ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Μαριάνθη Γεωργίτση

Επίκουρη Καθηγήτρια Ιατρικής Βιολογίας-Ιατρικής Γενετικής

Α' Εργαστήριο Ιατρικής Βιολογίας-Γενετικής

Τομέας Βιολογικών Επιστημών και Προληπτικής Ιατρικής

ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

Σχολή Επιστημών Υγείας

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

E-mail: margeorgitsi@auth.gr

Application of Population Genetics in the study of complex disorders in humans

14.01.2021

We are quite similar, but we are different…

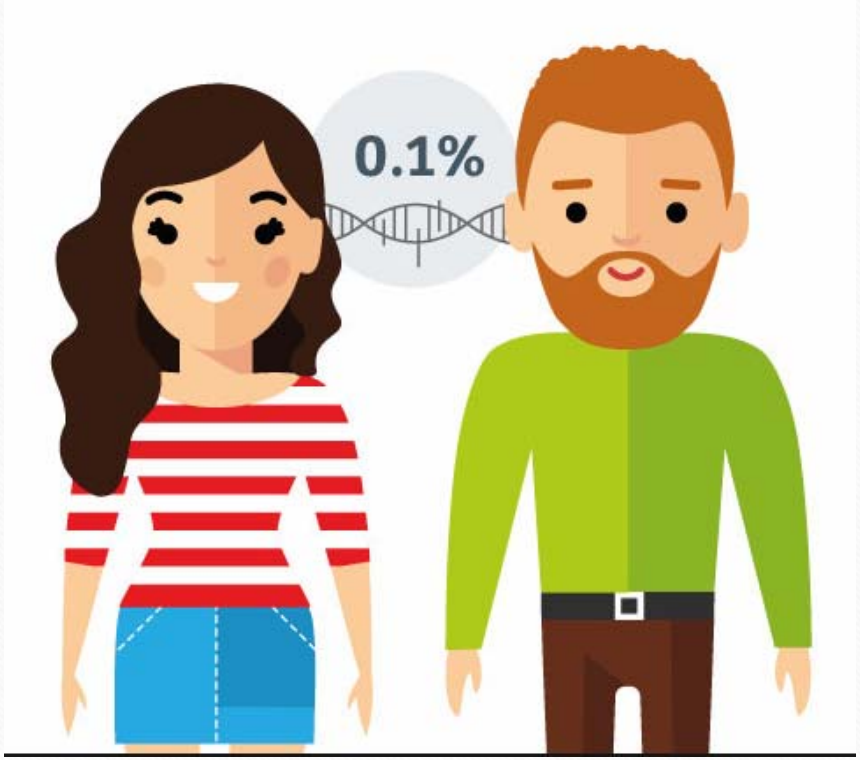
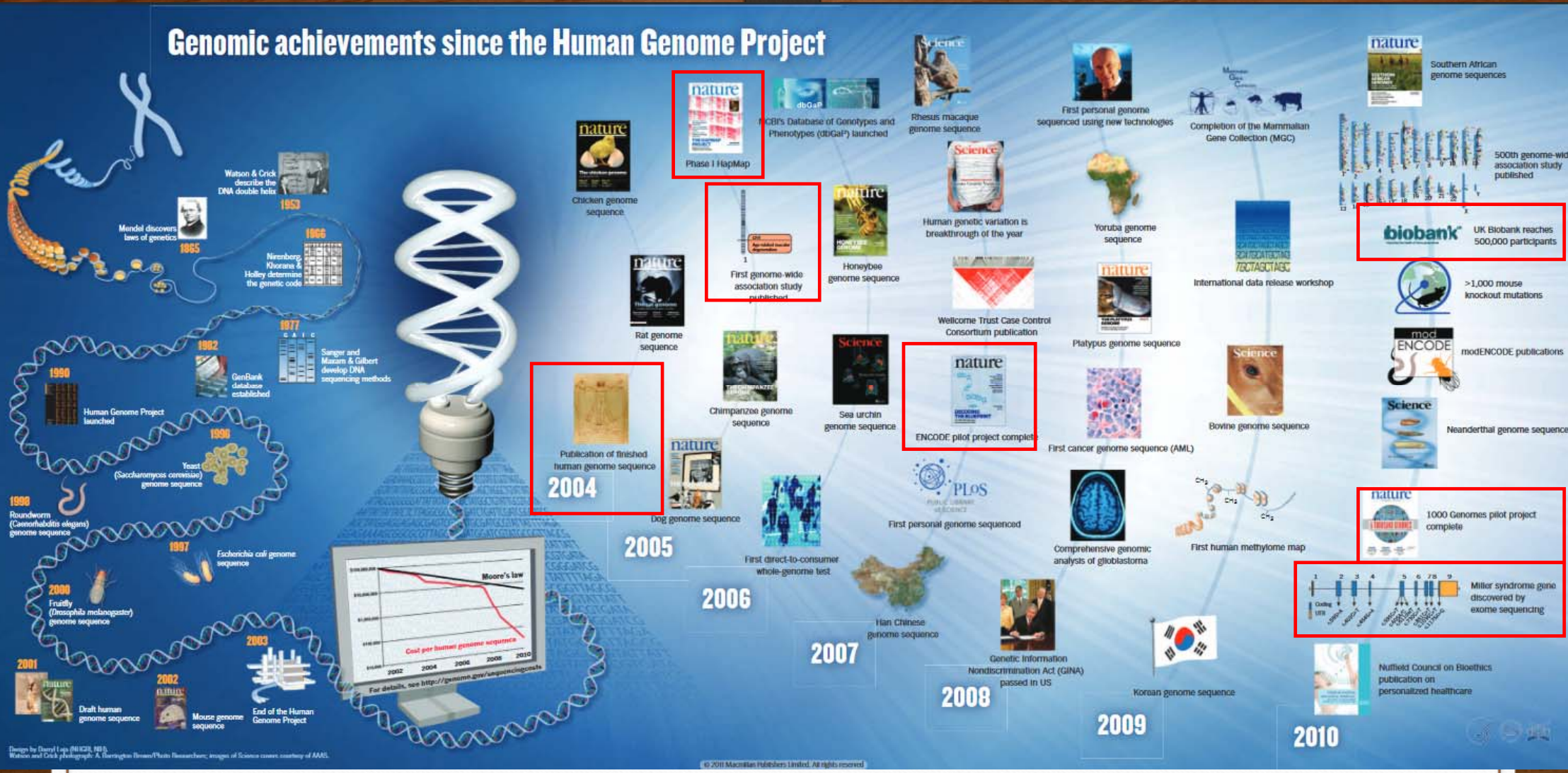


Table 2. SNPs Identified through Whole-Genome Sequencing of DNA from the Proband.*

SNP Type	No. of SNPs
Nongene	2,255,102
Gene	1,165,204
Intron	1,064,655
Promoter	60,075
3′ UTR	16,350
5′ UTR	3,517
Splice regulatory site	2,089
Splice site	112
Synonymous	9,337
Stop→stop	17
Nonsynonymous	9,069
Stop→gain	121
Stop→loss	27
Total	3,420,306

Lupski *et al.*, *New Eng J Med*, 2010

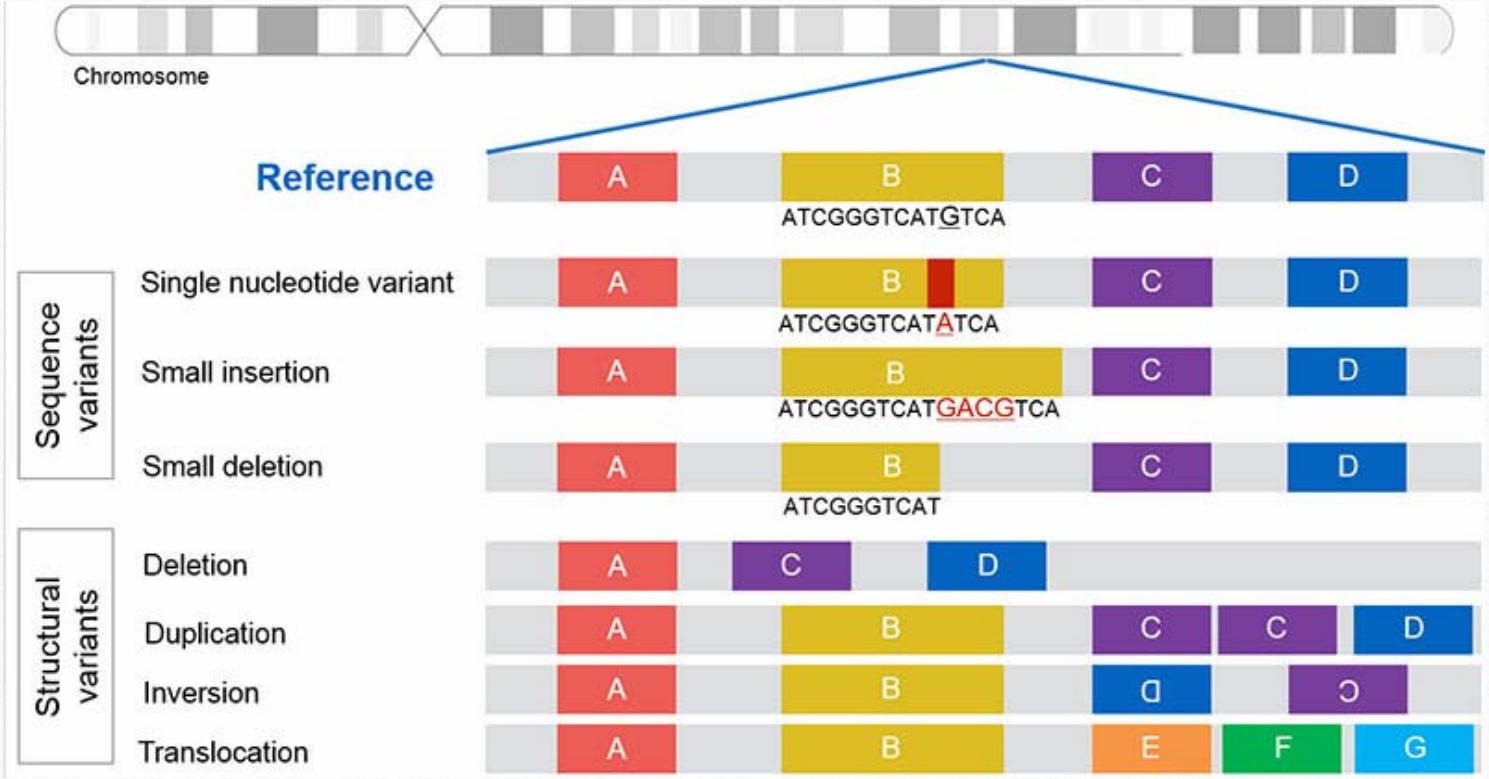


Green, Guyer & NHGRI, *Nature*, 470:204-213, 2011

Human Genome Variation

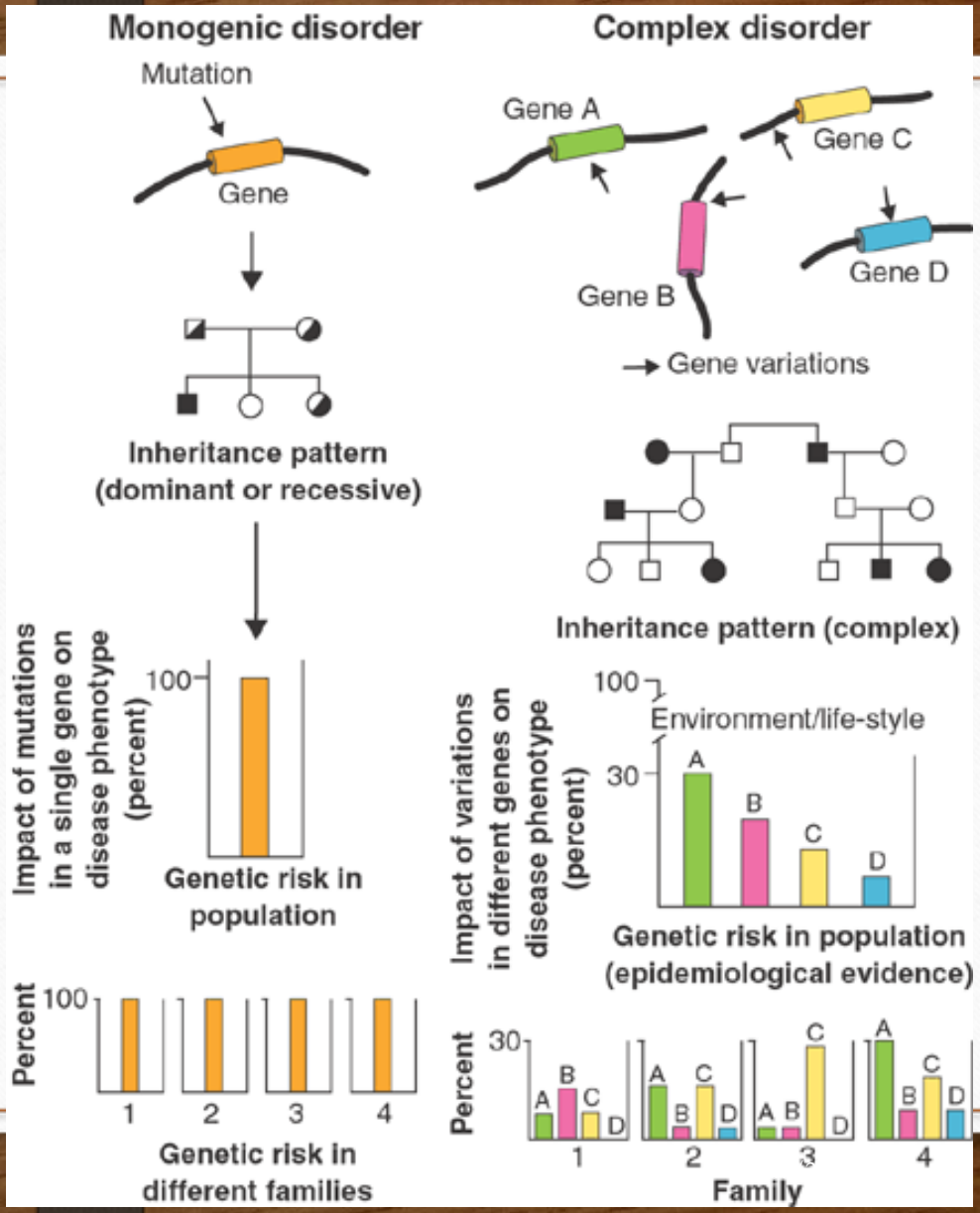
- Single Nucleotide Polymorphisms (SNPs) – diallelic or multiallelic base-pair changes
- Short Tandem Repeats (STRs) – microsatellite DNA: di-, tri-, tetra-, penta-, hexa-, etc nucleotide repeats, typically 5-35 repeats in tandem)
- Larger repeats (10-100 bp/repeat), ie Variable Number of Tandem Repeats (VNTRs)
- Insertions/Deletions (Indels) (variable length)
- Segmental duplications (variable size)
- Structural variants (from >200bp to ≥1 kb), ie Copy Number Variants (CNVs)

Human Genome Variation



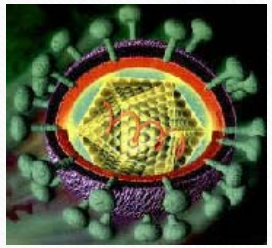
http://www.labspaces.net/blog/1627/When_Whole_Genome_Sequencing_Doesn___t_Give_Us_the_Whole_Genome

Monogenic vs Complex Disorders



UV

infections



radiation

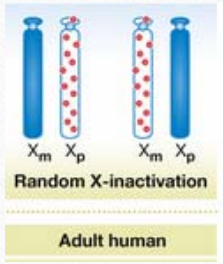
nutrition

ENVIRONMENT

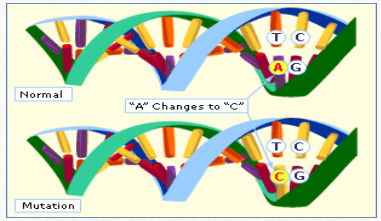
smoking

drugs

stress

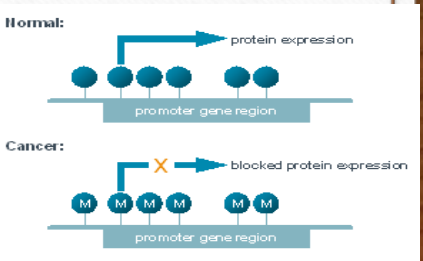


X-chr inactivation (females)



