PUBH 8445: Lecture 1

Saonli Basu, Ph.D.

Division of Biostatistics
School of Public Health
University of Minnesota
saonli@umn.edu

Statistical Genetics

- It can broadly be classified into three sub categories:
  - **Mendelian Genetics:** studies the transmission of alleles in pedigrees.
  - **Population Genetics:** the rules of how genes behave in population.
  - **Quantitative Genetics:** the rules of transmission of complex quantitative traits, those with both a genetic and environmental basis.

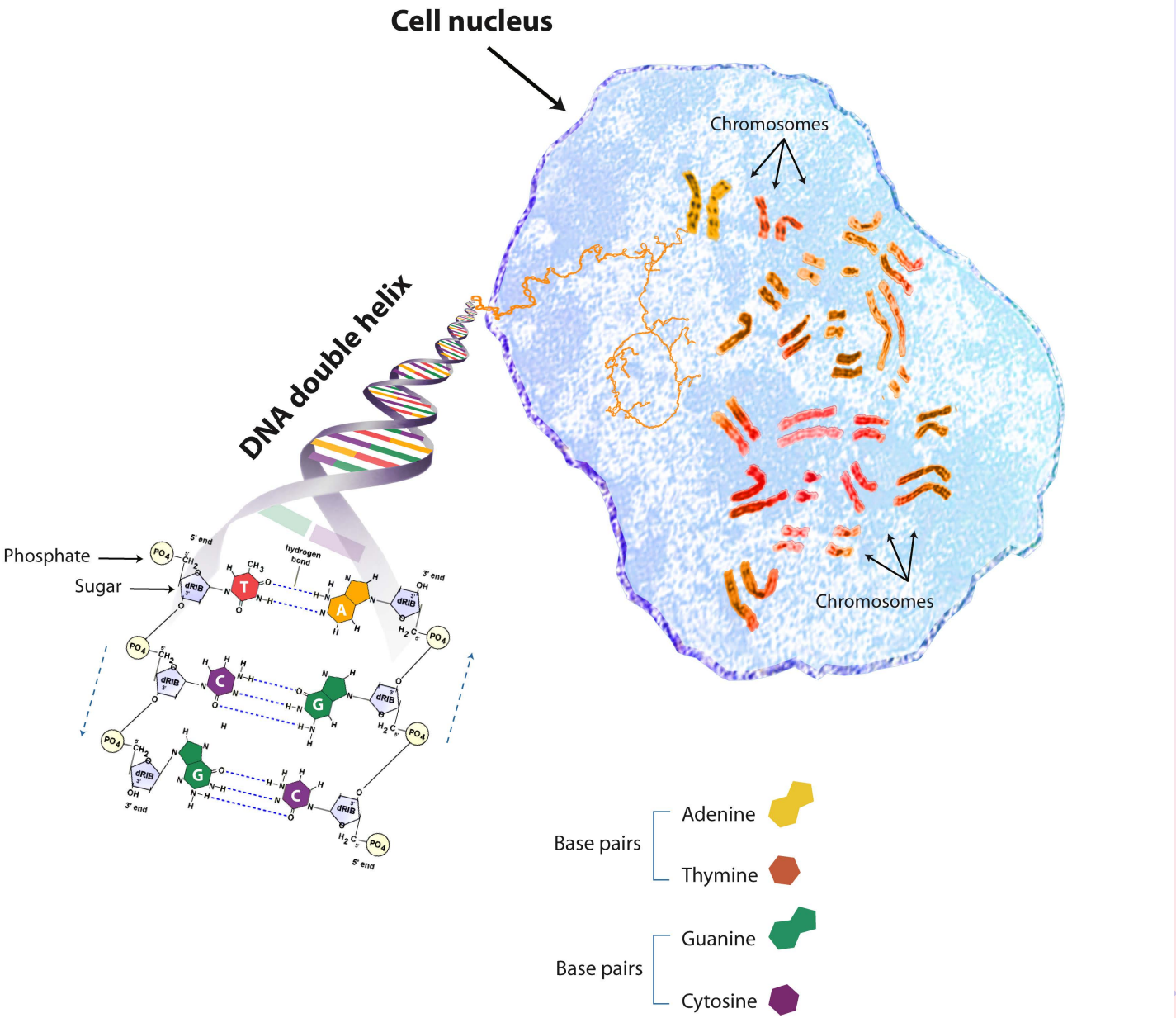
Journals useful for our course

- Nature Genetics
- American Journal of Human Genetics
- Frontier in Genetics
- Genetic Epidemiology
- PLoS Genetics
- **Nature Reviews Genetics** (for example: `http://www.nature.com/nrg/focus/stats/index.html`)

Genetic Terminologies

- DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called mitochondrial DNA or mtDNA).
- The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism.
- DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix.

