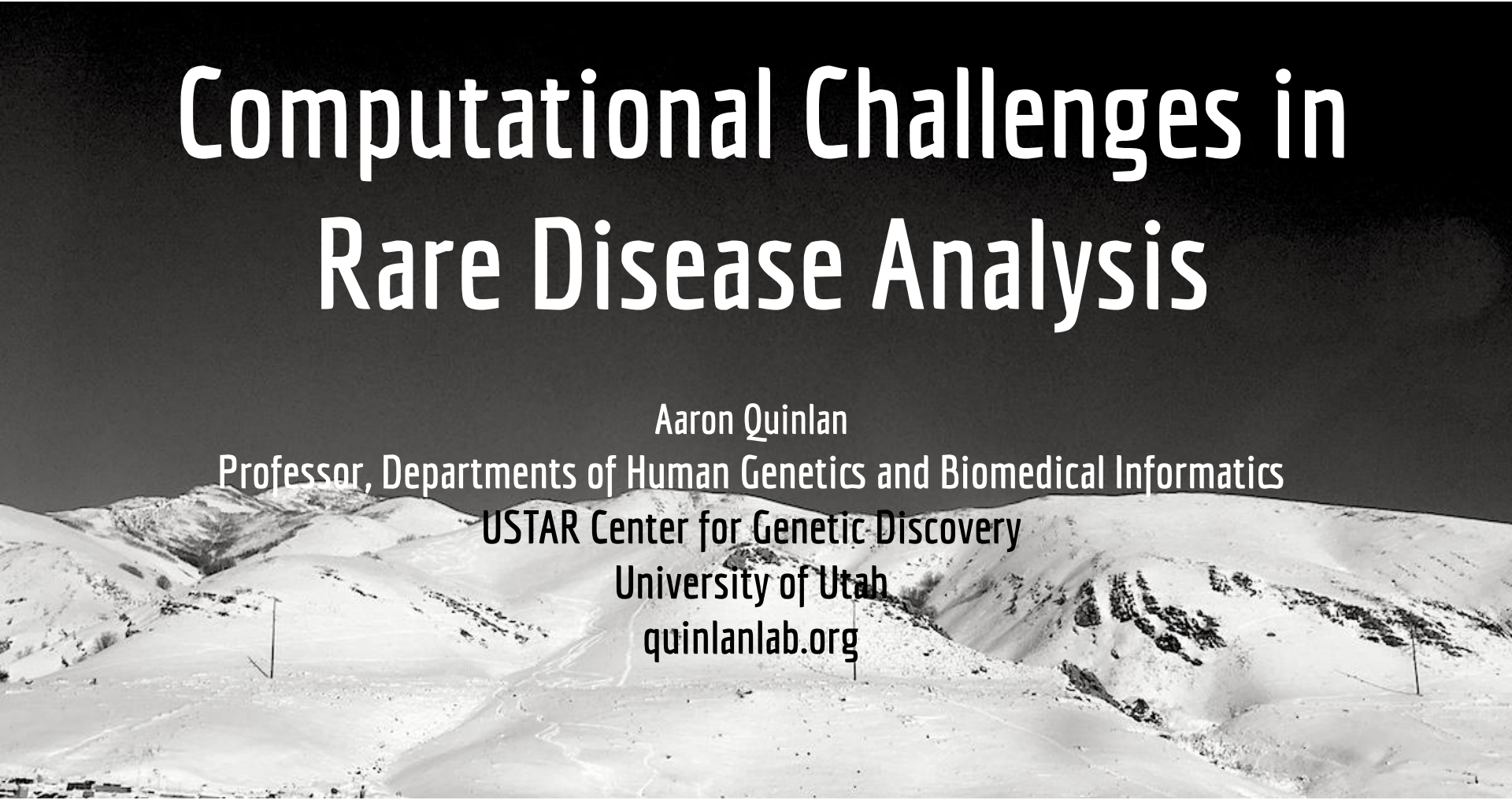
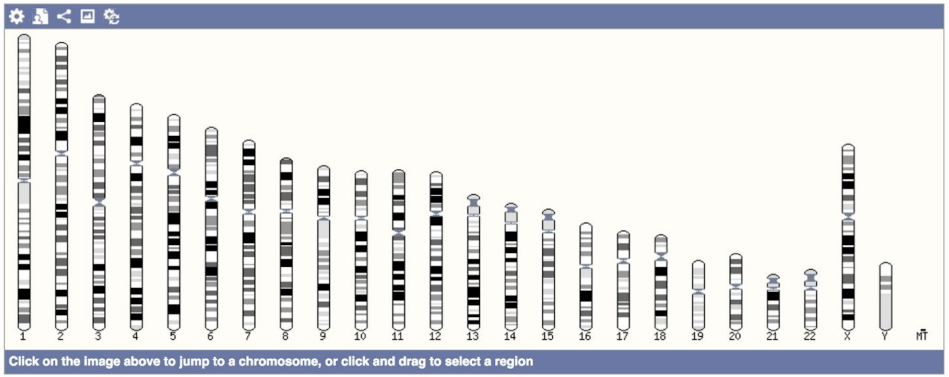
Computational Challenges in Rare Disease Analysis



Aaron Quinlan
Professor, Departments of Human Genetics and Biomedical Informatics
USTAR Center for Genetic Discovery
University of Utah
quinlanlab.org

The human genome - basic stats



- 3.096 billion base pairs (haploid)
- ~20,000 protein coding genes
- 198,002 coding transcripts (isoforms of a gene that each encode a distinct protein product)

Summary

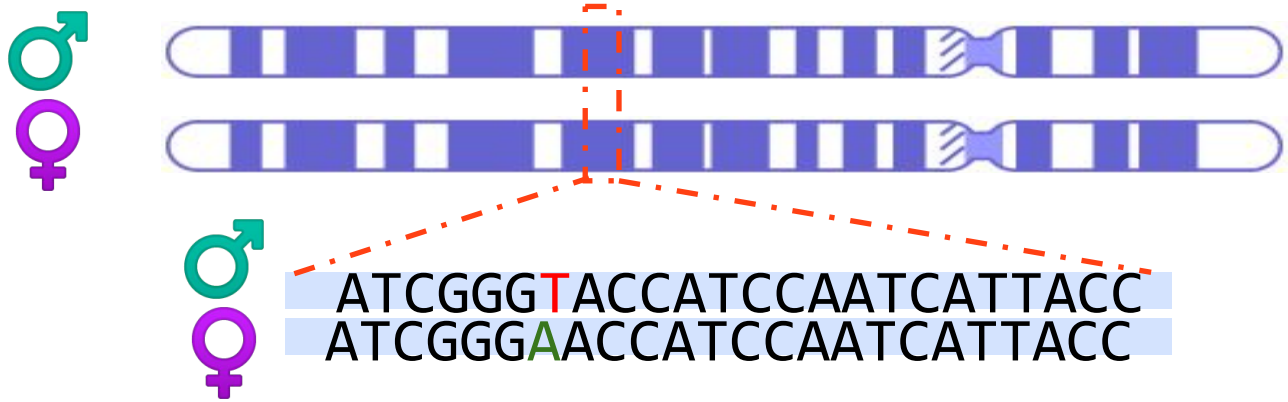
Assembly	GRCh38.p7 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.22, Dec 2013
Database version	87.38
Base Pairs	3,547,762,741
Golden Path Length	3,096,649,726
Genebuild by	Ensembl
Genebuild method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jun 2016
Gencode version	GENCODE 25

Gene counts (Primary assembly)

Coding genes	20,441 (incl 526 readthrough)
Non coding genes	22,219
Small non coding genes	5,052
Long non coding genes	14,727 (incl 214 readthrough)
Misc non coding genes	2,222
Pseudogenes	14,606 (incl 5 readthrough)
Gene transcripts	198,002

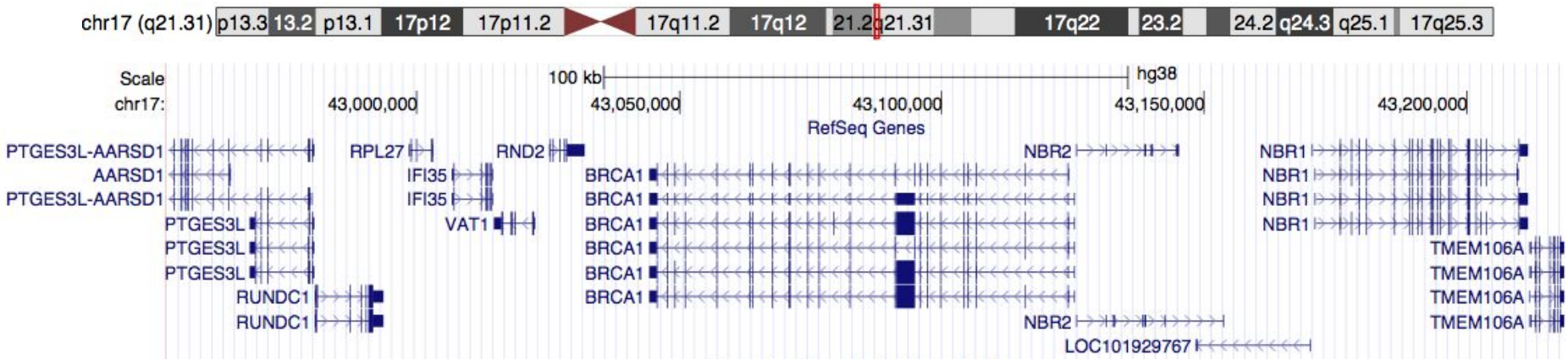
http://uswest.ensembl.org/Homo_sapiens/Location/Genome

Humans are diploid.



