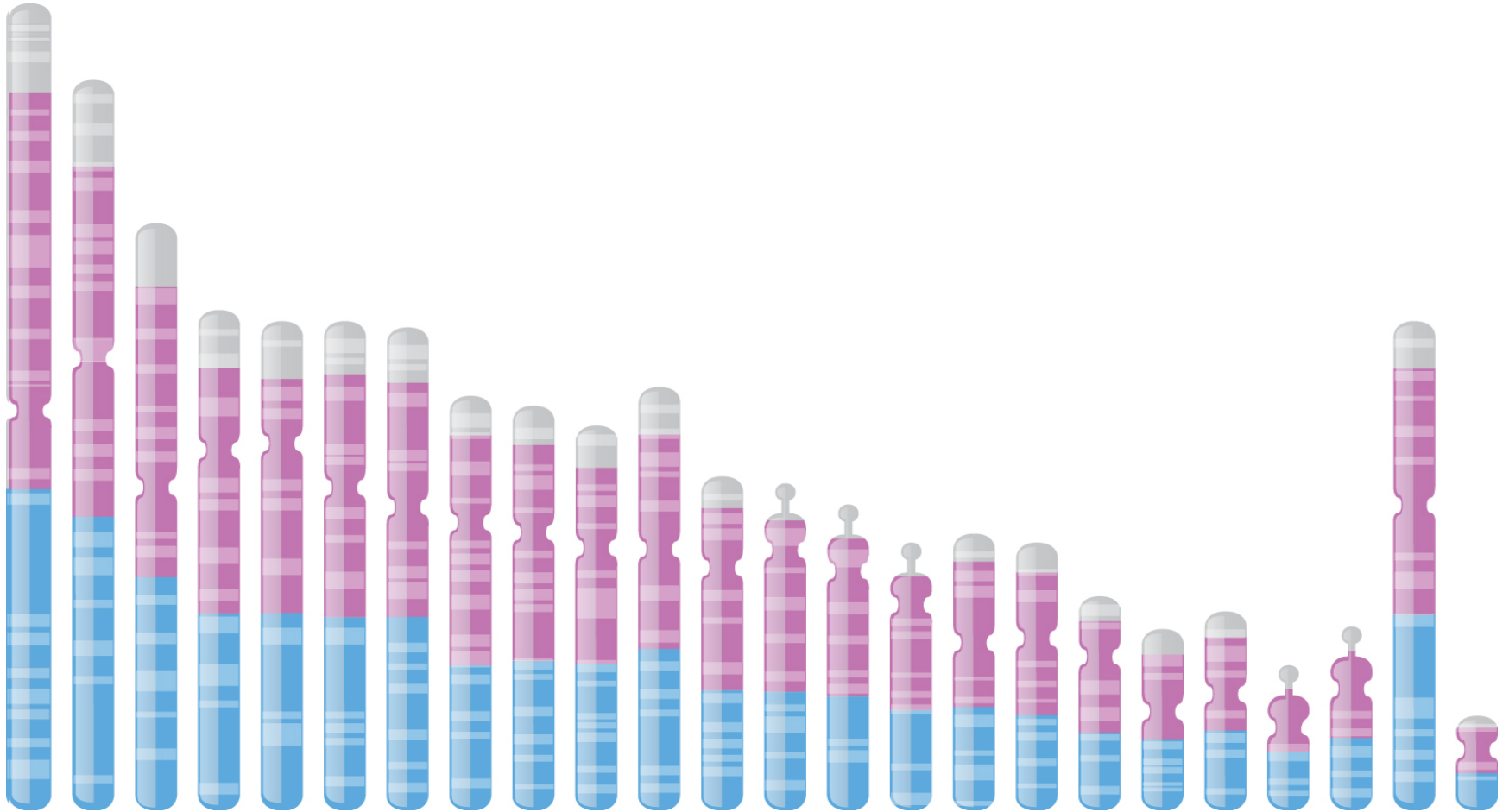
Project	Year	Approx. Cumulative SNPs found	Tag SNPs needed for max coverage	Lower limit of allele frequency targeted	% variation tagged ( $r^2 > 0.8$ )
HapMap	2003-2007	3M	0.6M	5%	>90%
1kG Pilot Project	2008-2009	17M	2.5M	2.5%	~80%
1kG Full Project	2010	38M	5.0M	1%	>90%