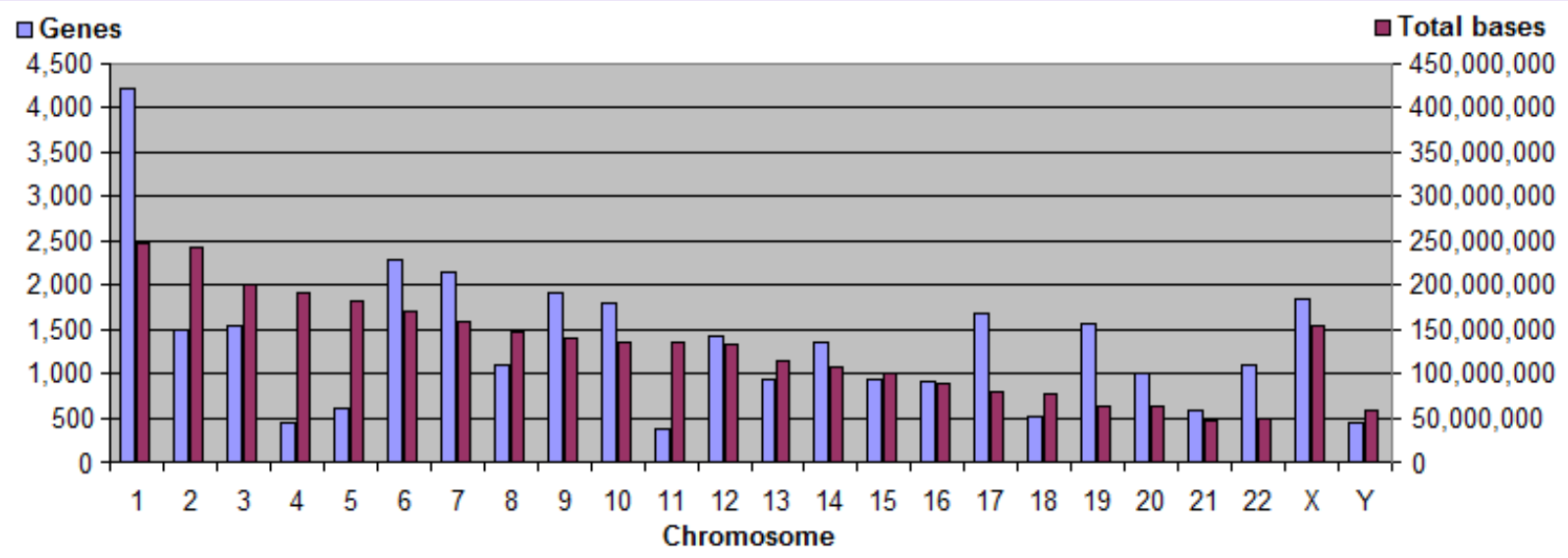
Genetic Terminologies



Terminologies(contd)

- **Chromosome:** The entire genome (complete set of nuclear DNA) is arranged in pairs of chromosomes.
- There are 22 autosomes and 2 sex chromosomes.
- For every pair of chromosomes, one is inherited from the mother of an individual and one is inherited from the father of an individual.
- Chromosomes that are of the same pair and carry the same set of genes and are called homologous.

Terminologies(contd)



terminologies(contd)

- **Locus:** Each position of the genome is called a "locus" ("loci" for multiple locations). A locus could represent a single base position or a collection of bases.
- **Allele:** The variations observed in the human population at a locus are called the "alleles" for that locus. If the locus represents a single base, there could be at most two variations. This type of locus is called "Single Nucleotide Polymorphism" (SNP). There are markers called microsattelites (Short Tandem Repeats: GTAGTAGTAGTAGTA...)
- **Gene:** A gene is the basic physical and functional unit of heredity. Genes, which are made up of DNA, act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. The Human Genome Project has estimated that humans have about 20,000 genes.

What is Heritability?

- **Reference:** Tenesa A and Haley CS (2013), The heritability of human disease: estimation, uses and abuses, Nature Reviews Genetics 14: 139-149
- The proportion of the phenotypic variance attributable to genetic differences is termed the heritability.
- We need to be aware of two statistics: the narrow-sense heritability (h^2), which refers to differences among the additive genetic values, and the broad-sense heritability (H^2), which refers to genetic differences as differences between genotypic values.

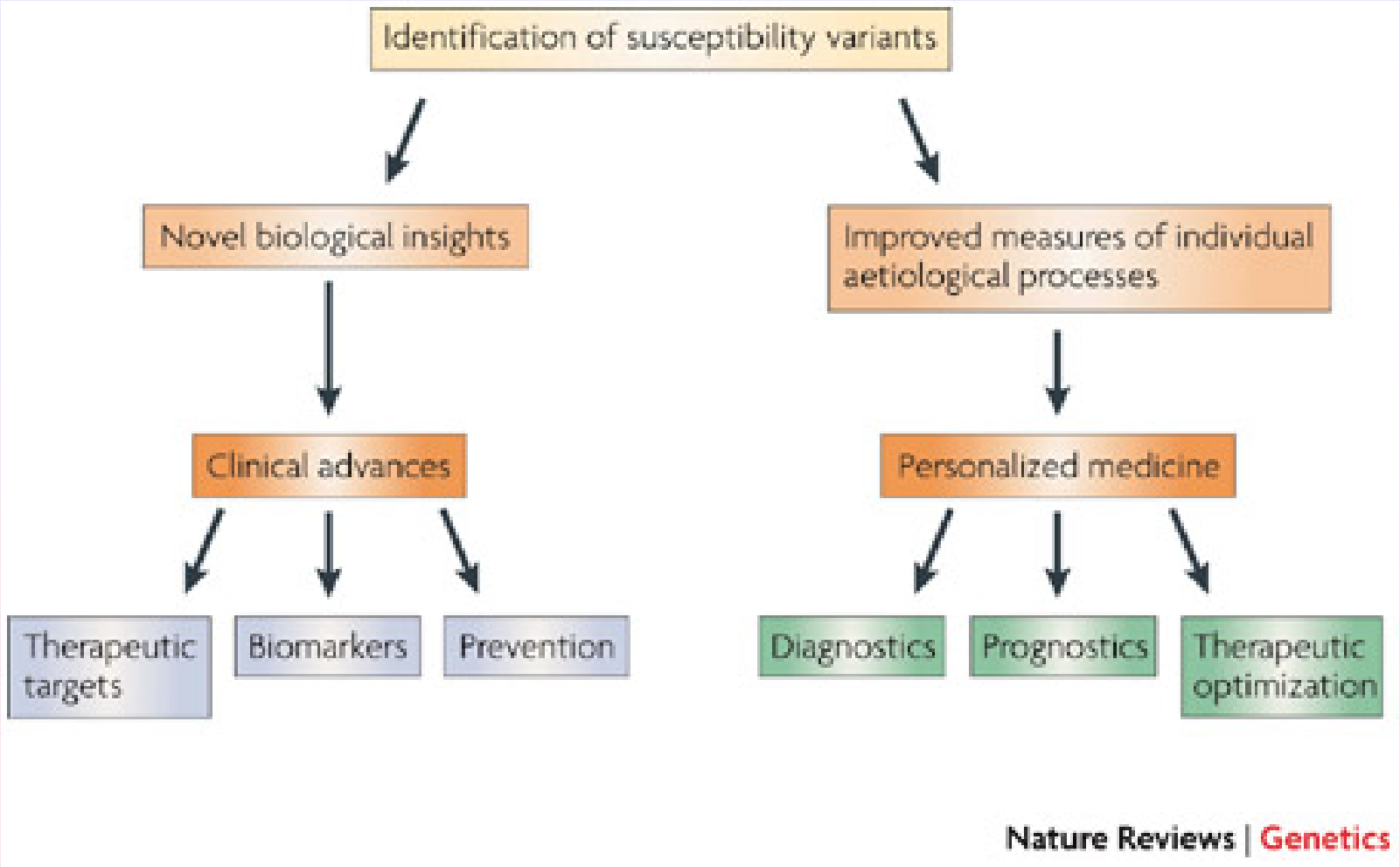
Examples of heritable Traits

- http://www.snpedia.com/index.php/Heritability
- http://tga.nig.ac.jp/h2db/
- Polderman et al. (2015), Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet 47(7):702-9.