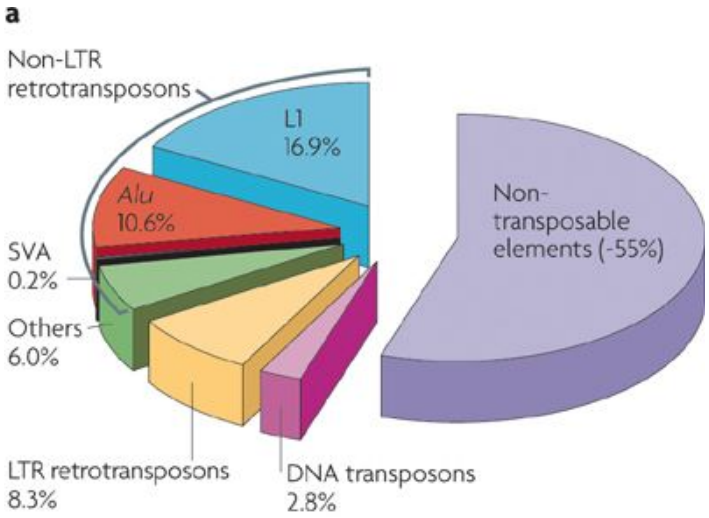
Our genome is comprised of a paternal and a maternal "haplotype". Together, they form our "genotype"

A measly 2% of the human genome encodes proteins.

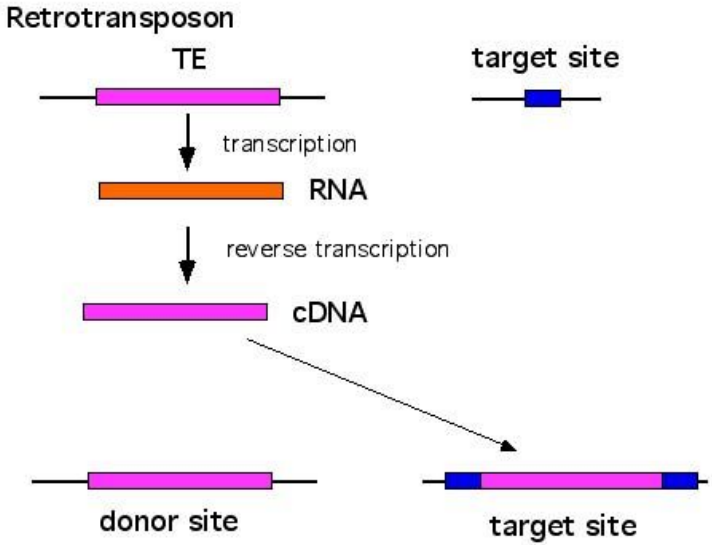


https://genome.ucsc.edu

Half of the human genome is comprised of repeats



McClintock's "jumping genes" in maize



Retrotransposons use a "copy/paste" mechanism
DNA transposons use a "cut/paste" mechanism

http://www.nature.com/nrg/journal/v10/n10/pdf/nrg2640.pdf

Problem: Half of the human genome is comprised of repeats

(first bit of human chromosome 1)

Alignment is central to most genomic research

FASTQ format

Align to genome

Quality control

Find peaks

Quantify transcripts

Find SVs

.
.
.

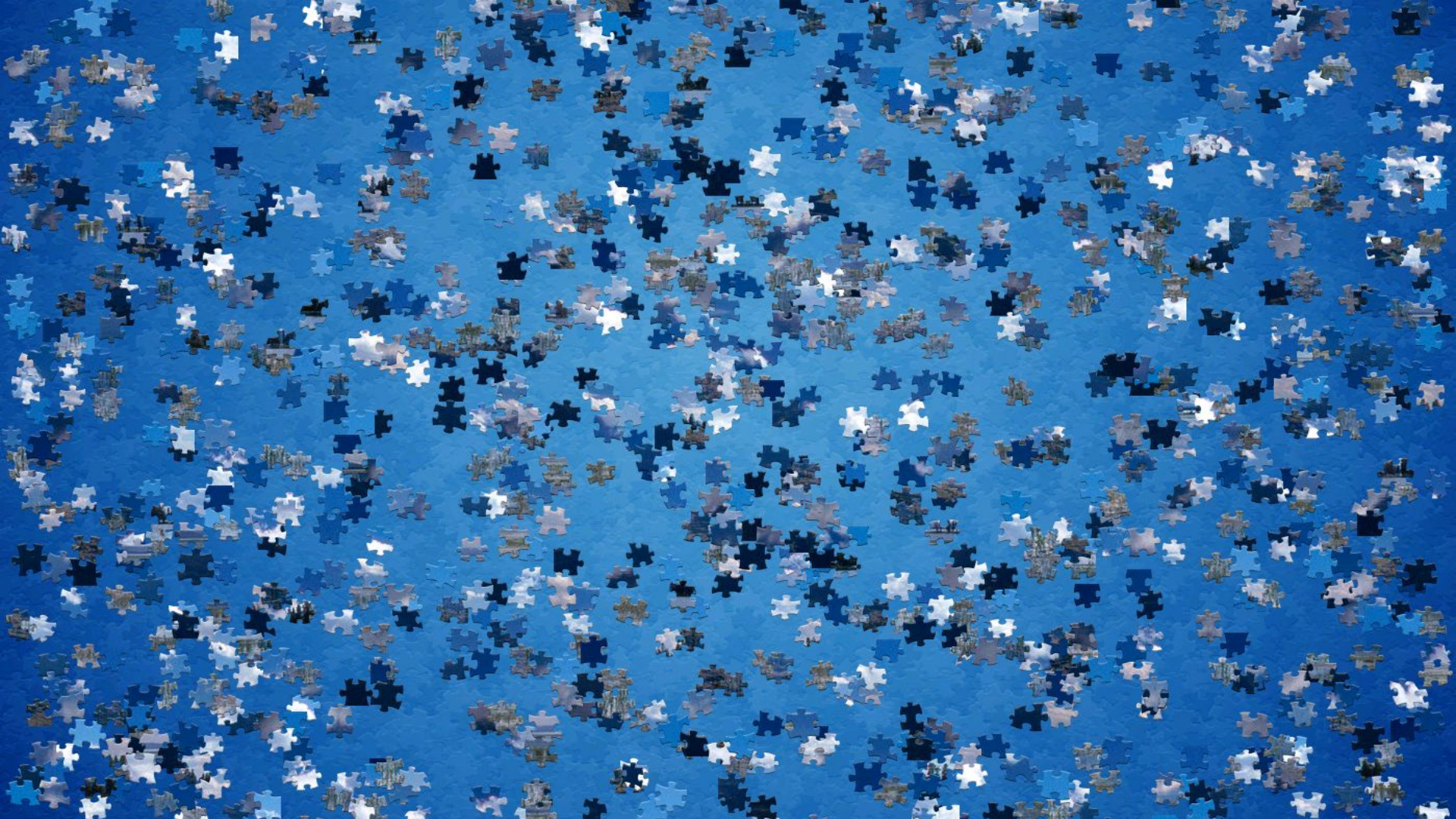
etc.

Analysis, Interpretation, Discovery

Phase 0

Phase1

Phase 2-100



Best case scenario: an error-free sequencing technology

ATTCGAAACA
TTCGCGCAAT
CTGGACTCAA

ATTCGAAACA
TTCGCGCAAT
CTGGACTCAA

Aligner

TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTCACGAA
AAAAAAGTCTGGACTCAACTAGGATACACACA
TTCGGCTACAGATACCAAAAAAAAAAAAAAA
AAATTTTCACCATTGAGGCACCACCTTCTCGT
CGCTGCGTCGCTCTGCTCGCTTCGGCTAAAAA
TTCGCGCAATACATTCGGCTACAGATACCAAA
AAAA

Computers are rather good at finding ***exact*** matches.
Think Google.

Reality check. Errors happen. Frequently.