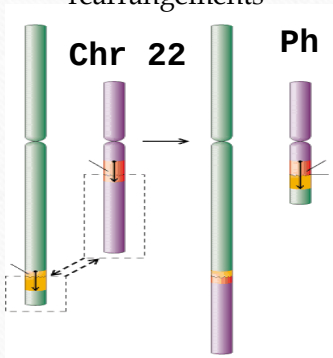
DNA variation

Epigenetics

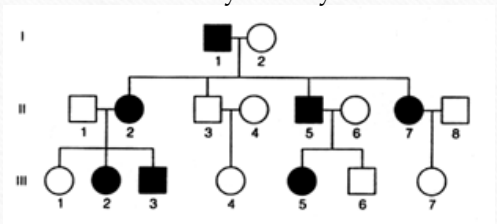


GENETICS & DNA

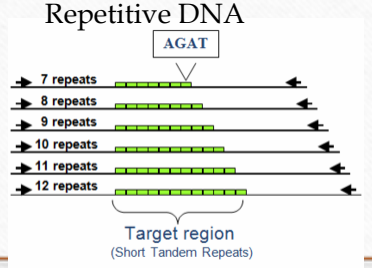
Chromosomal rearrangements

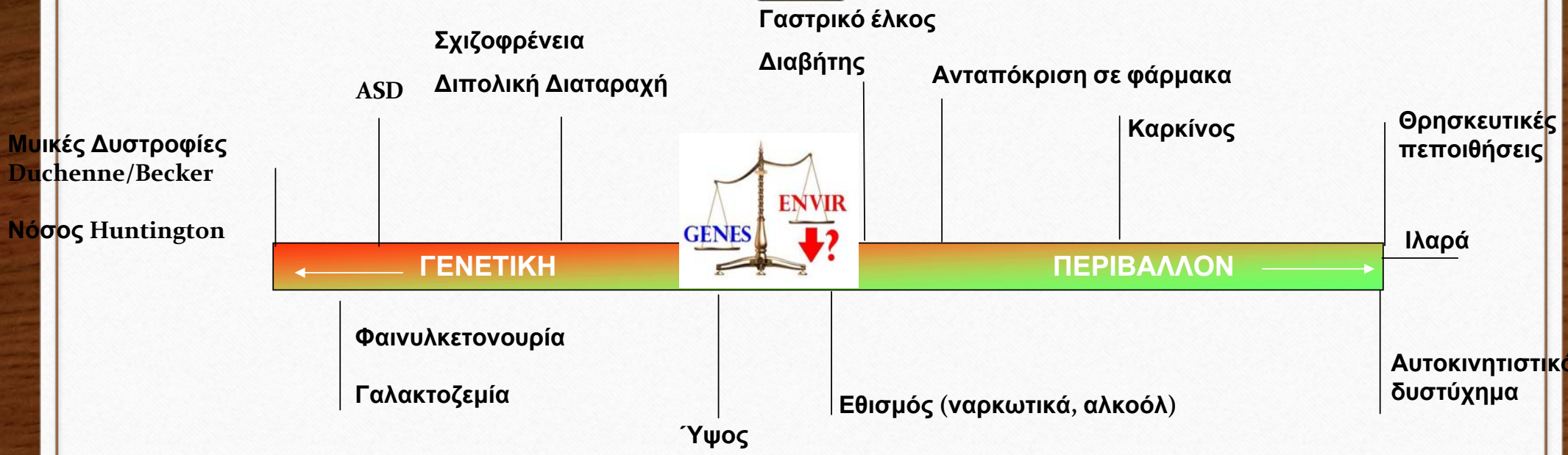


Family history



Repetitive DNA





Σπάνια νοσήματα
Σχετικά απλή γενετική βάση
Μονοπαραγοντικά (σχεδόν)
"Rare disease-rare variant" hypothesis

Κοινά νοσήματα
Σύνθετη γενετική βάση
Πολυπαραγοντικά
"Common disease-common variant" hypothesis

Single gene disorders

- Low impact on public health cost
- One or a few gene(s)
- Mendelian inheritance (dominant/recessive)
- Rare variants
- Classical genetics approaches
- Examples:
  - Huntington's disease
  - Cystic fibrosis
  - Muscular dystrophy Duchenne/Becker
  - Rett Syndrome
  - Fragile X
  - . . .

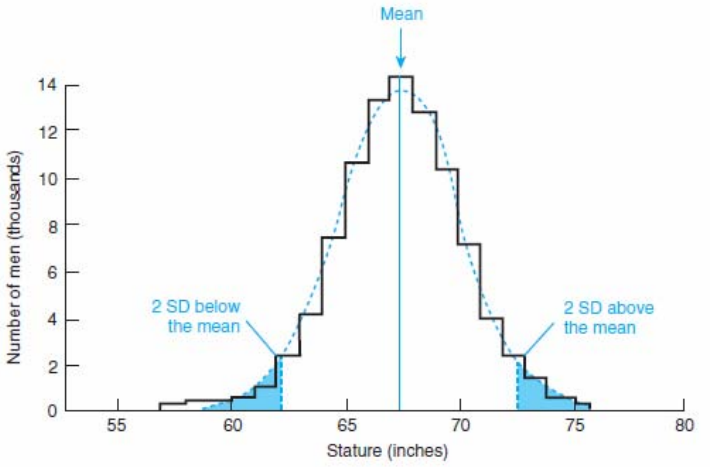
Multifactorial disorders (Complex traits)

- Serious impact on public health cost
- Multiple genes and loci
- Complex pattern of inheritance (additive)
- Variable heritability (h^2)
- Common and rare genetic variants
- Genome scans – new technologies
- Examples:
  - Stroke/CVD
  - Diabetes (Type 2)
  - Schizophrenia/Bipolar Disorder
  - ADHD/OCD
  - Osteoarthritis
  - Alzheimer's/Dementia
  - Autism Spectrum Disorder (ASD)

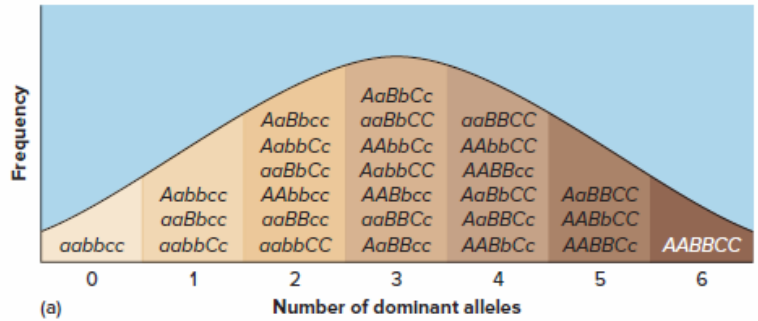
Quantitative traits and Quantitative Trait Loci (QTLs)

"All-or-none" vs "Shades of grey"

Figure 7.3 **The inheritance of height.** Genetics students at the University of Notre Dame lined up by height in inches, revealing the continuously varying nature of height. *David Hyde/Wayne Falda/McGraw-Hill Education*



Genetics in Medicine, Thompson & Thompson, 7th ed., 2007



(b)

Figure 7.4 **Variations in skin color. (a)** A model of three genes, with two alleles each, can explain broad hues of human skin. In actuality, this trait likely involves many more than three genes. **(b)** Humans come in a great variety of skin colors. Skin color genes can assort in interesting ways. These beautiful young ladies, Marcia and Millie, are twins! Their father is Jamaican with dark skin and tight dark curls and their mother is European with fair skin and golden-brown hair. *(b): SWNS/South West News Service Ltd.*

Human Genetics: Concepts and Applications, McGraw Hill, 13th ed., 2021

The genetic component of complex disorders

Twin correlations for 17.804 traits from 2.748 publications including 14.558.903 twin pairs (virtually all published twin studies of complex traits between 1958-2012) from 39 countries → heritability ~50% (all human traits are heritable)

(Polderman *et al.*, *Nat Genet*, 2015)

Table 8-2

Risk Ratios λ_r for Siblings of Probands with Diseases with Familial Aggregation and Complex Inheritance

Disease	Relationship	λ_r
Schizophrenia	Siblings	12
Autism	Siblings	150
Manic-depressive (bipolar) disorder	Siblings	7
Type 1 diabetes mellitus	Siblings	35
Crohn's disease	Siblings	25
Multiple sclerosis	Siblings	24

Data from Rimoin DL, Connor JM, Pyeritz RE: Emery and Rimoin's Principles and Practice of Medical Genetics, 3rd ed. Edinburgh, Churchill Livingstone, 1997; and King RA, Rotter JI, Motulsky AG: The Genetic Basis of Common Diseases, 2nd ed. Oxford, England, Oxford University Press, 2002.

Genetics in Medicine, Thompson & Thompson, 7th ed., 2007

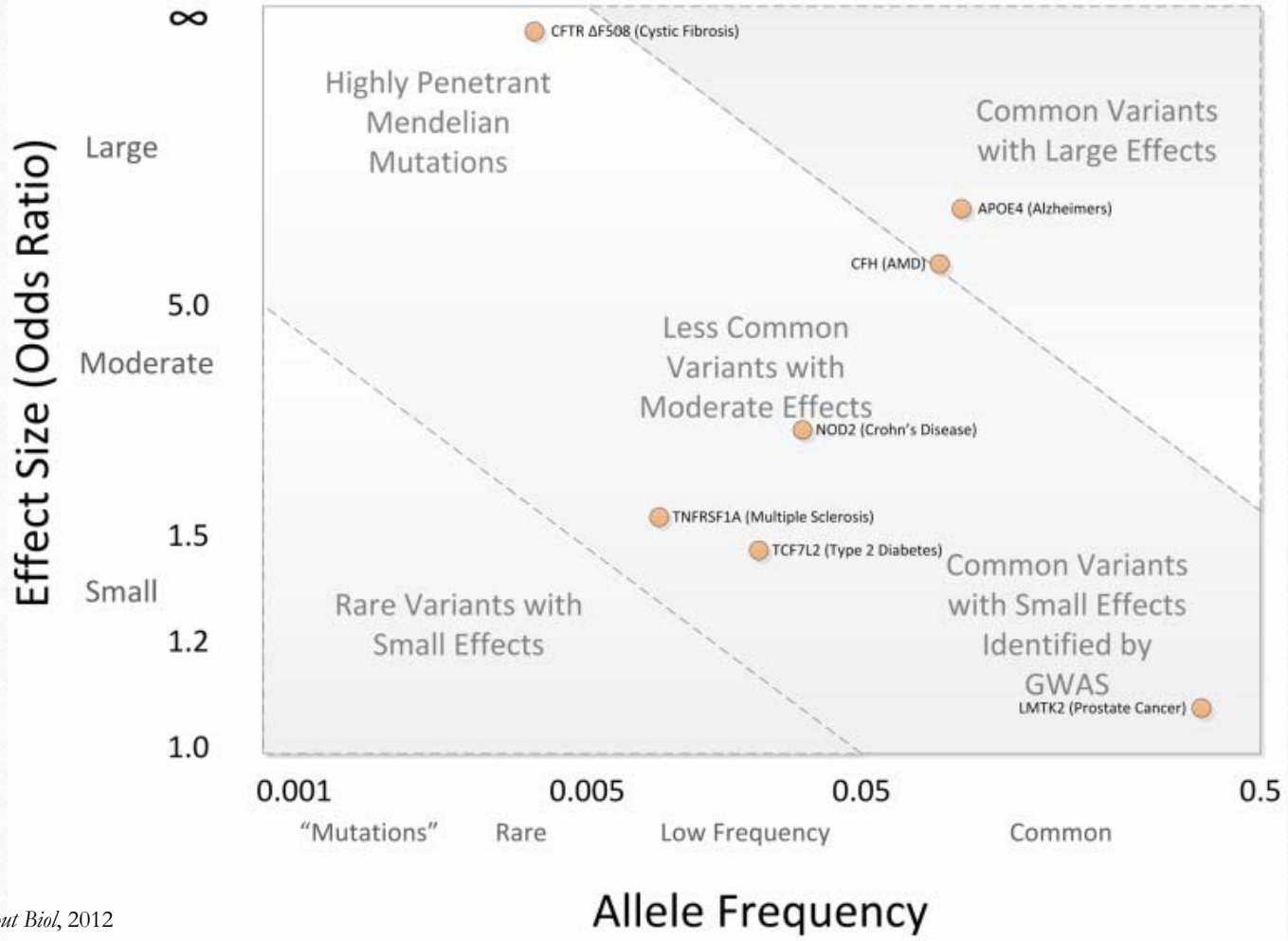
Table 8-4

Concordance Rates in MZ and DZ Twins

	Concordance (%)	
Disorder	MZ	DZ
Nontraumatic epilepsy	70	6
Multiple sclerosis	17.8	2
Type 1 diabetes	40	4.8
Schizophrenia	46	15
Bipolar disease	62	8
Osteoarthritis	32	16
Rheumatoid arthritis	12.3	3.5
Psoriasis	72	15
Cleft lip with or without cleft palate	30	2
Systemic lupus erythematosus	22	0

Data from Rimoin DL, Connor JM, Pyeritz RE: Emery and Rimoin's Principles and Practice of Medical Genetics, 3rd ed. Edinburgh, Churchill Livingstone, 1997; King RA, Rotter JI, Motulsky AG: The Genetic Basis of Common Diseases. Oxford, England, Oxford University Press, 1992; and Tsuang MT: Recent advances in genetic research on schizophrenia. J Biomed Sci 5:28-30, 1998.

Human Genetics: Concepts and Applications, McGraw Hill, 13th ed., 2021



Bush & Moore, *PLoS Comput Biol*, 2012

Heritability
(Κληρονομησιμότητα)

$$P = G + E$$

$$V_P = V_G + V_E$$

$$V_P = V_A + V_D + V_I + V_E$$

Phenotypic Variance | Additive Genetic Variance | Non-Additive Genetic Variance | Environmental Variance

$$H^2 = V_G/V_P$$

Broad-sense heritability

$$h^2 = \frac{V_A}{V_P}$$

Narrow-sense heritability

A heritability close to 1 indicates a large portion of the phenotype (variation) is due to genetic factors

$$V_G = V_A + V_D + V_I$$

- The total *genetic* variance for a character (V_G) is a function of:
- **Additive genetic variance** (V_A) – variation due to the additive effects of alleles
- **Dominance genetic variation** (V_D) – variation due to dominance relationships among alleles
- **Epistatic genetic variation** (V_I) – variation due to interactions among loci

Heritability
(Κληρονομησιμότητα)

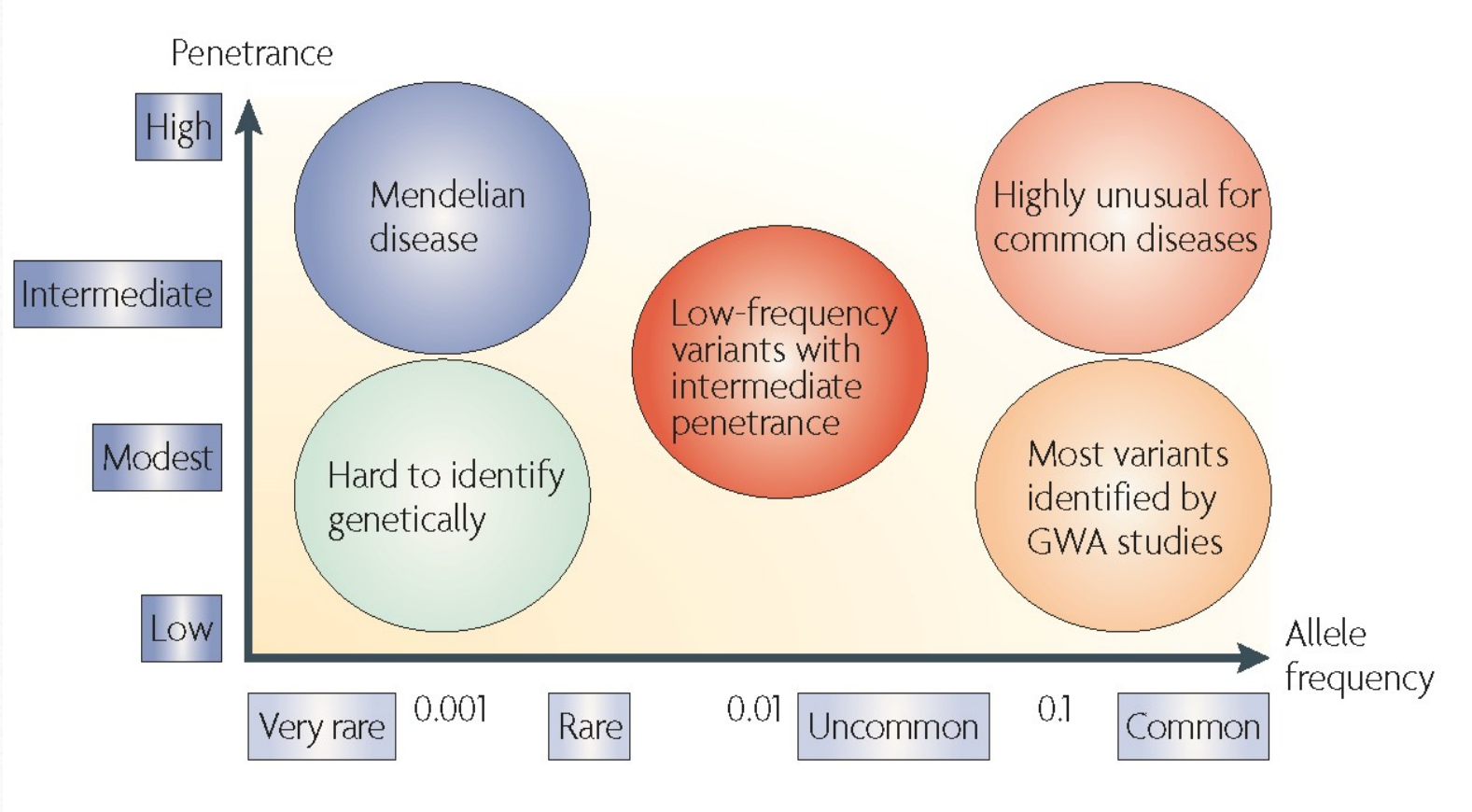
Quantitative Trait (QT)	h^2
Height (humans)	0.65
Milk production (cows)	0.35
Number of offspring (pigs)	0.05
Egg production (chicken)	0.10
Tail length (mice)	0.40
Body size (Drosophila)	0.40

Table 7.2 Heritabilities for Some Human Traits

Trait	Heritability
Clubfoot	0.8
Height	0.8
Blood pressure	0.6
Body mass index	0.4–0.7
Verbal aptitude	0.7
Mathematical aptitude	0.3
Spelling aptitude	0.5
Total fingerprint ridge count	0.9
Intelligence	0.5–0.8
Total serum cholesterol	0.6