Motivation

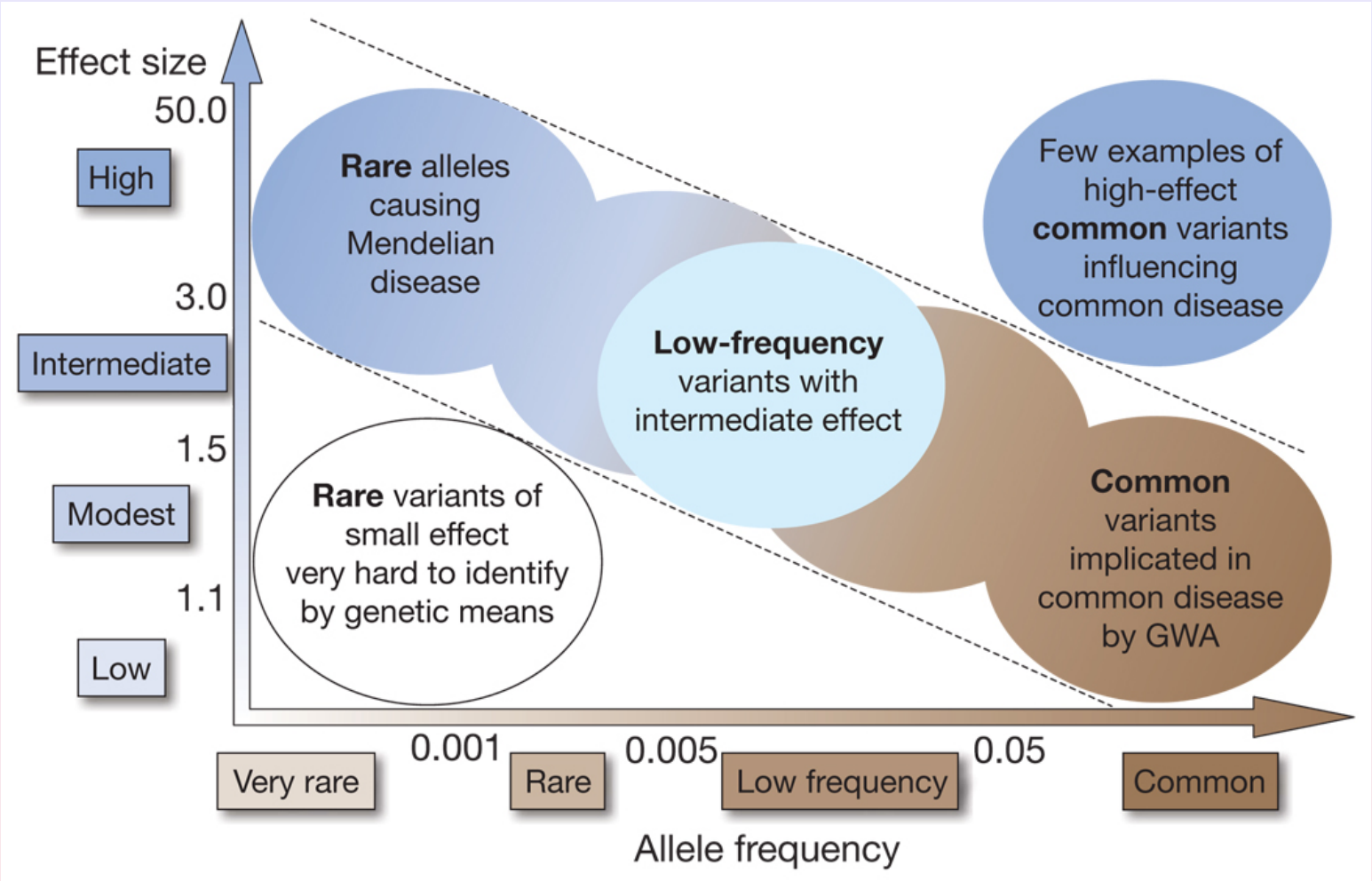
- Population Genetics: Understand evolution and speciation.
- Quantitative Genetics: Improve plant and animal breeding. How can we increase agricultural production?
- Genetic Epidemiology: Understand and Control human diseases.

Motivation



Source: McCarthy et al. (2008), Genome-wide association studies for complex traits: consensus, uncertainty and challenges, Nature Reviews Genetics 9, 356-369.

