

Strategies for finding disease genes.



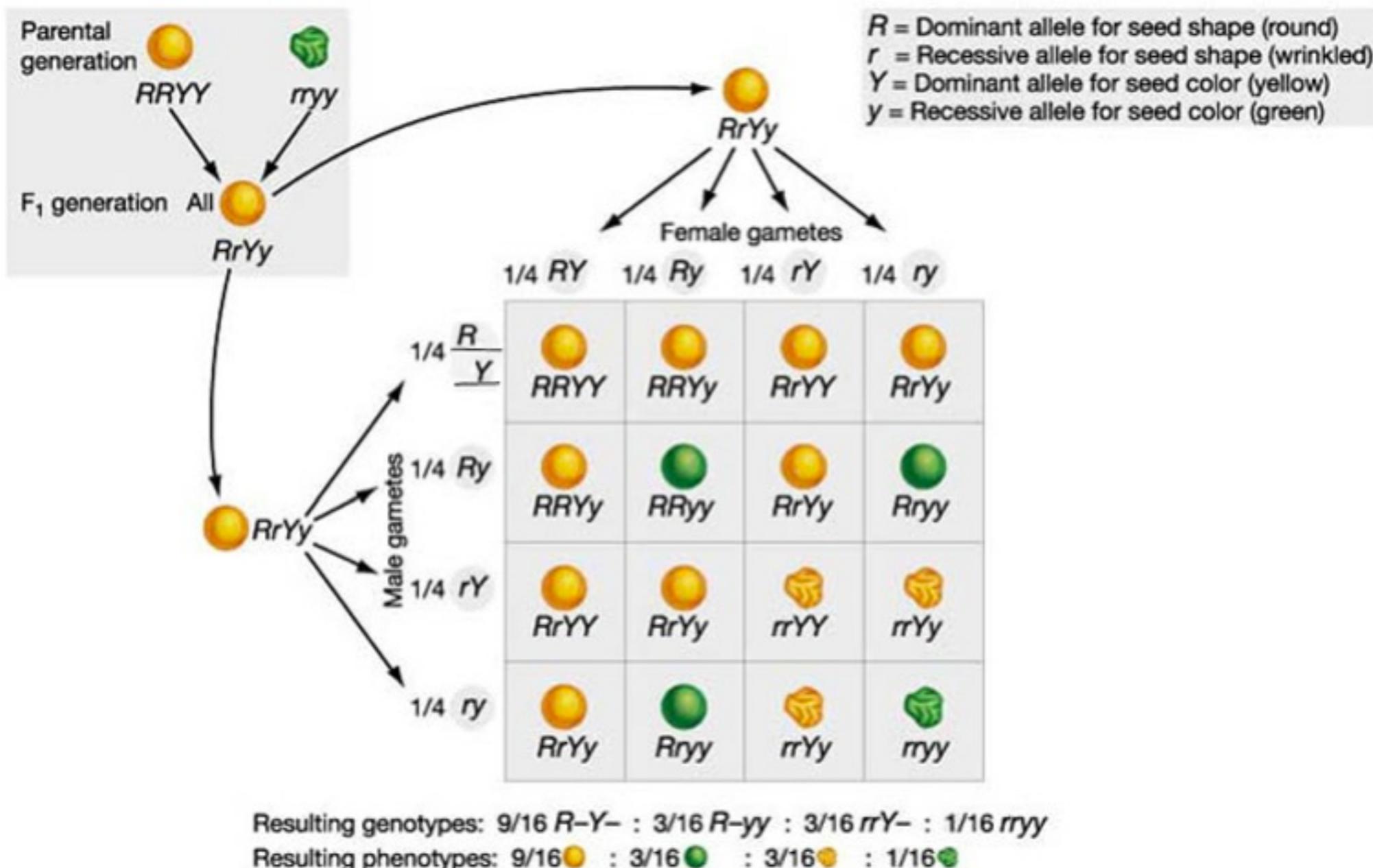
Aaron Quinlan

EGAG

3-Feb-2016

quinlanlab.org | aaronquinlan@gmail.com

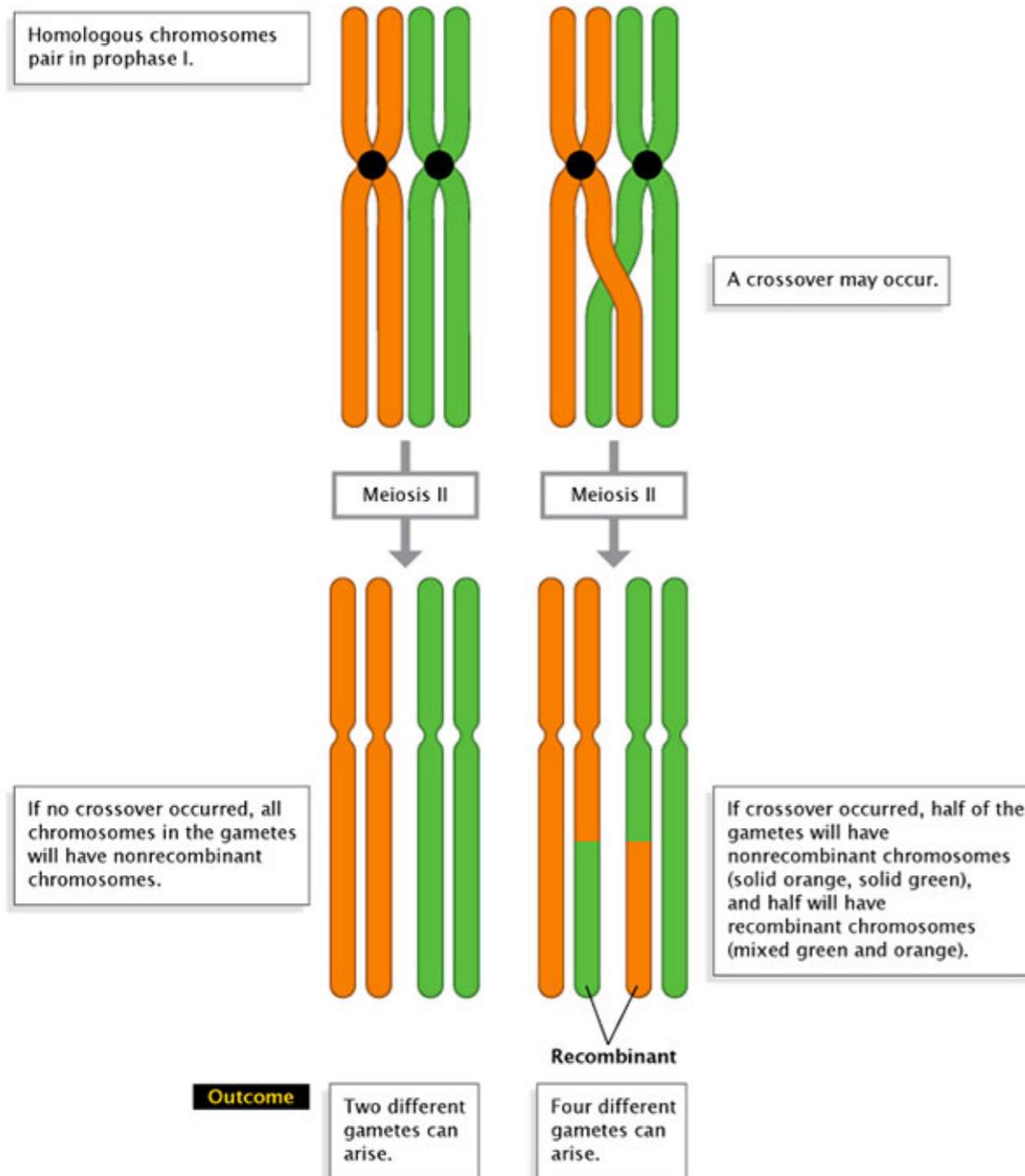
Mendel's second law “independent assortment of traits”



We are lucky that Mendel happened to select two traits that were not physically linked!



T.H. Morgan: chromosomal theory of inheritance

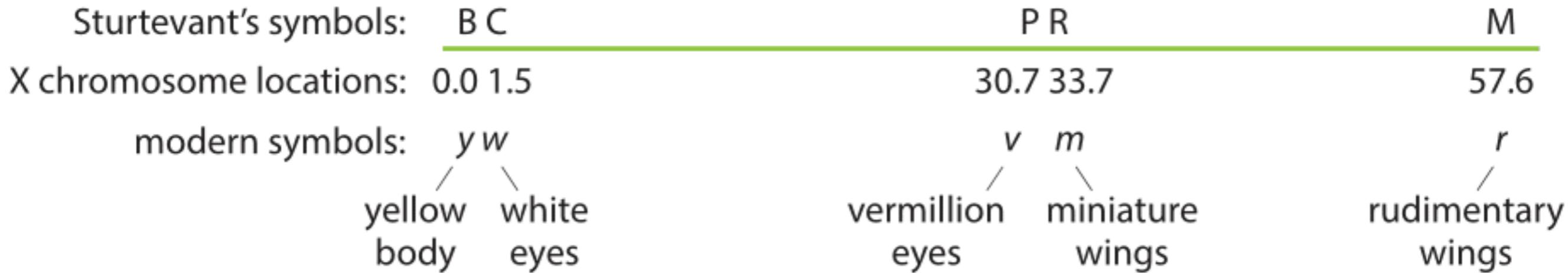


chromosomal theory of inheritance: that genes are located on chromosomes like beads on a string, and that some genes are linked (meaning they are on the same chromosome and always inherited together)



Alfred Henry Sturtevant
Photo courtesy of Cold Spring Harbor Laboratory Archives.

Sturtevant (19 yr old undergrad!): the first genetic map



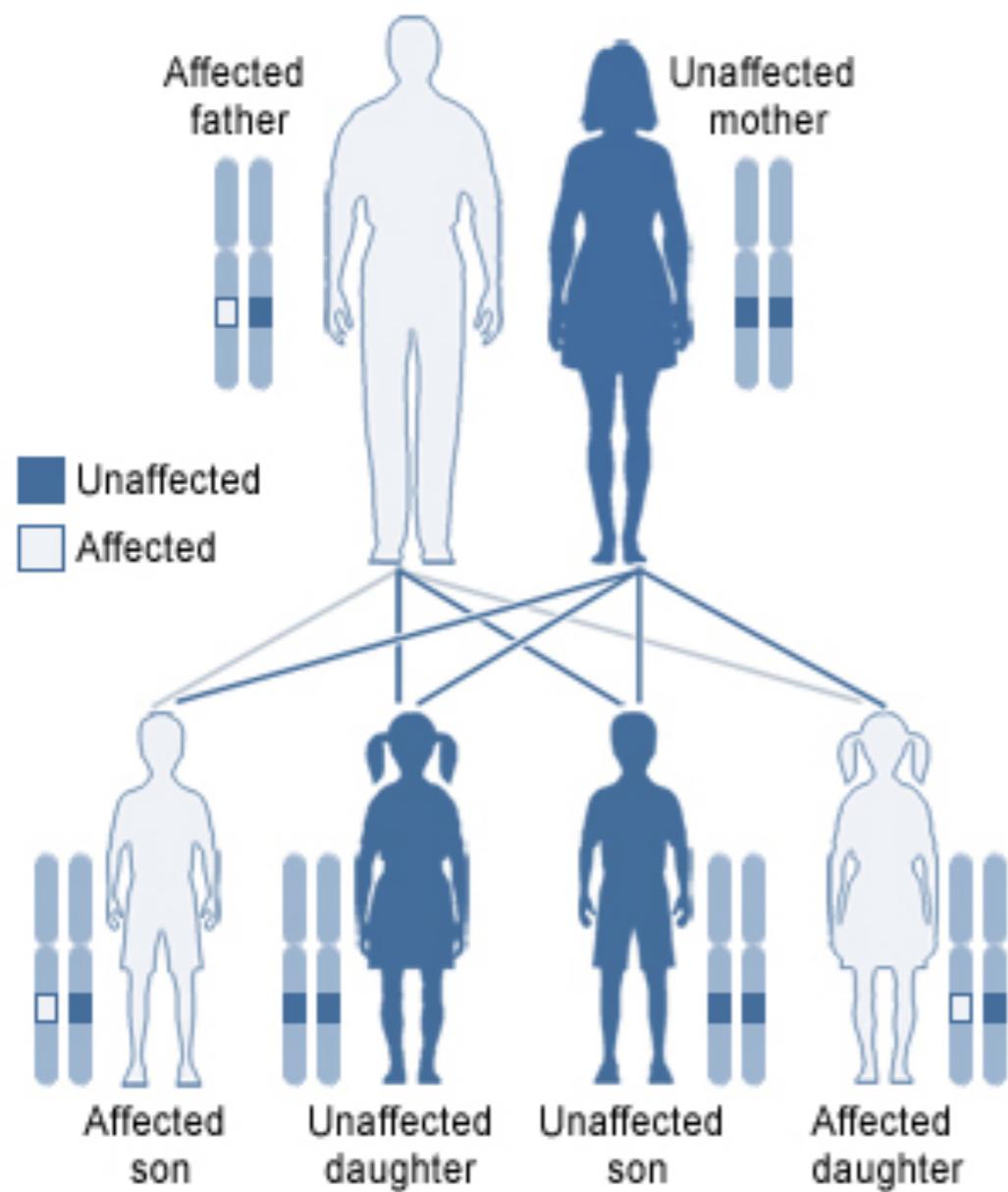
Sturtevant realized that if the frequency of crossing over was related to distance, one could use this information to map out the genes on a chromosome. After all, the **farther apart two genes were on a chromosome, the more likely it was that these genes would separate during recombination.**

Therefore, as Sturtevant explained it, the "**proportion of crossovers could be used as an index of the distance between any two factors**" (Sturtevant, 1913). Collecting a stack of laboratory data, Sturtevant went home and spent most of the night drawing the first chromosomal linkage map for the genes located on the X chromosome of fruit flies (Weiner, 1999).