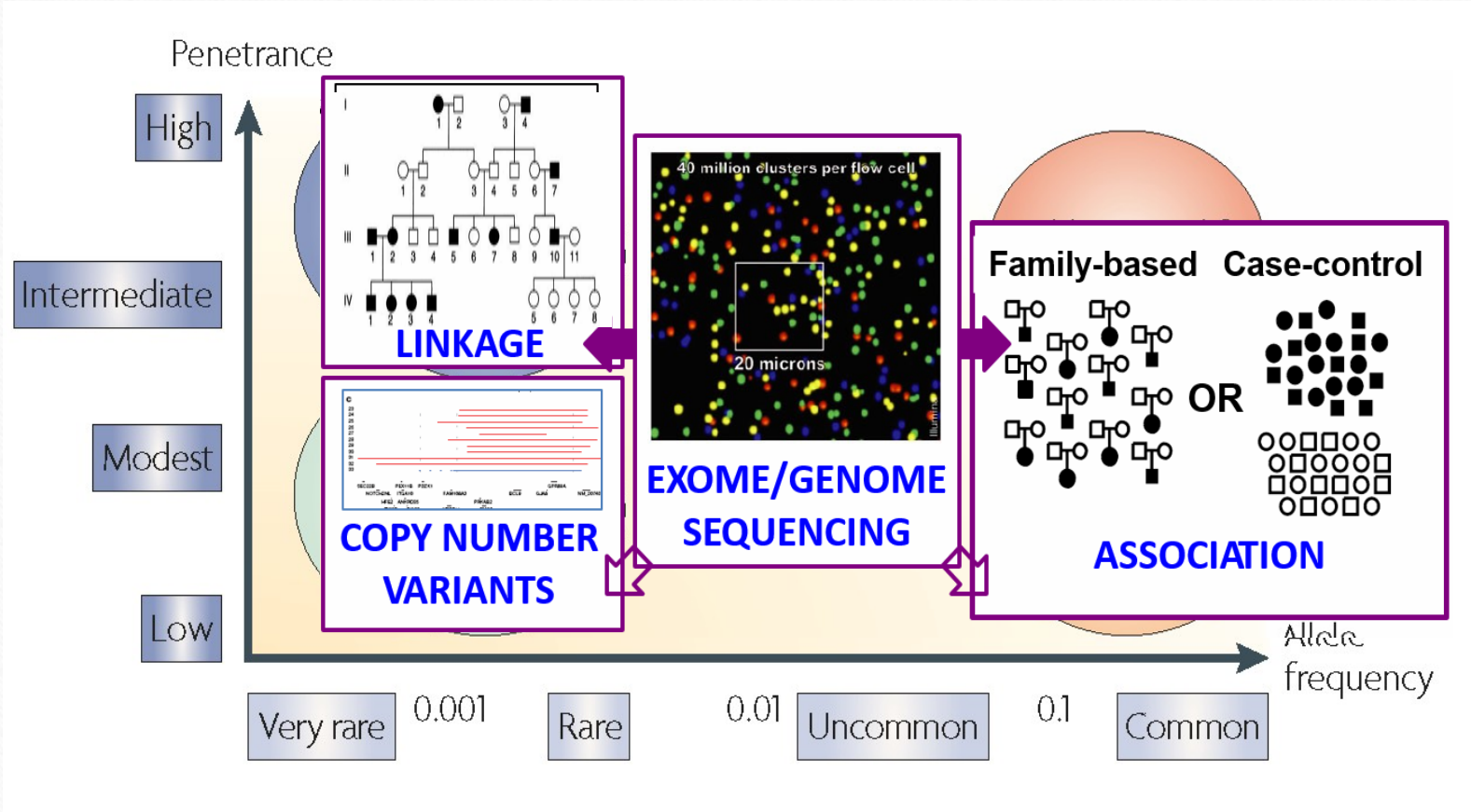
Human Genetics: Concepts and Applications, McGraw Hill, 13th ed., 2021

The landscape of human genome variation



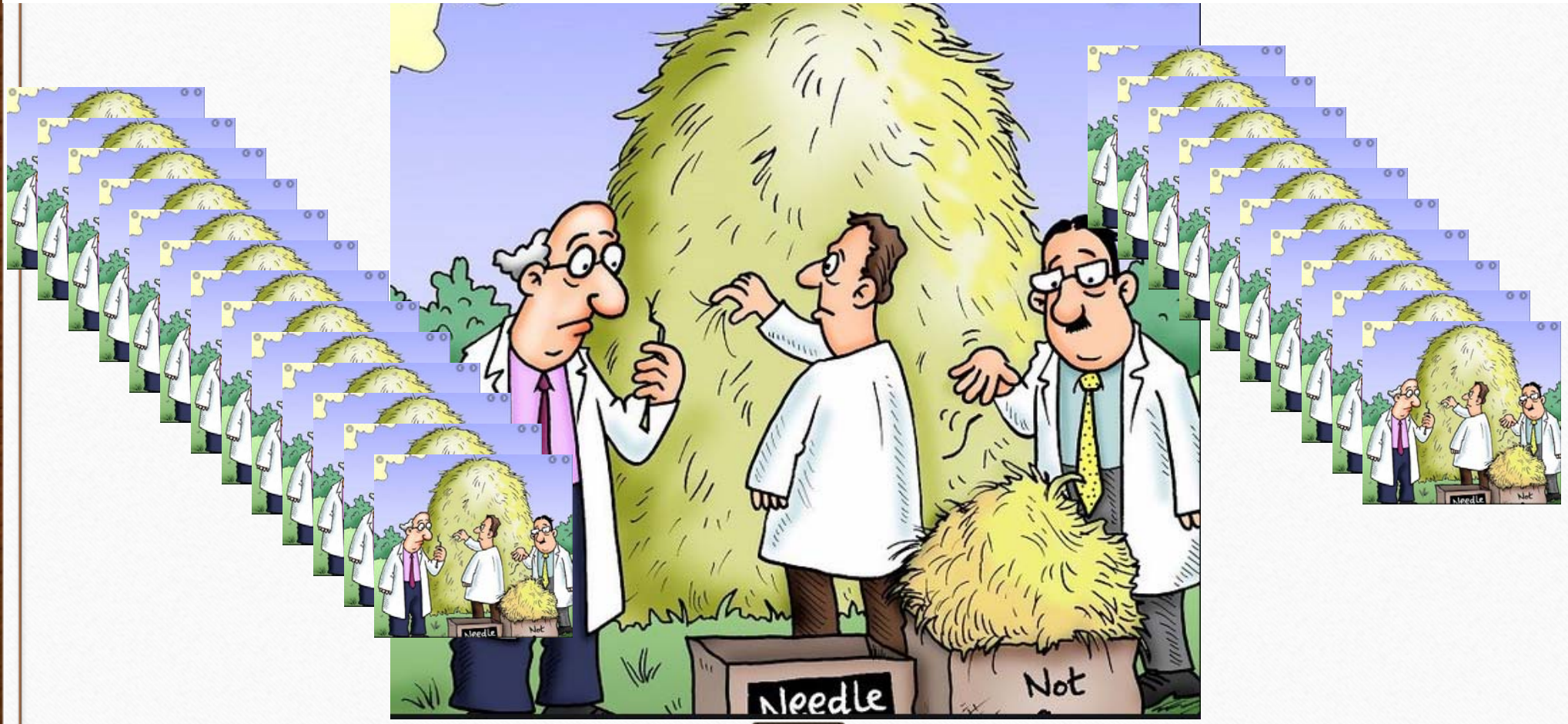
McCarthy *et al.*, *Nat Rev Genet*, 2008

The landscape of human genome variation



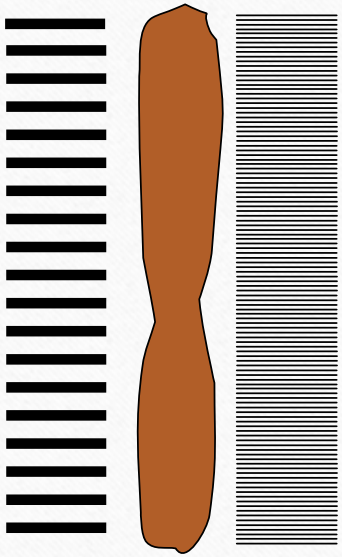
McCarthy *et al.*, *Nat Rev Genet*, 2008

We should not expect common variants to have large effects, for evolutionary reasons. Large effects are bad, as most variants with large effects are consequently selected against, thus never become fixed and common.

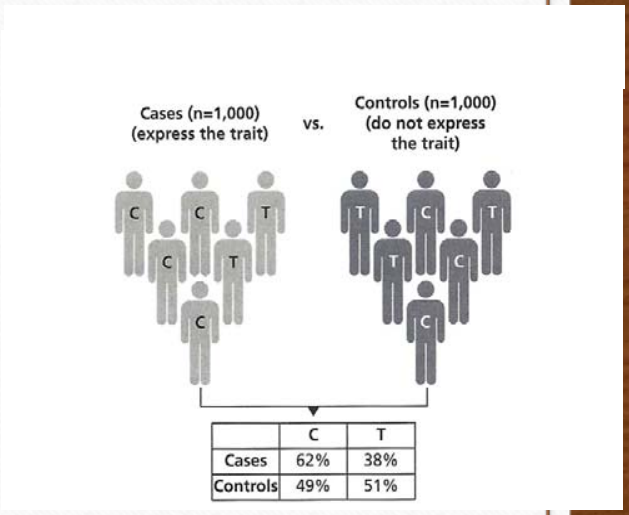


SNPs can be used to create dense marker maps

microsatellites



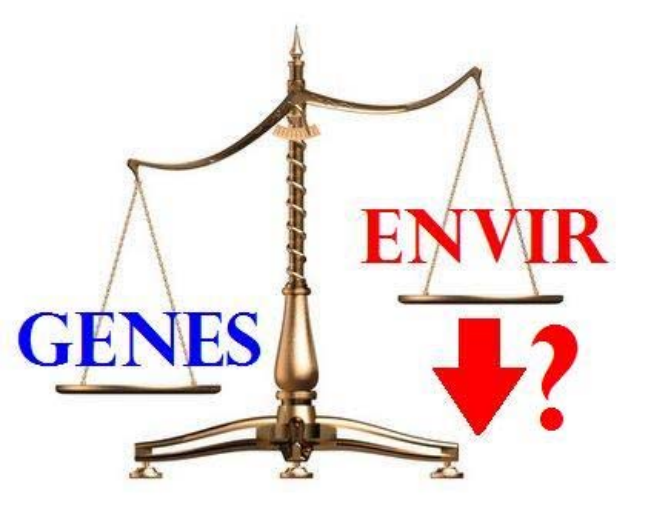
SNPs



Recent genome-wide association studies use millions of SNPs in a case-control design

Genetic Association Studies

Aim: To unravel associations between **genetic data** (ie genotypes) of commonly occurring genetic variants with information regarding a trait or a medical phenotype (ie disease) under study, using **statistical analyses** and a **large enough sample size** (typically cases versus controls), in order to support the statistics that these variants contribute to trait/disease risk.



Examples of complex diseases

Type II Diabetes Mellitus
Obesity
Cardiovascular diseases
Cancer (non-hereditary)
Osteoarthritis
Autoimmune disorders
Alzheimer's disease
….

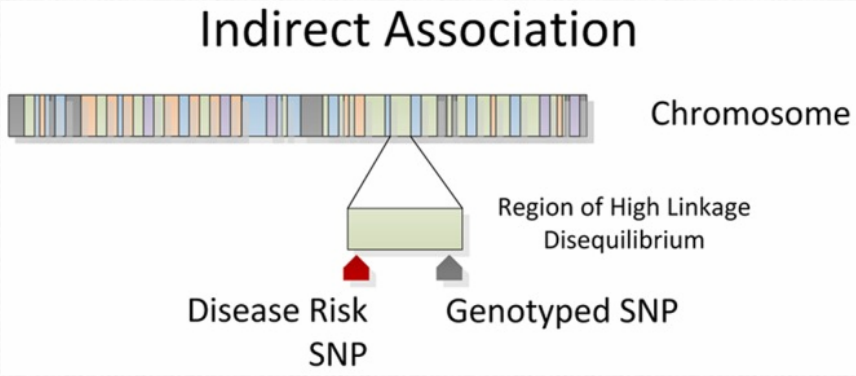
Schizophrenia
Autism
Bipolar Disorder
Obsessive Compulsive Disorder
Learning disabilities (Dyslexia)
….

Genetic Association Studies

- → Appropriate for complex phenotypes (quantitative or dichotomous traits)
- → Increased genetic (locus) heterogeneity (ie multigenic variance) – many genes, many variants
- → Common variance ("common disease – common variant" hypothesis) – modest/low disease risk per variant
- → Large numbers of **cases and controls** (healthy individuals), or family-based associations (**trios**, ie affected child and both parents) or **sibship-based** associations (ie two affected siblings) or **extreme phenotypes** to detect associations

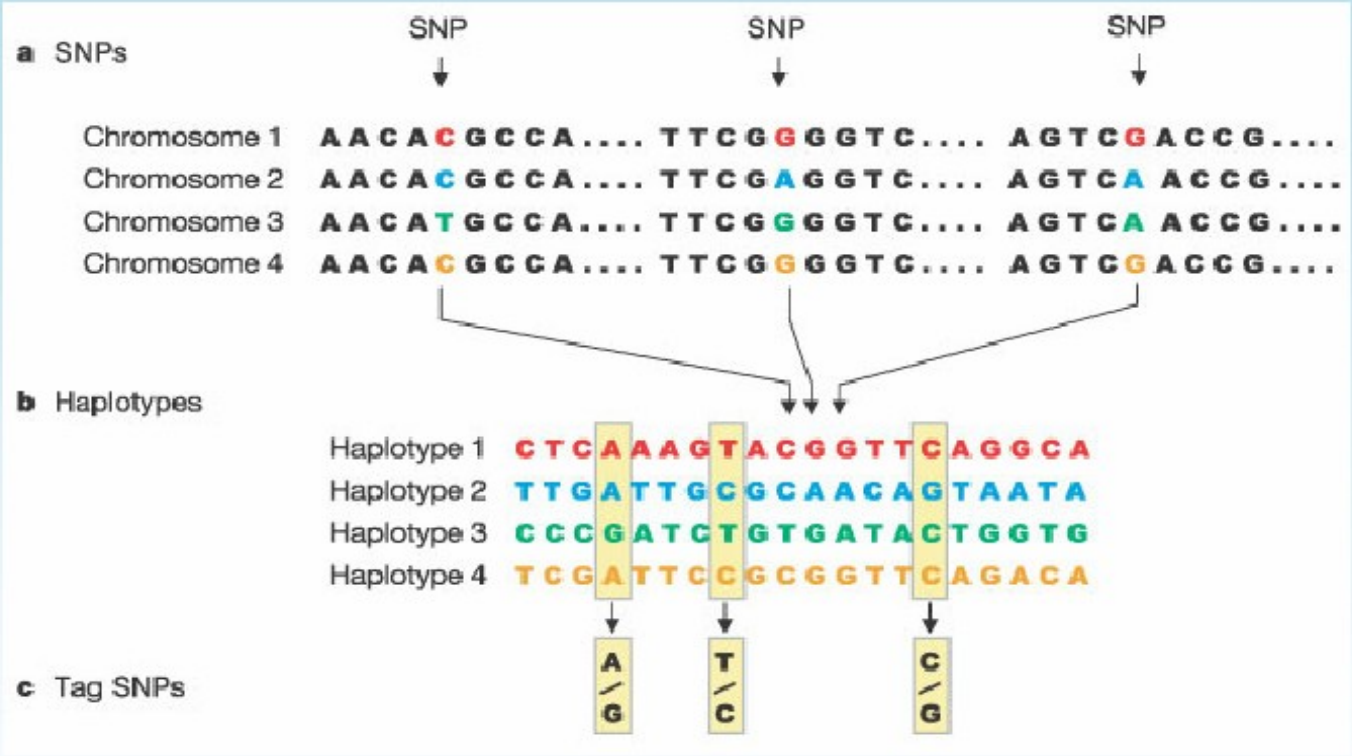
- → Mostly SNPs (**tagSNPs**): A single or a few SNPs within a chromosomal region that capture(s) (ie "tags") most of the common DNA variation in this particular region, owing to the effect of Linkage Disequilibrium (**LD**).

Note: LD is defined as the phenomenon of co-inheritance (non-random association) of genetic marker (SNP) alleles, unlikely to be separated by recombination (aka "linked" markers) within a population.