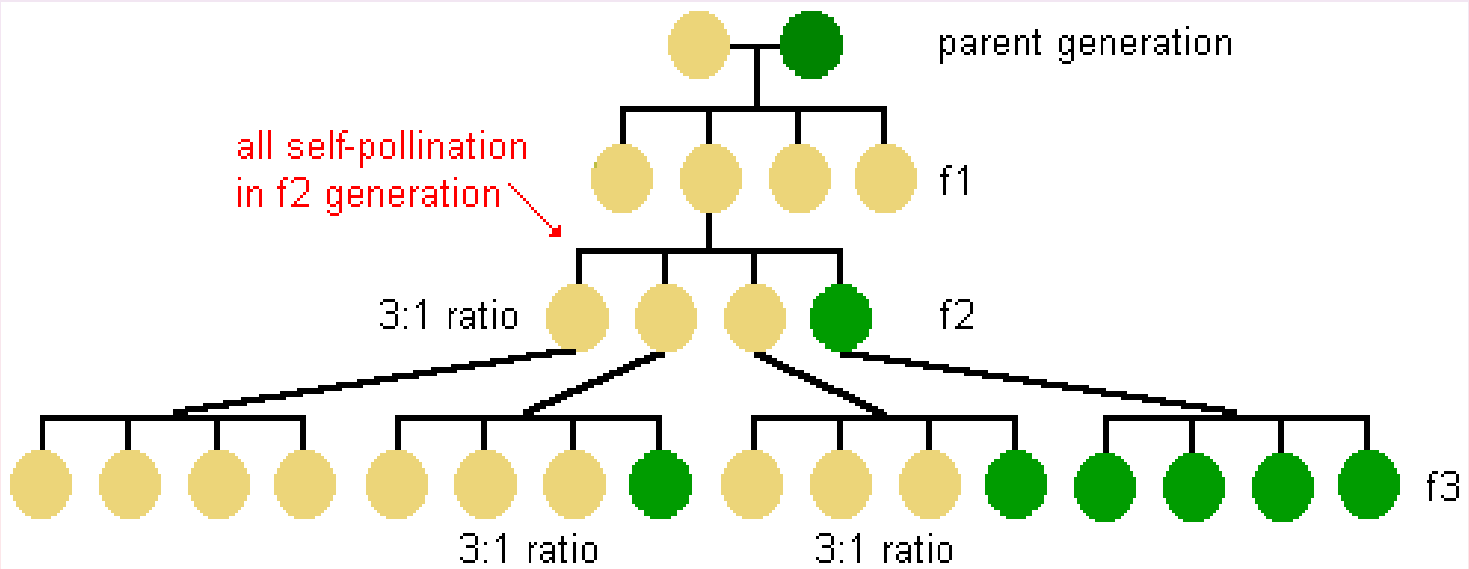
Genetic architecture



Source: Manolio et al. (2009), Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

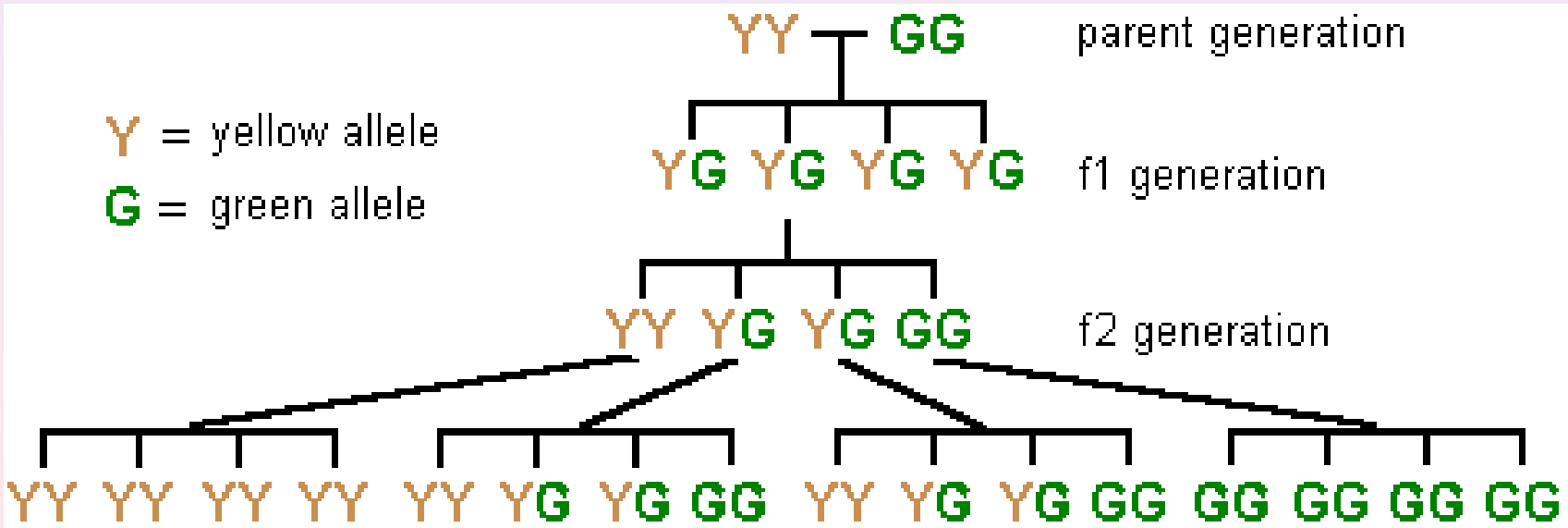
Mendelian Inheritance

- In cross-pollinating plants that either produce yellow or green pea seeds exclusively, Mendel found that the first offspring generation (f1) always has yellow seeds. However, the following generation (f2) consistently has a 3:1 ratio of yellow to green.
- This 3:1 ratio occurs in later generations as well.



Mendelian Inheritance

- **Law of Segregation:** for any particular trait, the pair of alleles of each parent separate and only one allele passes from each parent on to an offspring. Which allele in a parent's pair of alleles is inherited is a matter of chance. We now know that this segregation of alleles occurs during the process of **meiosis**.



Terminologies(contd)

- **Genotype:** its specific combination of alleles for a given locus/gene. Going back to Mendel's plants, we can now say that all of his true-breeding plants contained two of the same alleles for each of the observed genes. Yellow plants in this P generation had two alleles foryellow color (YY), and green P generation plants had two alleles for green color (GG). Anytime an organisms two alleles for a specific trait are identical, that the individual is said to be homozygous for that trait.
- On the other hand, crossing the two color plants to produce F1 hybrids created a generation of plants with one Y allele and one G allele (YG). An organism with two opposing alleles for a single gene is said to be heterozygous for that trait.

terminologies(contd)

- **Phenotype:** The genetic makeup of a certain trait (e.g., YY, YG and GG) is called its genotype, while the physical expression of these traits (e.g., yellow or green) is called a phenotype.
    - Dominant/Recessive: For the pea plants, if the Y allele is **dominant** and the G allele is**recessive**, only two phenotypes are possible. Both the plants with YG and YY genotypes will have the yellow color phenotype, while the plants with the GG genotype will have the green color phenotype. A **trait** is the general aspect of physiology being shown in the phenotype. So, for example, the trait here is the pea seed-color of the pea plant. The phenotype can be either yellow or green color, depending on the genotype.