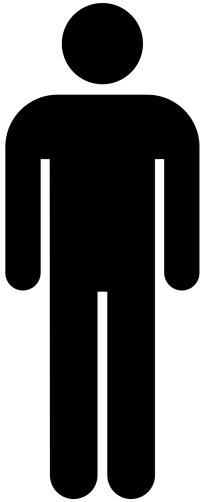
ATTCGAAACA

ATTTGAAACA → Aligner →

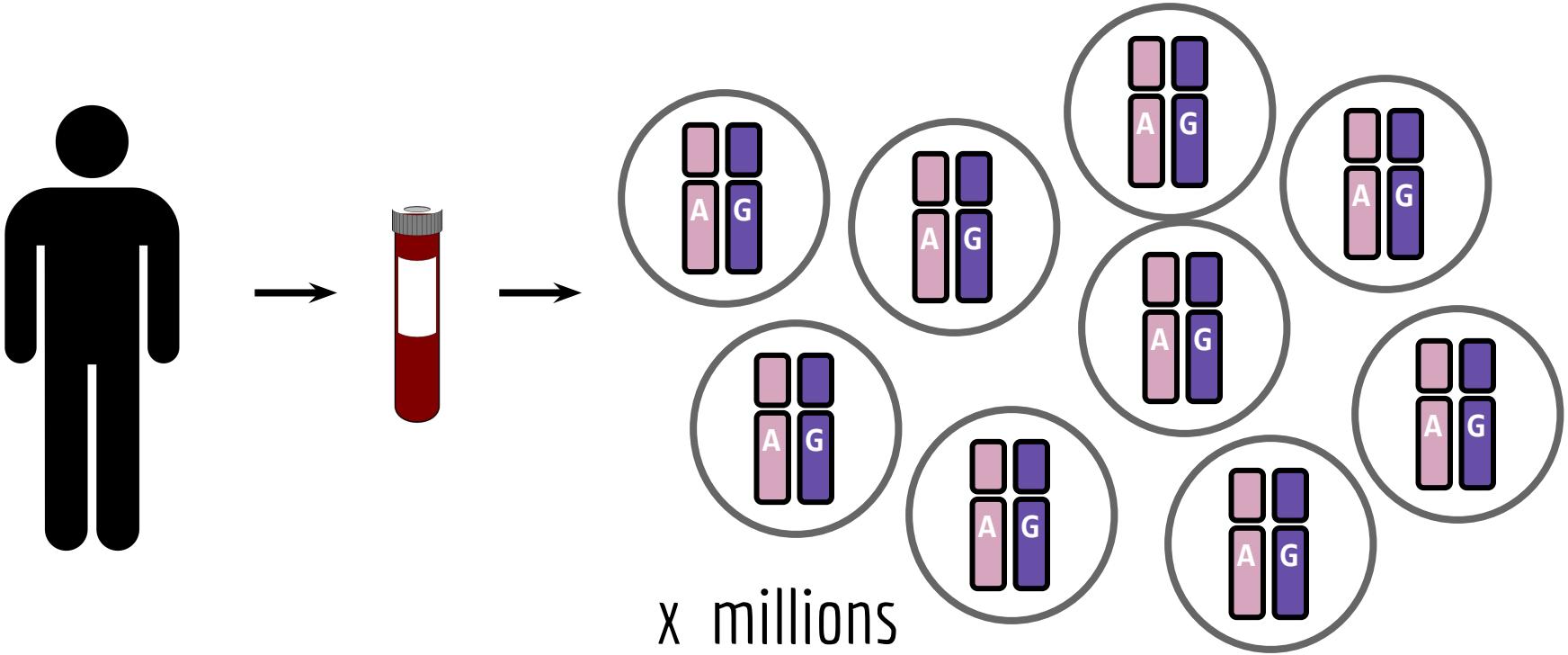
TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTGGAAAC
AAAAAAGTGCTACAGATACCACTAGGATACAC
ACATTCGGCTACAGATACCAAAAAAAAAAAAA
AAAAAATTTTCACCATTGAGGCACCACCTTCT
CGTCGCTGCGTCGCTCTGCTCGCGGCTAAAAA
ATTAGAAACAACATTCGGCTACAGATACCAAA
ATTT

"Fuzzy" matching is much more computationally expensive.
Think Google's "Did you mean..."

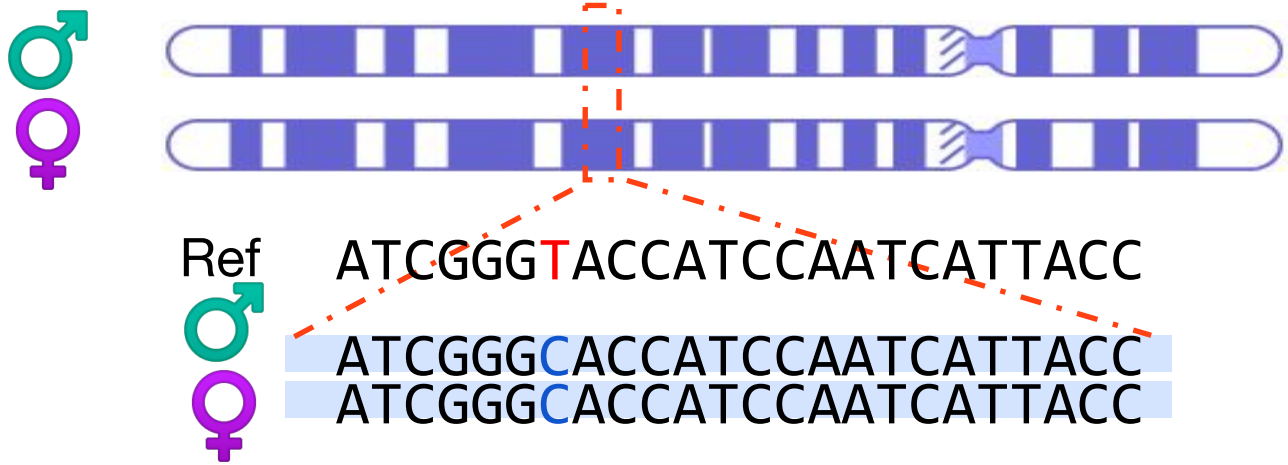
Goal: find all variants in an individual's diploid genome.



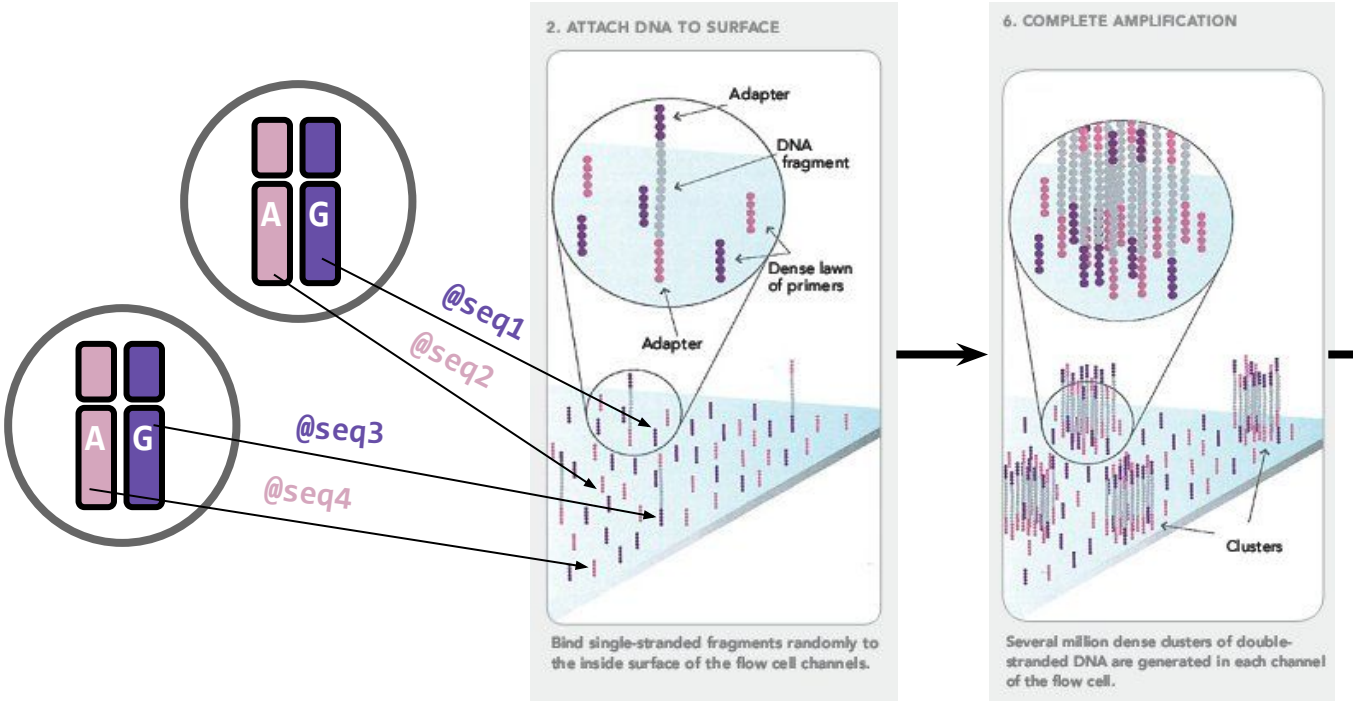
Find inherited genetic variation by sequencing DNA from millions of cells



What if an individual is homozygous for an "alternate" allele?