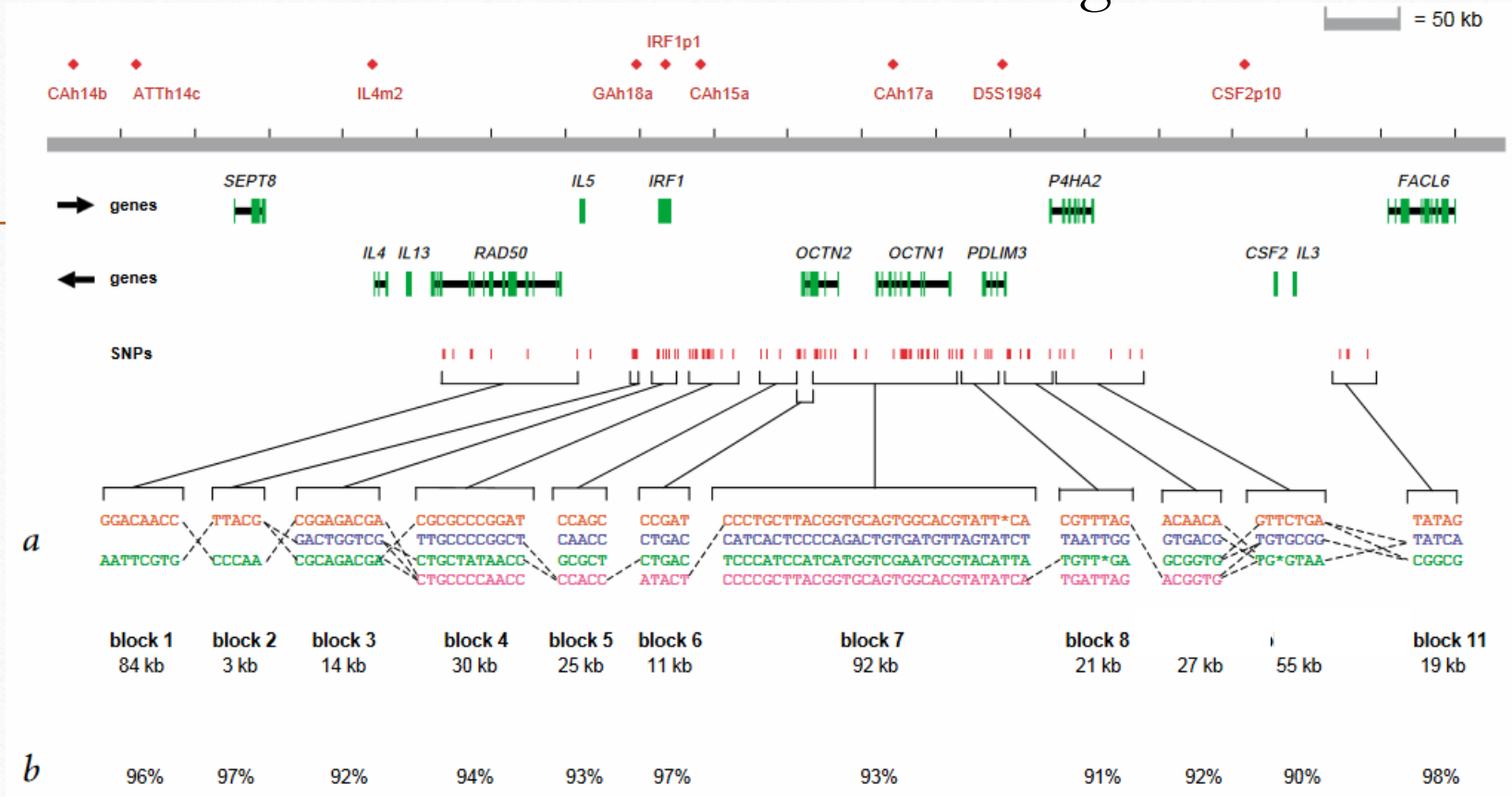
Bush & Moore, *PLoS Comput Biol*, 2012

Human Sequence Variation



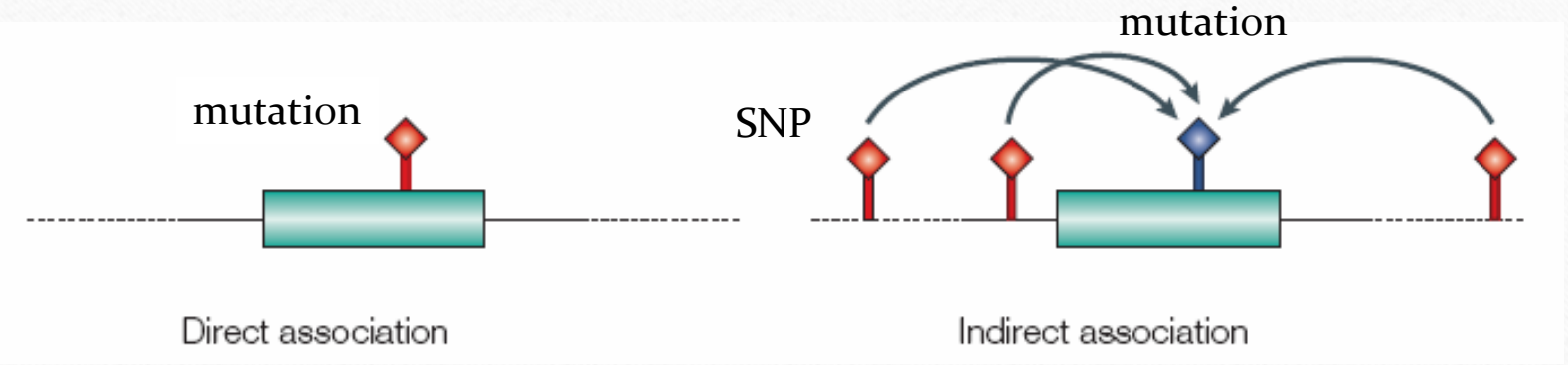
The International HapMap Project, Nature 2003

A block-like structure of the human genome…

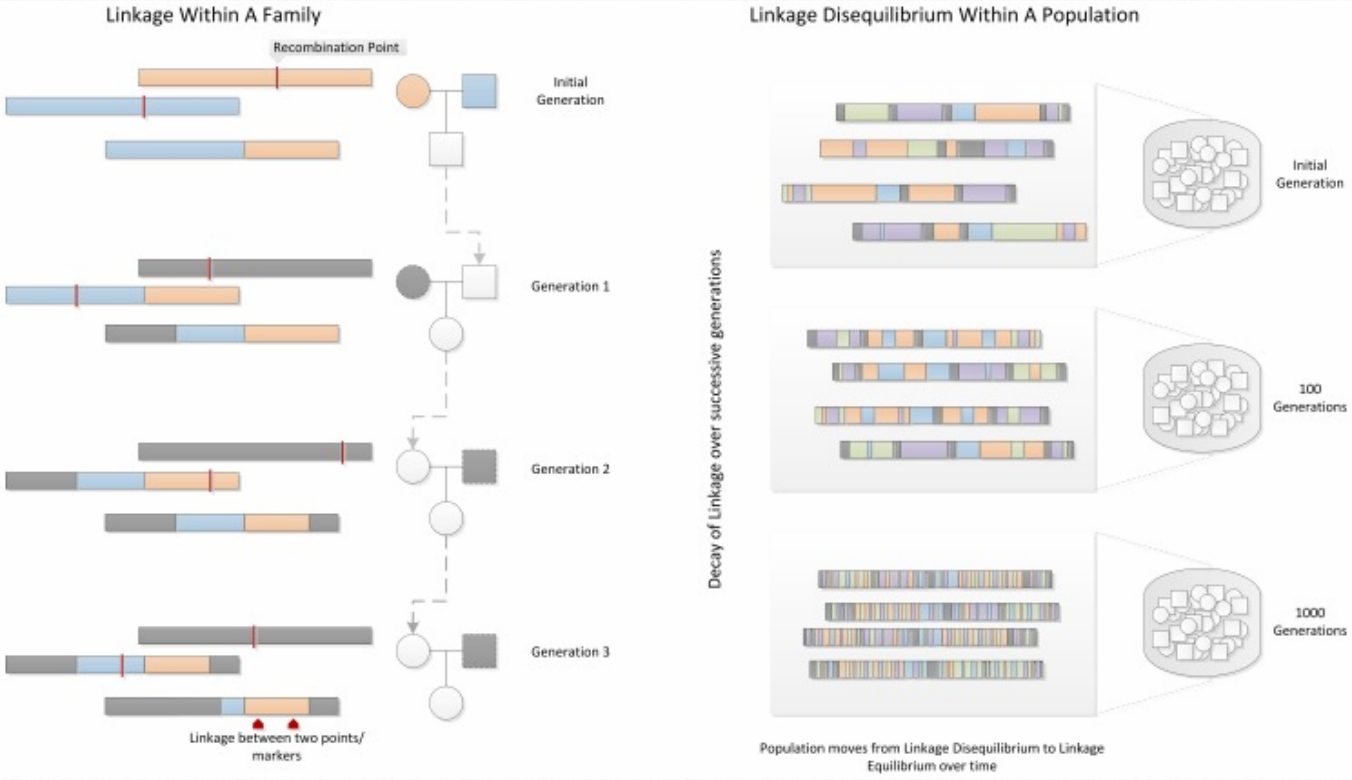


Daly et al. (2001). Nature Genet 29: 229-232

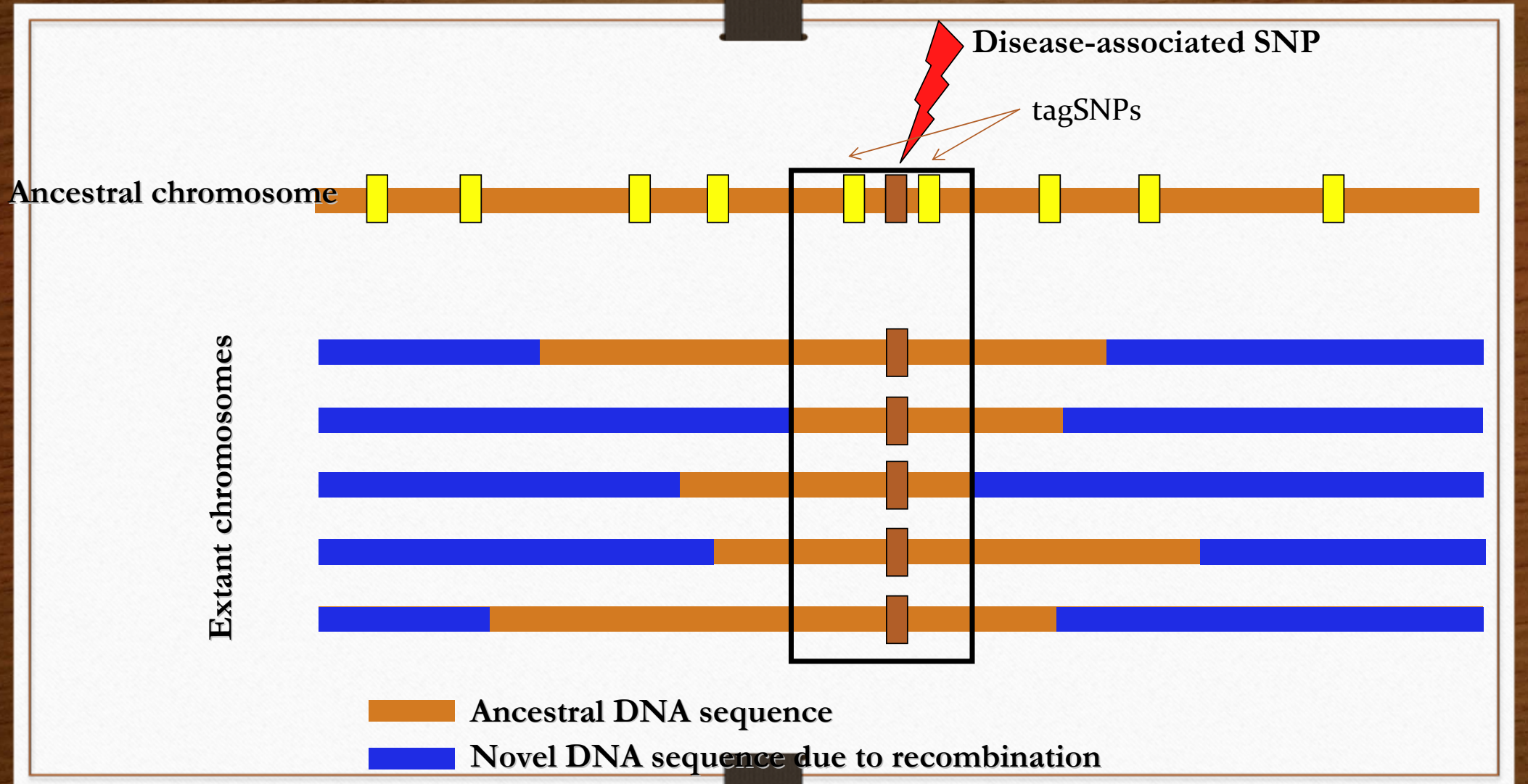
Genome structure allows the selection of tagging SNPs



Hirschhorn & Daly, *Nat Rev Genet*, 2005



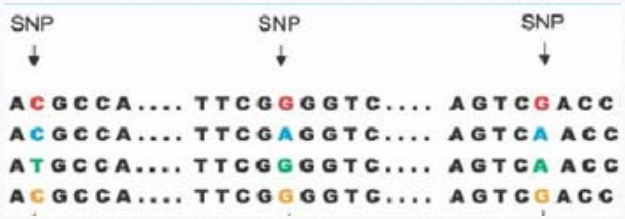
Bush & Moore, *PLoS Comput Biol*, 2012


Disease-associated SNP
tagSNPs
Ancestral chromosome
Extant chromosomes
Ancestral DNA sequence
Novel DNA sequence due to recombination

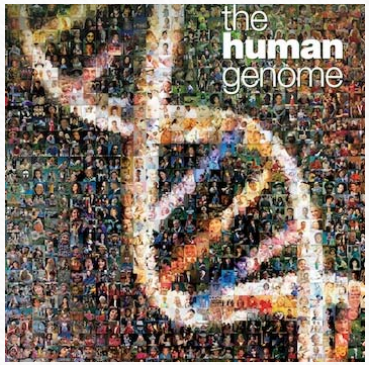
Genetic architecture of a complex disorder

Population

Genetic variants



- Number
- Frequency
- Effects
- Linkage Disequilibrium



https://www.genome.gov/10001772

NOTE: The first genetic association studies focused on **candidate gene analysis** and, therefore, were not suited for novel genetic risk loci identification.

The completion of the **Human Genome Project**, the **HapMap Project** and **1000Genomes Project**, along with many technological and conceptual advances, have paved the way to array (chip)-based GWAS and Next-Generation Sequencing

→ **High-throughput targeted genotyping and NGS**

http://hapmap.ncbi.nlm.nih.gov/

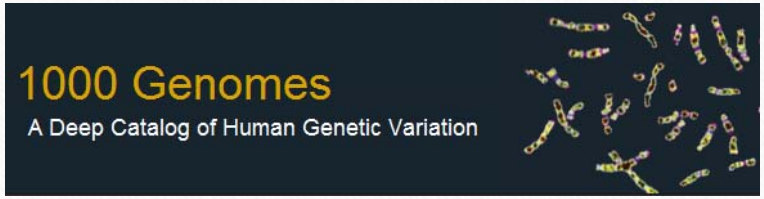
GENOME 10K

Autism Genome 10K

Genomics england

The 100,000 genomes project

THE PRECISION MEDICINE INITIATIVE



https://www.internationalgenome.org/

Genome-wide association studies (GWAS)

Patients

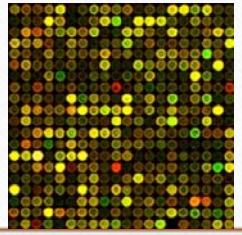
Thousands of samples
– The more the better !

Controls

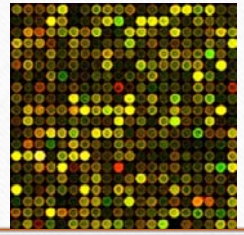
- ✓ Better phenotyping of cases – Inclusion/Exclusion criteria
- ✓ More careful selection of controls
- ✓ Investigators joining forces – Data sharing
- ✓ Funding of large-scale projects
- ✓ International data repositories

- ✓ Catalogued SNPs (Minor Allele Frequencies – MAFs) and tagSNPs
- ✓ Population-specific variation (AIMs)
- ✓ Patterns of LD per population
- ✓ Frameworks to analyze enormous datasets - bioinformatics

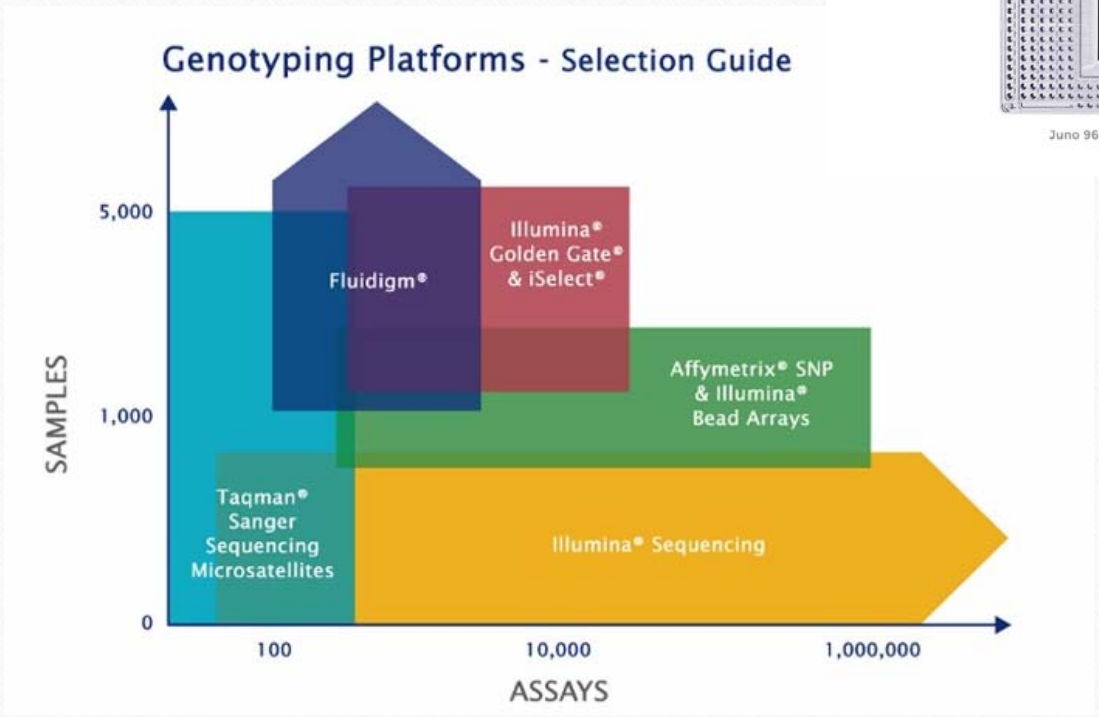
Genotyping of eg 500k-6M SNPs



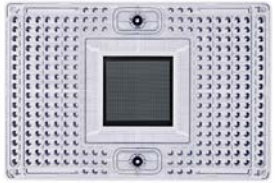
Comparison of alleles
–
Statistical analysis



