Introduction to personal genomics

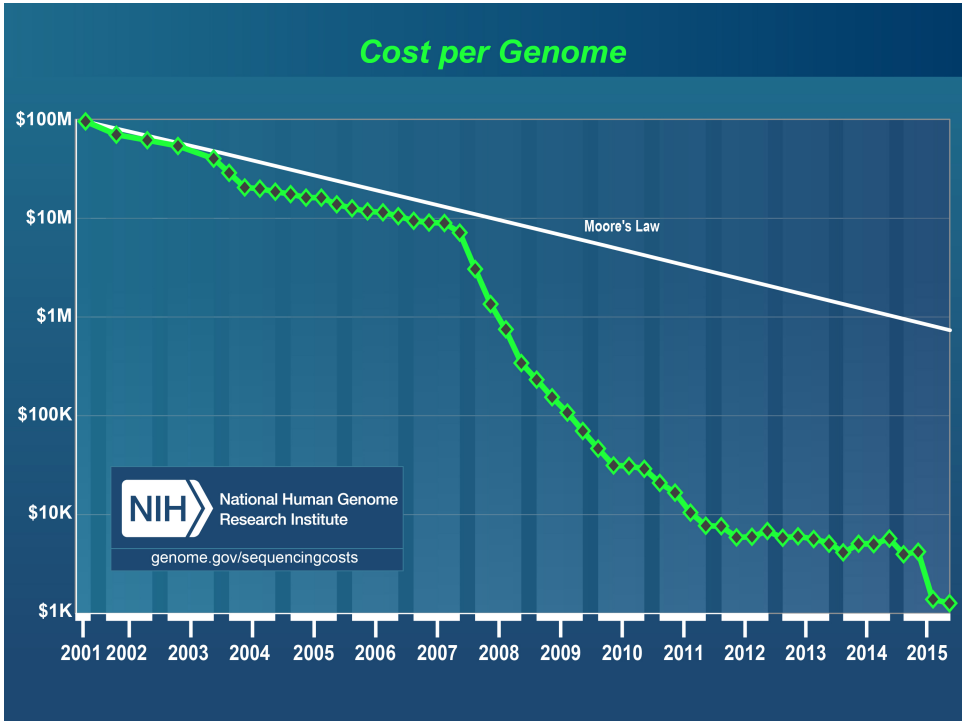
Mengting Gu
Paul Muir

Human Genome Project

- Project formally launched in 1990
- Declared complete in 2003
- International, collaborative research program to complete mapping and understanding of all the genes of human beings
- Very costly!

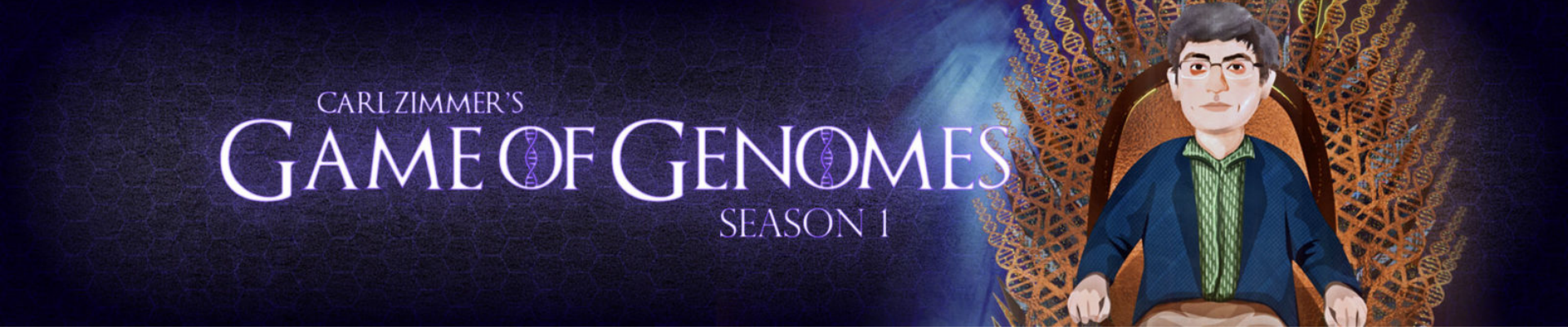
The cost of whole genome sequencing drops

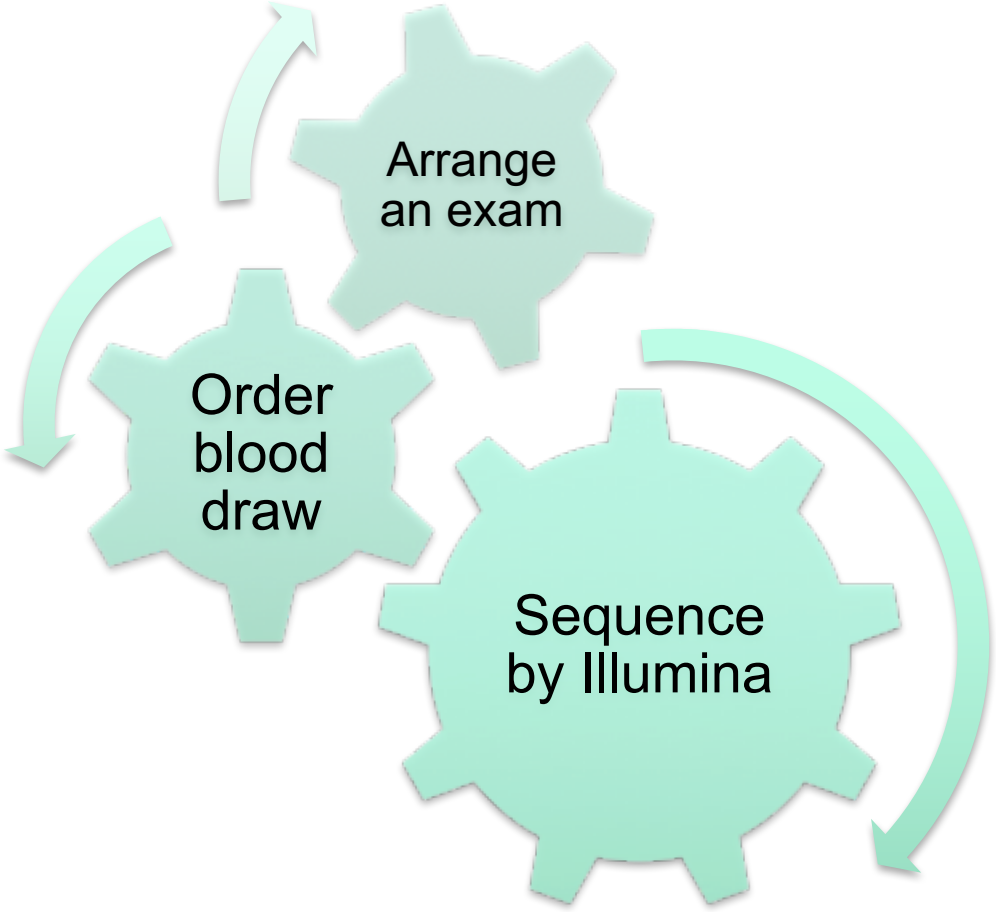
- The genome of Craig Venter by Sanger sequencing reportedly cost approximately $100 million
- In subsequent study, Dr. Watson's genome is sequenced for less than $1 million