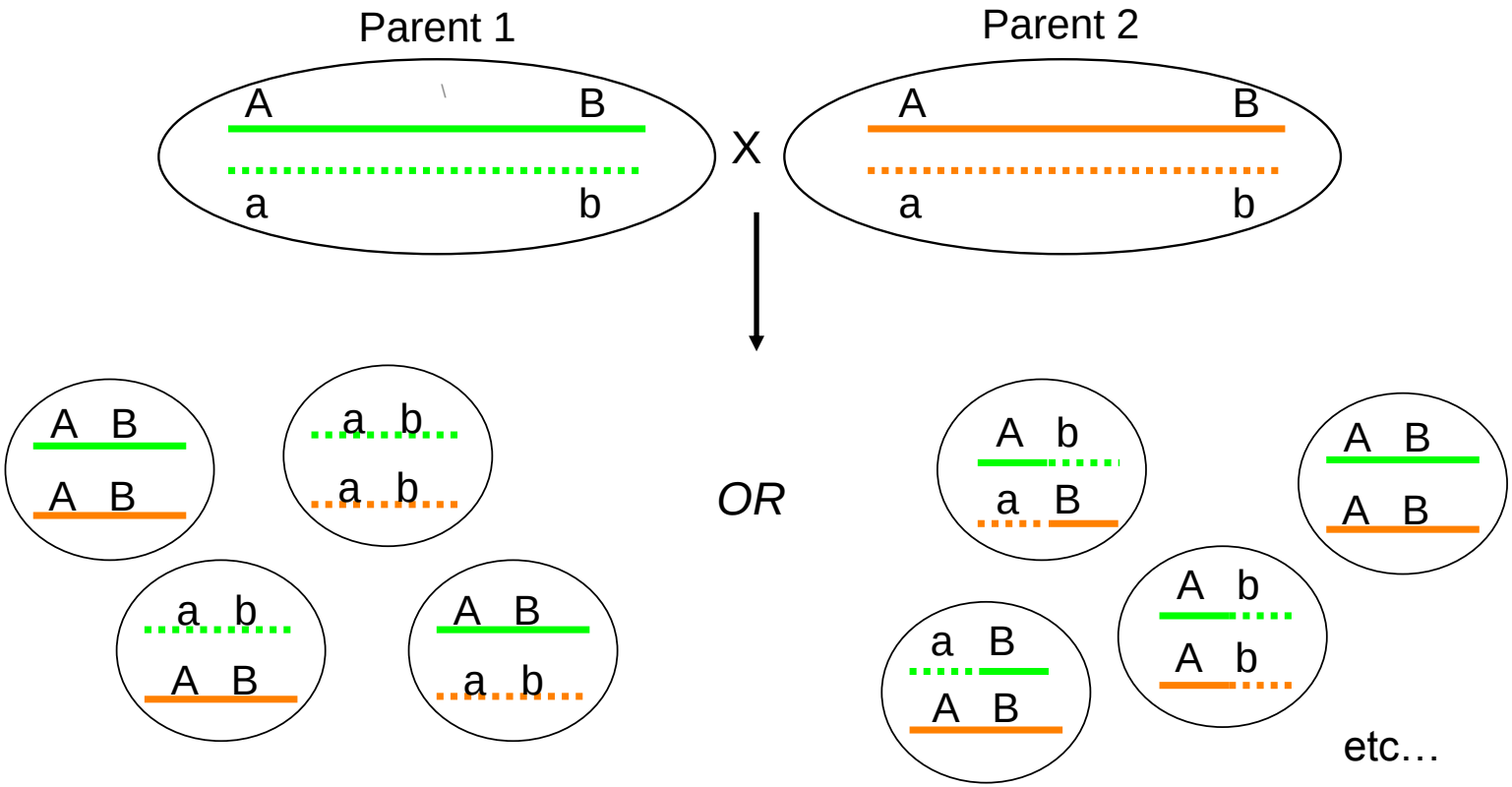
Example: Genotype vs Phenotype

- The ABO locus is on chromosome 9
- The (main) alleles at the locus are A, B, and O.
- The 6 genotypes are AA, AO, BB, BO, AB and OO
- Homozygotes are AA,BB,OO; Heterozygotes are AO,BO and AB.
- The 4 phenotypes are blood types A, B, AB and O
- O allele is recessive to A and to B; A and B are each dominant to O
- AO and AA are blood type A; BB and BO are blood type B.
- A and B are codominant: AA, AB and BB are distinguishable.

Mendel's Principles of Genetic Inheritance

- **Law of Independent Assortment:** In the gametes, alleles of one gene separate independently of those of another gene, and thus all possible combinations of alleles are equally probable.
- **Law of Dominance:** Each trait is determined by two factors (alleles), inherited one from each parent. These factors each exhibit a characteristic dominant, co-dominant, or recessive expression, and those that are dominant will mask the expression of those that are recessive.

Basic Concepts



High LD -> No Recombination
(r^2 = 1) SNP1 “tags” SNP2

Low LD -> Recombination
Many possibilities

Quantitative Genetics

- Quantitative genetics is the study of these polygenic traitsQuantitative genetic variation can be described in three ways:
  - Traits are influenced by multiple genes, i.e. theyre polygenic.
  - They are usually influence more easily by environmental factors than simple Mendelian traits.
  - Both of the factors above usually lead to a continuous distribution of the particular trait. For example, you can see the near normal distribution when comparing a sample population by their height.