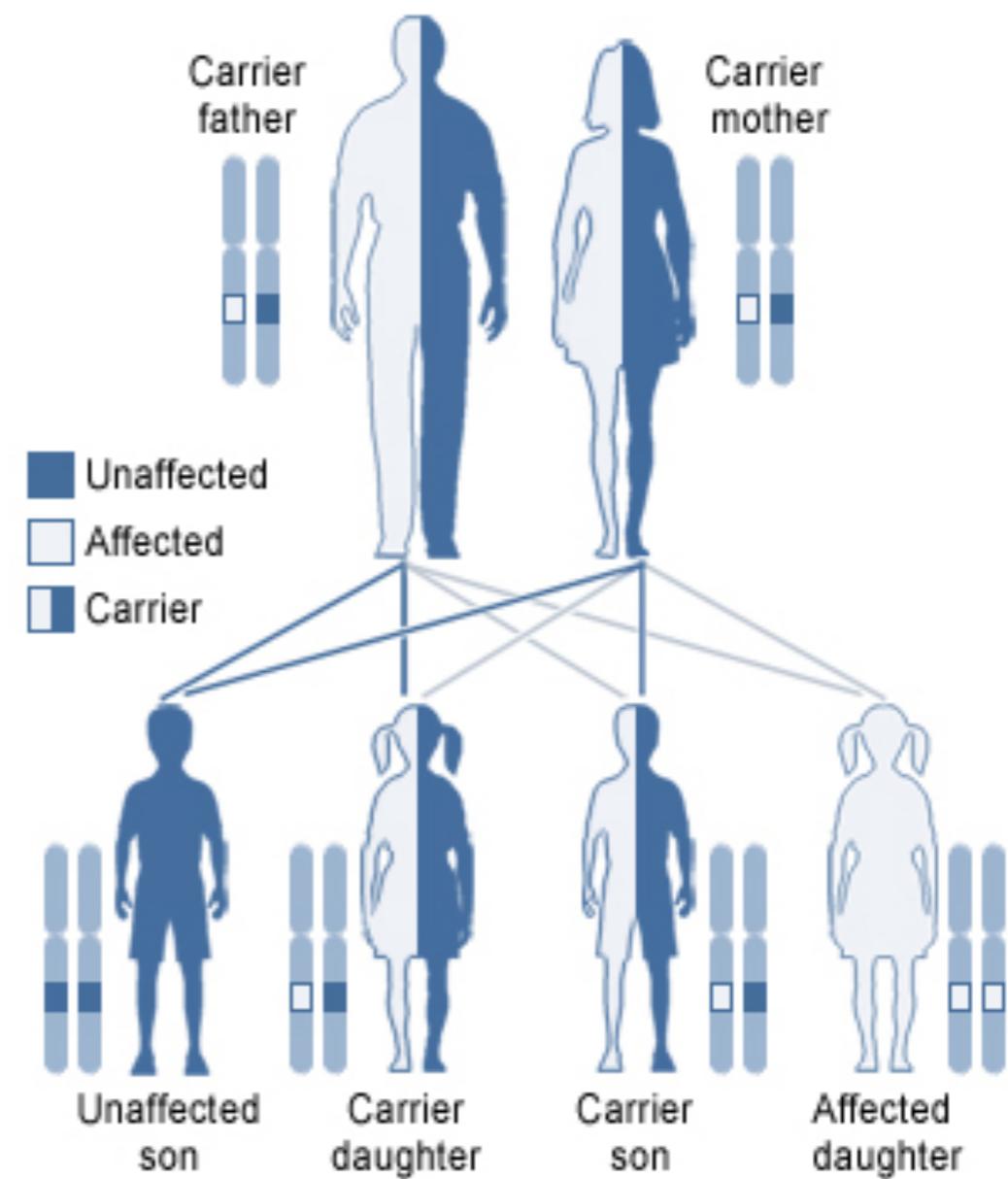
Sturtevant then worked out the order and the linear distances between these linked genes, thus forming a linkage map. **In doing so, he computed the distance in an arbitrary unit he called the "map unit," which represented a recombination frequency of 0.01, or 1%.** Later, the map unit was renamed the **centimorgan** (cM), in honor of Thomas Hunt Morgan.