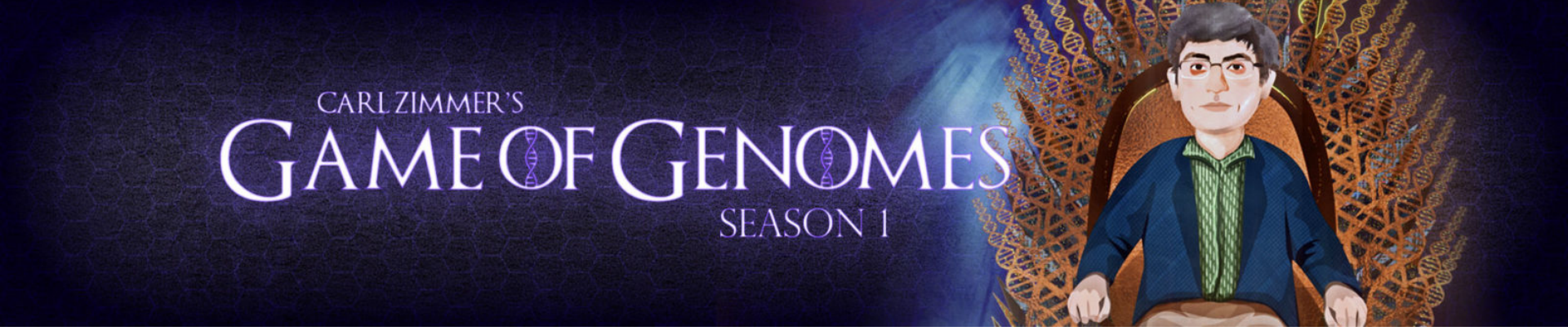
- A decade ago, the cost of whole genome sequencing has dropped to around 10 million (~2006)
- The cost to generate a high-quality whole human genome sequence in mid-2015 was just above $4000
- By late in 2015, the number had fallen below $1500

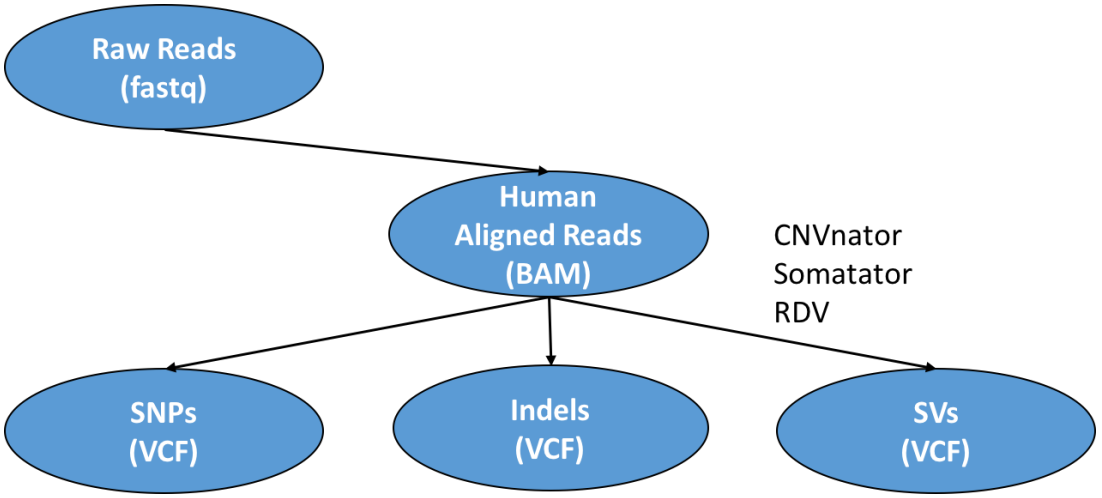
- Illumina introduces the NovaSeq series
- Designed to Usher in the $100 Genome





- **Cost: $3100**
- **Illumina briefly review the sequencing data, evaluating the risk for 1200 disorders, from familiar ones like lung cancer to obscure ones like cherubism**


CARL ZIMMER'S
GAME OF GENOMES
SEASON 1


Raw Reads
(fastq)
Human
Aligned Reads
(BAM)
CNVnator
Somatator
RDV
SNPs
(VCF)
Indels
(VCF)
SVs
(VCF)

Genome Variation

TP53 Sequence:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

Single Nucleotide Polymorphism (SNP) –1nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