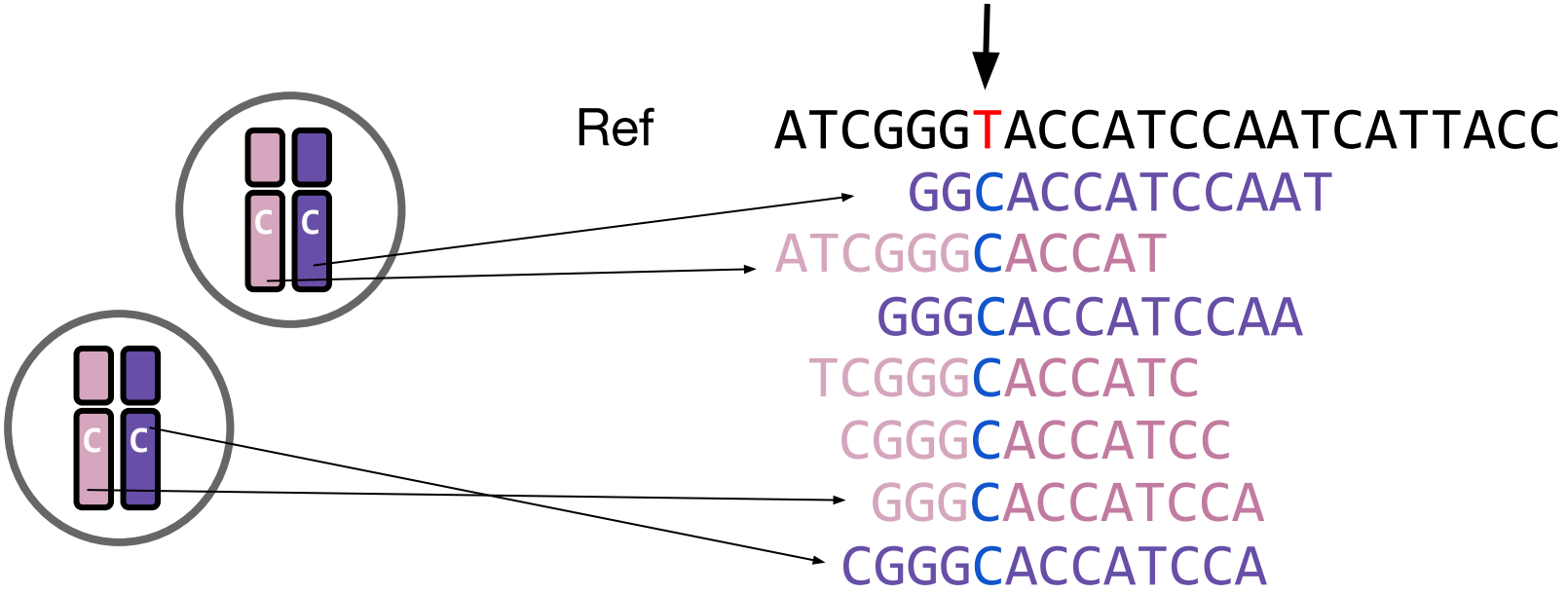


Each DNA cluster is amplified from a single strand from a single haploid chromosome from a single cell.

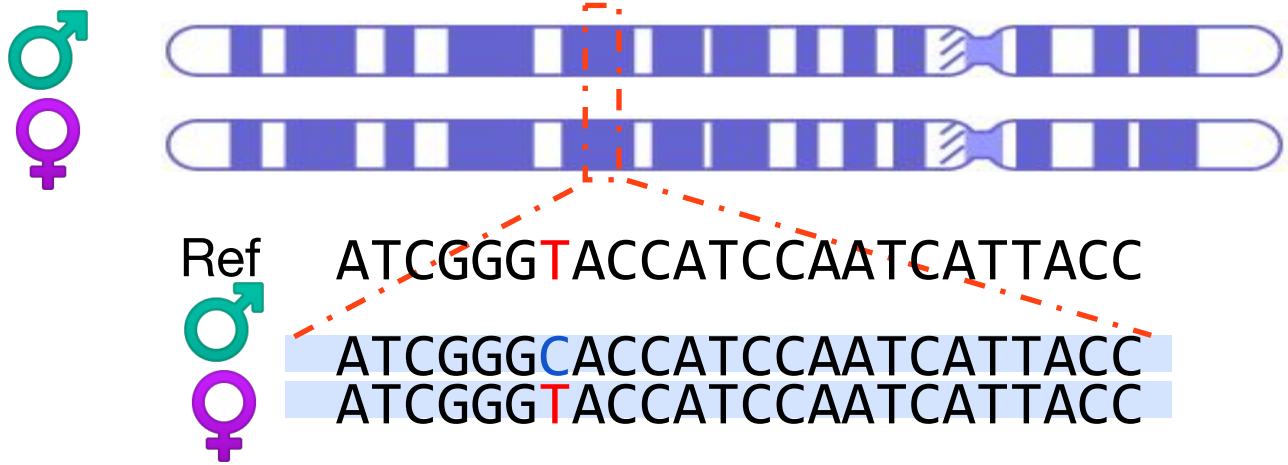



```
@seq1
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
@seq3
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq4
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
```

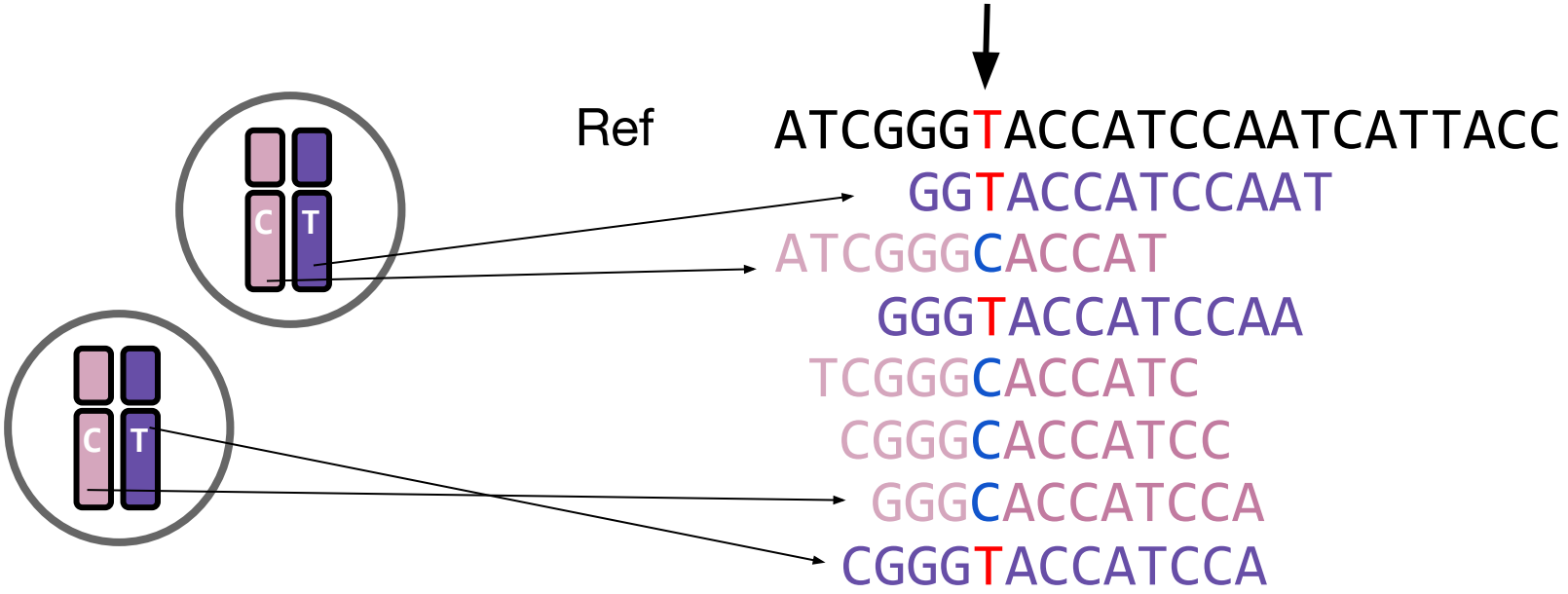
What if an individual is homozygous for an "alternate" allele?



What if an individual is heterozygous for an "alternate" allele?



What if an individual is heterozygous for an "alternate" allele?



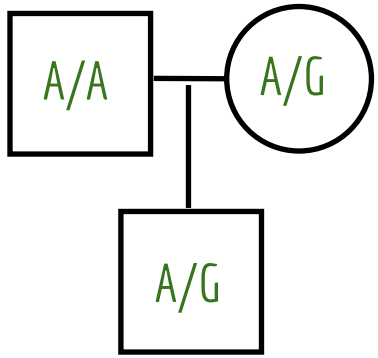
Variant Calling Overview

→ FASTQ → Align (BWA) → BAM → Detect SNP/INDELs (GATK or FreeBayes) → VCF

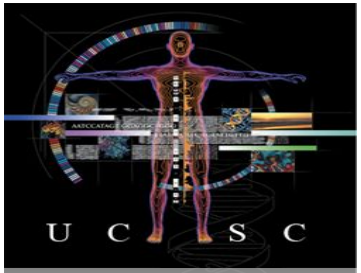
Raw VCF files are naked. Interpretation requires annotation.

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=variantcallerXYZ
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	DAD	KID
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3



Annotation provides context for interpretation.



Conservation
Repeat elements
Genome Gaps
Cytobands
Gene annotations
"Mappability"
DeCIPHER
ISGA

dbSNP
Short Genetic Variations

gnomAD browser

ClinVar

OMIM
Online Mendelian Inheritance in Man

1000 Genomes
A Deep Catalog of Human Genetic Variation

ENCODE

Chromatin marks
DNA methylation
RNA expression
TF binding

Pfam

KEGG
Kyoto Encyclopedia of Genes and Genomes

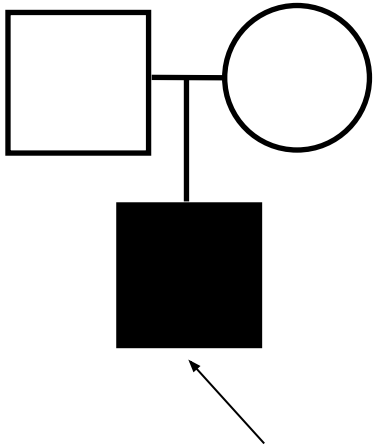
HGMD®

Human Protein Reference Database

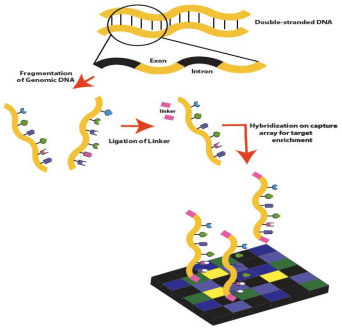
Genetic variation

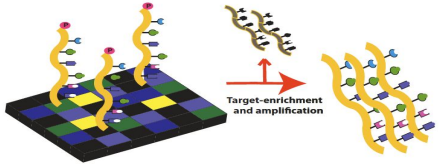
...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...