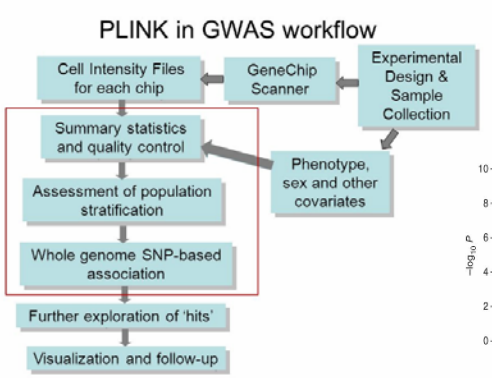
Available genotyping platforms – A comparison

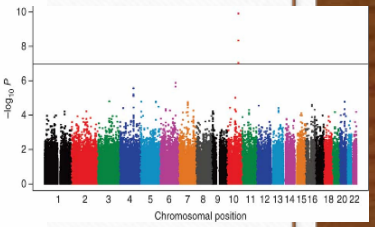


http://www.lifesciences.sourcebioscience.com/media/426305/genotyping%20image.png


Juno 96.96 Genotyping IFC

Workflow of a typical GWAS study



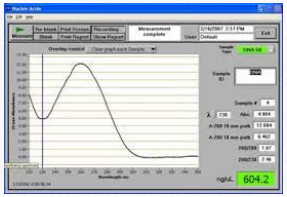


Sample collection

Peripheral blood
Buccal cells
…

DNA extraction and Quantification

Manual purification
Kit-based purification



Fluorospectrophotometry
Spectrophotometry

DNA Plating

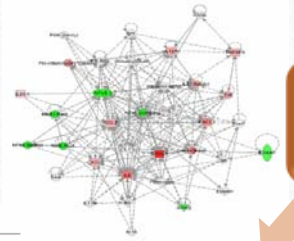
96-well plates
384-well plates

Array run

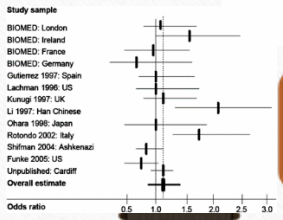
QC

Statistical analysis (trait-dependent)

- ***ANOVA***
- ***Chi-square***
- **Covariate adjustment**
- **Multiple testing correction**
- **Population stratification**



Pathway analyses



Meta-analyses

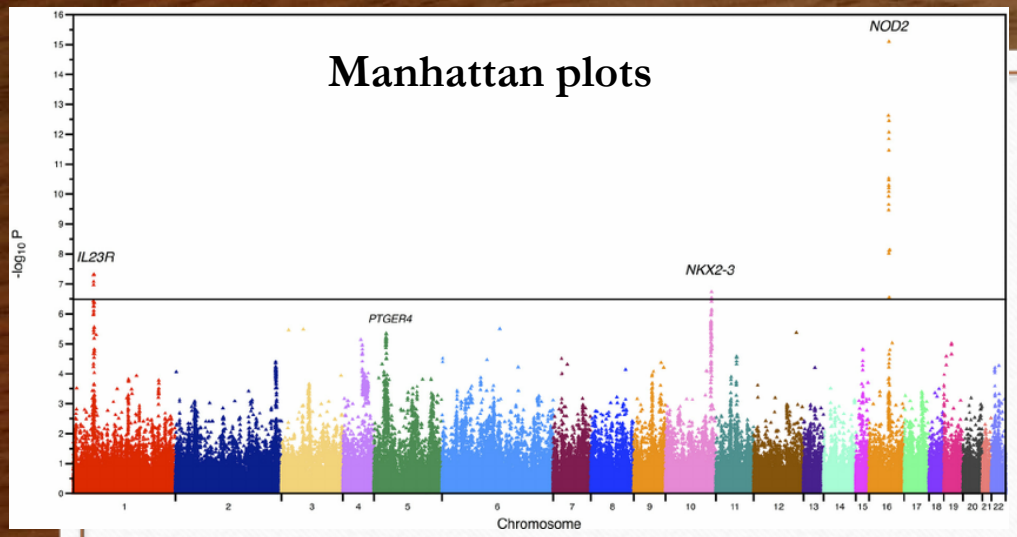
Replication study (stage 2)

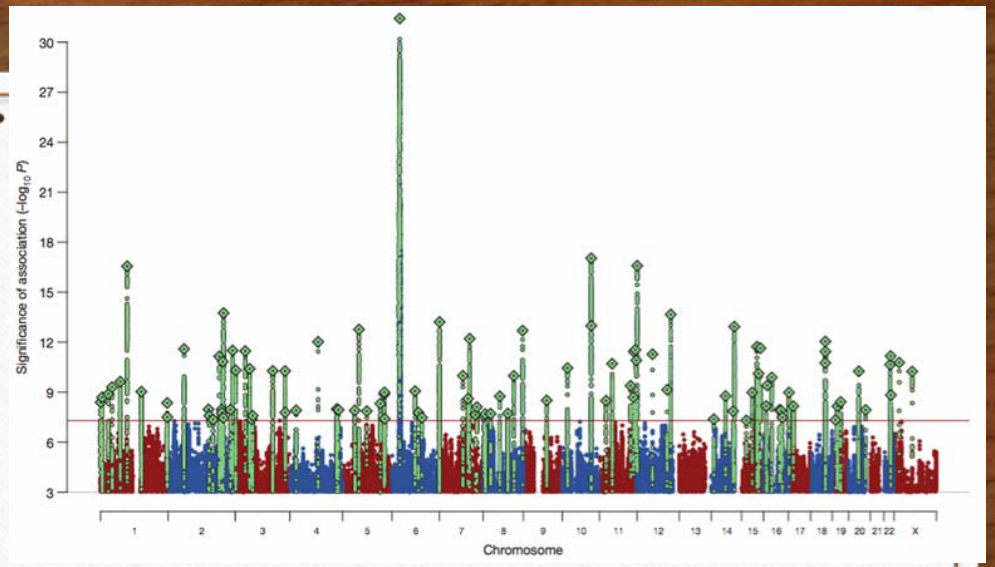
- **Sample size (large)**
- **Population**
- **Phenotypic criteria**

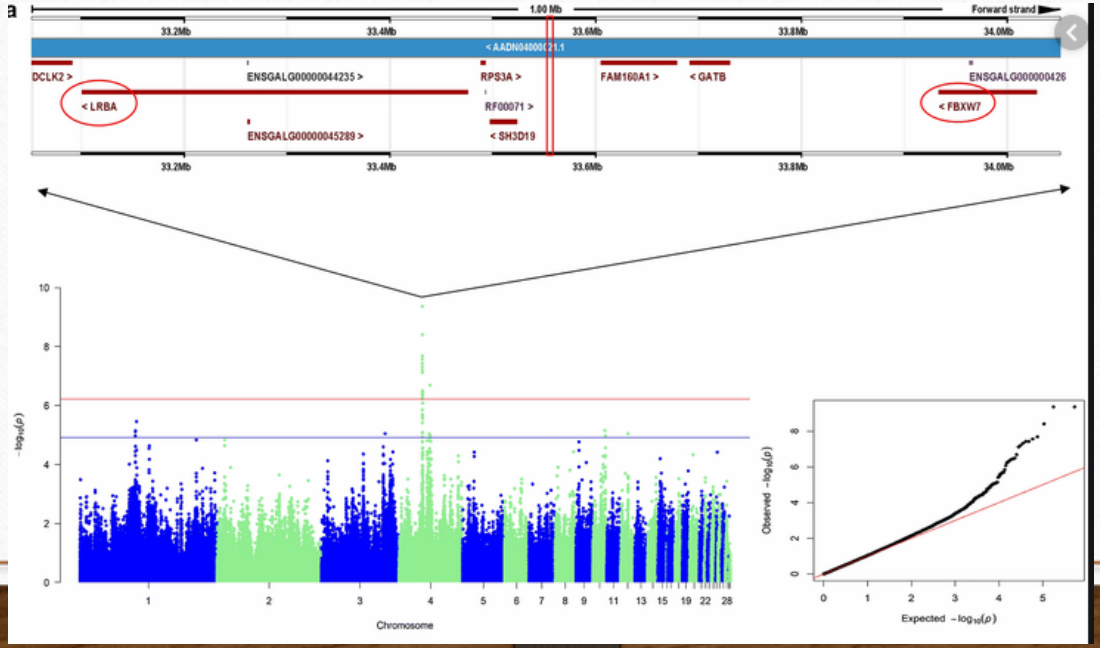
Validation (technical, functional)

- **Other genotyping platform**
- ***In silico* analysis**
- **Sequencing**
- **Gene expression analysis**
- **Cell-based analysis**
- **Animal models**

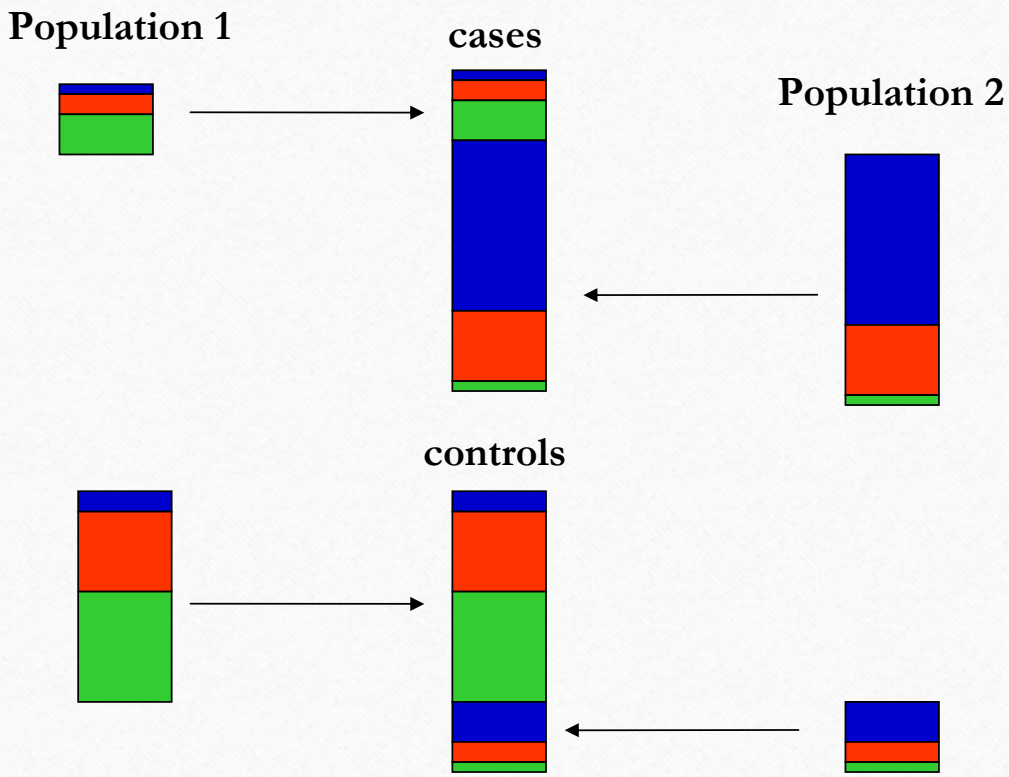
Manhattan plots







Population stratification as a confounding factor in genetic association studies can lead to false-positive associations and wrong assumptions



Published Genome-Wide Associations as of May 2018

p≤5X10-8 for 17 trait categories

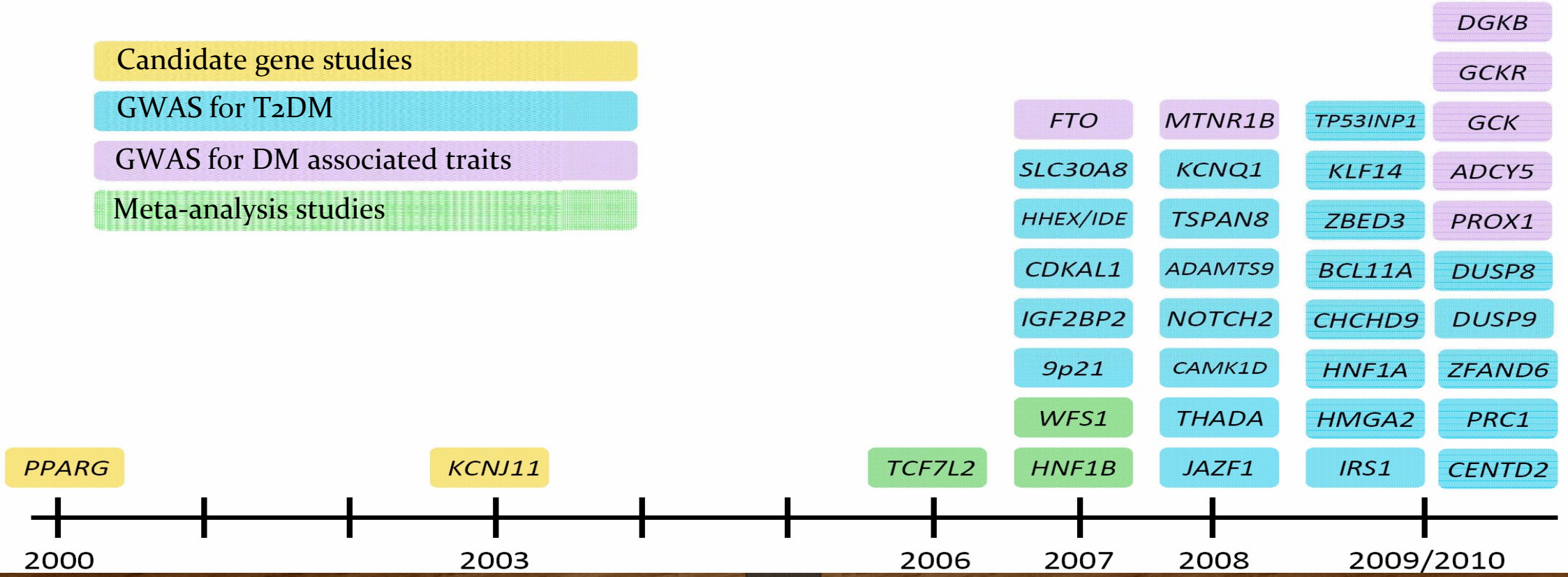
>10.000 SNPs

NIH National Human Genome Research Institute

EMBL-EBI

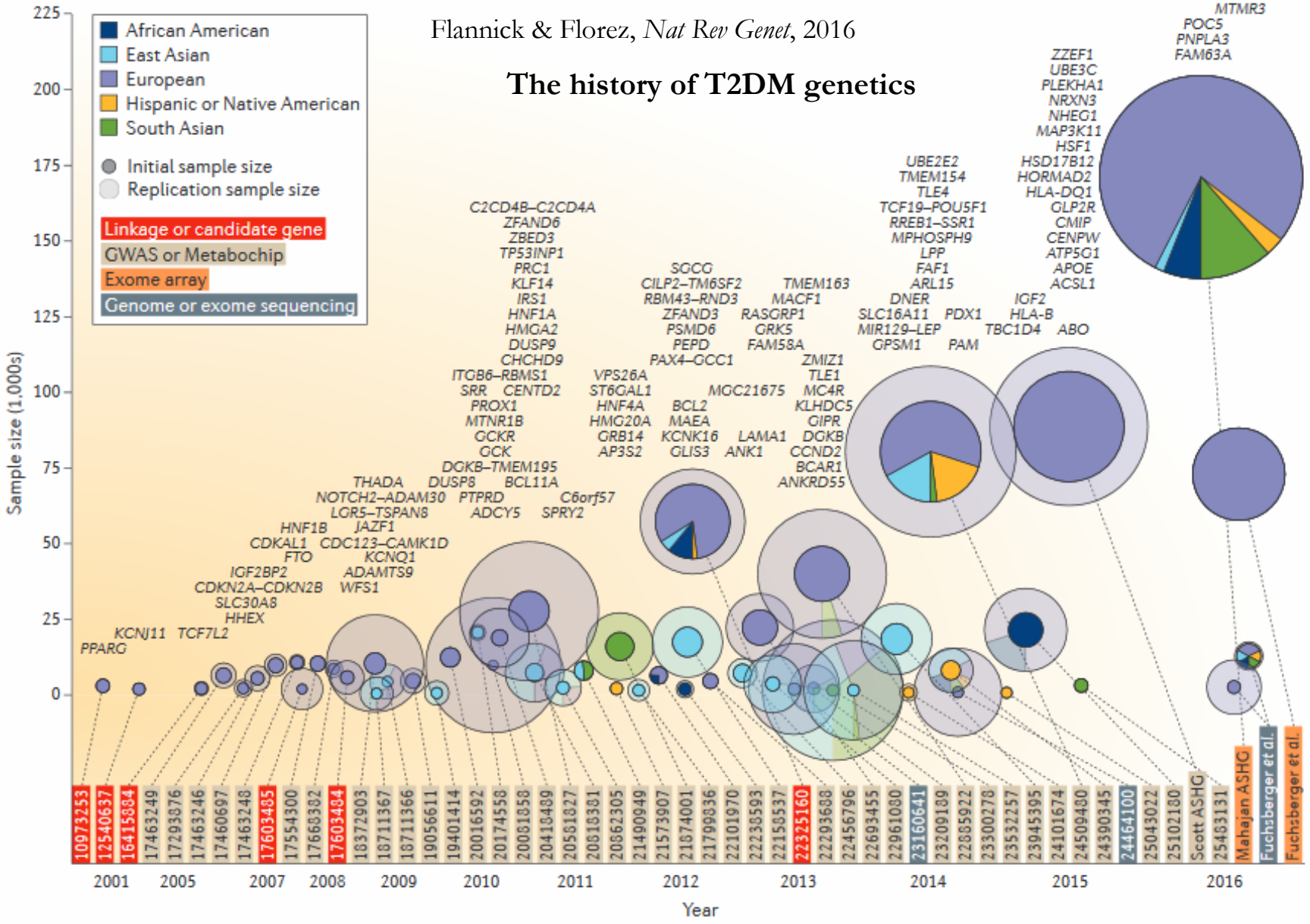
NHGRI-EBI GWAS Catalog
www.ebi.ac.uk/gwas