# The spectrum of known variation is expanding at an unprecedented rate...



# ...and resetting the reference point for GWAS.





# So maybe it's in the rare(r) variants

OPEN ACCESS Freely available online

PLOS BIOLOGY

## Synopsis

### Common Disease, Multiple Rare (and Distant) Variants

Richard Robinson\*

Freelance Science Writer, Sherborn, Massachusetts, United States of America

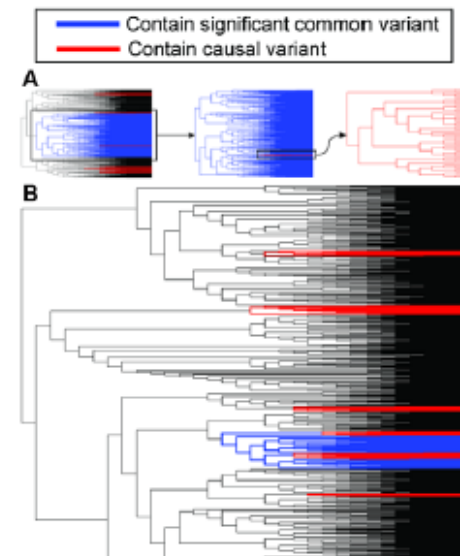
Genome-wide association (GWA) studies have emerged as a potentially powerful tool for discovery of new genes for common diseases, such as Alzheimer's disease and stroke. But the common interpretation of GWA findings might be incorrect in many cases, according to a new study by Samuel Dickson, David Goldstein, and colleagues in this issue of *PLoS Biology*. Their results suggest that the signals in these studies may not always be pointing to a few common gene variants, as assumed by most researchers, but instead to many rare variants, each of which causes relatively few cases, and each of which may be relatively far away from the site identified in the GWA study.

A GWA study compares DNA sequence

variant" hypothesis), and the difficulty in finding culprit genes was that these modest effects make the genes very difficult to recognize.