Studying disease in families

Autosomal dominant



Autosomal recessive



U.S. National Library of Medicine

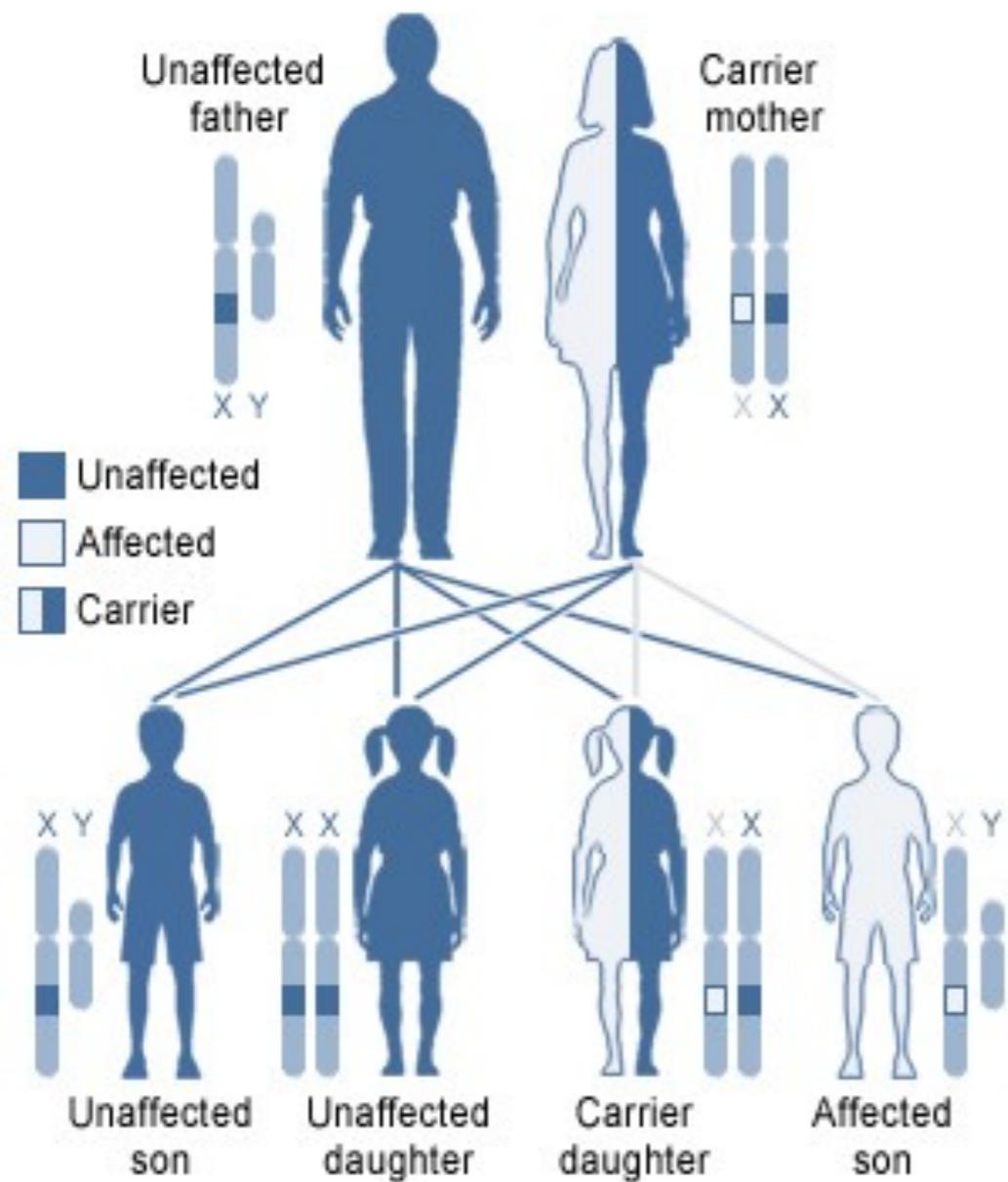
e.g., Huntington's disease

U.S. National Library of Medicine

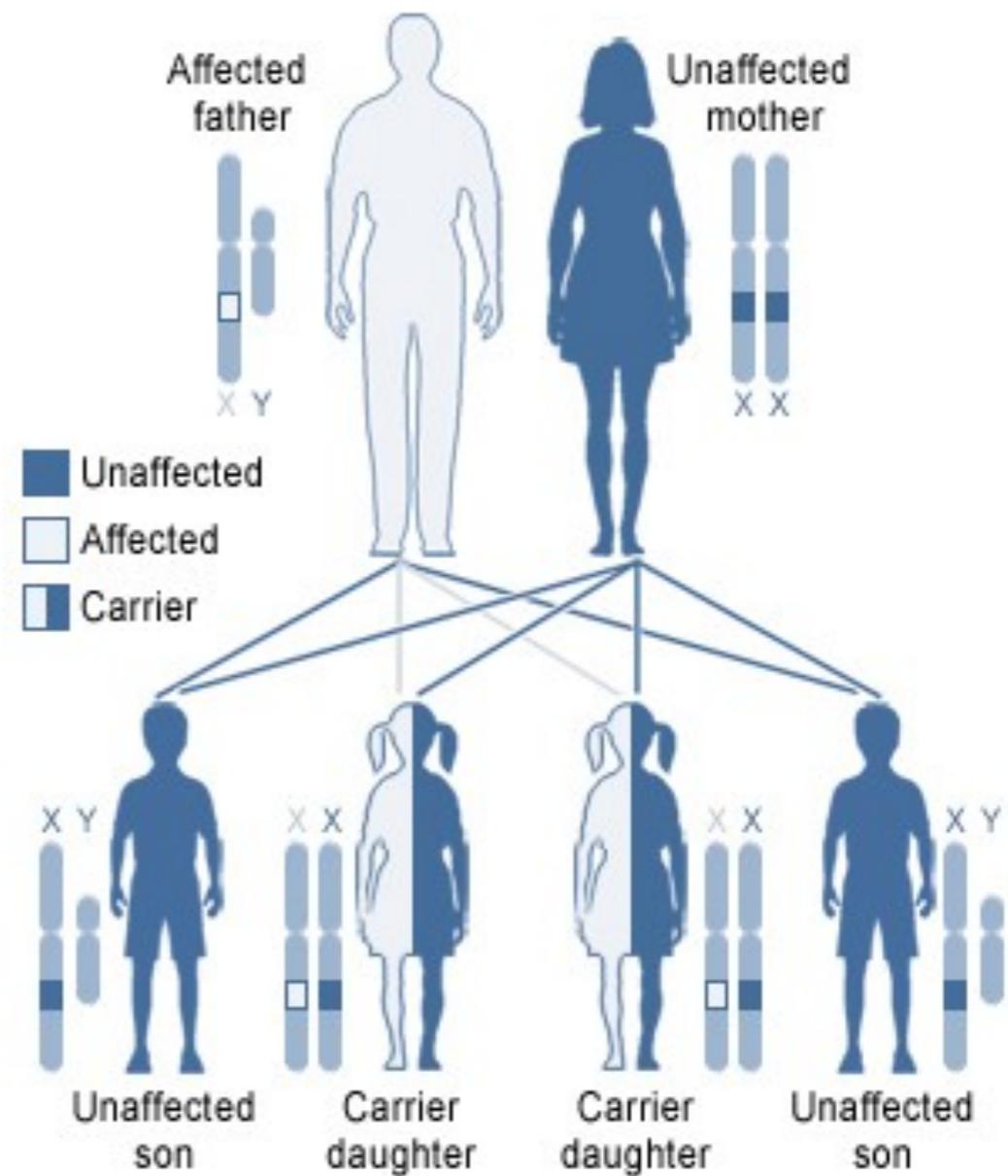
e.g., cystic fibrosis

Studying disease in families

X-linked recessive, carrier mother



X-linked recessive, affected father



Linkage mapping: use linkage to track down disease genes in families

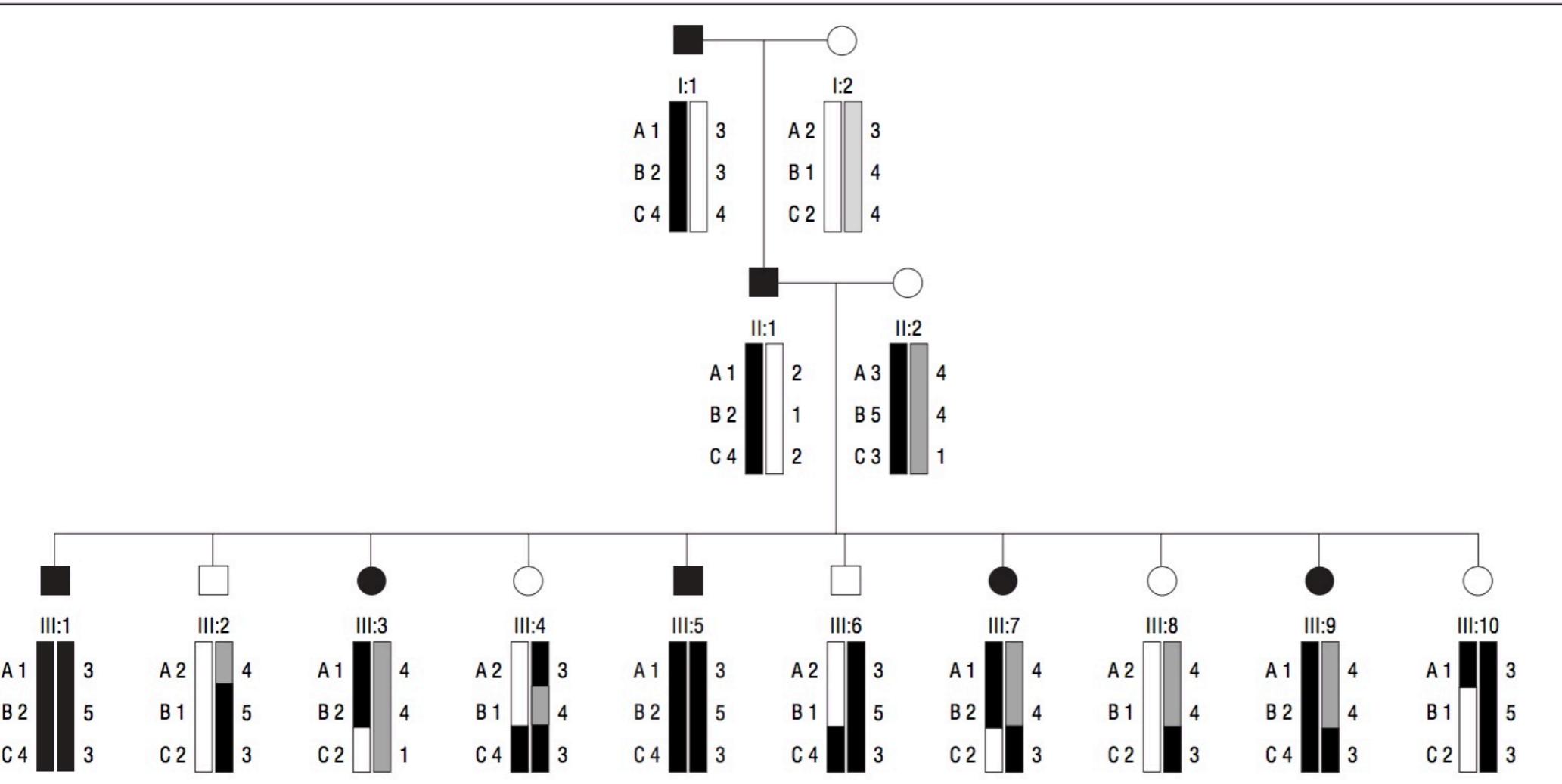


Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.

Linkage mapping: use linkage to track down disease genes in families

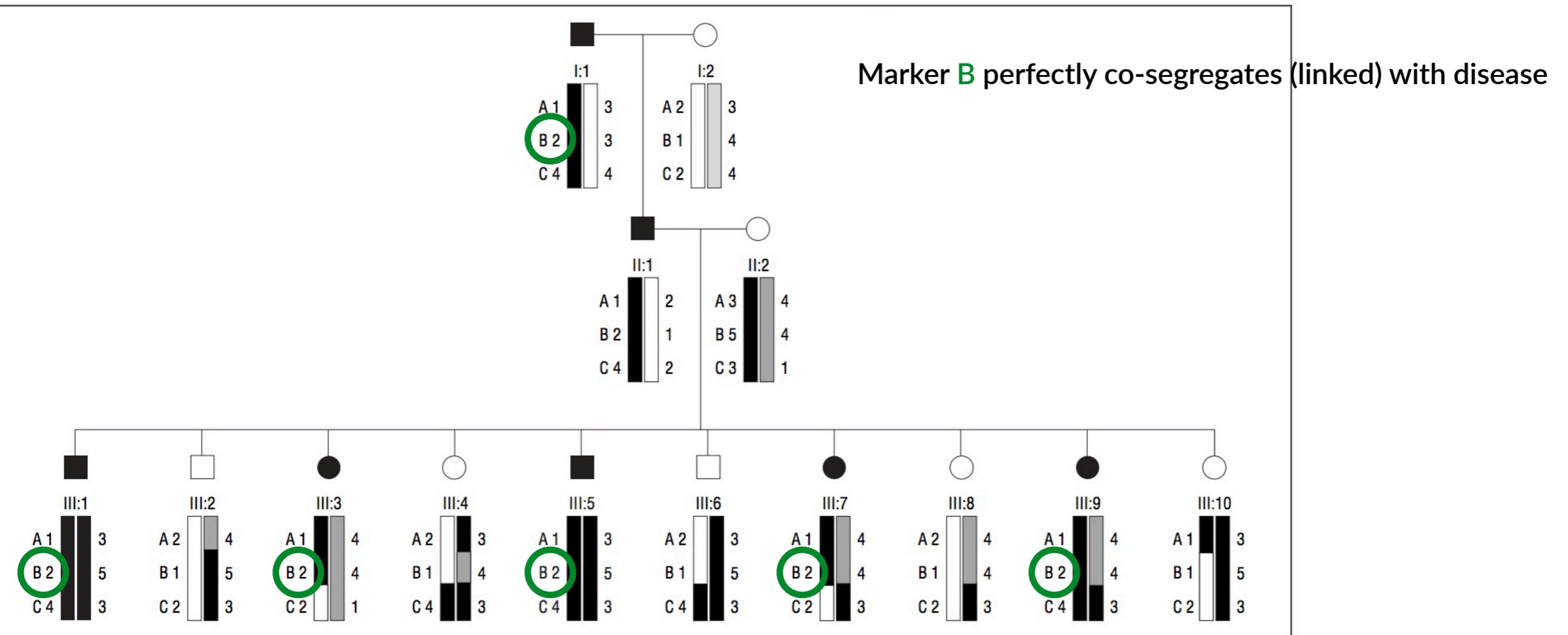


Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.

Linkage mapping: use linkage to track down disease genes in families

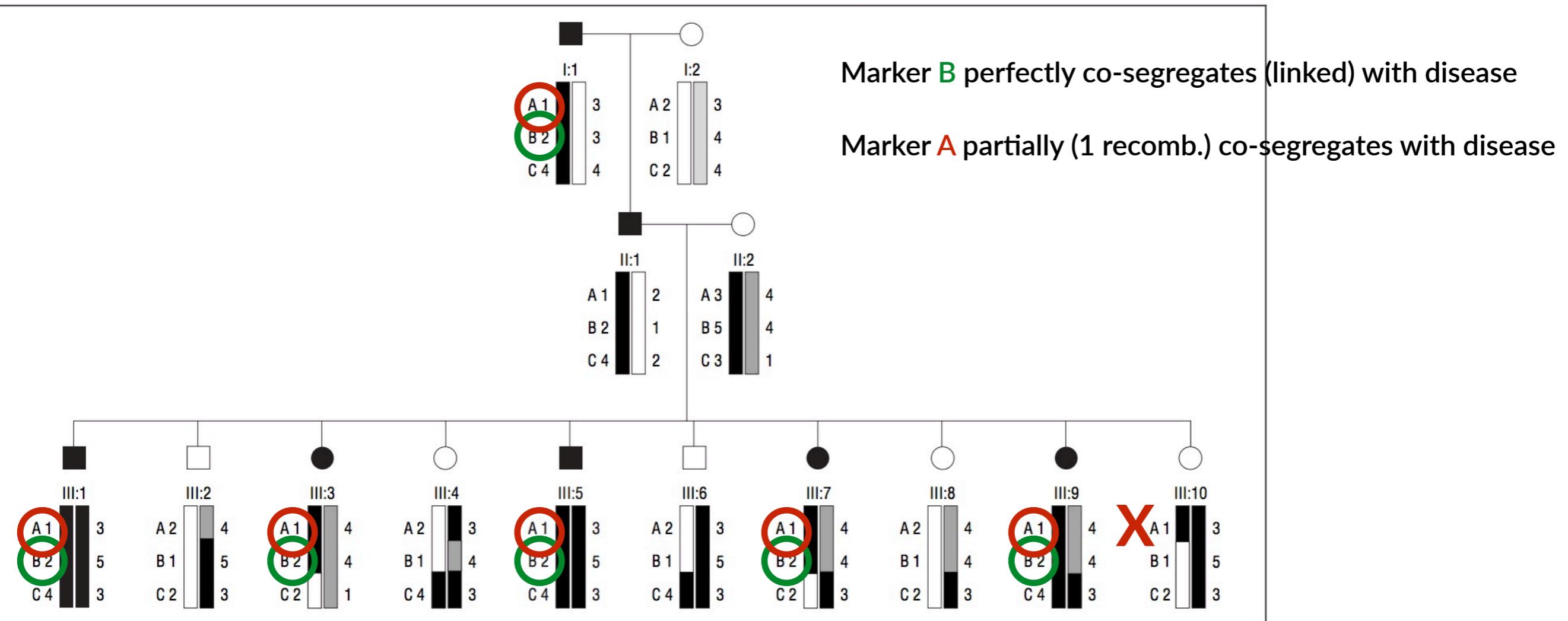


Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.