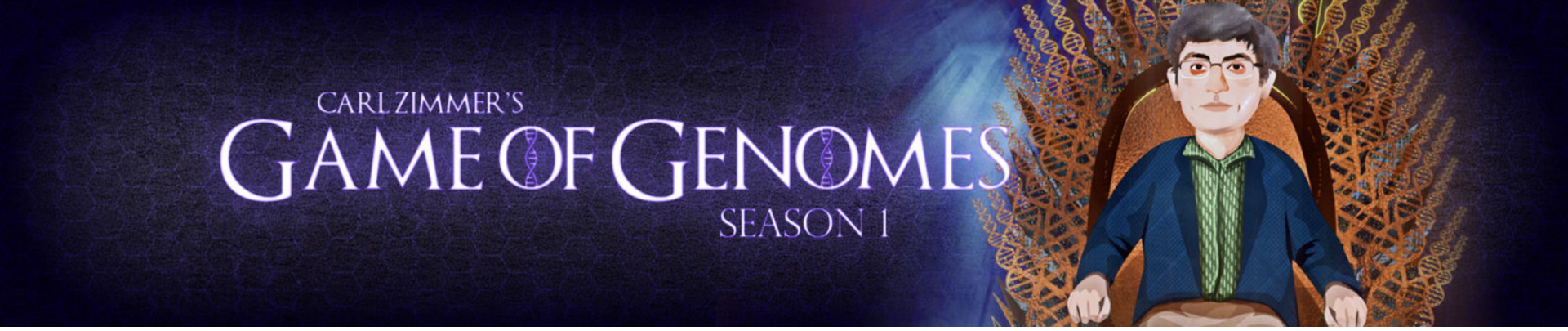
T or A or C

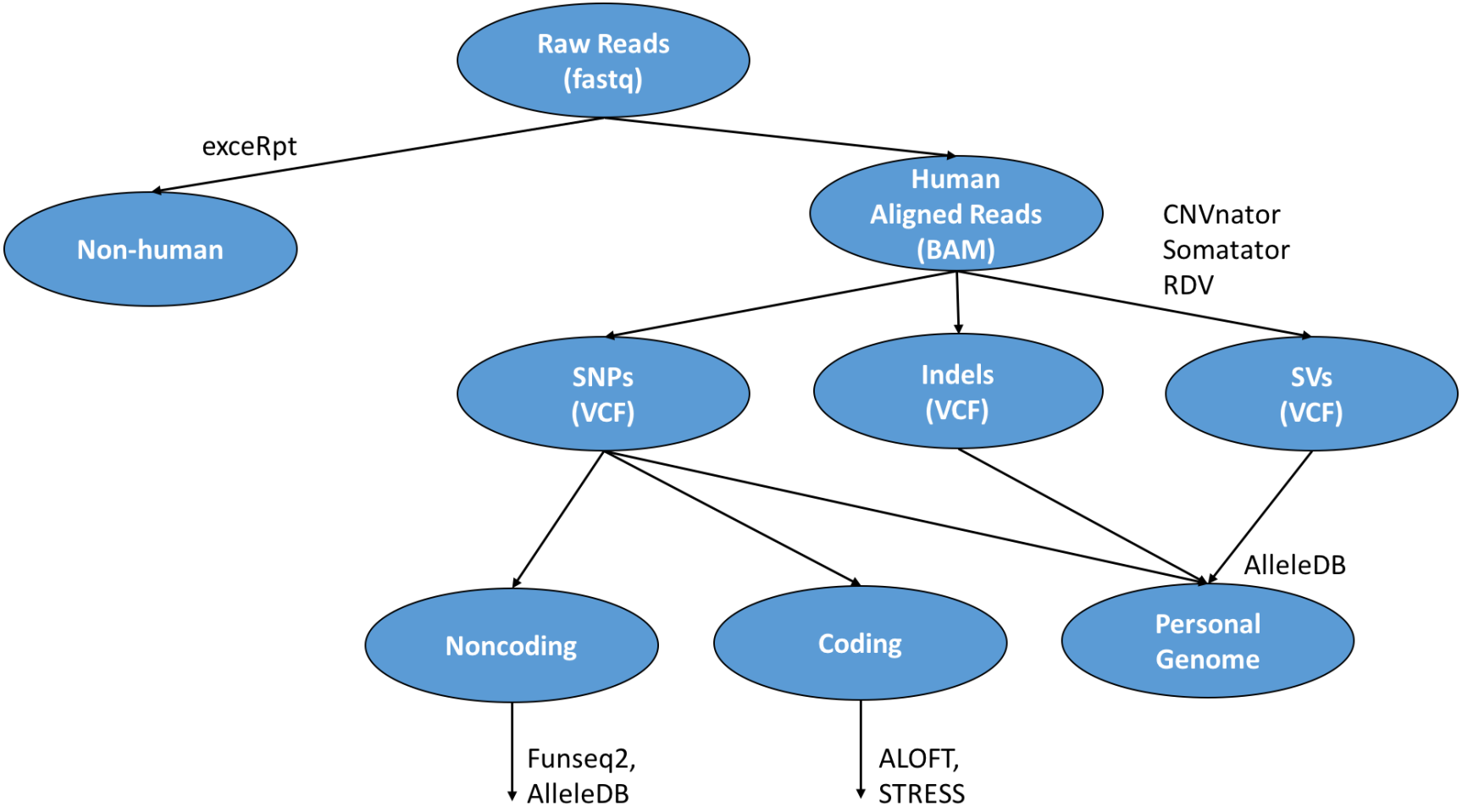
Small Insertions and DELetions (INDEL) – 1-10nt:

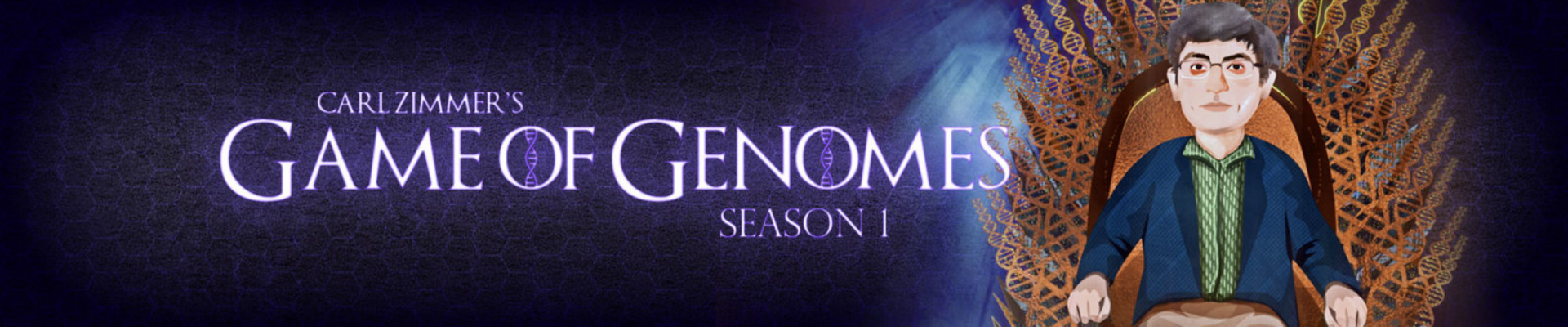
…GGAGTCTTCCAGTGTGATGATGGT~~GAGGATG~~GGCCTCCGGTT…

Large Structural Variations (SV) -- >100nt:

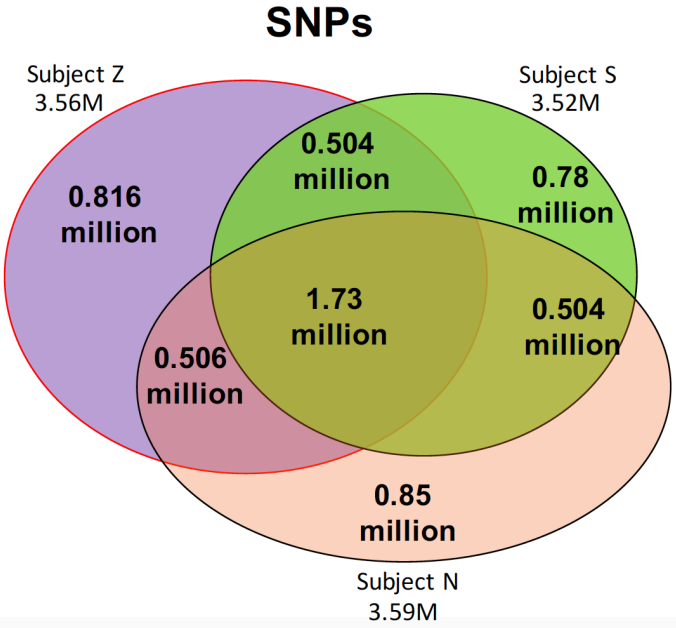
…GGAGTC~~TTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…~~