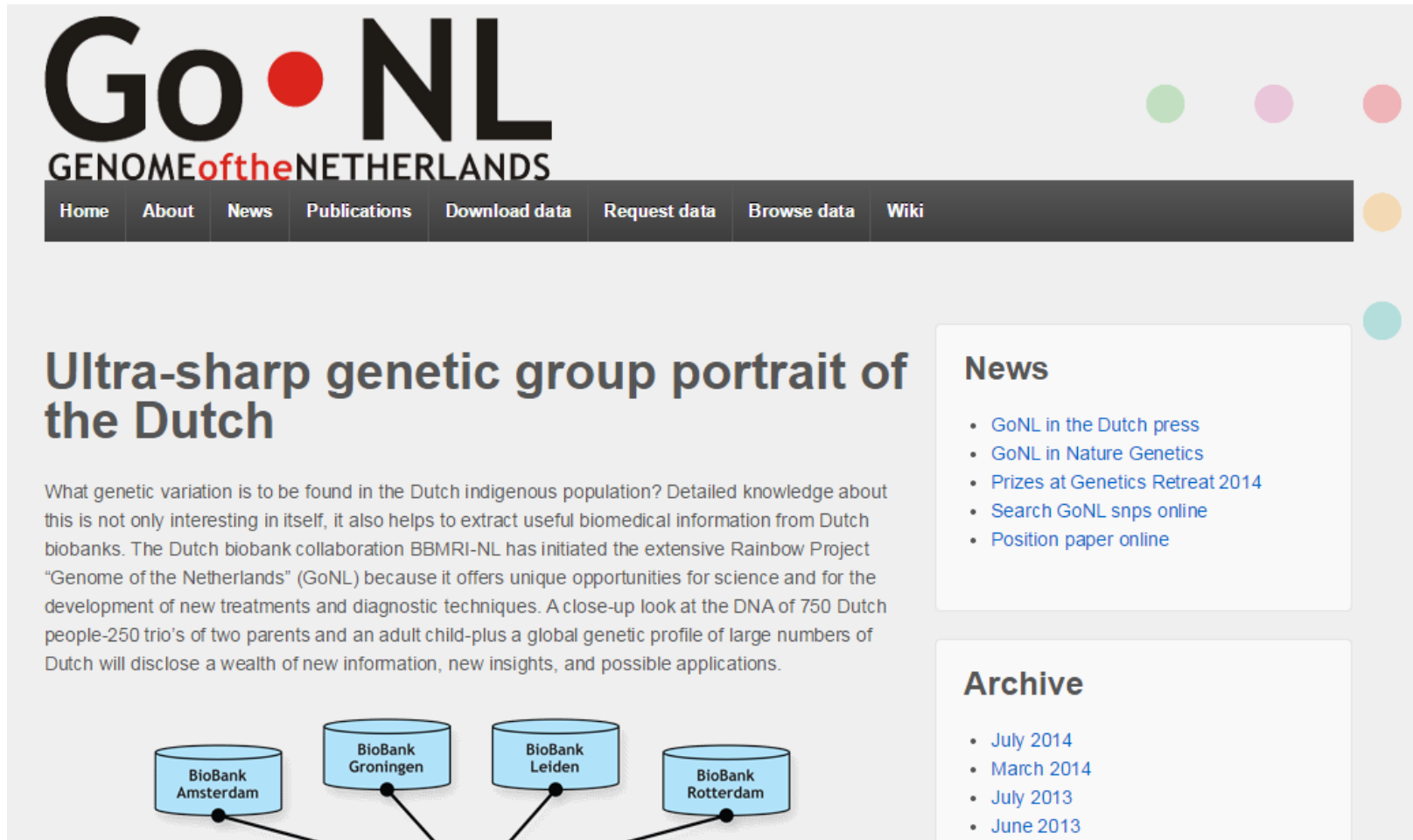
But an alternative explanation is also possible, that the disease is caused by multiple strong-effect variants, each of which is found in only a few people (the "common disease, many rare variants" hypothesis). Instead of the common signpost pointing to a common weak-effect variant, it might be pointing to many strong-effect variants. To distinguish this scenario from the common interpretation, the authors refer to associations between rare higher-impact variants and common markers as "synthetic associations".

In the world of synthetic associations, the





# Other reference panels



The screenshot shows the GoNL (Genome of the Netherlands) website. The header features the GoNL logo with a red dot between 'Go' and 'NL', and the text 'GENOME of the NETHERLANDS' below it. A navigation bar contains links: Home, About, News, Publications, Download data, Request data, Browse data, and Wiki. The main content area has a large heading 'Ultra-sharp genetic group portrait of the Dutch' followed by a paragraph about genetic variation in the Dutch population. Below the text is a diagram showing four biobanks: BioBank Amsterdam, BioBank Groningen, BioBank Leiden, and BioBank Rotterdam, each represented by a blue cylinder and connected by lines. On the right side, there are two sections: 'News' with a list of links and 'Archive' with a list of dates.

## GoNL

GENOME of the NETHERLANDS

Home About News Publications Download data Request data Browse data Wiki

### Ultra-sharp genetic group portrait of the Dutch

What genetic variation is to be found in the Dutch indigenous population? Detailed knowledge about this is not only interesting in itself, it also helps to extract useful biomedical information from Dutch biobanks. The Dutch biobank collaboration BBMRI-NL has initiated the extensive Rainbow Project "Genome of the Netherlands" (GoNL) because it offers unique opportunities for science and for the development of new treatments and diagnostic techniques. A close-up look at the DNA of 750 Dutch people-250 trio's of two parents and an adult child-plus a global genetic profile of large numbers of Dutch will disclose a wealth of new information, new insights, and possible applications.

BioBank Amsterdam BioBank Groningen BioBank Leiden BioBank Rotterdam


#### News

- [GoNL in the Dutch press](#)
- [GoNL in Nature Genetics](#)
- [Prizes at Genetics Retreat 2014](#)
- [Search GoNL snps online](#)
- [Position paper online](#)


#### Archive

- July 2014
- March 2014
- July 2013
- June 2013

# Other reference panels

 Google Cloud Platform


Google Genomics X Search this site

My console 

Why Google Products ▾ Solutions Pricing Customers Documentation Support Partners


Free Trial Contact Sales

Products > Documentation > Google Genomics

Google Genomics  211

What is Google Genomics?