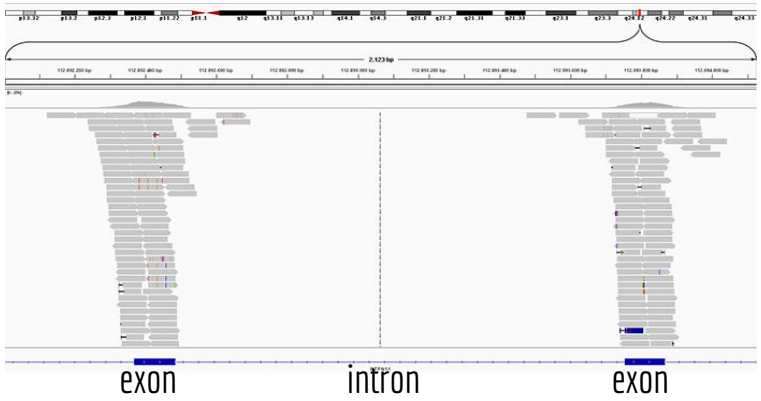
Case study: exome sequencing for rare disease



Mom and dad are unaffected, kid (proband) is affected with rare, Mendelian disease phenotype







>40% of developmental disorders caused by de novo mutation

Prevalence and architecture of *de novo* mutations in developmental disorders

Deciphering Developmental Disorders Study*

The genomes of individuals with severe, undiagnosed developmental disorders are enriched in damaging *de novo* mutations (DNMs) in developmentally important genes. Here we have sequenced the exomes of 4,293 families containing individuals with developmental disorders, and meta-analysed these data with data from another 3,287 individuals with similar disorders. We show that the most important factors influencing the diagnostic yield of DNMs are the sex of the affected individual, the relatedness of their parents, whether close relatives are affected and the parental ages. We identified 94 genes enriched in damaging DNMs, including 14 that previously lacked compelling evidence of involvement in developmental disorders. We have also characterized the phenotypic diversity among these disorders. We estimate that 42% of our cohort carry pathogenic DNMs in coding sequences; approximately half of these DNMs disrupt gene function and the remainder result in altered protein function. We estimate that developmental disorders caused by DNMs have an average prevalence of 1 in 213 to 1 in 448 births, depending on parental age. Given current global demographics, this equates to almost 400,000 children born per year.

https://www.nature.com/articles/nature21062.pdf

Goal: find variants that disrupt gene function

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

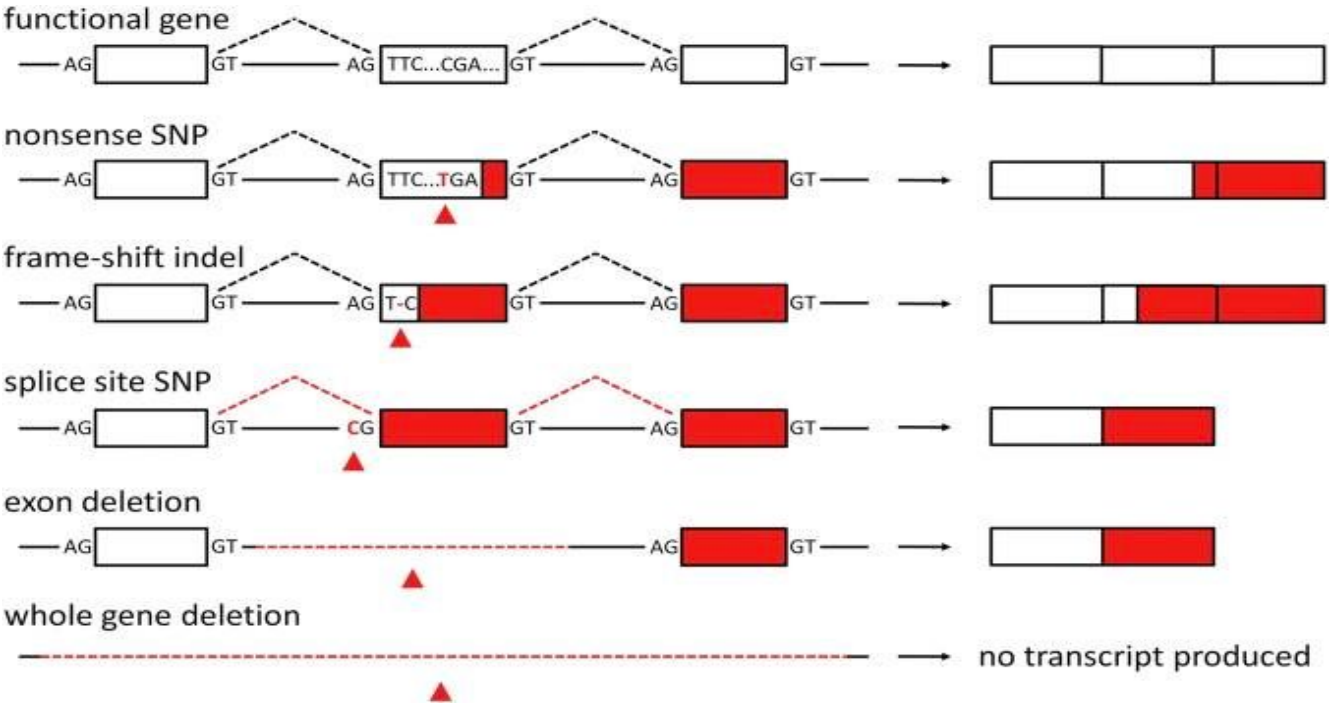
stop-gain (nonsense)

	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

Loss of function variants



MacArthur et al, Science. 2012 February 17; 335(6070): 823–828.

Uh oh. Needle in a haystack problem.

Table 1 | **Median number of protein-coding variants and effects among world super-populations***

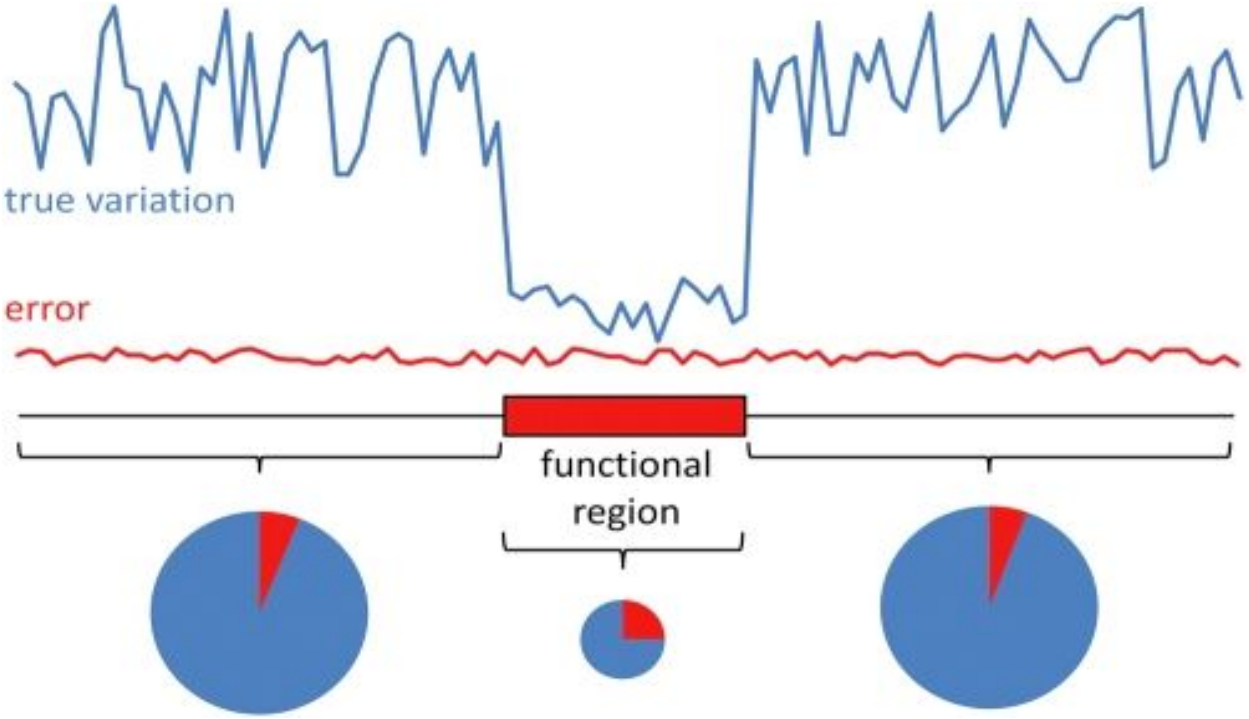
Super-population code	Synonymous (het; hom alt)	Missense (het; hom alt)			Frameshift (het; hom alt)	Stop gain (het; hom alt)	Start lost (het; hom alt)	Splice donor (het; hom alt)	Splice acceptor (het; hom alt)
		Total	SIFT Del	PP Del					
EUR	6961; 4317	7220; 4452	116; 55	116; 38	151; 146	93; 35	61; 52	184; 99	114; 72
AFR	9296; 4673	9347; 4820	163; 56	156; 31	196; 150	123; 32	78; 51	231; 116	150; 80
AMR	7257; 4314	7449; 4479	121; 56	121; 38	154; 145	96; 34	62; 50	187; 101	117; 76
SAS	7180; 4397	7366; 4550	123; 56	121; 39	159; 148	93; 36	68; 49	186; 103	117; 78
EAS	6502; 4759	6802; 4908	105; 66	113; 45	143; 149	89; 38	62; 54	171; 112	115; 86