CARL ZIMMER'S
GAME OF GENOMES
SEASON 1

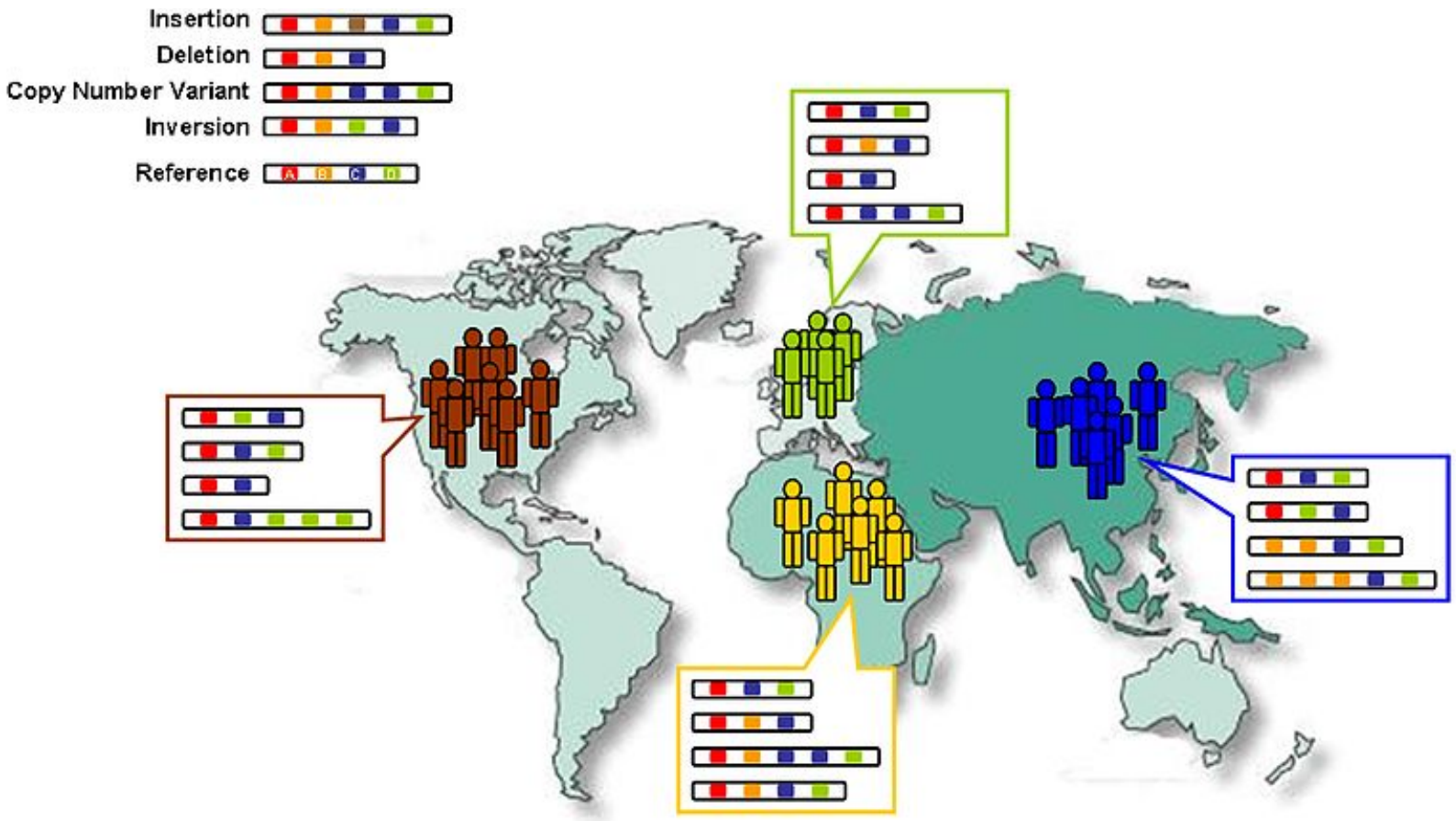

Raw Reads (fastq)
exceRpt
Non-human
Human Aligned Reads (BAM)
CNVnator
Somatator
RDV
SNPs (VCF)
Indels (VCF)
SVs (VCF)
AlleleDB
Noncoding
Coding
Personal Genome
Funseq2, AlleleDB
ALOFT, STRESS



- Normal range of number of SNPs
- Carl's case: more than 3M SNPs
- How do we know if the SNP is harmful?



- Thousand genome project
- Common SNP data base found in the population

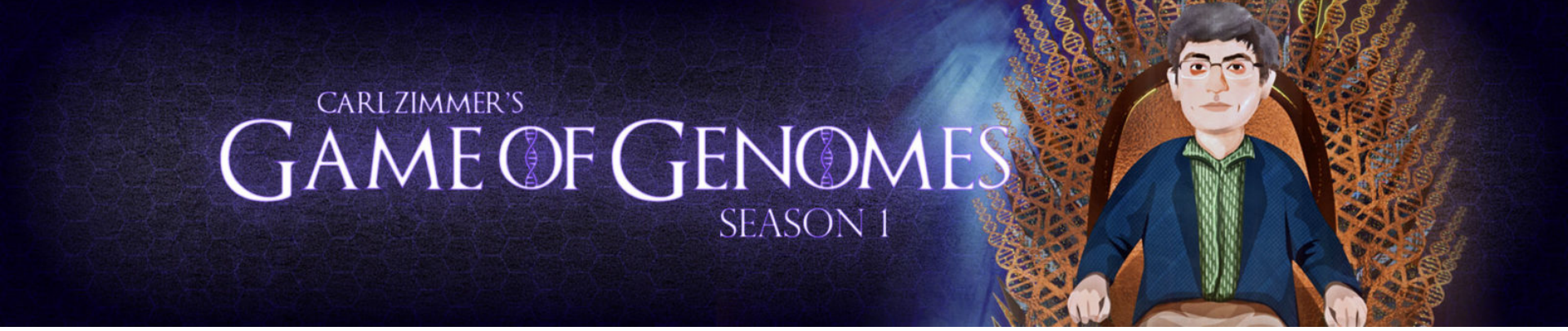


https://en.wikipedia.org/wiki/1000_Genomes_Project

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.