T2DM associated genes after 2007 – The GWAS era

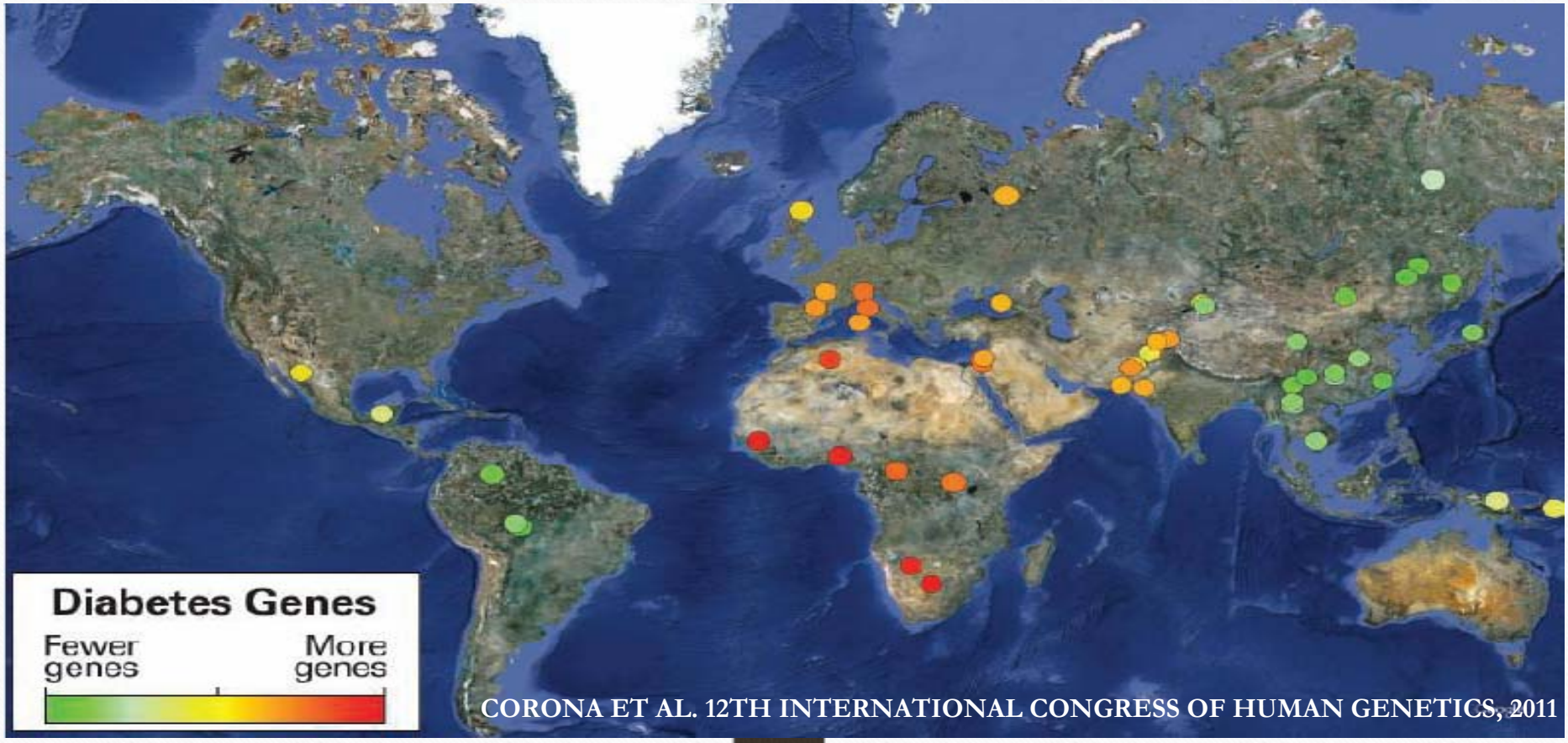


Flannick & Florez, *Nat Rev Genet*, 2016

The history of T2DM genetics

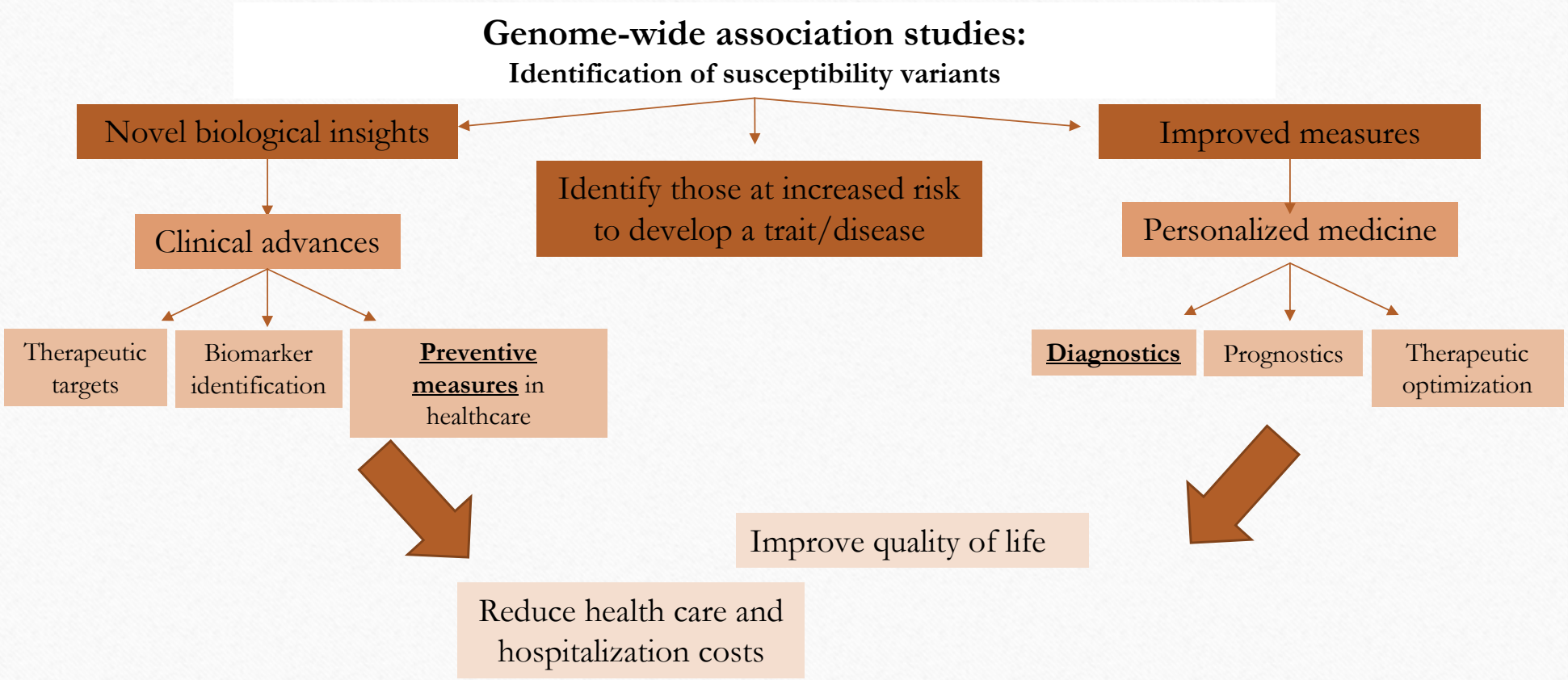


Alleles associated with complex disorders differ in frequency around the world



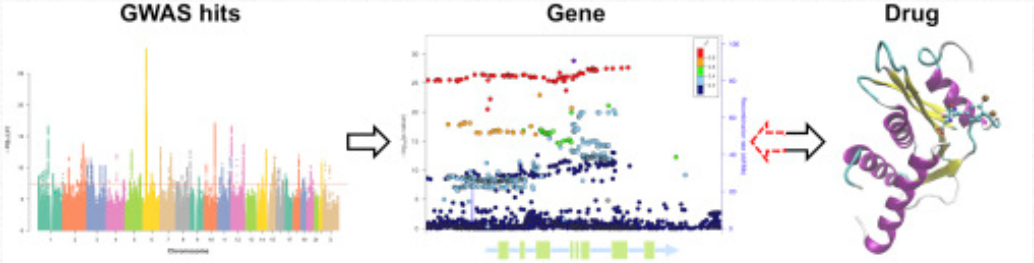
CORONA ET AL. 12TH INTERNATIONAL CONGRESS OF HUMAN GENETICS, 2011

What GWAS studies have to offer – Translational research



Adapted from McCarthy *et al.*, *Nat Rev Genet*, 9:357-369, 2008

Examples of links between GWAS discoveries and drugs



Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

Visscher *et al., Am J Hum Genet,* 2017

What GWAS studies yet have to offer

Pros (+)

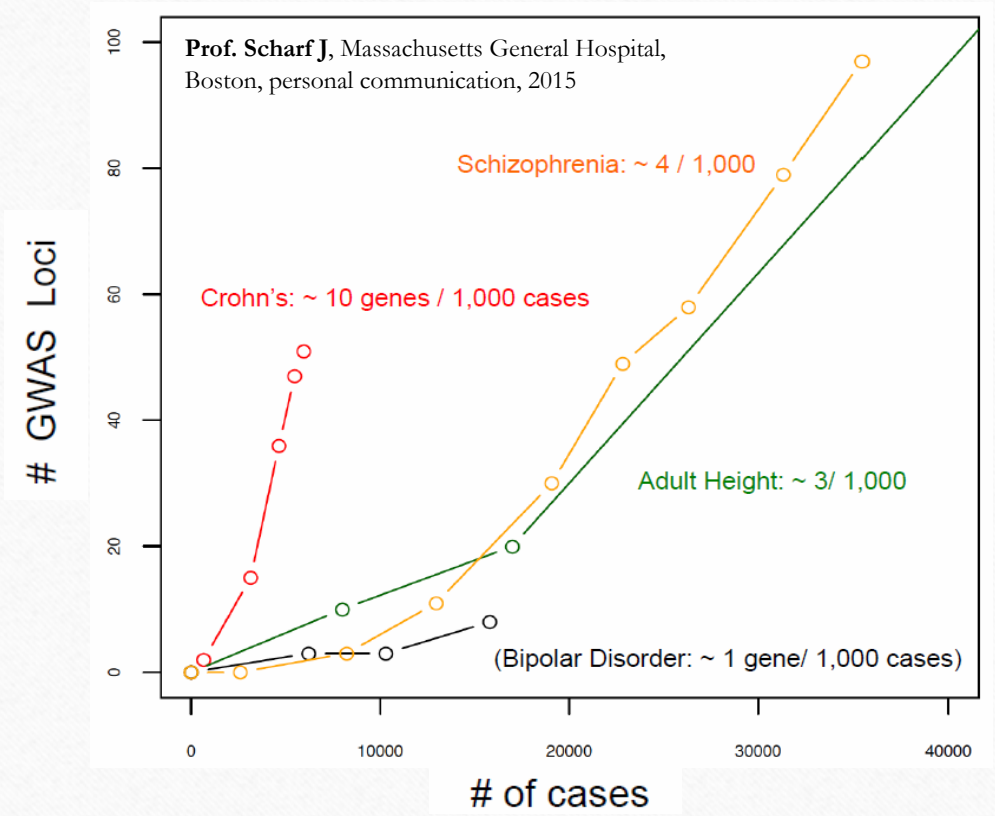
- High-throughput analysis (million variants)
- Large-scale projects with thousands participants
- Suitable for complex, non-Mendelian disorders, quantitative traits, eQTLs, sQTLs, …
- Variety of software tools for computerized analysis
- Dataset repositories for meta-analyses continuously curated and updated
- Pathway analyses
- Certified service providers worldwide
- ...

"Missing heritability" of common diseases

Cons (-)

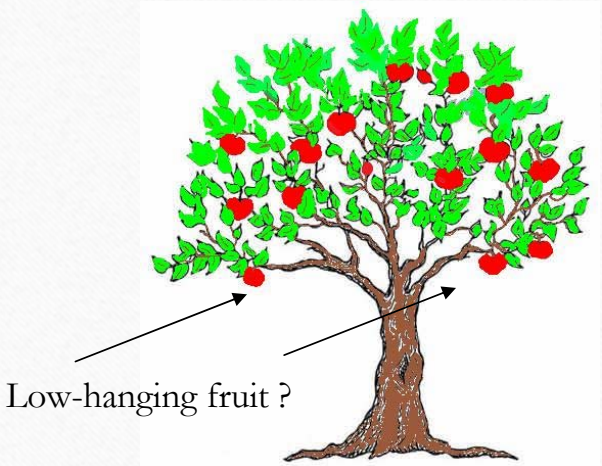
- They target pre-defined markers (biased)
- Not suitable for identification (*de novo*) studies
- Alleles confer modest (OR<1.5) or even small effect sizes [small Odds Ratios (typically 1.05<OR<1.2)]
- False-positive (population stratification, genotyping errors, selection bias, etc) or false-negative results (insensitivity to rare variants, lack of genetic variants from platforms, lack of variation in a SNP in the population under study)
- The richer they are in context, the more expensive
- Their analysis requires special training in bioinformatics and computationally intensive analyses and infrastructure
- Functional approaches to interpret the data are needed (gene expression analysis, cell and animal model manipulations, etc)
- ...

What GWAS studies yet have to offer



Missing heritability not targeted by GWAS studies

- Rare variants (MAF<1%)
- Structural variants
- Gene-gene interactions
- Gene-environment interactions
- Population isolates and population extremes

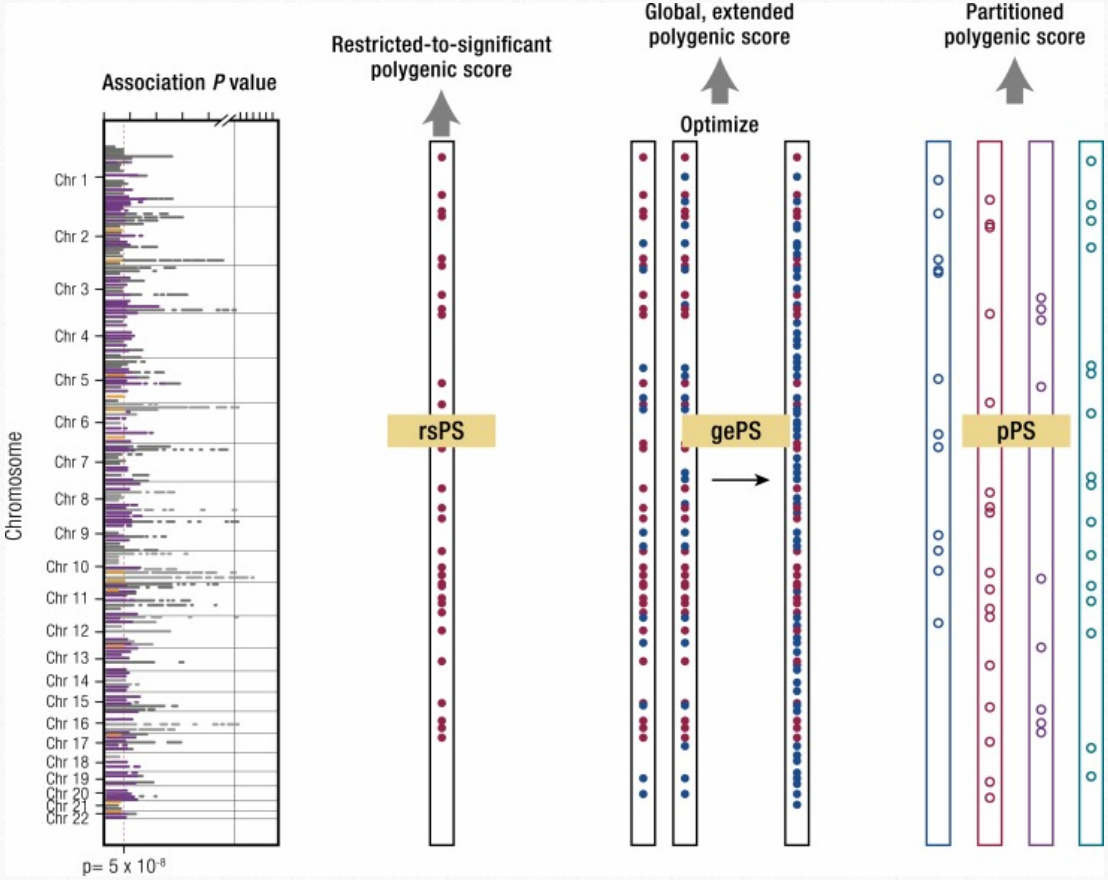


(high penetrance, high MAF)

Polygenic Risk Scores (PRSs)

(Πολυγονιδιακοί Δείκτες Κινδύνου)

Udhler *et al.*, *Endocr Rev*, 2019



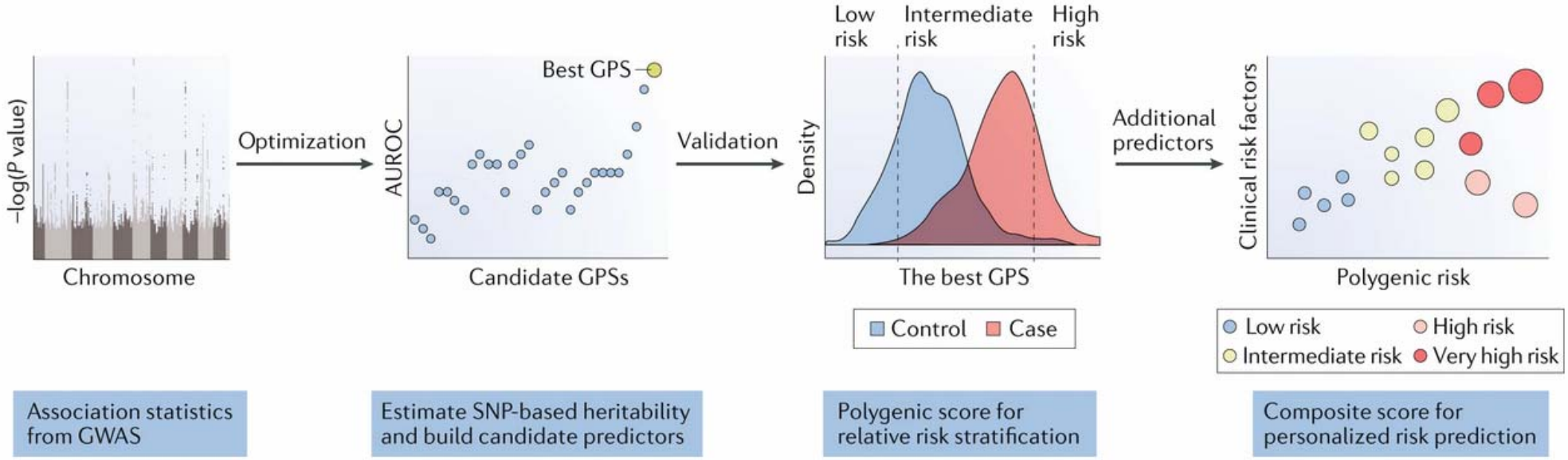
Restricted-to-significant polygenic scores (rsPSs): scores composed of variants at the extreme of a statistical distribution, most usually those that pass the genome-wide significant threshold for the trait concerned.