Linkage mapping: use linkage to track down disease genes in families

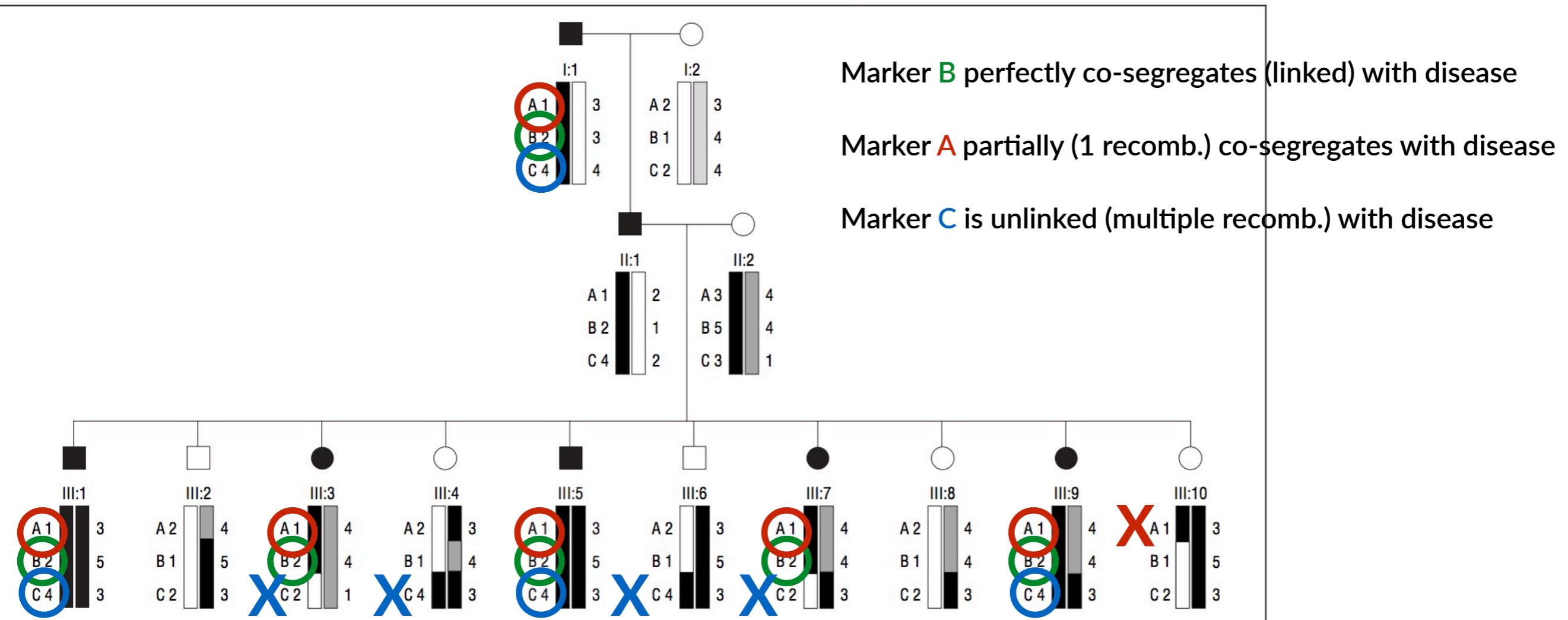
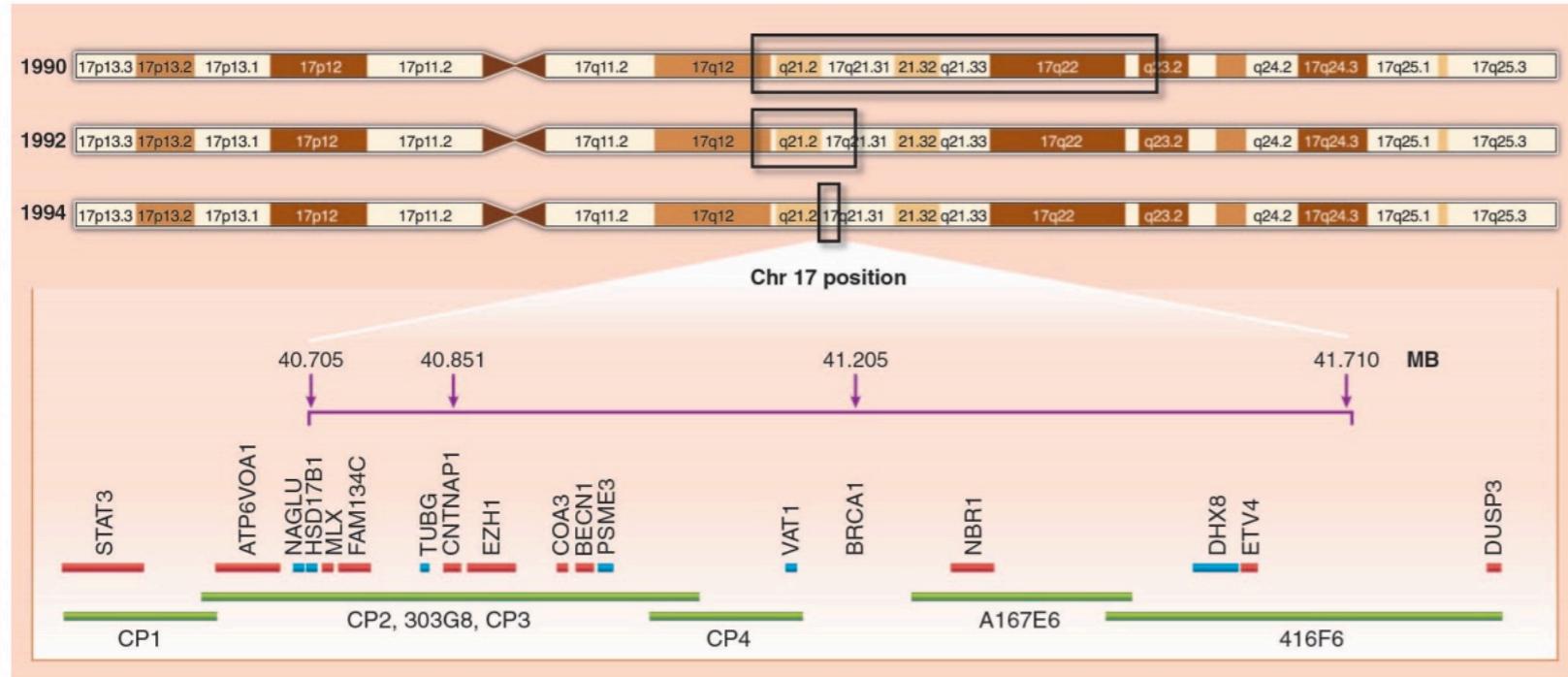


Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.

Linkage mapping successes



Science. 1994 Sep 30;265(5181):2088-90.

Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13.

Wooster R¹, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, et al.

⊕ Author information

Abstract

A small proportion of breast cancer, in particular those cases arising at a young age, is due to the inheritance of dominant susceptibility genes conferring a high risk of the disease. A genomic linkage search was performed with 15 high-risk breast cancer families that were unlinked to the BRCA1 locus on chromosome 17q21. This analysis localized a second breast cancer susceptibility locus, BRCA2, to a 6-centimorgan interval on chromosome 13q12-13. Preliminary evidence suggests that BRCA2 confers a high risk of breast cancer but, unlike BRCA1, does not confer a substantially elevated risk of ovarian cancer.

BRCA1, ~1994
M.C. King, Ray White, others

BRCA2, ~1994

Genetic Linkage Map of Six Polymorphic DNA Markers around the Gene for Familial Adenomatous Polyposis on Chromosome 5

M. G. Dunlop,* A. H. Wyllie,† Y. Nakamura,‡§ C. M. Steel,* H. J. Evans,* R. L. White,§ and C. C. Bird†

APC, ~1990
Ray White, C. C. Bird

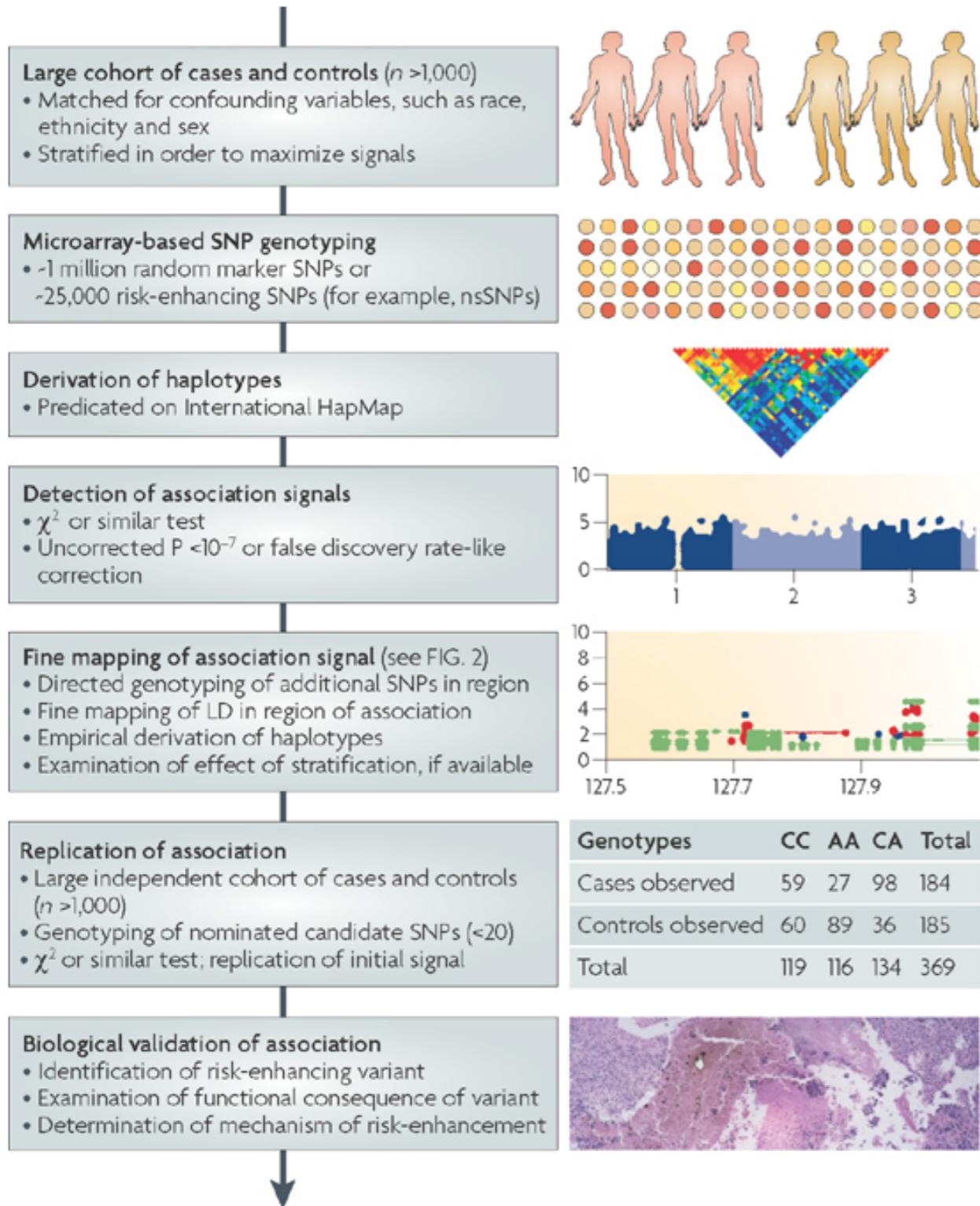
*Medical Research Council Human Genetics Unit, Western General Hospital; and †C. R. C. Laboratories, Department of Pathology, Edinburgh University Medical School, Edinburgh; ‡Division of Biochemistry, Cancer Institute, Tokyo; and §Howard Hughes Medical Institute and Department of Human Genetics, University of Utah Medical Center, Salt Lake City

Linkage mapping strengths and weaknesses

Pedigree diagrams illustrating inheritance of a trait:		
Property of mapping approach	Linkage analysis	Association analysis
Data type studied	Relatives	Unrelated or related individuals
Relevant parameter	Recombination fraction	Association statistic
Range of effect detected (linkage or association)	Long (≤ 5 Mb)	Short (≤ 100 kb)
Number of markers required for genome-wide coverage	Moderate (500–1,000)	Large (>100,000)
Statistics used	Cumbersome (requires tailor-made likelihood methods)	Elegant; can use the range of classical statistical tools
Dealing with correlated markers	Pose problems in presence of ungenotyped individuals	Can be handled efficiently
Biological basis of approach	Observe (or infer) recombination in pedigree data	Exploit unobserved recombination events in past generations
Dealing with allelic heterogeneity	Not a problem	Reduces power
Detecting genotyping errors	Potentially detected as Mendelian inconsistencies	Potentially detected only in family data, but not in case-control data
Most suitable application	Rare, dominant traits	Common traits