Steps in Positional Cloning

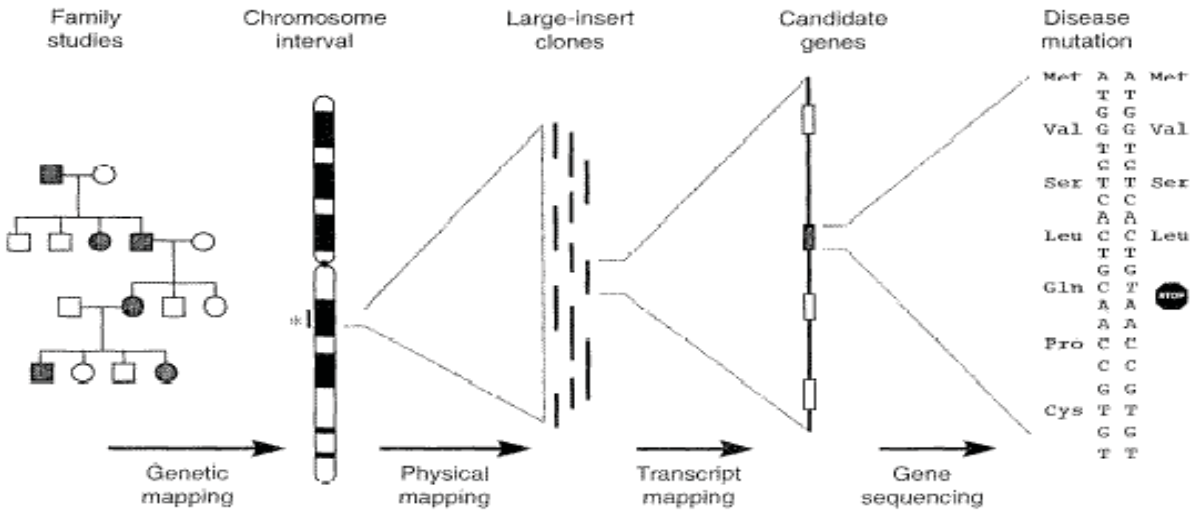


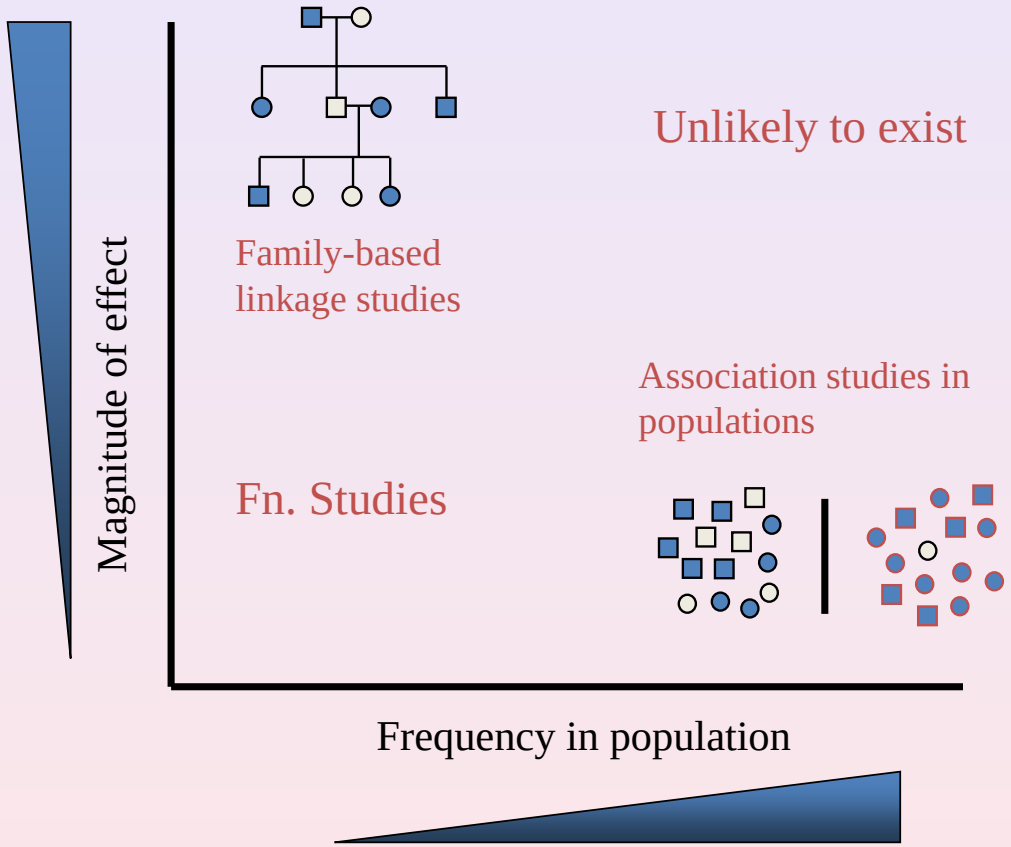
Fig. 1. Steps in positional cloning. Positioning of disease loci to chromosomal regions with genetic markers has become increasingly straightforward, particularly given the recent release of the Généthon genetic map containing 5264 markers (*17*). However, identification and evaluation of the genes within the implicated region remains a major stumbling block.

Schuler (1996) *Science*

Broad Genetic Epidemiology Study Design Categories:

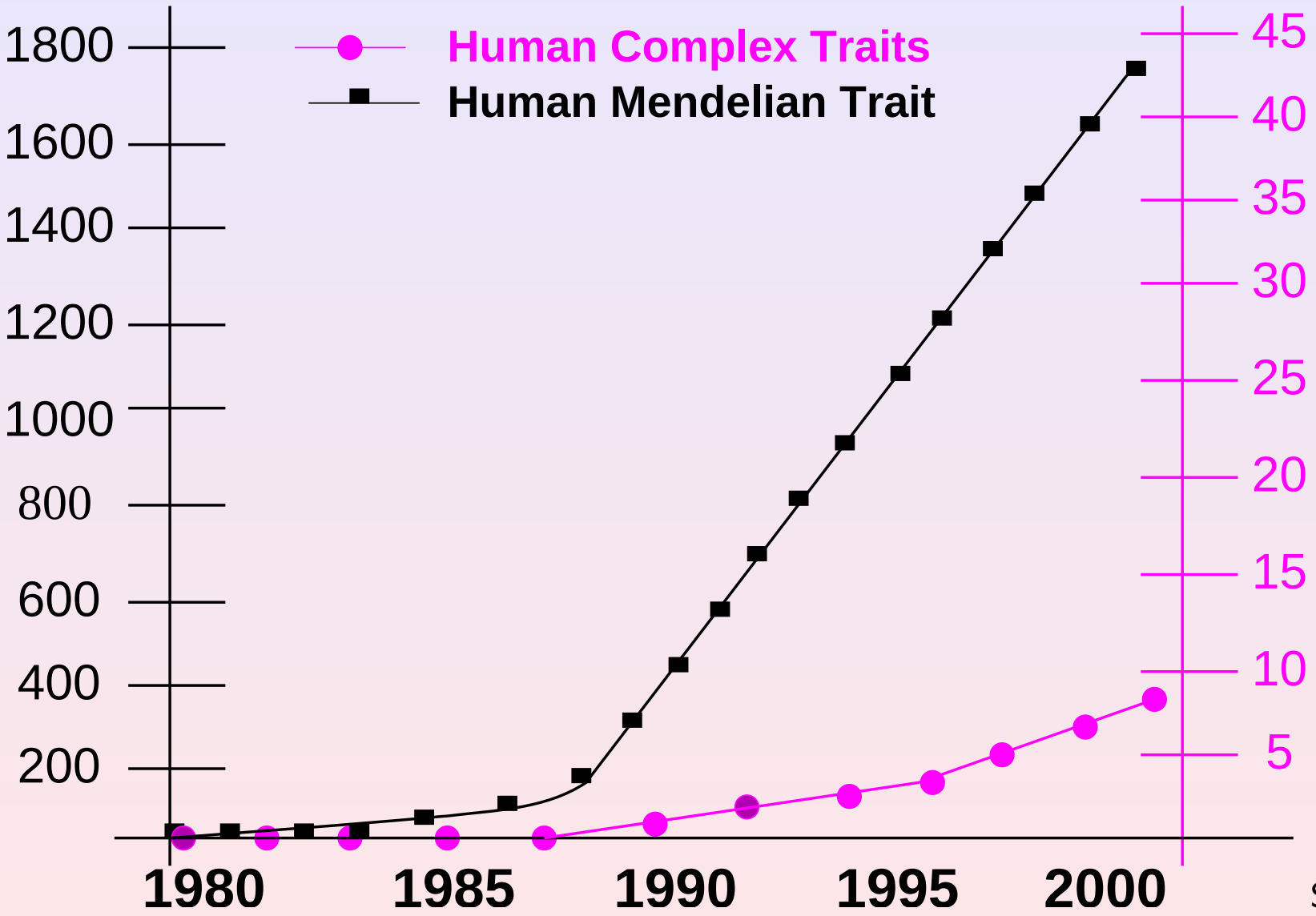
- Linkage Analysis
  Follows meiotic events through families for co-segregation of disease and particular genetic variants
  - Large Families
  - Sibling Pairs (or other family pairs)
  - Works VERY well for "Mendelian" diseases
- Association Studies Detect association between genetic variants and disease across families: exploits linkage disequilibrium.
  - Case-Control designs
  - Cohort designs
  - Parents with affected child trios (TDT)
  - May be more appropriate for complex diseases