Global extended polygenic scores (gePSs): scores generated from a deeper set of variants generated from genome-wide analyses, typically involving large numbers of subthreshold significant variants.

Partitioned or process-specific polygenic scores (pPSs): scores composed of variants grouped according to some common biological process (*e.g.*, association with a related endophenotype, tissue expression of related genes, chromatin state).

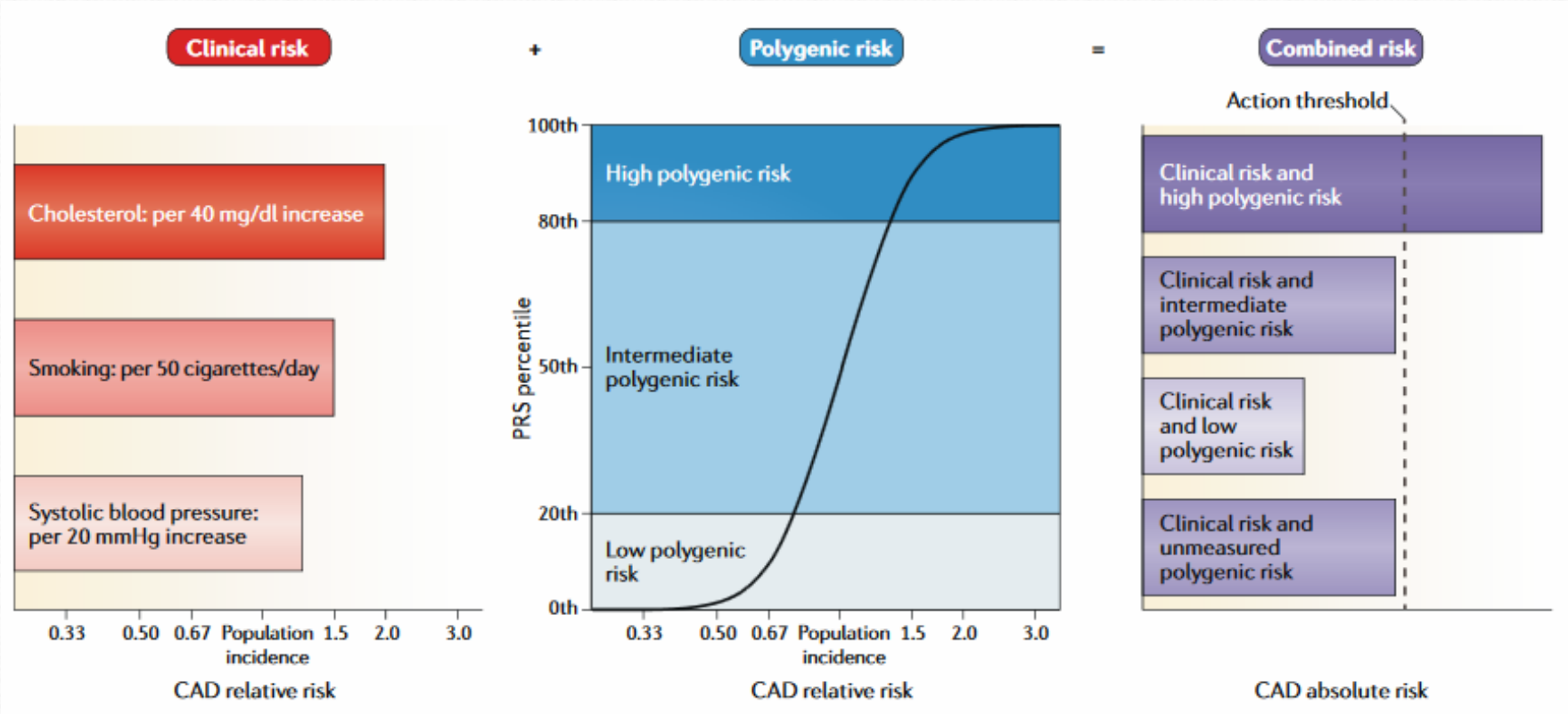
Polygenic Risk Scores (PRSs)

Liu & Kiryluk, *Nat Rev Nephrol*, 2018



A genome-wide polygenic risk score (GPS) is based on genome-wide association study (GWAS) summary statistics. The optimization step enables selection of the best method according to the genetic architecture of a disease under study. The validation step requires an external cohort and is critical to obtaining reliable metrics of performance. Clinical predictors of absolute risk will require incorporation of additional demographic, clinical or lifestyle factors into composite risk models. AUROC, area under receiver operating characteristic. SNP, single-nucleotide polymorphism.

Polygenic Risk Scores (PRSs)



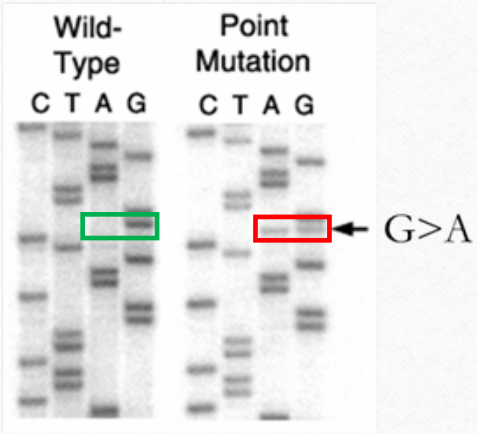
Torkamani *et al., Nat Rev Genet,* 2018

Sanger Sequencing (1977 – presently)

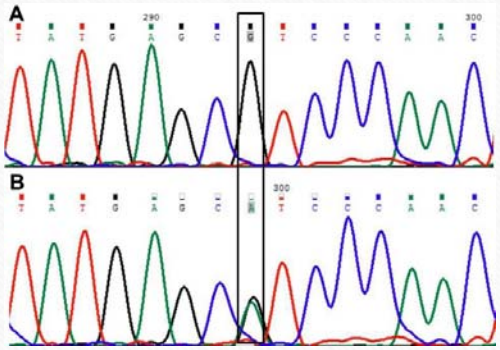
The first generation sequencing technology

Polyacrylamide gel electrophoresis and autoradiography

In the past



At present



Capillary electrophoresis in automated genetic analyzers with laser-based detection of fluorochromes

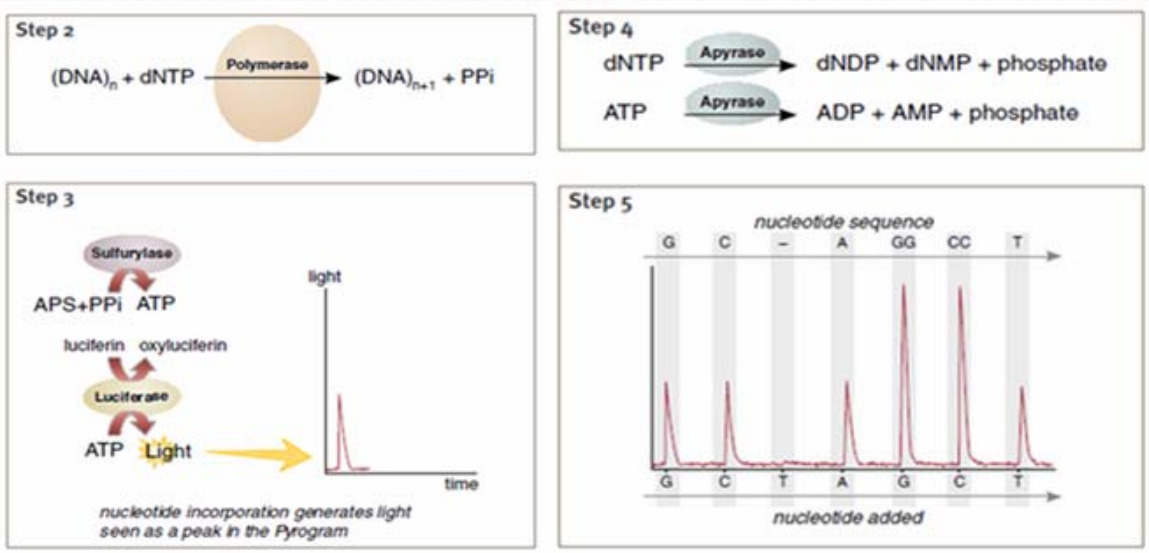
Sanger sequencing is still the method of choice for:

✓Small-scale projects
✓Targeted genotyping (exons, splice-sites, SNPs, indels, repeats)
✓Limited budgets
✓Long DNA fragments (up to 1000 bp)

✓Validation of NGS results !

Pyrosequencing (~1996 – 2013 discontinued)

The second generation sequencing technology



https://www.adelaide.edu.au/saef/new/whatis/

- Relies on the detection of PPi release (light emission) upon nucleotide incorporation and not on di-deoxynucleotide-based chain termination
- No gel electrophoresis or fragment separation procedure necessary → Faster
- Accuracy, flexibility, multiple processing
- But! Shorter sequences can be read (amplicon 100-300 bp, Pyrogram ~100 bp)

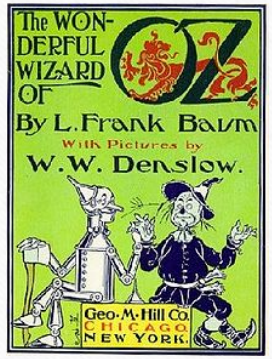
"Sequencing by synthesis" principle

Next Generation Sequencing

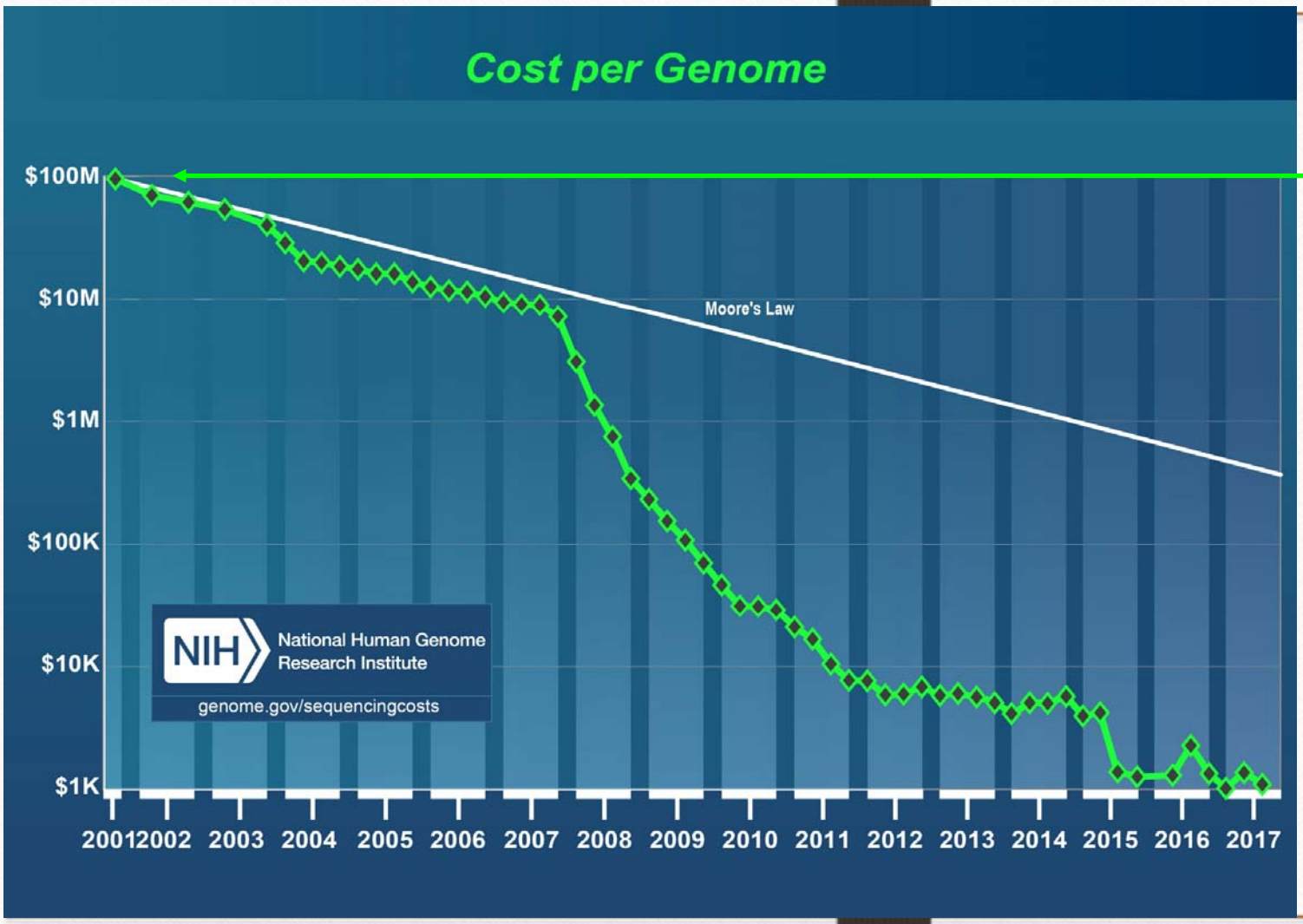
The next generation sequencing technology

FEELING WE'RE NOT
TOTO, I HAVE A FEE
ELING WE'RE NOT IN KANSAS ANYMORE
A FEELING
AVE A FEELING WE'RE NOT IN KANSAS A
ING WE'R
'RE NOT IN KANSAS ANYMORE
ELING WE'RE NOT IN KANS
I HAVE A FEELING WE
FEELING WE'R
NG WE'RE NOT IN KANSAS ANYMO