VS

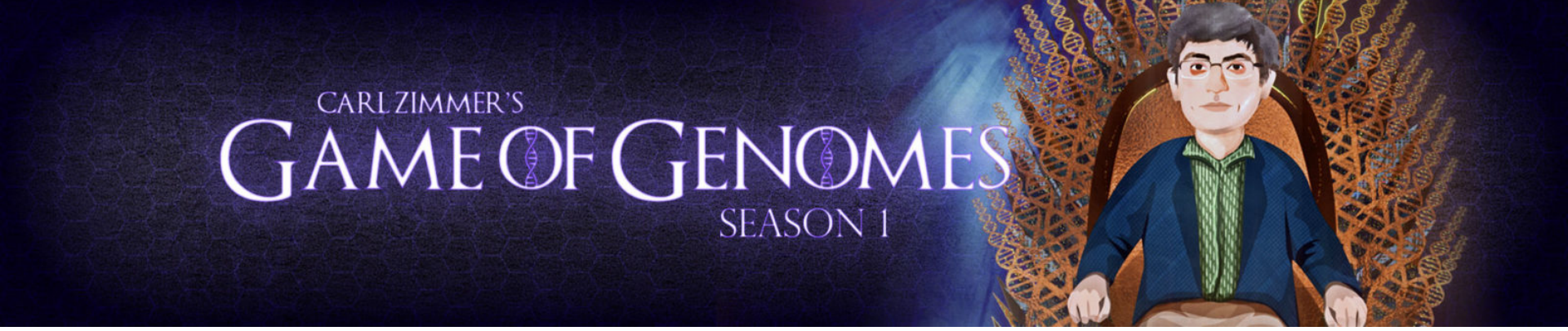


- **Got a variant in a gene for heart muscles, called DSG2**

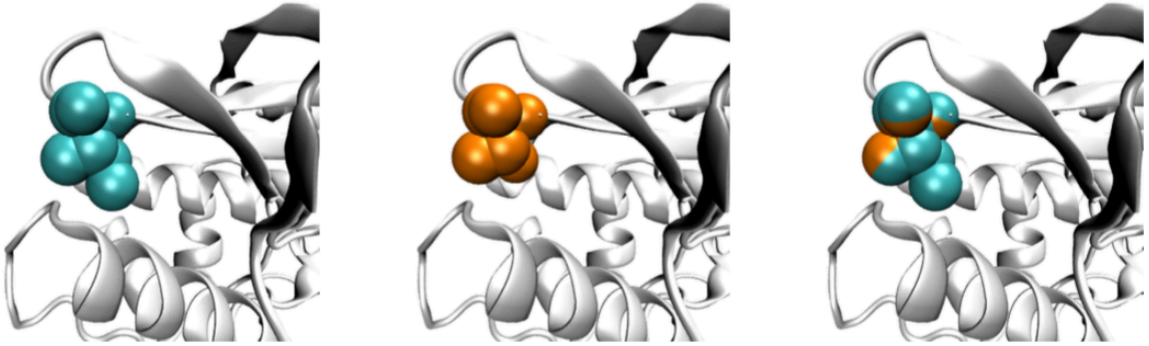
- **DSG2 gene encodes a protein in humans called Desmoglein-2**
- **Mutations in desmoglein-2 have been associated with arrhythmogenic right ventricular cardiomyopathy**

1 in 200 People of European descent carry this variant

We're all different in our DNA. We're finally starting to understand when those differences matter ---- Carl Zimmer



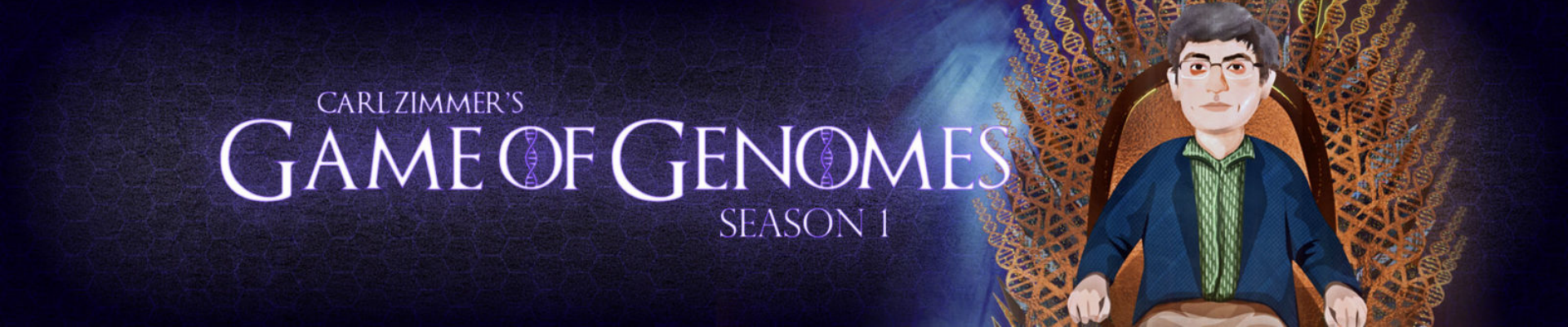
SNP changing protein structure



Wild-type *Mutated* *(superimposed)*

114: I->T

- NAT2, an enzyme in the liver that breaks down caffeine and other toxins with a similar molecular structure.
- NAT2 helps break down certain medicines too. The variant puts people at risk of bad side effects from those drugs.



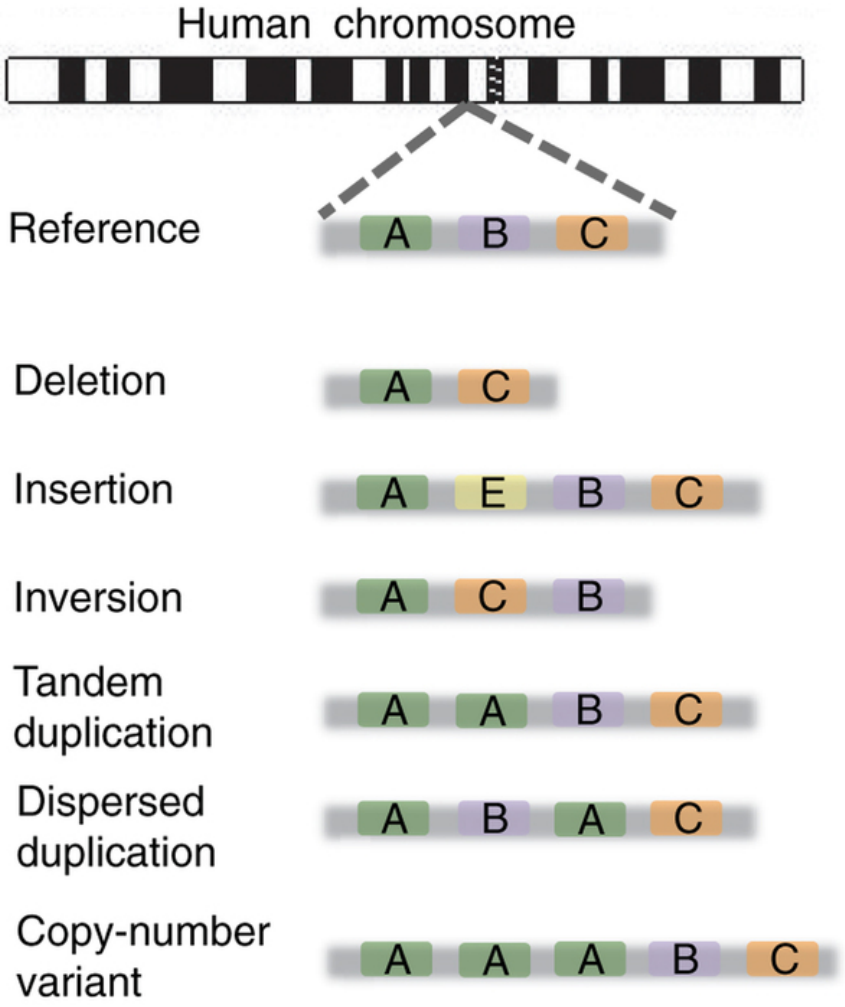
Indels (Insertions/deletions)