AFR, individuals of African descent; AMR, individuals of admixed descent from the Americas; EAS, individuals of East-Asian descent; EUR, individuals of European descent; PP Del, PolyPhen2 predicted the missense variant to be deleterious; SAS, individuals of South-Asian descent; SIFT Del, SIFT predicted the missense variant to be deleterious. *We measured the average number of heterozygous (het) and homozygous alternate (hom alt) genotype counts among the 2,504 individuals sequenced by the 1000 Genomes Project. All genetic variants affecting genes were annotated with the Variant Effect Predictor and categorized by their most deleterious predicted effect.

Many predicted loss of function variants observed in a typical human. Even as a homozygote!!!

https://www.nature.com/articles/nrg.2017.52.pdf

Apparent loss-of-function variants are enriched for error.



MacArthur DG, Tyler-Smith C. (2010) Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet*. 19(R2):R125-130.

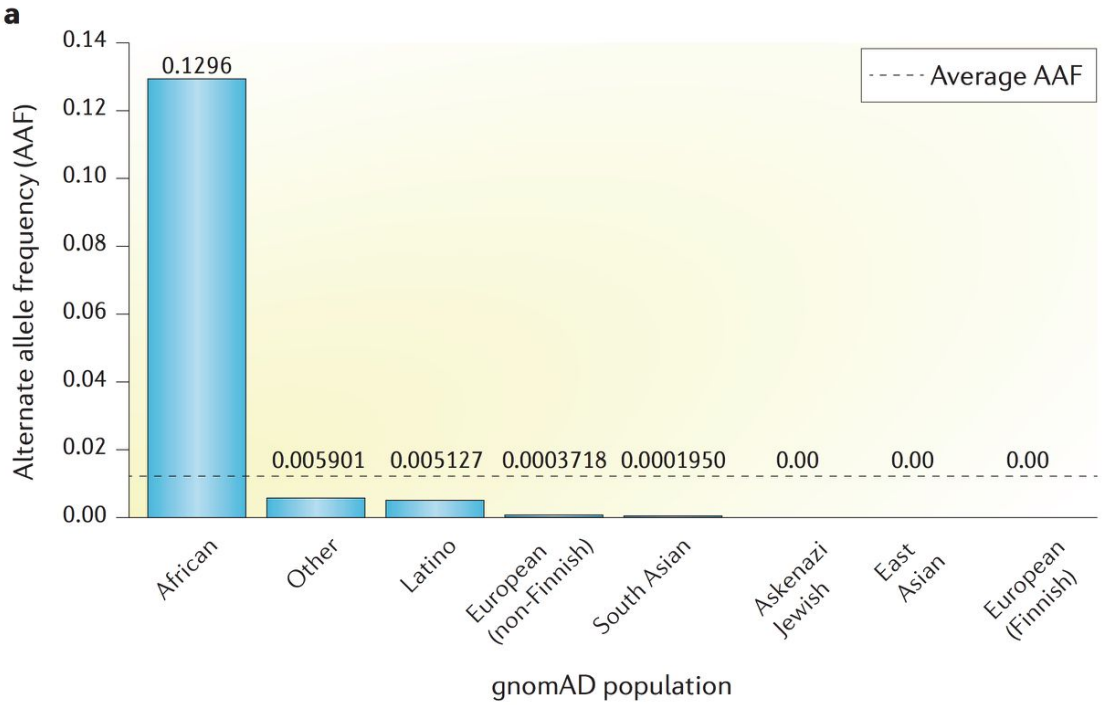
If a disease phenotype is rare, the causal variant should also be similarly rare

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P. Birnbaum[2], Jack A. Kosmicki[1,2,6,13], Laramie E. Duncan[1,2,6], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2], Emma Pierce-Hoffman[1,2], Joanne Berghout[14,15], David N. Cooper[16], Nicole Deflaux[17], Mark DePristo[18], Ron Do[19,20,21,22], Jason Flannick[2,23], Menachem Fromer[1,6,19,20,24], Laura Gauthier[18], Jackie Goldstein[1,2,6], Namrata Gupta[2], Daniel Howrigan[1,2,6], Adam Kiezun[18], Mitja I. Kurki[2,25], Ami Levy Moonshine[18], Pradeep Natarajan[2,26,27,28], Lorena Orozco[29], Gina M. Peloso[2,27,28], Ryan Poplin[18], Manuel A. Rivas[2], Valentin Ruano-Rubio[18], Samuel A. Rose[6], Douglas M. Ruderfer[19,20,24], Khalid Shakir[18], Peter D. Stenson[16], Christine Stevens[2], Brett P. Thomas[1,2], Grace Tiao[18], Maria T. Tusie-Luna[30], Ben Weisburd[2], Hong-Hee Won[31], Dongmei Yu[6,25,27,32], David M. Altshuler[2,33], Diego Ardissino[34], Michael Boehnke[35], John Danesh[36], Stacey Donnelly[2], Roberto Elosua[37], Jose C. Florez[2,26,27], Stacey B. Gabriel[2], Gad Getz[18,26,38], Stephen J. Glatt[39,40,41], Christina M. Hultman[42], Sekar Kathiresan[2,26,27,28], Markku Laakso[43], Steven McCarroll[6,8], Mark I. McCarthy[44,45,46], Dermot McGovern[47], Ruth McPherson[48], Benjamin M. Neale[1,2,6], Aarno Palotie[1,2,5,49], Shaun M. Purcell[19,20,24], Danish Saleheen[50,51,52], Jeremiah M. Scharf[2,6,25,27,32], Pamela Sklar[19,20,24,53,54], Patrick F. Sullivan[55,56], Jaakko Tuomilehto[57], Ming T. Tsuang[58], Hugh C. Watkins[44,59], James G. Wilson[60], Mark J. Daly[1,2,6], Daniel G. MacArthur[1,2] & Exome Aggregation Consortium†

Large-scale reference data sets of human genetic variation are critical for the medical and functional interpretation of DNA sequence changes. Here we describe the aggregation and analysis of high-quality exome (protein-coding region) DNA sequence data for 60,706 individuals of diverse ancestries generated as part of the Exome Aggregation Consortium (ExAC). This catalogue of human genetic diversity contains an average of one variant every eight bases of the exome, and provides direct evidence for the presence of widespread mutational recurrence. We have used this catalogue to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation; identifying 3,230 genes with near-complete depletion of predicted protein-truncating variants, with 72% of these genes having no currently established human disease phenotype. Finally, we demonstrate that these data can be used for the efficient filtering of candidate disease-causing variants, and for the discovery of human 'knockout' variants in protein-coding genes.

An allele underlying a rare disease should be rare in all ancestries!