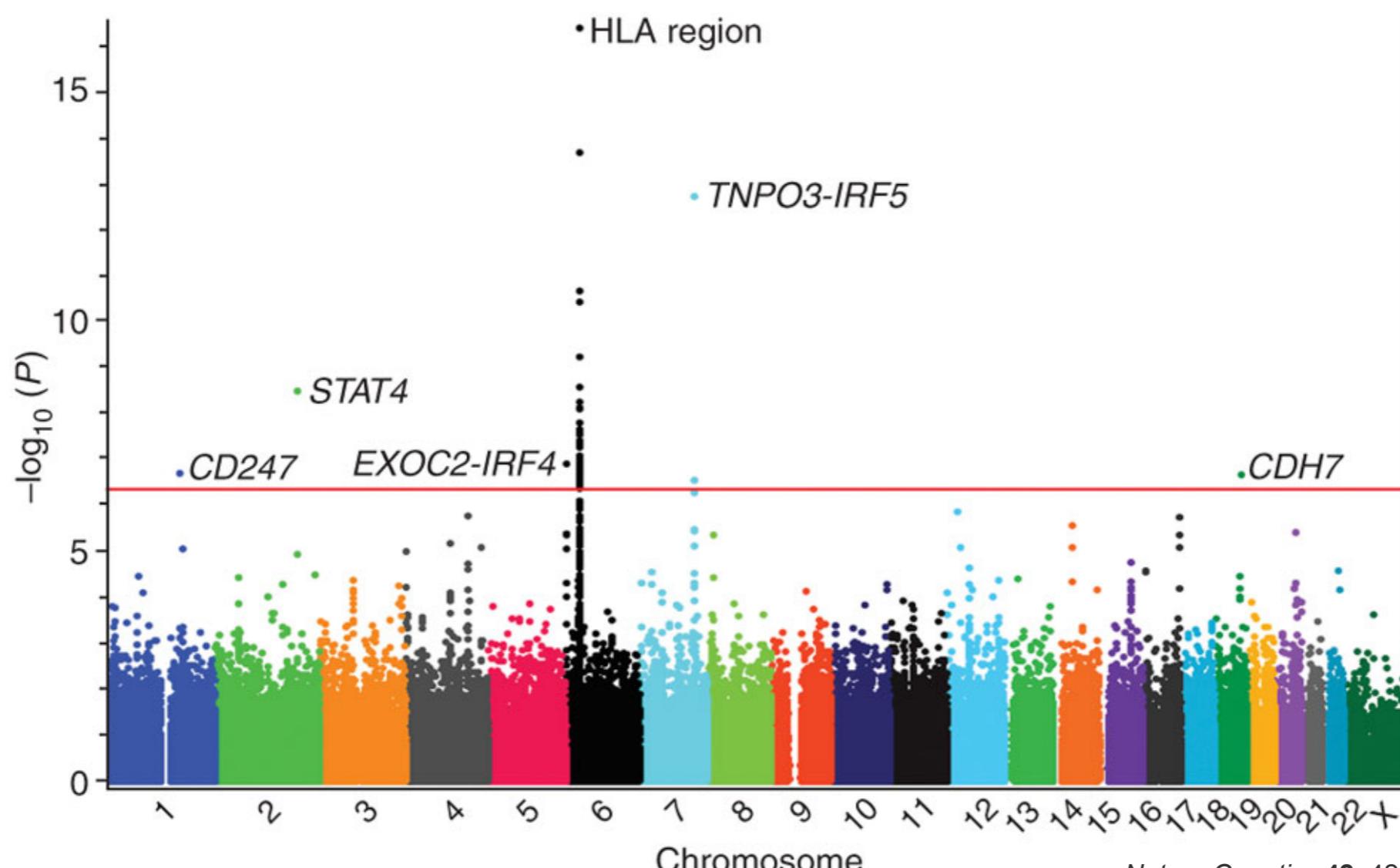
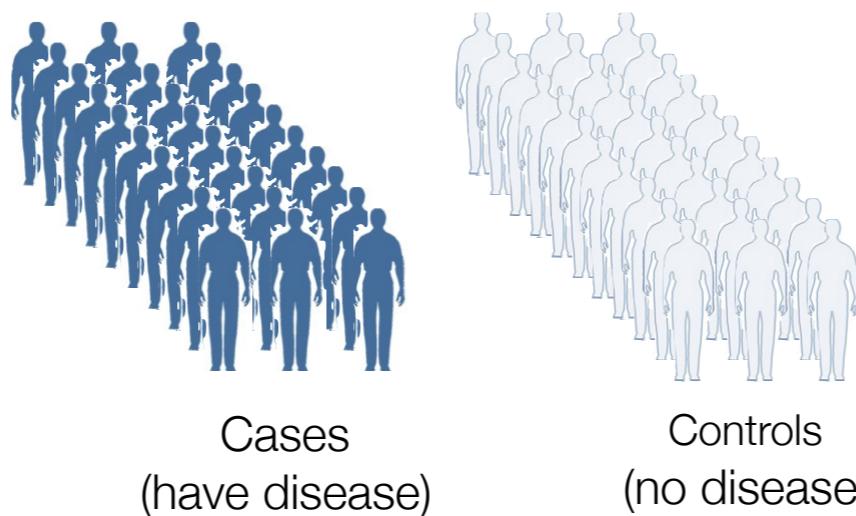
Nature Reviews | Genetics

Association Mapping



Common disease / common variant hypothesis

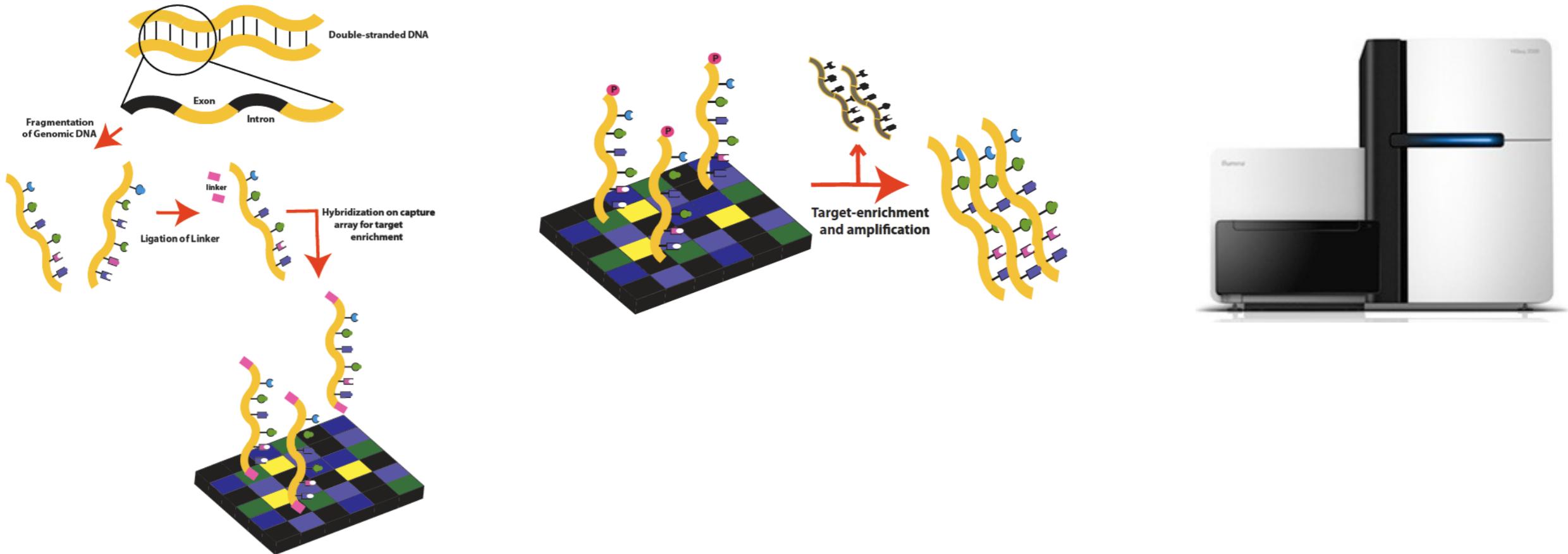
Genome-wide association studies



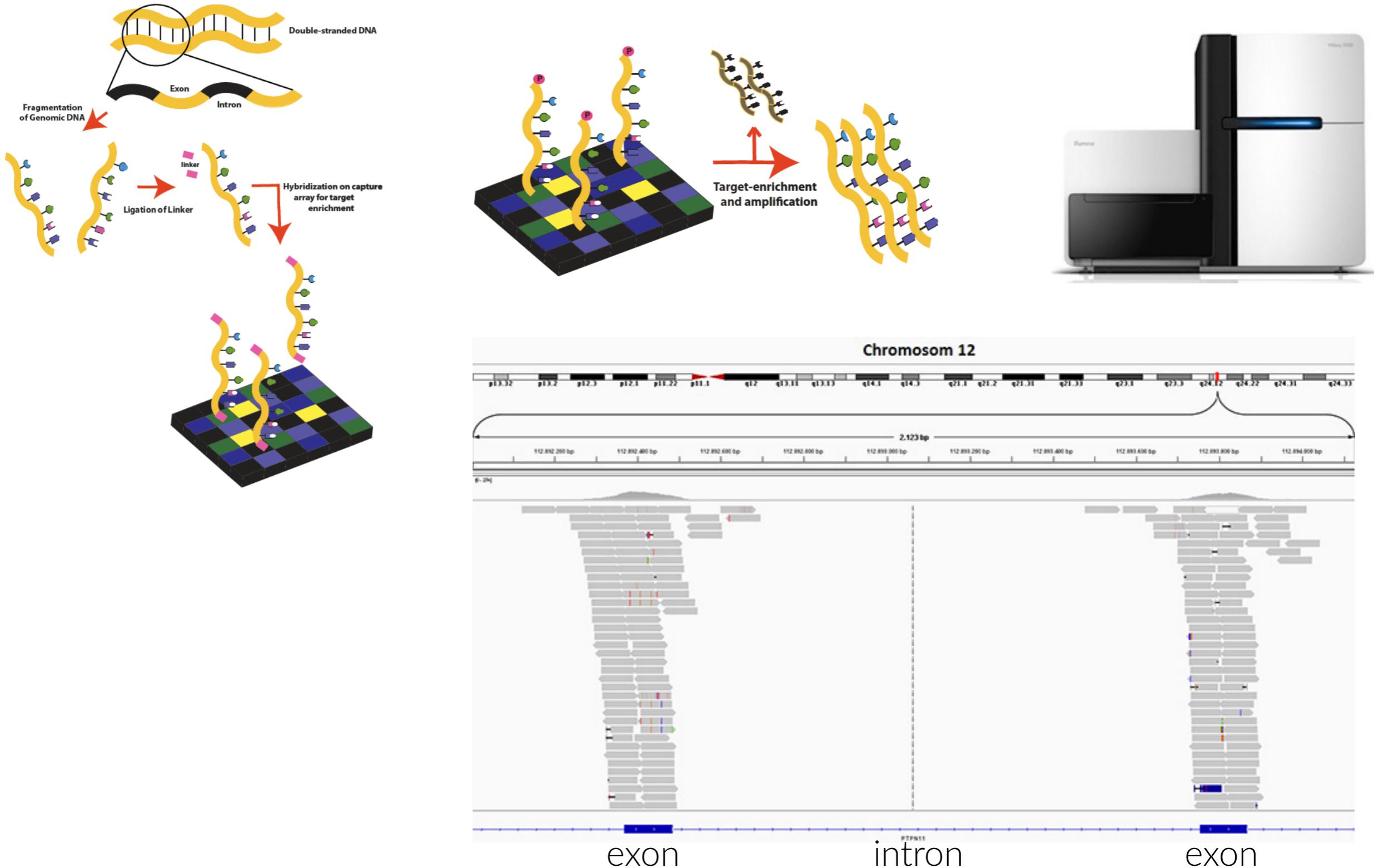
The advent of high-throughput DNA sequencing



Exome sequencing: sequence the protein-coding portion (2%) of the genome



Exome sequencing: sequence the protein-coding portion (2%) of the genome



Exome sequencing to solve Mendelian diseases

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng^{1,10}, Kati J Buckingham^{2,10}, Choli Lee¹, Abigail W Bigham², Holly K Tabor^{2,3}, Karin M Dent⁴, Chad D Huff⁵, Paul T Shannon⁶, Ethylin Wang Jabs^{7,8}, Deborah A Nickerson¹, Jay Shendure¹ & Michael J Bamshad^{1,2,9}

Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome

Sarah B Ng^{1,7}, Abigail W Bigham^{2,7}, Kati J Buckingham², Mark C Hannibal^{2,3}, Margaret J McMillin², Heidi I Gildersleeve², Anita E Beck^{2,3}, Holly K Tabor^{2,3}, Gregory M Cooper¹, Heather C Mefford², Choli Lee¹, Emily H Turner¹, Joshua D Smith¹, Mark J Rieder¹, Koh-ichiro Yoshiura⁴, Naomichi Matsumoto⁵, Tohru Ohta⁶, Norio Niikawa⁶, Deborah A Nickerson¹, Michael J Bamshad¹⁻³ & Jay Shendure¹



Now also the Broad Institute (MacArthur)

Identifying rare variation in the human genome and exome

1000 Genomes
A Deep Catalog of Human Genetic Variation

~2500 WGS samples



NHLBI Grand Opportunity Exome Sequencing Project (ESP)

~6500 WES samples

ExAC Browser (Beta) | Exome Aggregation Consortium

~65000 WES samples.