- Pricing and Quotas
- Storing Genomic Data
- Processing Genomic Data
- Exploring Genomic Data
- Sharing Genomic Data
- Developer's Guide
- Genomics API
- Genomics Tools
- Support

 Google Genomics

**Explore genetic variation interactively.** Compare entire cohorts in seconds with SQL-like queries. Compute transition/transversion ratios, genome-wide association, allelic frequency and more.

**Process big genomic data easily.** Run batch analyses like principal component analysis and Hardy-Weinberg equilibrium on as many samples as you like, in minutes or hours, with just a little code.

**Use Google's infrastructure and big data expertise.** Store one genome or a million using Google Genomics and take advantage of the same infrastructure that powers Search, Maps, YouTube, Gmail and Drive.



# Other reference panels



## UK10K

*Rare Genetic Variants in Health and Disease (2010-2013)*

### What is UK10K?

The UK10K project will enable researchers in the UK and beyond to better understand the link between low-frequency and rare genetic changes, and human disease caused by harmful changes to the proteins the body makes.

Although many hundreds of genes that are involved in causing disease have already been identified, it is believed that many more remain to be discovered. The UK10K project aims to help uncover them by studying the genetic code of 10,000 people in much finer detail than ever before.



Wellcome Library,  
London

### Project Design

Not all genetic changes are harmful or lead to disease, so the project is taking a two-pronged approach to identify rare variants and their effects:

- by studying and comparing the DNA of 4,000 people whose physical characteristics are well documented, the project aims to identify those changes that have no discernible effect and those that may be linked to a particular disease;



HOME

GOALS

RESEARCHERS

STUDY SAMPLES

DATA ACCESS

DATA & METHODS

ETHICS

FUNDING

VACANCIES

CONSORTIUM

PUBLICATIONS

POSTERS

CONTACT US

LOGIN

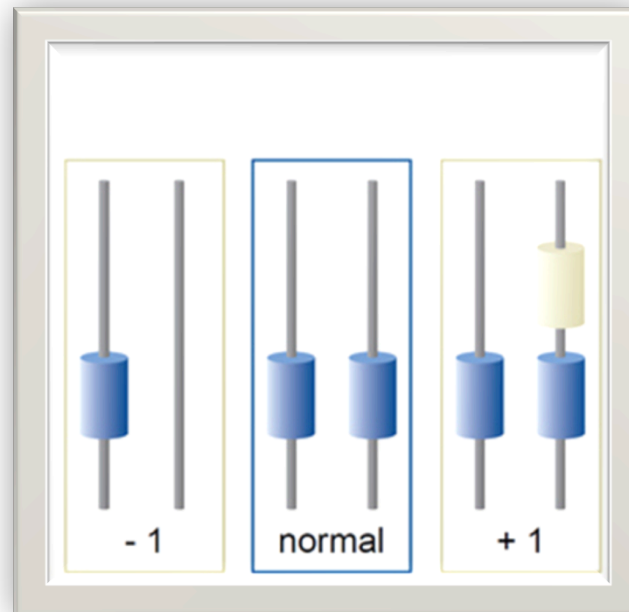
PENNSTATE



<http://www.uk10k.org/>

# I'm interested in other variation besides SNPs...

Perhaps we should look at CNVs



# CNVs are Associated with Various Phenotypes

## Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans

Timothy J. Aitman<sup>1</sup>, Rong Dong<sup>1\*</sup>, Timothy J. Vyse<sup>2\*</sup>, Penny J. Norsworthy<sup>1\*</sup>, Michelle D. Johnson<sup>1</sup>, Jennifer Smith<sup>1</sup>, Jonathan Mangion<sup>1</sup>, Cheri Robertson-Lowe<sup>1,2</sup>, Amy J. Marshall<sup>1</sup>, Enrico Petretto<sup>1</sup>, Matthew D. Hodges<sup>1</sup>, Gurjeet Bhargava<sup>3</sup>, Sheetal G. Patel<sup>1</sup>, Kelly Sheehan-Rooney<sup>1</sup>, Mark Duda<sup>1,2</sup>, Paul R. Cook<sup>1,2</sup>, David J. Evans<sup>5</sup>, Jan Domin<sup>5</sup>, Jonathan Flint<sup>4</sup>, Joseph J. Boyle<sup>5</sup>, Charles D. Pusey<sup>5</sup> & H. Terence Cook<sup>5</sup> [Nature](#) 2006