TOTO, I HAVE A FEELING WE'RE NOT IN KANSAS ANYMORE



The Wonderful Wizard of Oz, 1900, by L. Frank Baum



years

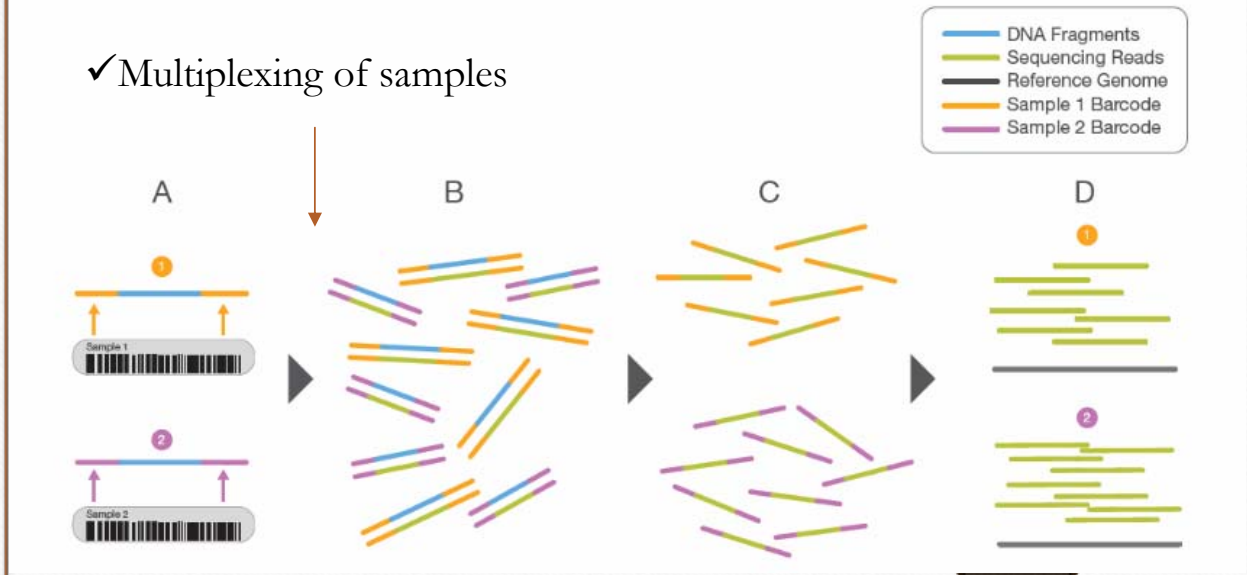
Genome sequencing

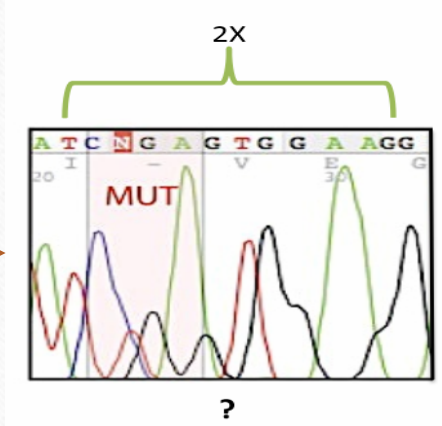
days

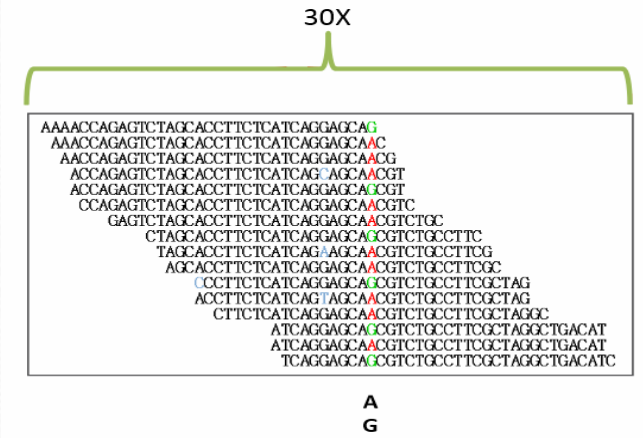
The next generation sequencing technology

NGS advantages:

- ✓High-throughput
- ✓Large amount of data
- ✓Rapid genome coverage
- ✓Depth of coverage (confidence of results – detection limit)
- ✓Deep-sequencing (eg tumor cells)
- ✓Multiplexing of samples



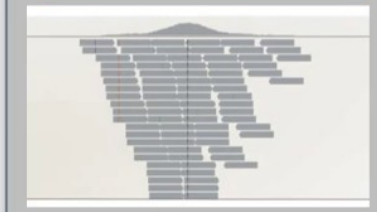




Illumina site: http://www.illumina.com/

Next Generation Sequencing

NGS PANEL



- Covers 10's-100's of genes
- Typically >99% completeness
- Exons +/- 15 bps into intron
- Requires capture before sequencing
- Higher density coverage of targeted regions
- Sanger fill-in

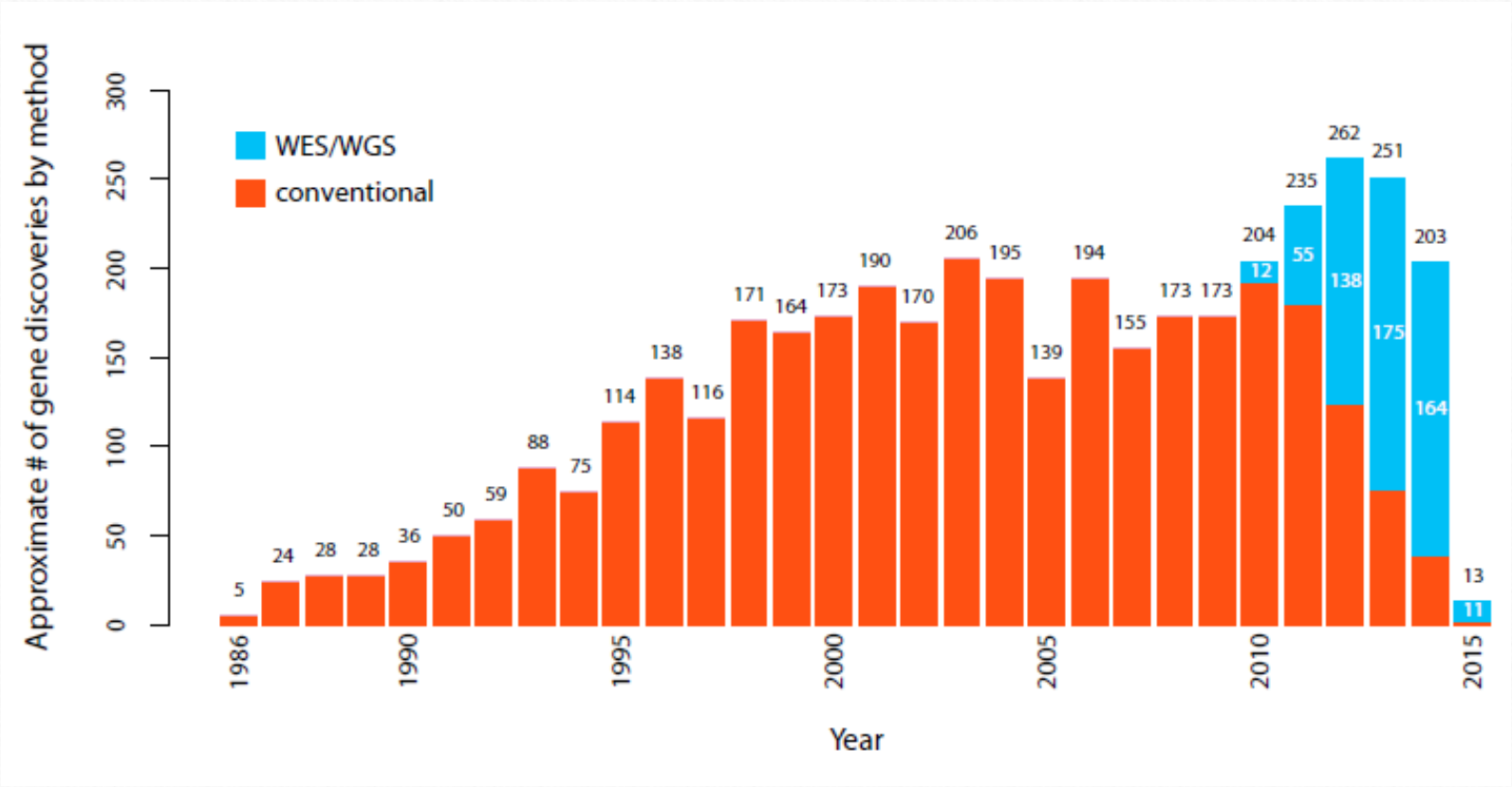
EXOME

- Covers ~1% of genome
- Typically ~97% completeness
- Exons +/- 15 bps into intron
- Requires capture before sequencing
- Moderate-High density coverage
- No Sanger fill-in

GENOME

- Covers ~95% of genome
- Exons, introns, promoters, 5'/ 3'UTR
- No capture required
- More complete and uniform coverage
- No Sanger fill-in

NGS has greatly improved our knowledge and understanding of the genetic basis of many disorders



Chong *et al.*, *Am J Hum Genet*, 2015

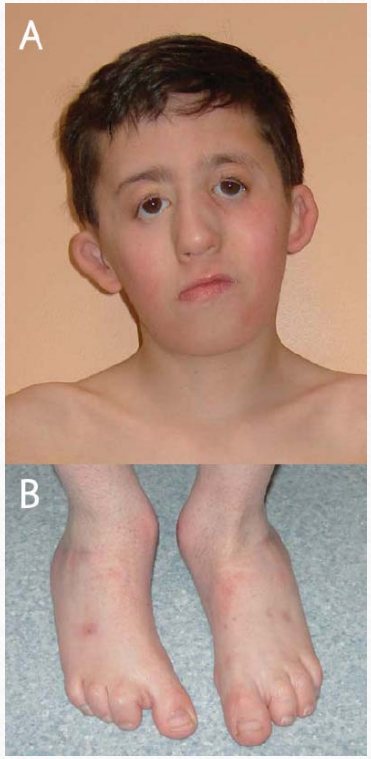
Exome sequencing identifies the cause of rare inherited syndromes

• *DHODH* encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway and harbors pathogenic variants associated with **Miller syndrome** (panels A, B).

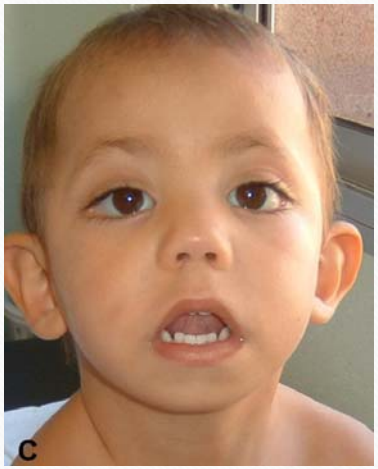
• *MLL2* encodes a Trithorax-group histone methyltransferase and harbors nonsense or frameshift mutations, associated with **Kabuki syndrome** (panel C).

• *PTPRJ* encodes a receptor-like PTP and harbors loss-of-function variants associated with **familial thrombocytopenia** (Marconi *et al.*, *Blood*, 2019)

• ……. many examples …….
https://media.nature.com/original/nature-assets/nrg/journal/v12/n11/extref/nrg3031-s1.pdf



Ng *et al.*, *Nat Genet*, 2010 (a)



Ng *et al.*, *Nat Genet*, 2010 (b)

Exome Sequencing Projects for complex traits and disorders

In the quest of the "missing heritability" of complex traits

NHLBI Grand Opportunity Exome Sequencing Project (ESP)

- Women's Health Initiative (WHI)
- Framingham Heart Study (FHS)
- Jackson Heart Study (JHS)
- Multi-Ethnic Study of Atherosclerosis (MESA)
- Atherosclerosis Risk in Communities (ARIC)
- Coronary Artery Risk Development in Young Adults (CARDIA)
- Cardiovascular Health Study (CHS)
- Genomic Research on Asthma in the African Diaspora (GRAAD)
- Lung Health Study (LHS)
- Pulmonary Arterial Hypertension (PAH) population
- Acute Lung Injury (ALI) cohort
- Cystic Fibrosis (CF) cohort

AGRE
Autism Genetic Resource Exchange

Autism Sequencing Consortium

T2D-GENES Consortium

Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples

CHARGE CONSORTIUM
COHORTS FOR HEART & AGING RESEARCH IN GENOMIC EPIDEMIOLOGY

PGC
Psychiatric Genomics Consortium

Making sense of NGS data

→ ~24.050 SNPs (exome-seq)

Bamshad *et al.*, *Nat Rev Genet*, 2011

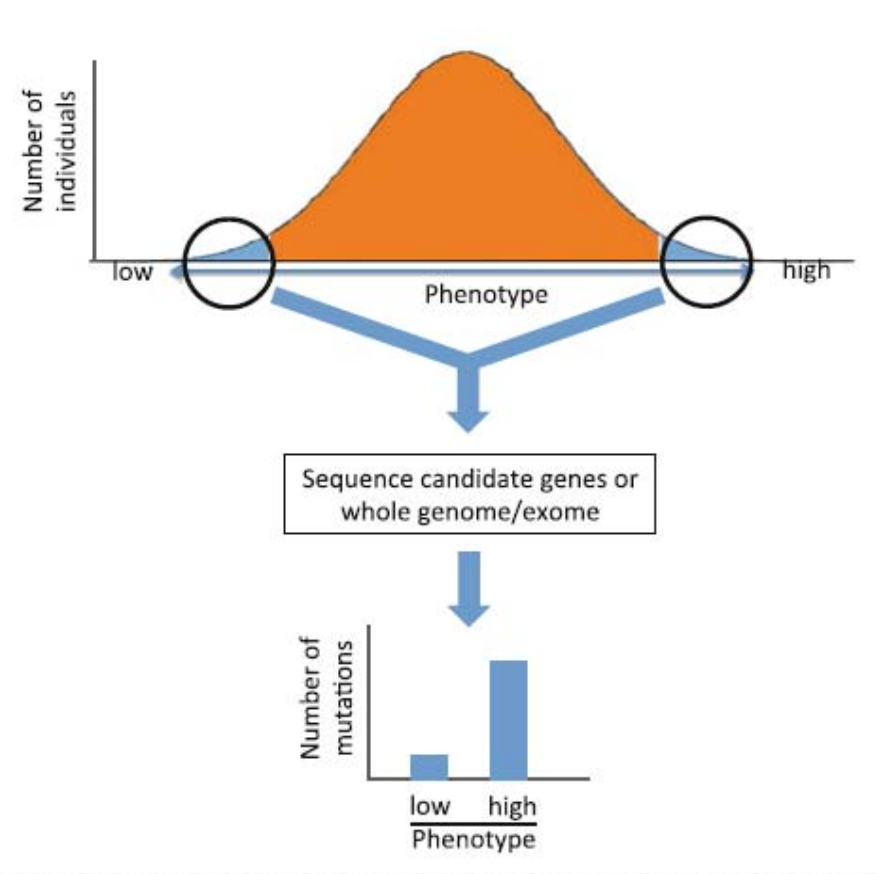
~2% of SNPs identified per individual by WES is **novel**

→ ~20.300 SNPs (exome-seq)

<u>Factors affecting the identification of causal alleles:</u>

- ❑ What is the **mode of inheritance** of a trait?
- ❑ Does **population structure** affect causal alleles?
- ❑ Does the phenotype arise ***de novo*** **or** due to **inherited** variants?
- ❑ **Locus heterogeneity** of a trait
- ❑ How large **sample size** in order to identify rare trait-associated alleles?
- ❑ What **analytical framework** to be used?

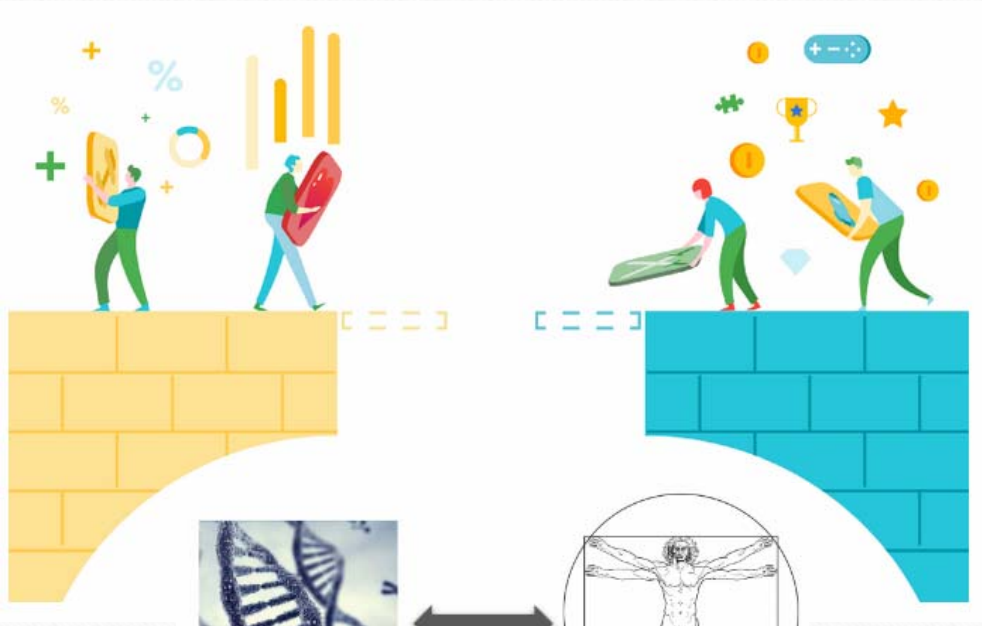
Alternative approach to explore rare variants using NGS



Brunham & Hayden, *Hum Genet*, 2013

- ✓ Some portion of the variability of common disease is likely due to rare variants in the same **genes that harbor common variants** associated with these traits.
- ✓ ***De novo* mutations** are the most extreme form of rare variants, since they may be "private" and have not been subject to selective pressure in previous generations.
- ✓ The "**extreme phenotype**" approach can be combined with WGS/WES to identify novel genes involved in complex traits.
- ✓ Similar to "extreme phenotype" is the "**endophenotype**" approach (mostly applicable in psychiatric and neurological research, ie bipolar disorder, schizophrenia, ADHD, Alzheimer's, etc): A trait within a phenotype/disorder with similar symptoms which gives the ability to differentiate between potential diagnoses, likely due to an underlying genetic component.

We need to bridge the knowledge gap from sequence to consequence



Genotype to Phenotype

Combining different methods to explore the genome … the way forward

‘Omics’ approaches

- Regulome → ChIP-seq/ATAC-seq
- Methylome → Methylation arrays
- Transcriptome (miRNAs, mRNAs) → RNA-seq & miRNA-seq
- Genome → Whole exome/genome sequencing

Comprehensive molecular signatures via integrated understanding of disease

Integrated ‘omics’ key benefits

- Disease-associated pathways
- Disease biomarkers
- Sub-phenotyping/disease subtyping
- Disease diagnosis/prognosis/ staging/survival
- Patient stratification
- Novel clinical interventions/ treatments
- Improved drug design
- Predictive/Precision medicine