- In coding regions, unless the length of an indel is a multiple of 3, it will produce frameshift mutation
- Likely to disrupt genes (loss-of-function variant)

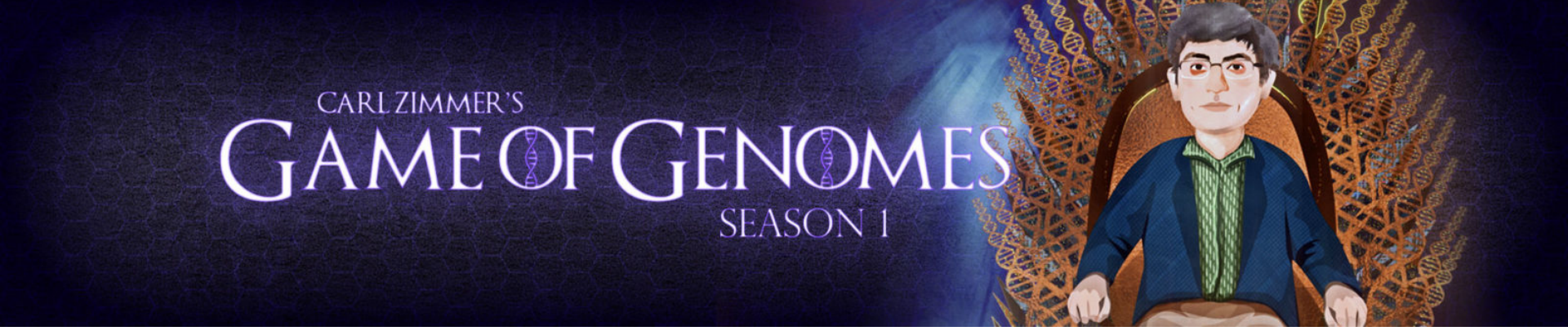
Example: Non-functional F8 gene

- Can't make essential clotting protein
- Get hemophilia and can bleed to death from a little cut

Structural Variation

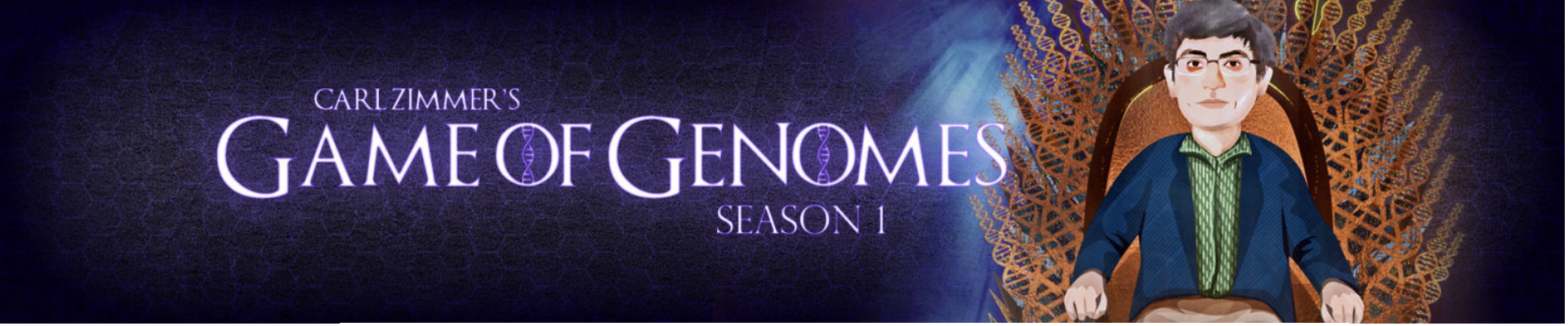


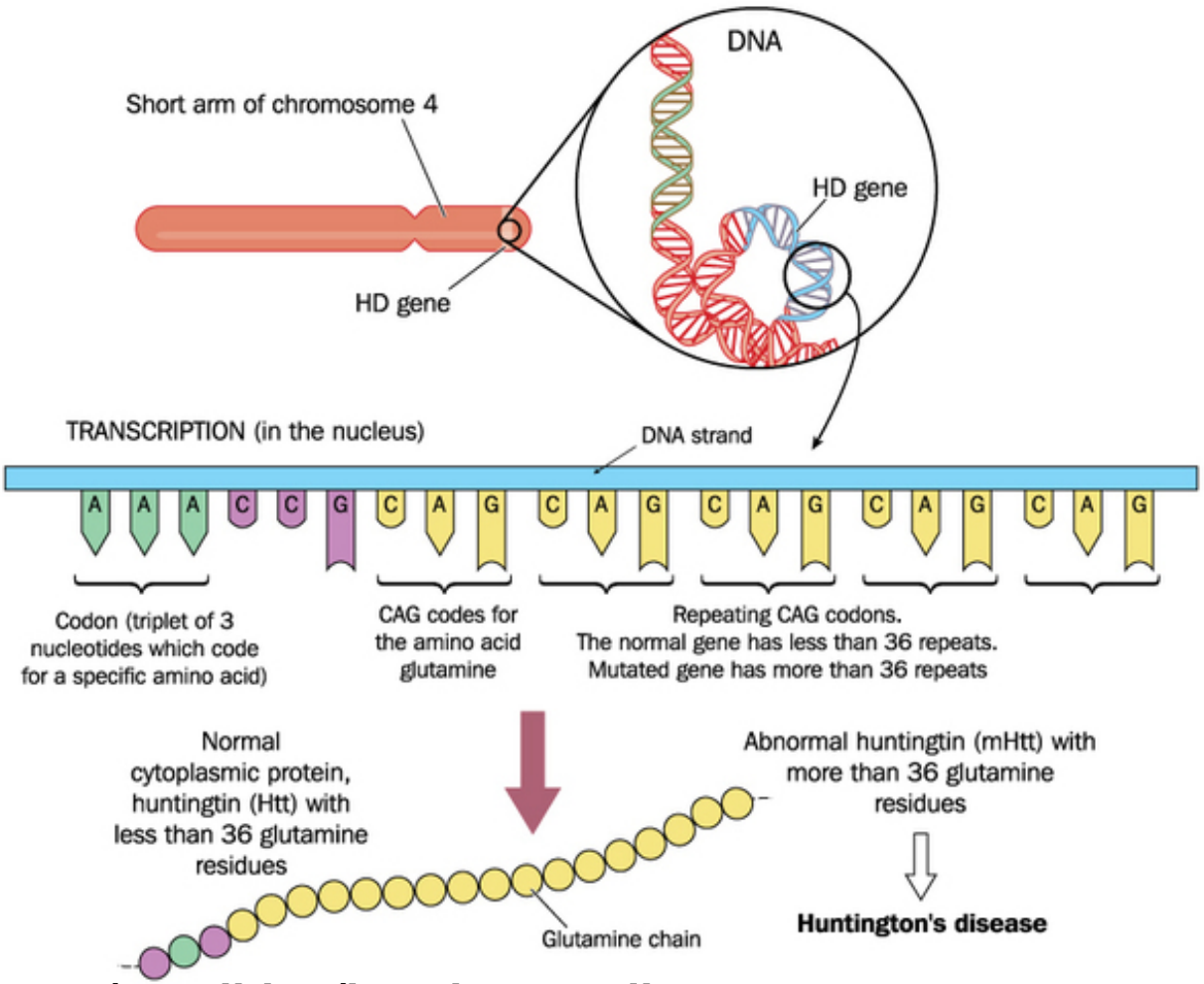
Baker, M., 2012. *Nature methods*, *9*(2), pp.133-137.