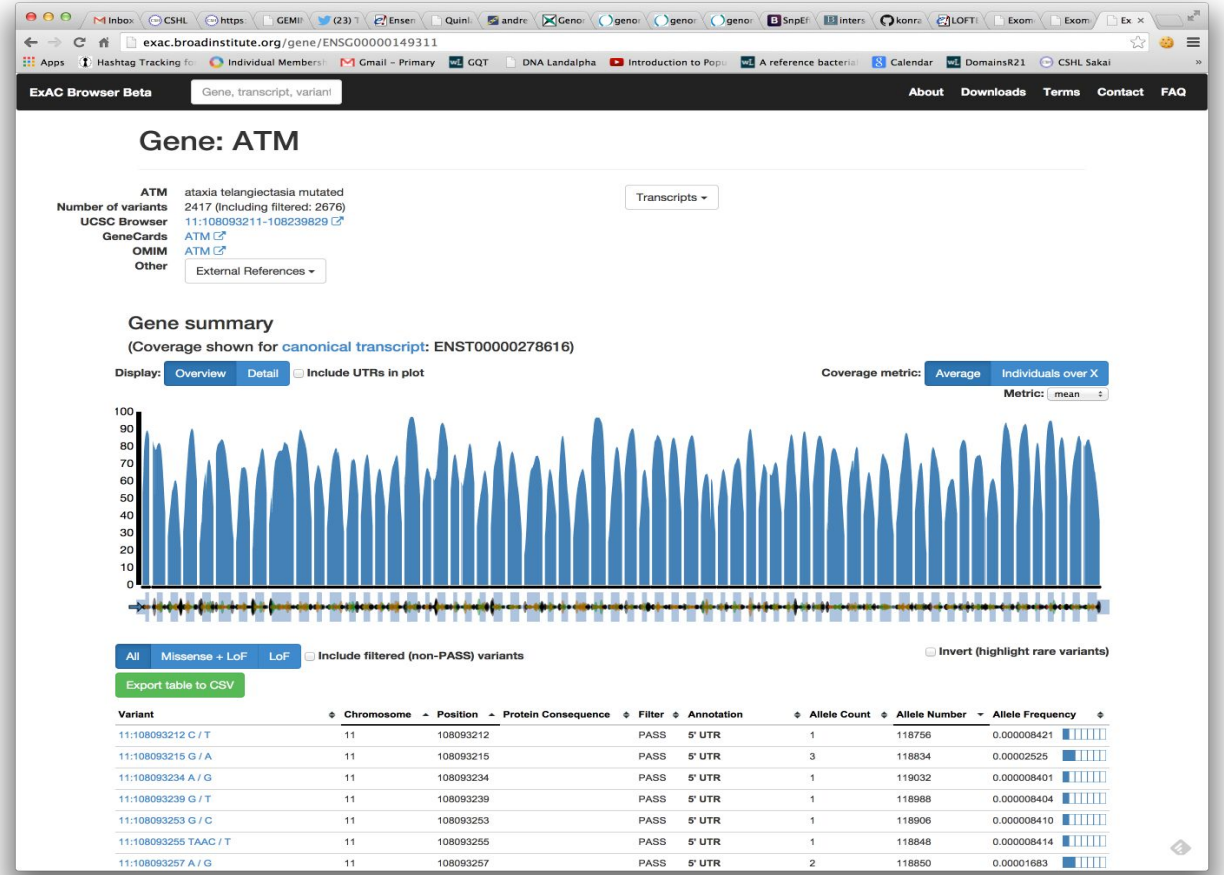
Settling the score: variant prioritization and Mendelian disease

Karen Eilbeck[1], Aaron Quinlan[1,2]* and Mark Yandell[2]*

Abstract | When investigating Mendelian disease using exome or genome sequencing, distinguishing disease-causing genetic variants from the multitude of candidate variants is a complex, multidimensional task. Many prioritization tools and online interpretation resources exist, and professional organizations have offered clinical guidelines for review and return of prioritization results. In this Review, we describe the strengths and weaknesses of widely used computational approaches, explain their roles in the diagnostic and discovery process and discuss how they can inform (and misinform) expert reviewers. We place variant prioritization in the wider context of gene prioritization, burden testing and genotype–phenotype association, and we discuss opportunities and challenges introduced by whole-genome sequencing.

Nature Reviews Genetics, 2017

gnomAD reports the allele frequency from diverse ancestries



exac.broadinstitute.org

gnomad.broadinstitute.org

Why are only 30% of rare disease cases diagnosed?

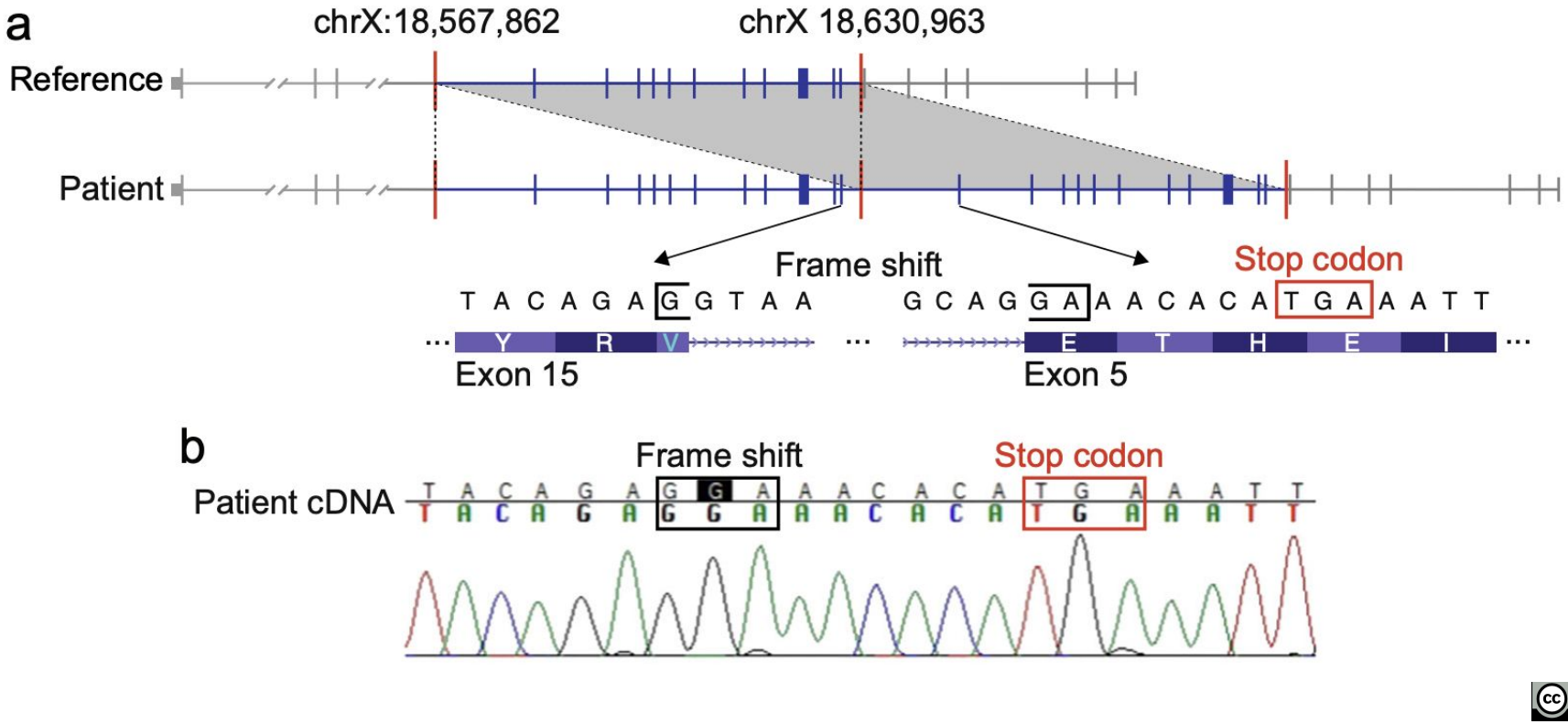
Hypothesis 1: we cannot detect the causal variant

ARTICLE OPEN

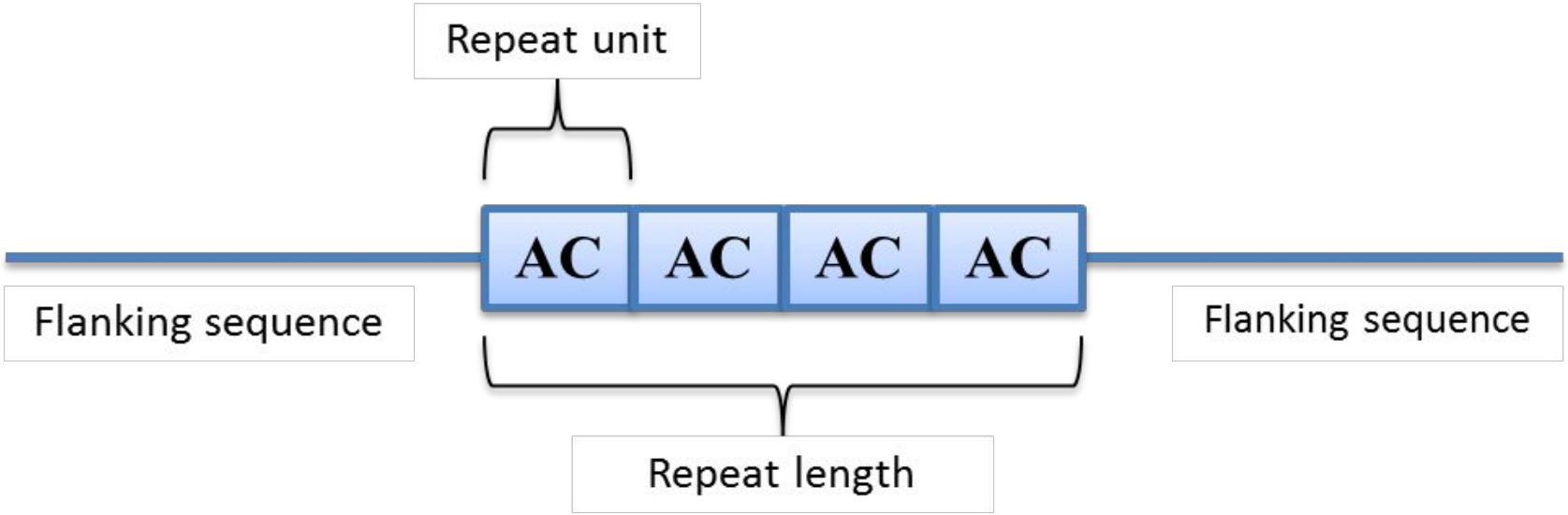
Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy

Betsy E. P. Ostrander[1], Russell J. Butterfield[1], Brent S. Pedersen[2], Andrew J. Farrell[2], Ryan M. Layer[2], Alistair Ward[2], Chase Miller[2], Tonya DiSera[2], Francis M. Filloux[1], Meghan S. Candee[1], Tara Newcomb[2], Joshua L. Bonkowsky[1], Gabor T. Marth[2] and Aaron R. Quinlan[1,2,3]

Hypothesis 1: e.g., structural variants



Hypothesis 1: e.g., Novel Short Tandem Repeats



- 1-6bp repeat units
- AKA microsatellites

- ~3% of the human genome
- High mutation rate, high polymorphism

Hypothesis 1: e.g., Novel Short Tandem Repeats

Disease	Gene	Repeat unit	Inheritance	Location	Pathogenic range	Reference
CANVAS	RFC1	AAGGG	AR	intronic	400-2000	Cortese 2019
Baratela-Scott Syndrome	XYLT1	CCG	AR	5’ non-coding		LaCroix 2019
BAFME1/FAME1	SAMD12	TTTCA	AD	intronic	225-458	Ishiura 2018
BAFME6/FAME6	TNRC6A	TTTCA	AD	intronic		Ishiura 2018
BAFME7/FAME7	RAPGEF2	TTTCA	AD	intronic		Ishiura 2018
SCA37	DAB1	ATTTC	AD	intronic	31-75	Seixas 2017
SCA31	BEAN	TGGAA	AD	intronic		Sato 2009

Hypothesis 2: We don't know what to make of a variant (VUS)

Predicting Splicing from Primary Sequence with Deep Learning

Kishore Jaganathan,[1,6] Sofia Kyriazopoulou Panagiotopoulou,[1,6] Jeremy F. McRae,[1,6] Siavash Fazel Darbandi,[2] David Knowles,[3] Yang I. Li,[3] Jack A. Kosmicki,[1,4] Juan Arbelaez,[2] Wenwu Cui,[1] Grace B. Schwartz,[2] Eric D. Chow,[5] Efstathios Kanterakis,[1] Hong Gao,[1] Amirali Kia,[1] Serafim Batzoglou,[1] Stephan J. Sanders,[2] and Kyle Kai-How Farh[1,7,*]

[1]Illumina Artificial Intelligence Laboratory, Illumina, Inc., San Diego, CA, USA

[2]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA

[3]Department of Genetics, Stanford University, Stanford, CA, USA

[4]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[5]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA

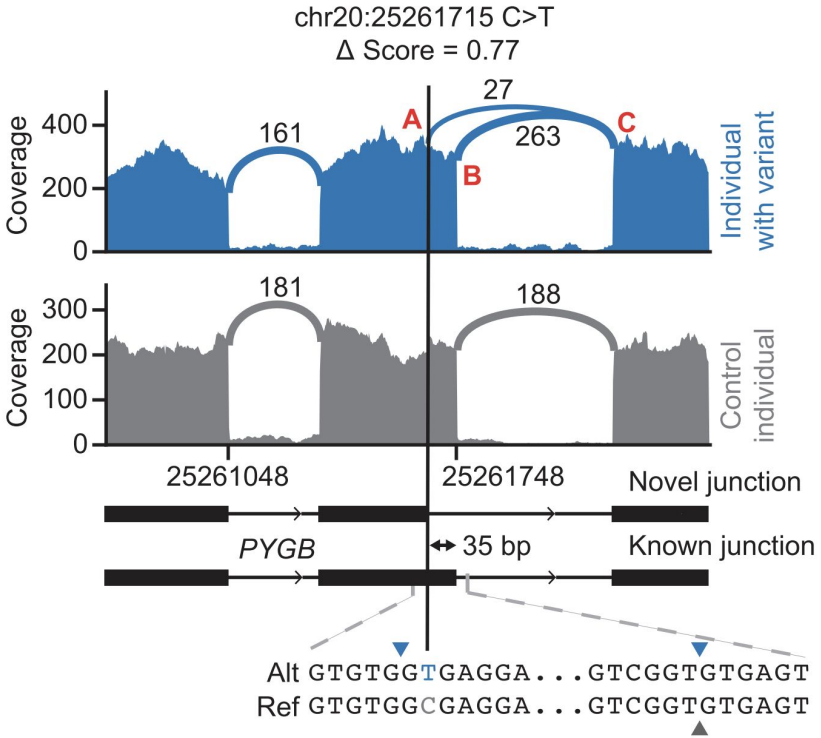
[6]These authors contributed equally

[7]Lead Contact

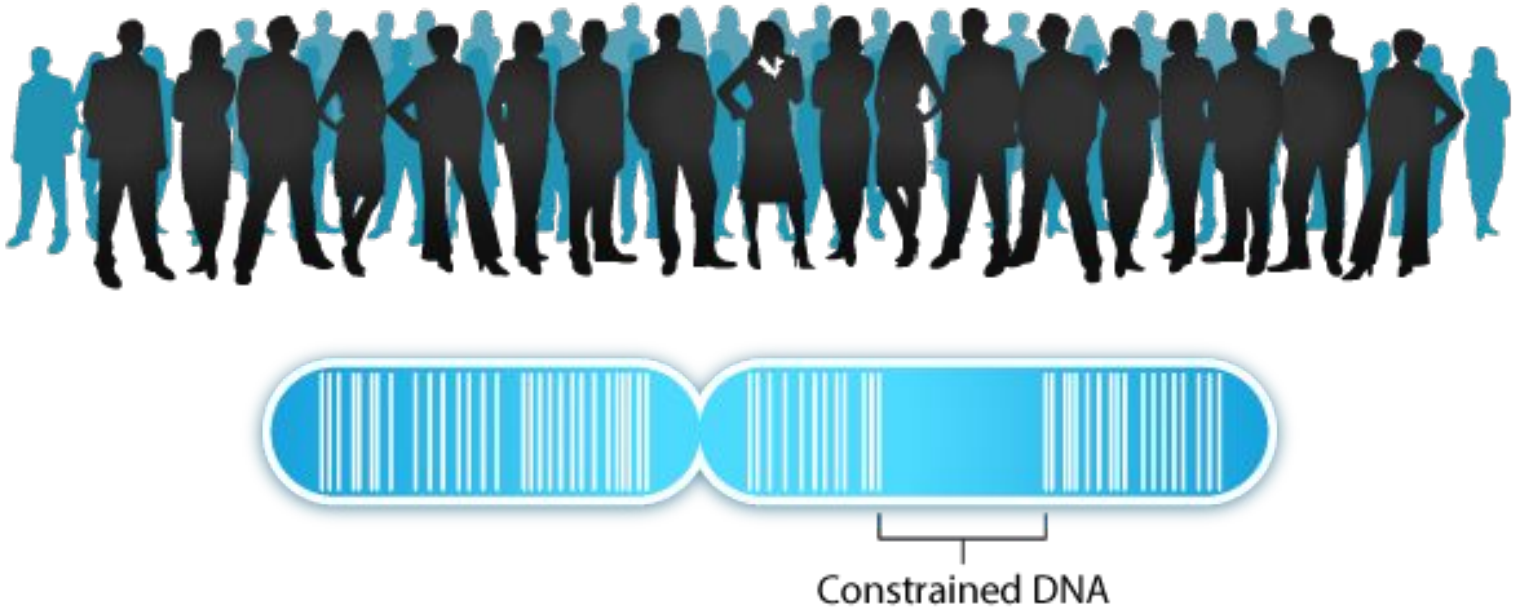
*Correspondence: kfarh@illumina.com

https://doi.org/10.1016/j.cell.2018.12.015

Hypothesis 2: e.g. spice-altering predictions



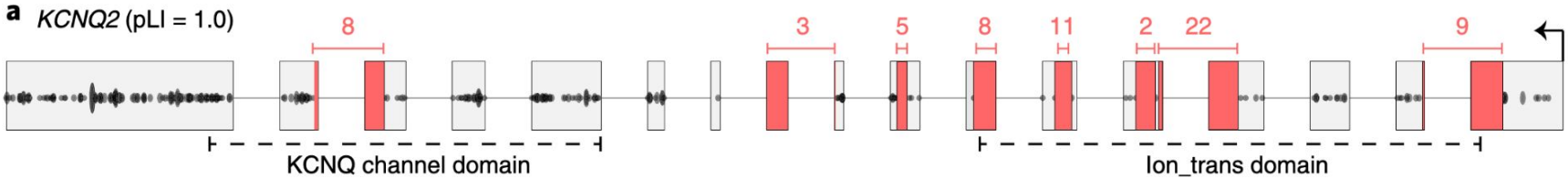
Hypothesis 2: genetic "constraint" (intolerance to mutation)



Hypothesis 2: genetic constraint

A map of constrained coding regions in the human genome

James M. Havrilla[1,2], Brent S. Pedersen[1,2], Ryan M. Layer[3,4] and Aaron R. Quinlan[1,2,5*]



Summary and things to ponder

- Modern DNA sequencing and advanced software can rapidly diagnose 33-50% of rare disease cases, often in just a few days
- The majority of cases are caused by loss-of-function variants in protein coding exons or splice sites.
- **Why are 50% of cases undiagnosable with the same assay, information?**
  - We are far from predicting phenotype directly from genotype.
  - We have a poor understanding of the consequence of genetic variation outside of exons
  - Often we have one interesting variant, but are missing the other (AR inheritance). How do we find it?
  - Some patients may in fact have multiple phenotypes. This complicates interpretation.