Search for a gene or variant or region

Examples - Gene: PCSK9, Transcript: ENST00000407236, Variant: 22-46615880-T-C, Multi-allelic variant: rs1800234, Region: 22:46615715-46615880

About ExAC

The [Exome Aggregation Consortium \(ExAC\)](#) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed [here](#).

All data here are released under a [Fort Lauderdale Agreement](#) for the benefit of the wider biomedical community - see the terms of use [here](#).

Sign up for our mailing list for future release announcements [here](#).

Recent News

January 13, 2015

- Version 0.3 ExAC data and browser (beta) is released! ([Release notes](#))

October 29, 2014

- Version 0.2 ExAC data and browser (beta) is released! Sign up for our mailing list for future release announcements [here](#).

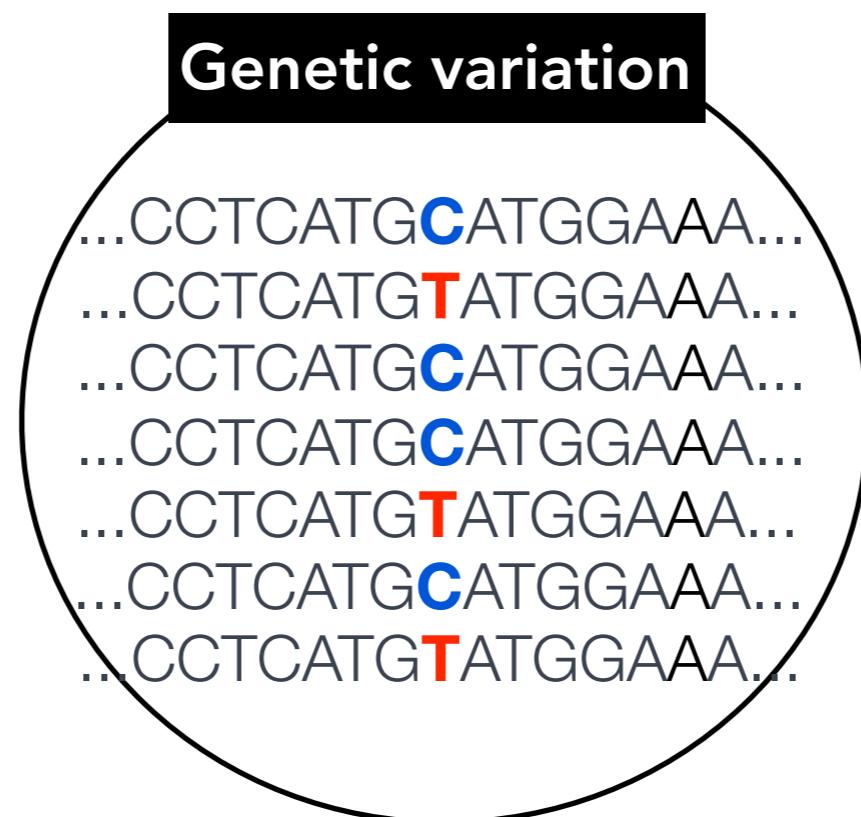
October 20, 2014

- Public release of ExAC Browser (beta) at ASHG!

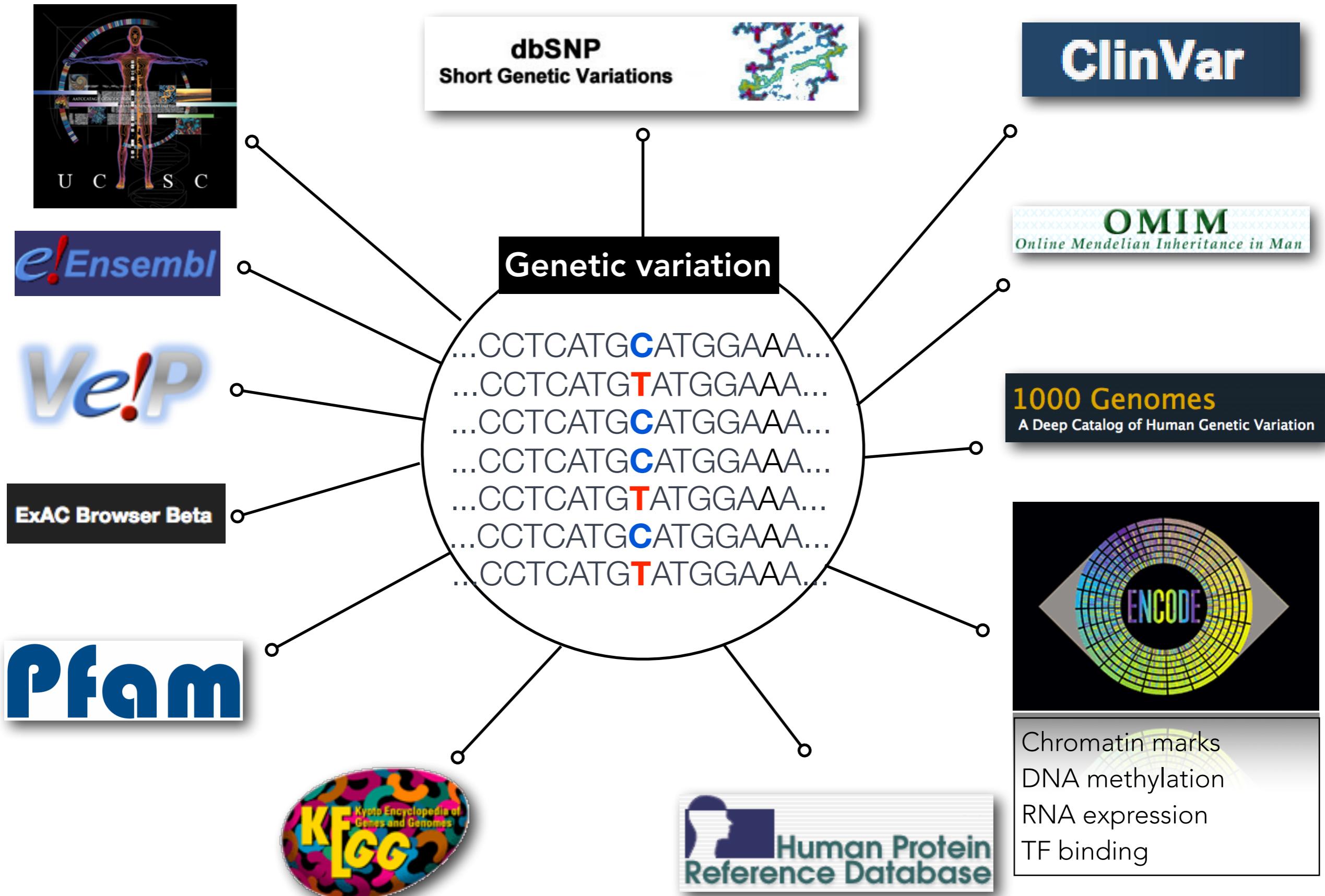
October 15, 2014

- Internal release to consortium now available!

Case study: exome sequencing of a familial disease



Case study: exome sequencing of a familial disease



Case study: exome sequencing of a familial disease

Step 1: annotated functional consequence

Impact sometimes hard to predict.

synonymous (silent)

	L	Q	T
Normal	ctg	cag	act
Mutated	ctg	caa	act
	L	Q	T

non-synonymous (missense)

	L	Q	T
Normal	ctg	cag	act
Mutated	ctg	cgg	act
	L	R	T

stop-gain (nonsense)

	L	Q	T
Normal	ctg	cag	act
Mutated	ctg	tag	act
	L	STOP	T

stop-loss

	L	STOP	T
Normal	ctg	tag	act
Mutated	ctg	cag	act
	L	Q	T

Case study: exome sequencing of a familial disease

Variant Effect Predictor

The VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. Simply input the coordinates of your variants and the nucleotide changes to find out the:

- **genes and transcripts** affected by the variants
- **location** of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- **consequence** of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- **known variants** that match yours, and associated minor allele frequencies from the **1000 Genomes Project**
- **SIFT** and **PolyPhen** scores for changes to protein sequence
- ... And [more!](#)

SnpEff

Genetic variant annotation and effect prediction toolbox.

[Download SnpEff](#)

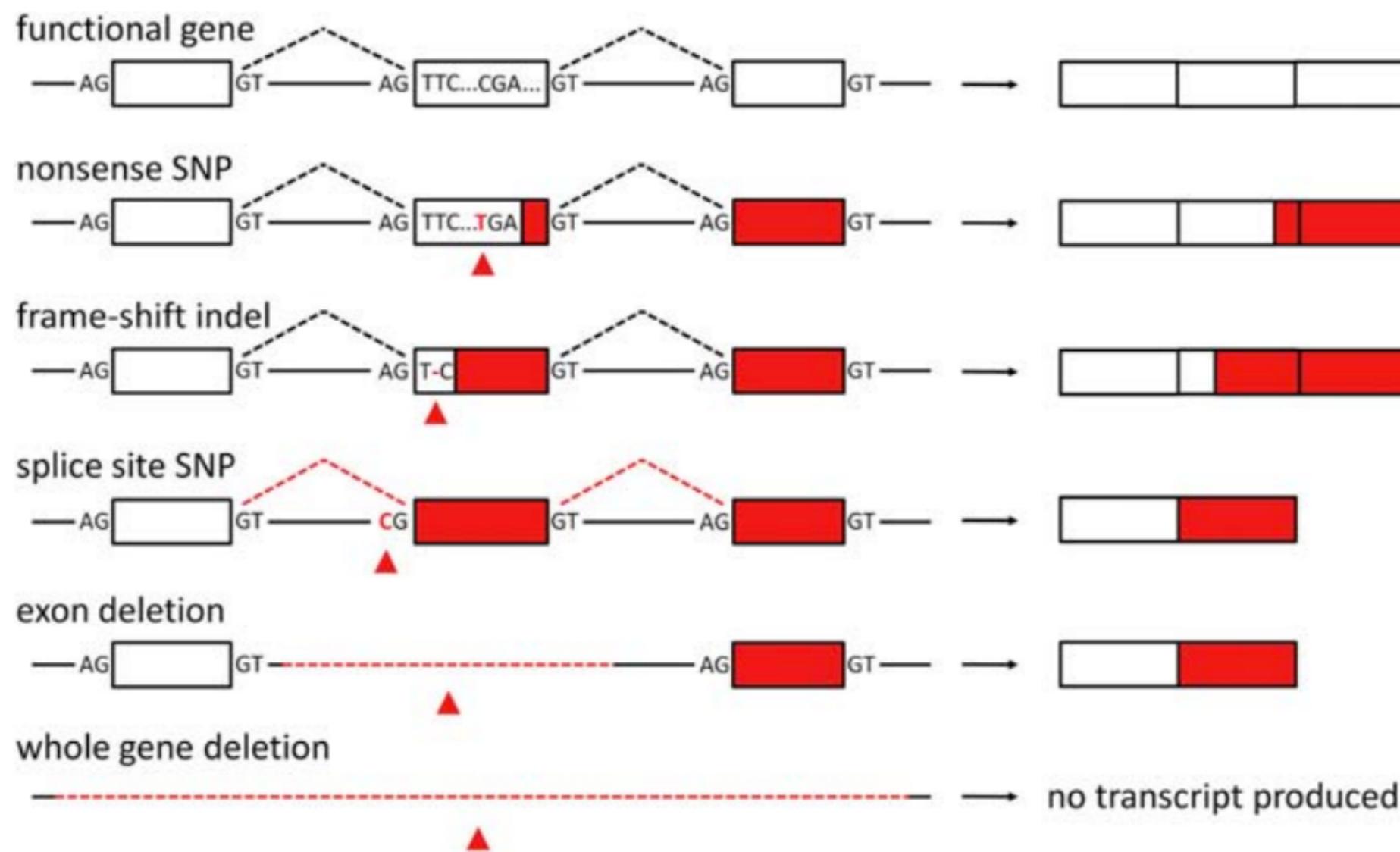
Important: This version implements the [new VCF annotation standard 'ANN' field](#).