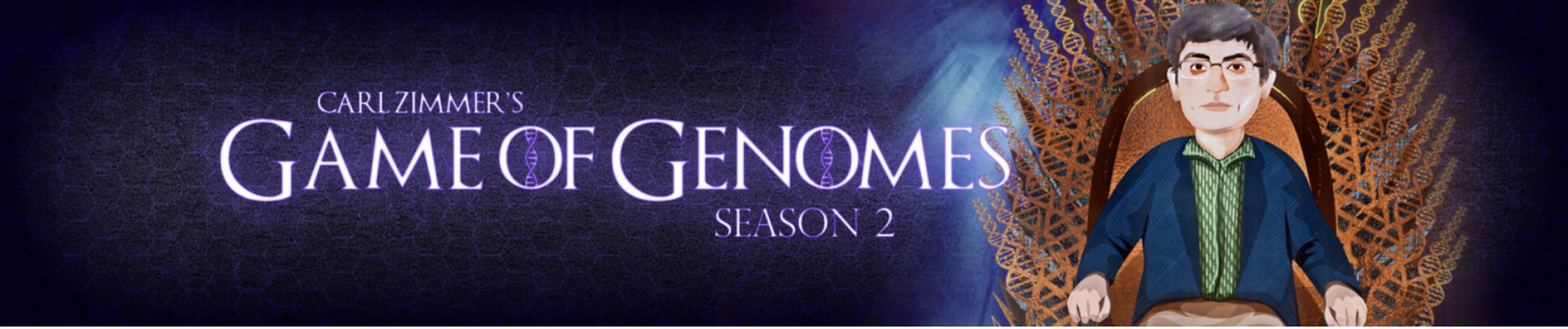
- Structural variation
- Example: HTT
- Certain mutations in HTT cause Huntington's disease.
- Healthy people have a wide range of CAG repeats. It's only when people get 37 or more CAG repeats in HTT that they are at risk of developing Huntington's disease.

- The reference genome has 19 CAG repeats. Carl has 17.