Allelic architecture and mapping strategy



Slide thanks to D. Altshuler

Gene Discovery



Source:
Glazier AM, Nadeau JH, Aitman TJ (2002), Finding genes that underlie complex traits. Science 298(5602):2345-9

Missing Heritability

- Over estimation of Heriitability
- Rare variants with big effects
- structural variation
- gene-gene and gene-environment interaction

Missing Heritability

Source: Manolio et al. (2009), Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

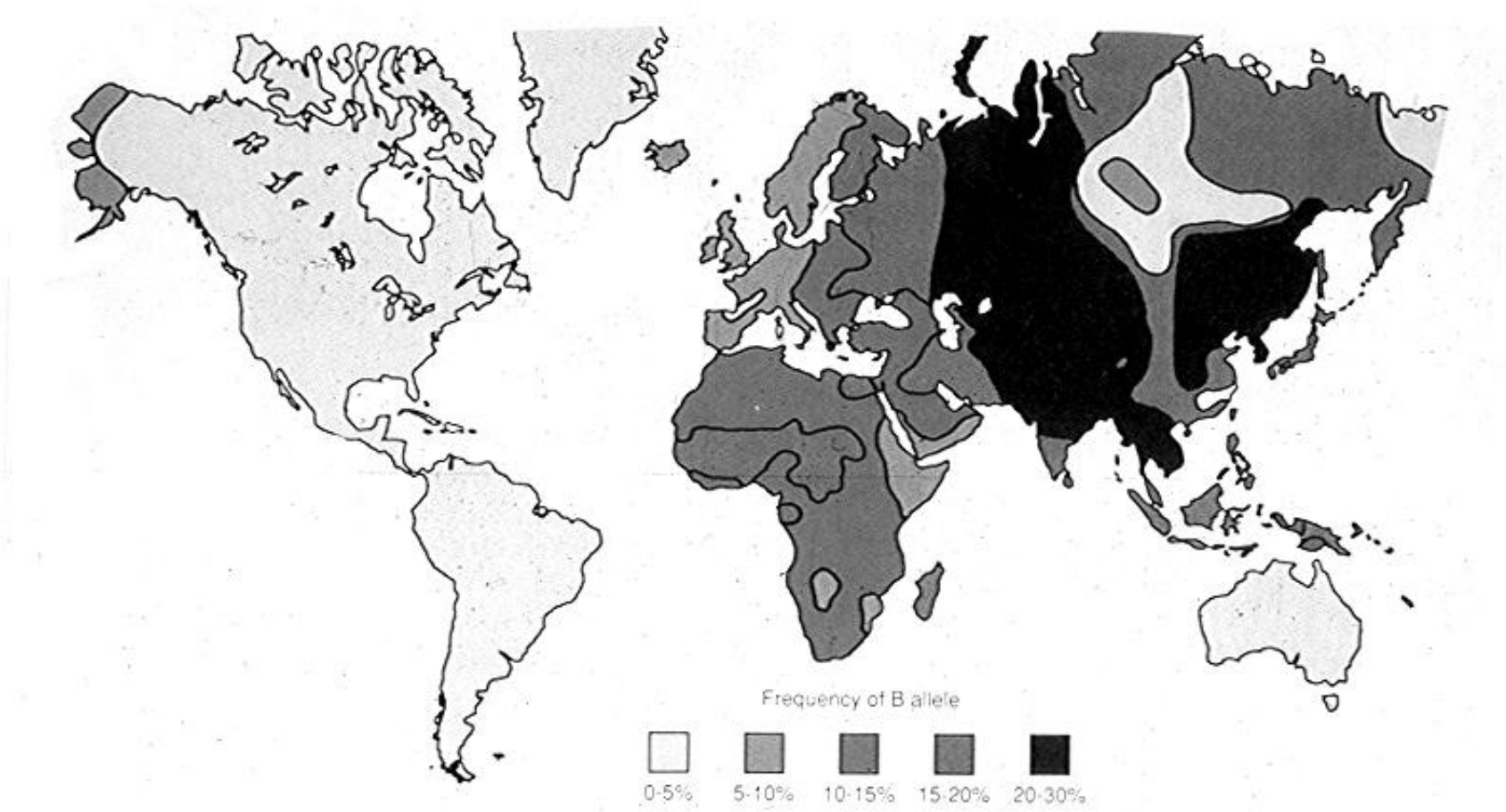
Disease	No. loci	heritability explained	Heritability measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn's disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.20%	Residual*phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.80%	Phenotypic variance
Fasting glucose	4	1.50%	Phenotypic variance

Missing Heritability

Source: Manolio et al. (2009), Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