## The Influence of *CCL3L1* Gene—Containing Segmental Duplications on HIV-1/AIDS Susceptibility

Enrique Gonzalez,<sup>1\*</sup> Hemant Kulkarni,<sup>1\*</sup> Hector Bolivar,<sup>1\*</sup> Andrea Mangano,<sup>2\*</sup> Racquel Sanchez,<sup>1,†</sup> Gabriel Catano,<sup>1,†</sup> Robert J. Nibbs,<sup>3,†</sup> Barry I. Freedman,<sup>4,†</sup> Marlon P. Quinones,<sup>1,†</sup> Michael J. Bamshad,<sup>5</sup> Krishna K. Murthy,<sup>6</sup> Brad H. Rovin,<sup>7</sup> William Bradley,<sup>8,9</sup> Robert A. Clark,<sup>1</sup> Stephanie A. Anderson,<sup>8,9</sup> Robert J. O'Connell,<sup>8,10</sup> Brian K. Agan,<sup>8,10</sup> Seema S. Ahuja,<sup>1</sup> Rosa Bologna,<sup>11</sup> Luisa Sen,<sup>2</sup> 2005 VOL 307 SCIENCE  
Matthew J. Dolan,<sup>8,10,12</sup> Sunil K. Ahuja<sup>1,‡</sup>

## Strong Association of De Novo Copy Number Mutations with Autism

Jonathan Sebat,<sup>1,\*</sup> B. Lakshmi,<sup>1</sup> Dheeraj Malhotra,<sup>1,\*</sup> Jennifer Troge,<sup>1,\*</sup> Christa Lese-Martin,<sup>2</sup> Tom Walsh,<sup>3</sup> Boris Yamrom,<sup>1</sup> Seungtae Yoon,<sup>1</sup> Alex Krasnitz,<sup>1</sup> Jude Kendall,<sup>3</sup> Anthony Leotta,<sup>1</sup> Deepa Pai,<sup>1</sup> Ray Zhang,<sup>1</sup> Yoon-Ha Lee,<sup>3</sup> James Hicks,<sup>4</sup> Sarah J. Spence,<sup>4</sup> Annette T. Lee,<sup>3</sup> Kaija Puura,<sup>6</sup> Terho Lehtimäki,<sup>7</sup> David Ledbetter,<sup>2</sup> Peter K. Gregersen,<sup>5</sup> Joel Bregman,<sup>8</sup> James S. Sutcliffe,<sup>9</sup> Valdehi Jobanputra,<sup>10</sup> Wendy Chung,<sup>10</sup> Dorothy Warburton,<sup>13</sup> Mary-Claire King,<sup>2</sup> David Skuse,<sup>11</sup> Daniel H. Geschwind,<sup>12</sup> T. Conrad Gilliam,<sup>13</sup> Kenny Ye,<sup>14</sup> Michael Wigler<sup>1</sup>

SCIENCE VOL 316 20 APRIL 2007

## A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon

Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome

Andrew J. Sharp<sup>1</sup>, Sierra Hansen<sup>1</sup>, Rebecca R. Selzer<sup>2</sup>, Ze Cheng<sup>1</sup>, Regina Regan<sup>3</sup>, Jane A. Hurst<sup>4</sup>, Helen Stewart<sup>4</sup>, Sue M. Price<sup>4</sup>, Edward Blair<sup>4</sup>, Rooul C. Henninkam<sup>5,6</sup>, Carrie A. Fitzpatrick<sup>7</sup>, Rick Sepraves<sup>8</sup>, Todd A. Richmond<sup>9</sup>, Cheryl Guiver<sup>9</sup>, Donna G. Albertson<sup>8,9</sup>, Daniel Pinkel<sup>8</sup>, Peggy S. Eis<sup>9</sup>, Stuart Schwartz<sup>7</sup>, Samantha J. L. Knight<sup>9</sup> & Ivan E. Eichler<sup>1</sup>

## Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia

Tom Walsh,<sup>1</sup> Jon M. McClellan,<sup>2,†</sup> Shane E. McCarthy,<sup>1,\*</sup> Anjene M. Addington,<sup>4</sup> Sarah D. Pierce,<sup>1</sup> Greg M. Cooper,<sup>1</sup> Alex J. Nord,<sup>3</sup> Mary Kuzenda,<sup>3,4</sup> Dheeraj Malhotra,<sup>4</sup> Abhishek Bhandari,<sup>1</sup> Sundeep M. Stray,<sup>1</sup> Caitlin F. Rippey,<sup>3</sup> Patricia Roccavara,<sup>5</sup> Vlad Makarov,<sup>6</sup> R. Lakshmi,<sup>4</sup> Robert L. Fireling,<sup>7</sup> Lianmarie Sikich,<sup>8</sup> Thomas Stromberg,<sup>9</sup> Barry Morrison,<sup>9</sup> Nitin Garg,<sup>9</sup> Philip Reilly,<sup>9</sup> Kristen Fichtner,<sup>9</sup> Laila Maay,<sup>9</sup> Peter Fischman,<sup>9</sup> Robert Long,<sup>9</sup> Pagen Chen,<sup>9</sup> Sean Davis,<sup>10</sup> Carl Baker,<sup>9</sup> Evan F. Fichler,<sup>9</sup> Paul S. Meltzer,<sup>10</sup> Stanley F. Nelson,<sup>9</sup> Andrew B. Stenington,<sup>11</sup> Ming K. Lee,<sup>1</sup> Judith L. Rapoport,<sup>4</sup> Mary-Claire King,<sup>3,5</sup> Jonathan Sebat<sup>1</sup>

## ARTICLES

## Origins and functional impact of copy number variation in the human genome

Donald F. Conrad<sup>1\*</sup>, Dalila Pinto<sup>2\*</sup>, Richard Redon<sup>1,3</sup>, Lars Feuk<sup>2,4</sup>, Omer Gokcumen<sup>5</sup>, Yujun Zhang<sup>1</sup>, Jan Aerts<sup>1</sup>, T. Daniel Andrews<sup>1</sup>, Chris Barnes<sup>1</sup>, Peter Campbell<sup>1</sup>, Tomas Fitzgerald<sup>1</sup>, Min Hu<sup>1</sup>, Chun Hwa Ihm<sup>5</sup>, Kati Kristiansson<sup>1</sup>, Daniel G. MacArthur<sup>1</sup>, Jeffrey R. MacDonald<sup>2</sup>, Ifejinelo Onyiah<sup>1</sup>, Andy Wing Chun Pang<sup>2</sup>, Sam Robson<sup>1</sup>, Kathy Stirrups<sup>1</sup>, Armand Valsesia<sup>1</sup>, Klaudia Walter<sup>1</sup>, John Wei<sup>2</sup>, Wellcome Trust Case Control Consortium†, Chris Tyler-Smith<sup>1</sup>, Nigel P. Carter<sup>1</sup>, Charles Lee<sup>5</sup>, Stephen W. Scherer<sup>2,6</sup> & Matthew E. Hurles<sup>1</sup>

## ARTICLES

## Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls

*The Wellcome Trust Case Control Consortium\**

# “Take-Home” Quotes from CNV Paper

- “We have demonstrated that high-confidence CNV calls can be assigned in large, real-world case-control samples for a substantial proportion of the common CNVs estimated to be present in the human genome.
- We have identified directly several CNV loci that are associated with common disease. Such loci could contribute to disease pathogenesis.
- However, the loci identified **are well tagged by SNPs and, hence, the associations can be, and were, detected indirectly via SNP association studies.**
- Among the CNVs that we could type well, those not well tagged by SNPs have the same overall association properties as those which are well tagged.”

**GWAS is fine, and CNVs are cool,  
but I want to detect ALL variation  
in my samples!**

- SEQUENCE is the SOLUTION



# Questions???