Latest version 4.2 (2015-12-05)
Requires Java 1.7

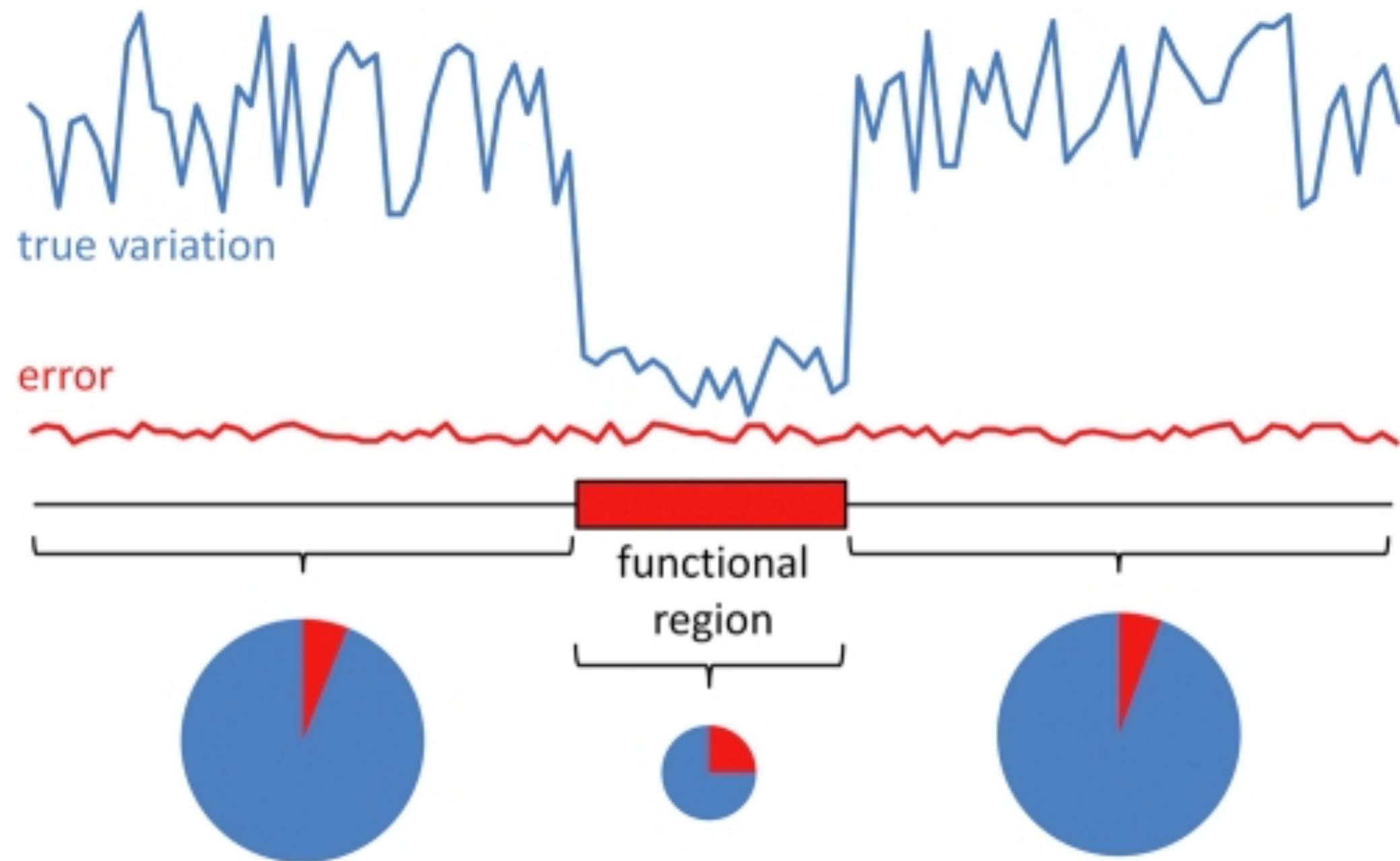
ANNOVAR Documentation

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast and many others). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

Loss of function mutations

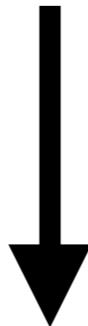


Interesting things are more likely to be wrong b/c they are rare

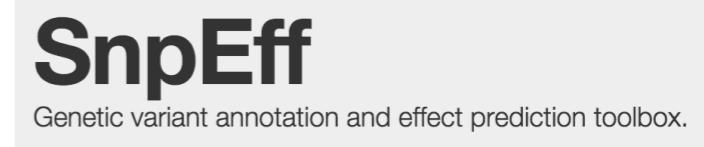


Many tools + many transcript annotations = many answers

Transcripts



Tools



Variant Annotation Tool

A computational framework to functionally annotate variants in personal genomes using a cloud-computing environment.



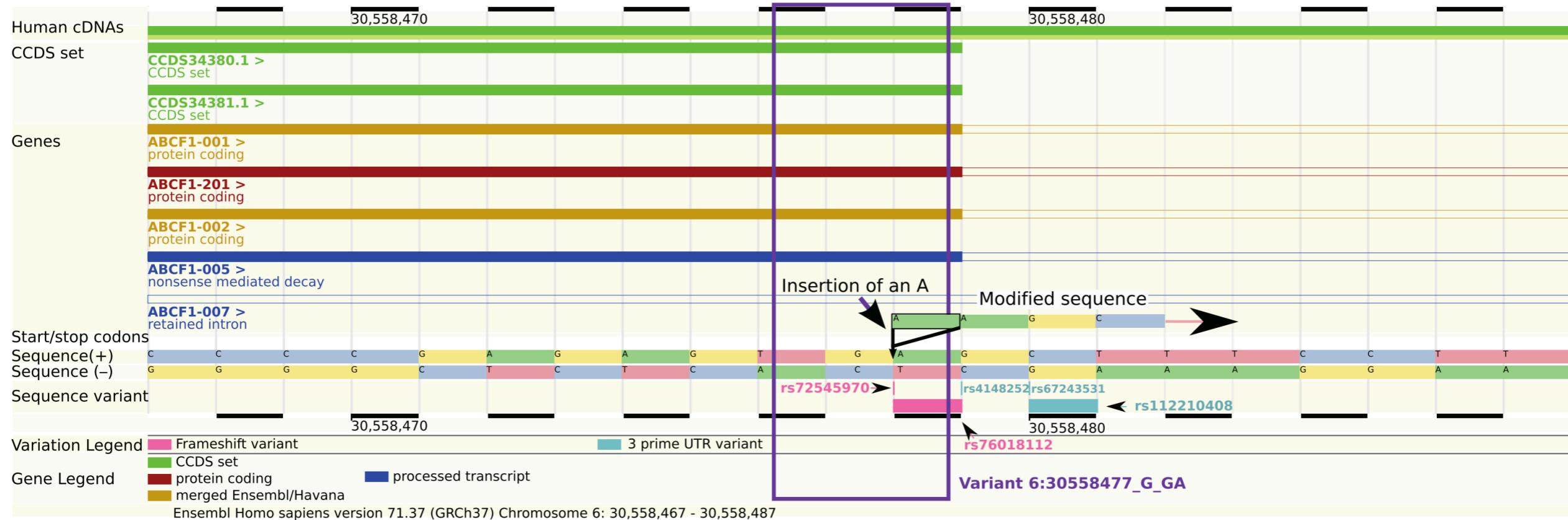
VAT (Yandell; part of VAAST)

Annotation software matters

Table 2 Same transcripts, different software: ANNOVAR and VEP annotations for exonic variants

	ANV+VEP	ANV	VEP	Exact match	Category match	ANV match rate (%)	VEP match rate (%)	Overall category match rate (%)
LOF total	104,915	77,527	96,761	68,284	69,373	88.08	70.57	66.12
Frameshift	19,021	15,822	16,685	13,486	-	85.24	80.83	-
Stop gained	16,758	14,960	16,146	14,348	-	95.91	88.86	-
Stop lost	1,113	906	1,077	870	-	96.03	80.78	-
All splicing	69,112	45,839	62,853	39,580	-	86.35	62.97	-
MISSENSE total	350,806	324,242	347,752	318,056	321,188	98.09	91.46	91.56
Inframe indel	9,455	8,650	6,600	5,795	-	66.99	87.80	-
Missense	343,284	315,592	339,953	312,261	-	98.94	91.85	-
Initiator codon	1,199	0	1,199	0	-	-	0.00	-
SYNONYMOUS and								
OTHER CODING total	182,120	172,463	175,483	165,643	165,826	96.05	94.39	91.05
Synonymous	181,873	172,463	175,053	165,643	-	96.05	94.62	-
Stop retained	203	0	203	0	-	-	0.00	-
Other coding	227	0	227	0	-	-	0.00	-
ALL LOF	104,915	77,527	96,761	68,284	69,373	88.08	70.57	66.12
ALL LOF and MISSENSE	455,721	401,769	444,513	386,340	390,561	96.16	86.91	85.70
ALL EXONIC	637,841	574,232	619,996	551,983	556,387	96.13	89.03	87.23

Example of annotation complexity



Insertion of a single A. What is the impact?

- a. frameshift
 - b. stopgain
 - c. synonymous
 - d. yes

An allele underlying a rare disease should be rare!

- Always filter by frequency separately in every available population
 - do NOT filter for frequency in only one population
 - do NOT filter on average worldwide frequency
- If variant causes severe phenotype, should ALWAYS be rare in every population



Rare disease discovery with **GEMINI** (Genome Mining)



OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations

Umadevi Paila¹, Brad A. Chapman², Rory Kirchner², Aaron R. Quinlan^{1*}



Uma
Paila



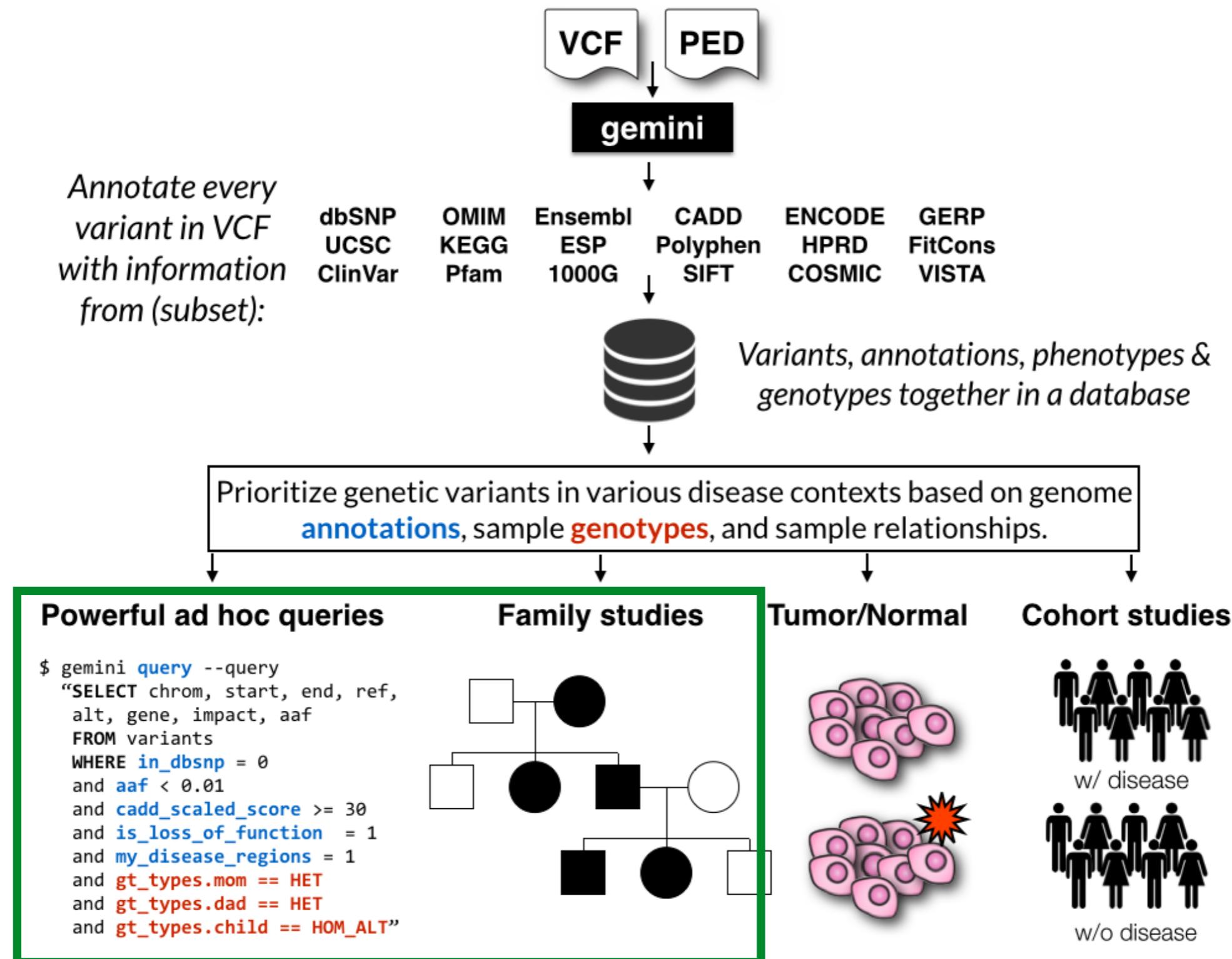
Brent
Pedersen

Code: github.com/arq5x/gemini

Docs: gemini.readthedocs.org

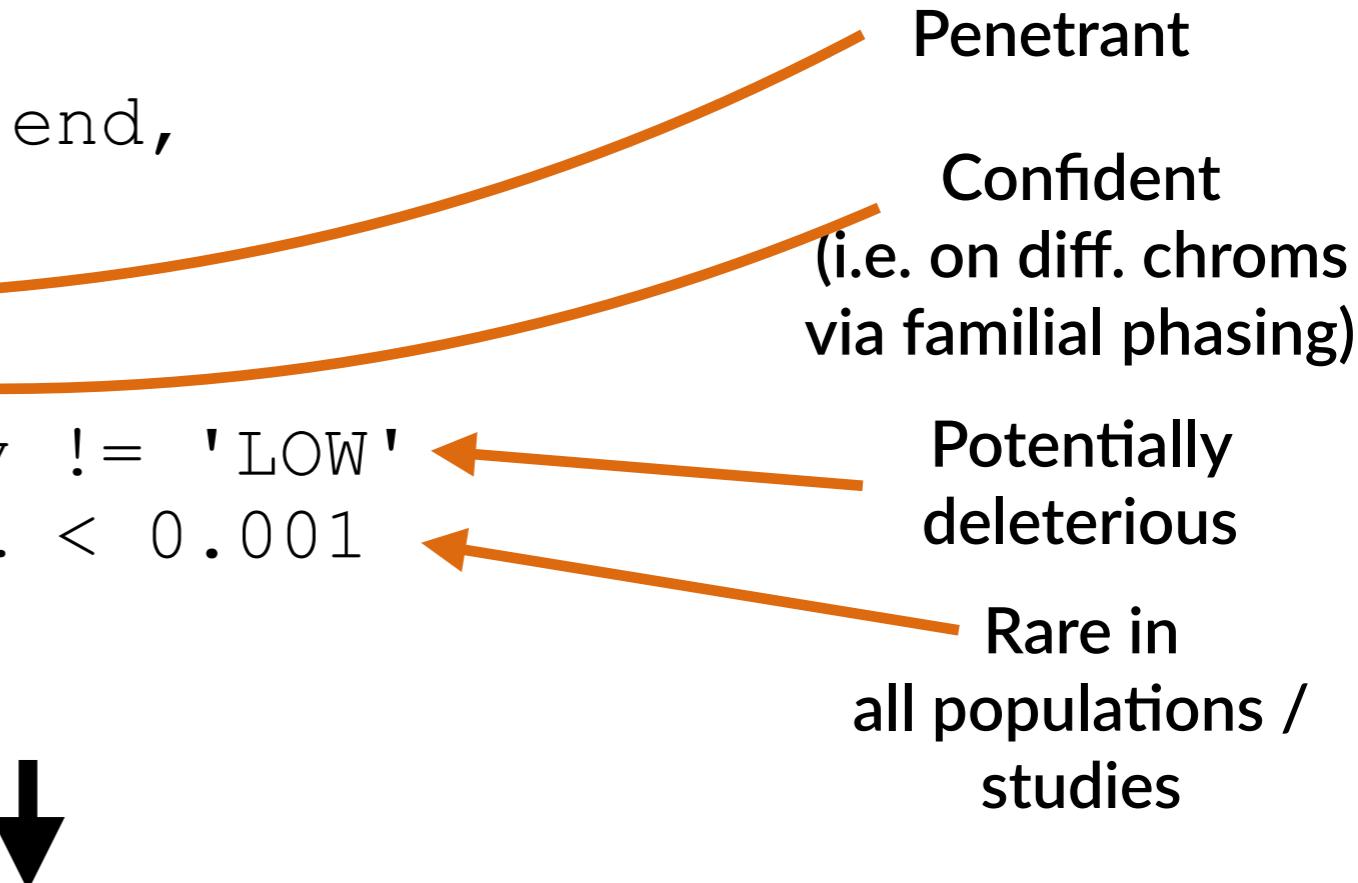
Goal: make rare disease research as simple and reproducible as possible

GEMINI integrates variants, annotations, relationships and genotypes into a simple database



Compound heterozygotes

```
$ gemini comp_hets  
  --columns "chrom, start, end,  
            gene, impact"  
  --min-kindreds 2  
  --max-priority 1  
  --filter "impact_severity != 'LOW'  
        and max_aaf_all < 0.001  
disease.db
```



chr17	78081692	78081693	GAA	non_syn_coding	4.33	14401	family1	1805;unaffected,1847;unaffected4805;affected	T/T,T/C,T C	4805	1	16_14401_14406	1
chr17	78084748	78084749	GAA	non_syn_coding	5.51	14406	family1	1805;unaffected,1847;unaffected4805;affected	G/A,G/G,A G	4805	1	16_14401_14406	1
chr2	129025757	129025758	HS6ST1	non_syn_coding	2.67	3657	family1	1805;unaffected,1847;unaffected4805;affected	C/A,C/C,A C	4805	1	45_3657_3660	1
chr2	129075743	129075744	HS6ST1	non_syn_coding	3.25	3660	family1	1805;unaffected,1847;unaffected4805;affected	G/G,G/C,G C	4805	1	45_3657_3660	1
chr22	45255643	45255644	PRR5-ARHGAP8	non_syn_coding	1.88	16838	family1	1805;unaffected,1847;unaffected4805;affected	G/G,G/T,G T	4805	1	169_16838_16839	1
chr22	45255687	45255688	PRR5-ARHGAP8	non_syn_coding	0.37	16839	family1	1805;unaffected,1847;unaffected4805;affected	C/T,C/C,T C	4805	1	169_16838_16839	1
chr15	71548994	71548995	THSD4	non_syn_coding	3.68	8777	family1	1805;unaffected,1847;unaffected4805;affected	C/C,C/T,C T	4805	1	181_8777_8780	1
chr15	71952898	71952899	THSD4	non_syn_coding	4.52	8780	family1	1805;unaffected,1847;unaffected4805;affected	G/A,G/G,A G	4805	1	181_8777_8780	1

GEMINI is popular for rare disease research.

Homozygous mutation of MTPAP causes cellular radiosensitivity and persistent DNA double-strand breaks

NT Martin^{1,2}, K Nakamura¹, U Paila³, J Woo¹, C Brown¹, JA Wright⁴, SN Teraoka⁴, S Haghayegh¹, D McCurdy⁵, M Schneider⁶, H Hu¹, AR Quinlan³, RA Gatti^{1,2,7} and P Concannon^{4,8}

Germline Mutations in MAP3K6 Are Associated with Familial Gastric Cancer

Daniel Gaston^{1,9}, Samantha Hansford^{2,9}, Carla Oliveira³, Mathew Nightingale¹, Hugo Pinheiro³, Christine Macgillivray⁴, Pardeep Kaurah², Andrea L. Rideout⁵, Patricia Steele⁵, Gabriela Soares⁶, Weei-Yuarn Huang⁷, Scott Whitehouse¹, Sarah Blowers⁸, Marissa A. LeBlanc¹, Haiyan Jiang⁹, Wenda Greer¹, Mark E. Samuels^{1,10}, Andrew Orr^{1,4}, Conrad V. Fernandez¹¹, Jacek Majewski¹², Mark Ludman^{5,13}, Sarah Dyack^{5,11}, Lynette S. Penney^{5,11}, Christopher R. McMaster¹⁴, David Huntsman², Karen Bedard^{1*}

New splicing mutation in the choline kinase beta (CHKB) gene causing a muscular dystrophy detected by whole-exome sequencing

Jorge Oliveira, Luís Negrão, Isabel Fineza, Ricardo Taipa, Manuel Melo-Pires, Ana Maria Fortuna, Ana Rita Gonçalves, Hugo Froufe, Conceição Egas, Rosário Santos and Mário Sousa

High-sensitivity sequencing reveals multi-organ somatic mosaicism causing DICER1 syndrome

Leanne de Kock,^{1,2} Yu Chang Wang,³ Timothée Revil,³ Dunarel Badescu,³ Barbara Rivera,^{1,2} Nelly Sabbaghian,² Mona Wu,^{1,2} Evan Weber,⁴ Claudio Sandoval,⁵ Saskia M J Hopman,⁶ Johannes H M Merks,⁶ Johanna M van Hagen,⁷ Antonia H M Bouts,⁸ David A Plager,⁹ Aparna Ramasubramanian,^{9,10} Linus Forsmark,¹¹ Kristine L Doyle,¹² Tonja Toler,¹³ Janine Callahan,¹⁴ Charlotte Engelenberg,¹⁵ Dorothée Bouron-Dal Soglio,¹⁶ John R Priest,¹⁷ Jiannis Ragoussis,³ William D Foulkes^{1,2,4,18}

REPORT

Mutations in NOTCH1 Cause Adams-Oliver Syndrome

Anna-Barbara Stittrich,^{1,9} Anna Lehman,^{2,9} Dale L. Bodian,^{3,9} Justin Ashworth,¹ Zheyuan Zong,² Hong Li,¹ Patricia Lam,² Alina Khromykh,³ Ramaswamy K. Iyer,³ Joseph G. Vockley,³ Rajiv Baveja,⁴ Ermelinda Santos Silva,⁵ Joanne Dixon,⁶ Eyby L. Leon,⁷ Benjamin D. Solomon,^{3,8} Gustavo Glusman,¹ John E. Niederhuber,^{3,10,*} Jared C. Roach,^{1,10} and Millan S. Patel^{2,10,*}

Diagnosis of an imprinted-gene syndrome by a novel bioinformatics analysis of whole-genome sequences from a family trio

Dale L. Bodian¹, Benjamin D. Solomon¹, Alina Khromykh¹, Dzung C. Thach¹, Ramaswamy K. Iyer¹, Kathleen Link², Robin L. Baker³, Rajiv Baveja³, Joseph G. Vockley¹ & John E. Niederhuber¹

¹Inova Translational Medicine Institute, Inova Health System, Falls Church, Virginia

²Department of Pediatric Endocrinology, Inova Children's Hospital, Falls Church, Virginia

³Fairfax Neonatal Associates PC, Inova Children's Hospital, Falls Church, Virginia

Rare disease genetics in Utah



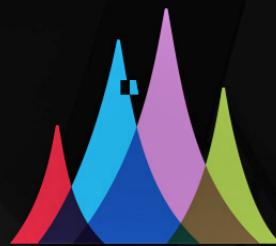
Family genetics

Clinical
collaborators



**Utah
Genome
Project**





**USTAR Center for
Genetic Discovery**

Colleagues



Mark Yandell



Gabor Marth



Lynn Jorde



Karen Eilbeck

Active Studies (primarily 60X WGS)

201	Congenital Heart Disease
96	Suicide
96	Longevity
80	Amyotrophic Lateral Sclerosis
78	Congenital Diaphragmatic Hernia
72	Autism
45	Early Infantile Epileptic Encephalopathy
40	Preterm Birth
37	Hereditary Hemorrhagic Telangiectasia
30	Tuberous Sclerosis
30	Chiari Malformation
10	Primary Ovarian Insufficiency
9	Familial Metatarsophalangeal Joint Osteoarthritis
9	Ataxia
8	Idiopathic Hypogonadotropic Hypogonadism
6	Brown Fat Metabolism
3	Mitochondrial Pyruvate Insufficiency

Summary

- Two basic and complementary approaches: family based and case-control based.
- Modern DNA sequencing has opened the flood gates for discovery.
- Our molecular and computation tools for disease genetics research have advanced dramatically in recent years. However, much of the low hanging fruit has been picked. Many of the unsolved rare diseases exhibit incomplete penetrance, phenotypic heterogeneity and germline and somatic mosaicism. These will be tough nuts to crack.