Disease	Locus	Size (kb)	Pop freq	Case frequency	Effect size (OR)
IMR					
Rare CNVs					
Autism/IMR	16p11.2	600	1x10-4	1%	100
Autism	16p11.2	600	3x10-4	0.50%	16
Schizophrenia	1q21.1	1,400	2x10-4	0.30%	15
IMR	1q21.1	1,400	2x10-4	0.47%	NP in controls
Schizophrenia	15q13.3	1,600	2x10-4	0.20%	12
Epilepsy	15q13.3	1,600	2x10-4	1.00%	NP in controls
IMR	15q13.3	1,600	2x10-4	0.30%	NP in controls
Schizophrenia	22q11.2	3,000	2.5x10-4	1%	40
Common CNPs					
Crohn's disease	IRGM	20	7%	11%	1.5
Body mass index	NEGR1	45	65%	Quantitative	<1
Psoriasis	LCE3C	30	55%	65%	1.3

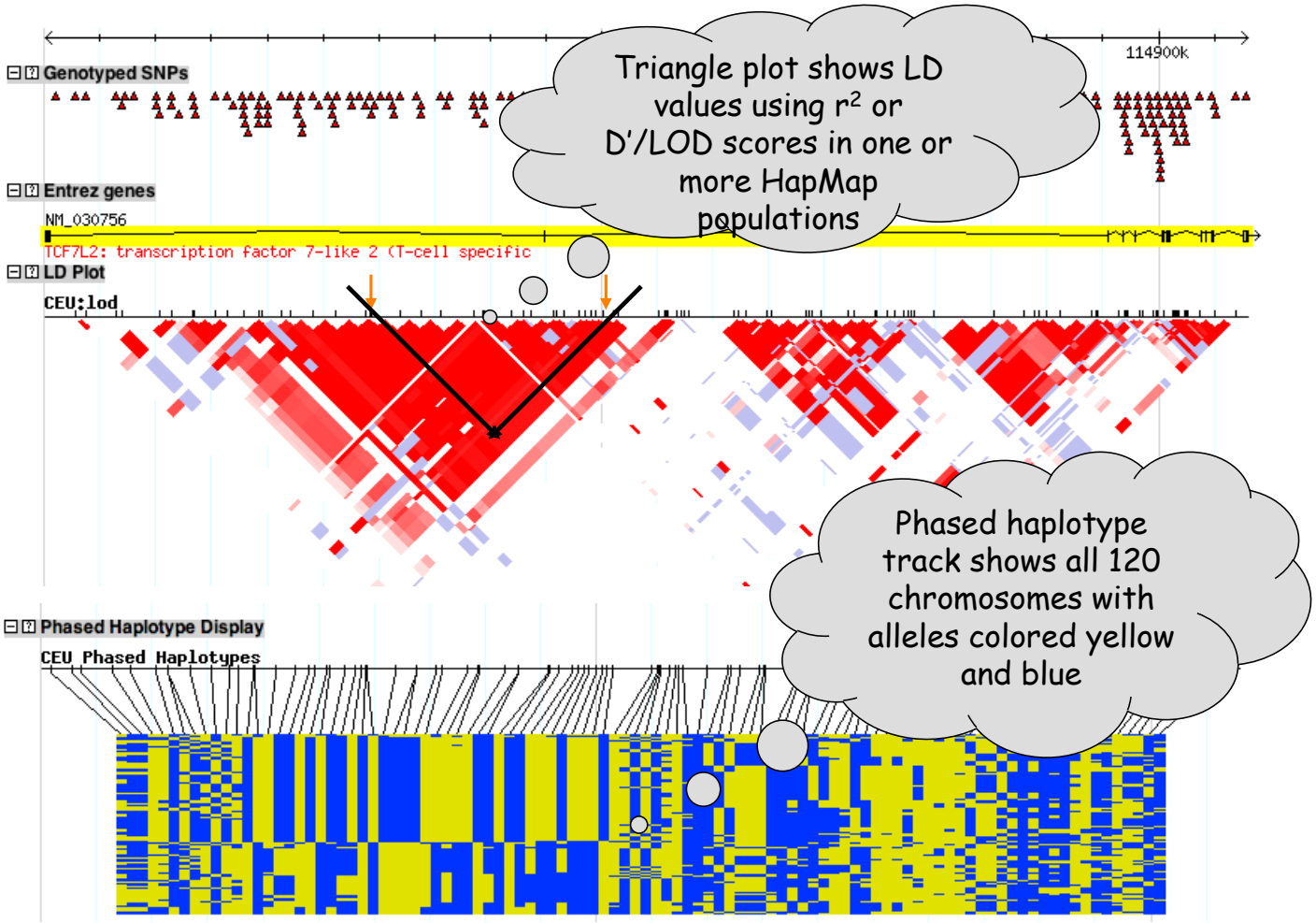
World-wide distribution of the I^B (ABO) allele



Hapmap Project

- Scientists thought the mutations that caused common diseases would themselves be common.
- They first identified the common mutations in the human population in a $100 million project called the HapMap. Then they compared patient's genomes with those of healthy genomes. The comparisons relied on ingenious devices called SNP chips, which scan just a tiny portion of the genome.
- These projects, called genome-wide association studies (GWAS), each cost several millions.
- The results of this costly international exercise have been disappointing. About 2,000 sites on the human genome have been statistically linked with various diseases, but in many cases the sites are not inside working genes, suggesting there may be some conceptual flaw in the statistics.

View variation patterns



1000 genome project

Disease Cause Is Pinpointed With Genome

By NICHOLAS WADE
Published: March 10, 2010

Two research teams have independently decoded the entire genome of patients to find the exact genetic cause of their diseases. The approach may offer a new start in the so far disappointing effort to identify the genetic roots of major killers like heart disease, diabetes and Alzheimer's.

Enlarge This Image

Michael Stravato for The New York Times

Dr. James R. Lupski, a medical geneticist with a nerve disease, had his whole genome decoded.

Multimedia

In the decade since the first full genetic code of a human was sequenced for some $500 million, less than a dozen genomes had been decoded, all of healthy people.

Geneticists said the new research showed it was now possible to sequence the entire genome of a patient at reasonable cost and with sufficient accuracy to be of practical use to medical researchers. One subject's genome cost just $50,000 to decode.

"We are finally about to turn the corner, and I suspect that in the next few years human genetics will finally begin to systematically deliver clinically meaningful findings," said