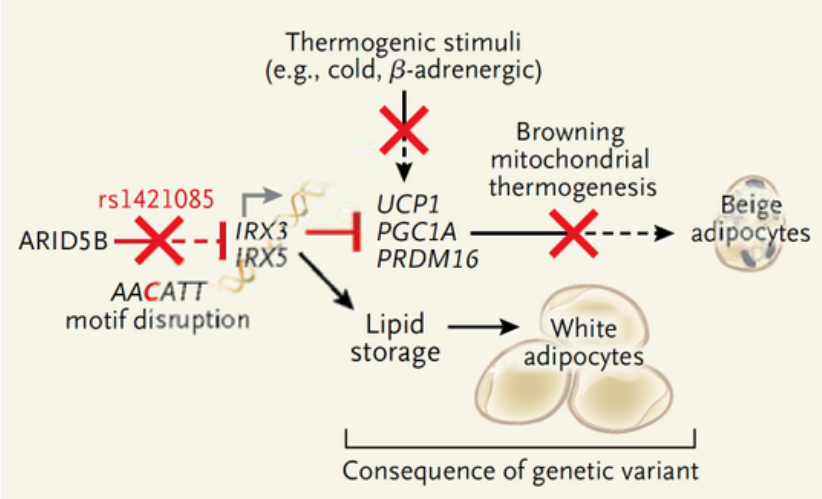


https://ghr.nlm.nih.gov/condition/huntington-disease

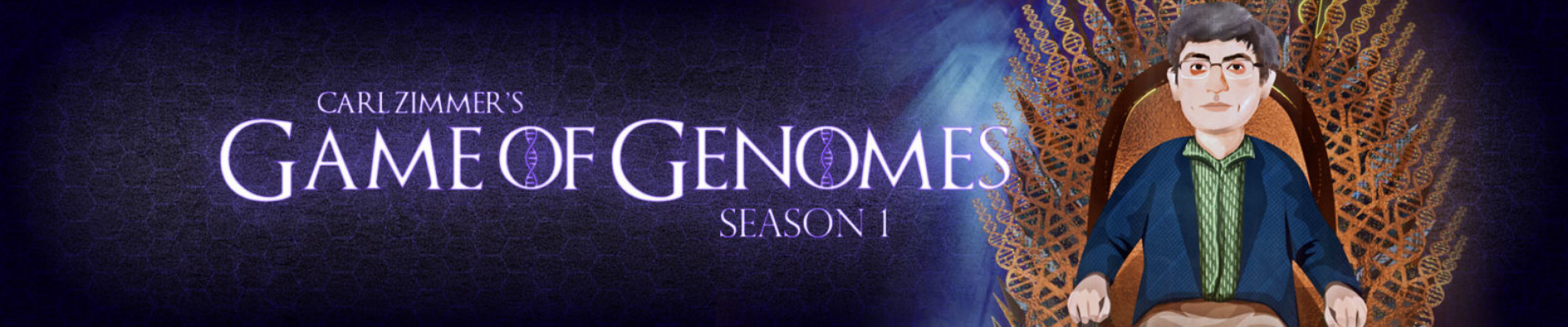


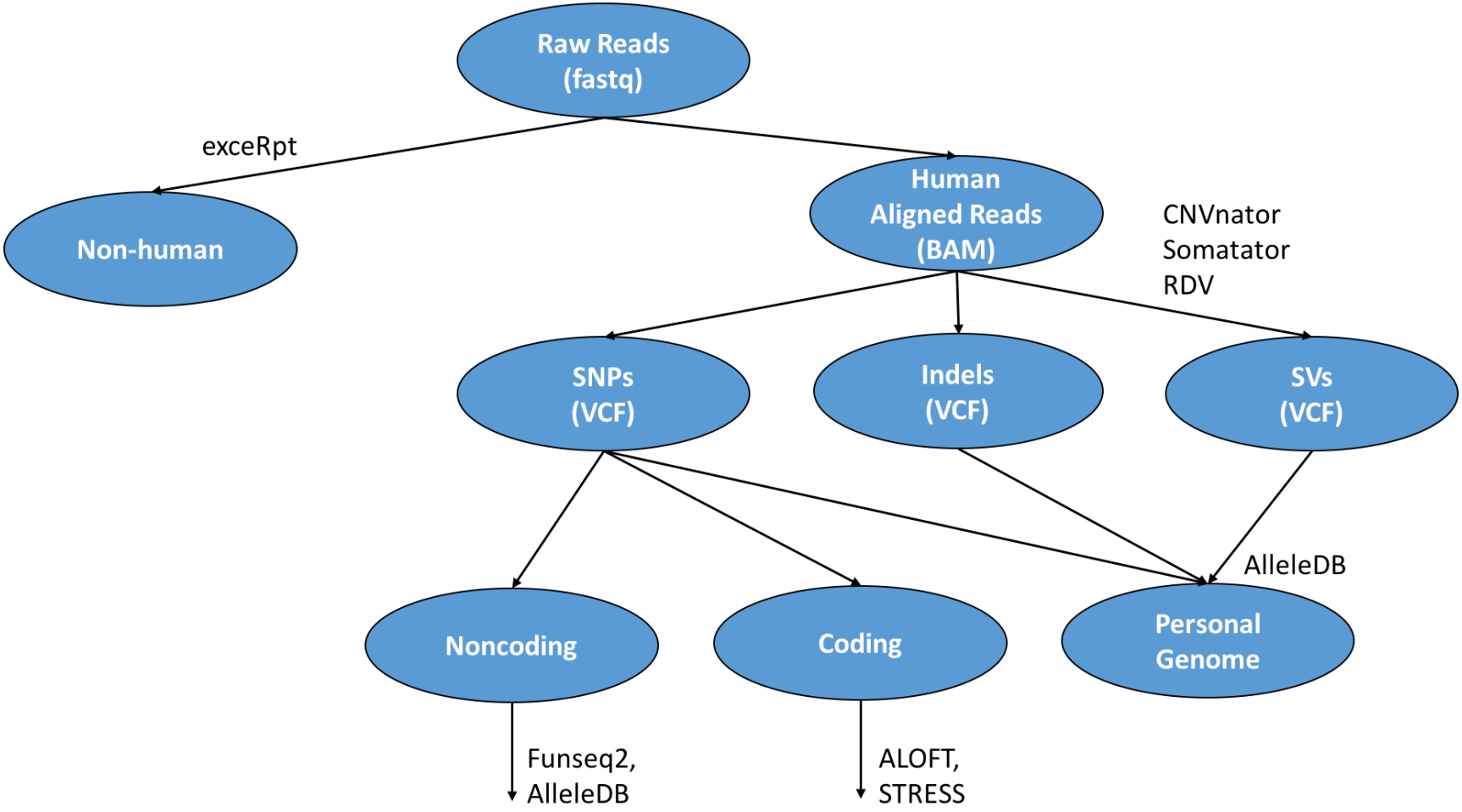
Non-coding variant

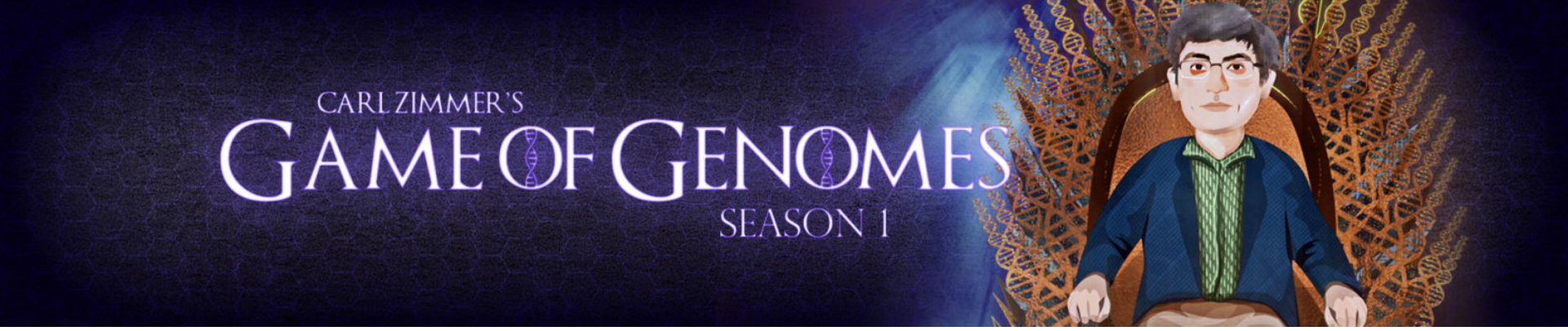
- Variant rs1421085
- Located in a genetic switch that activates several genes in fat cells
- The variant causes people to put on an average of 7 pounds



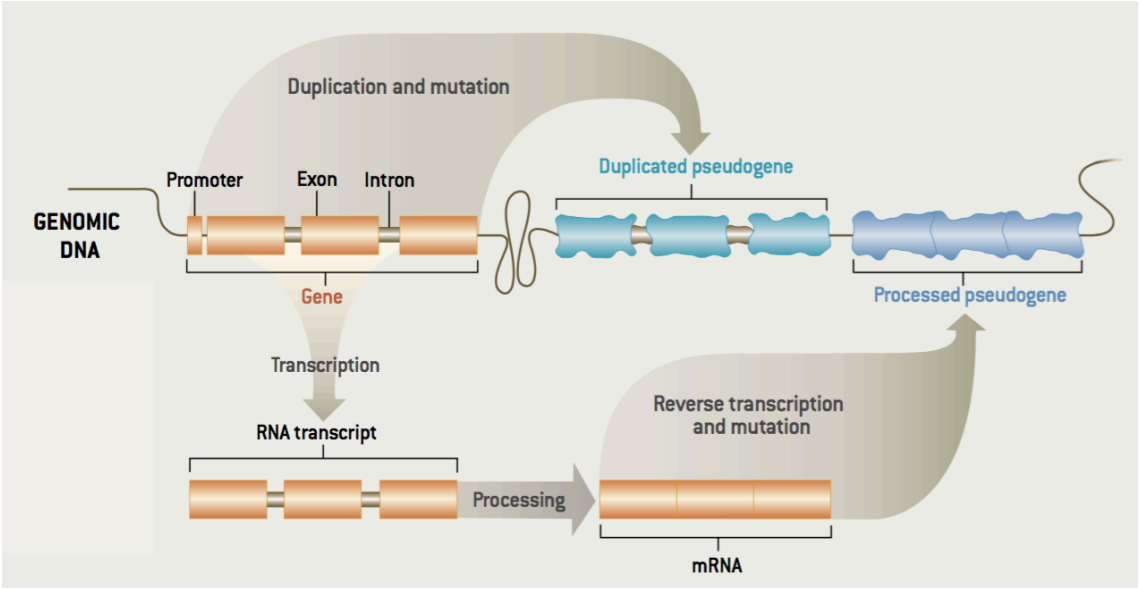
Claussnitzer, Melina, et al. *N Engl J Med* 2015.373 (2015): 895-907.


CARL ZIMMER'S
GAME OF GENOMES
SEASON 1


Raw Reads (fastq)
exceRpt
Non-human
Human Aligned Reads (BAM)
CNVnator
Somatator
RDV
SNPs (VCF)
Indels (VCF)
SVs (VCF)
AlleleDB
Noncoding
Coding
Personal Genome
Funseq2, AlleleDB
ALOFT, STRESS



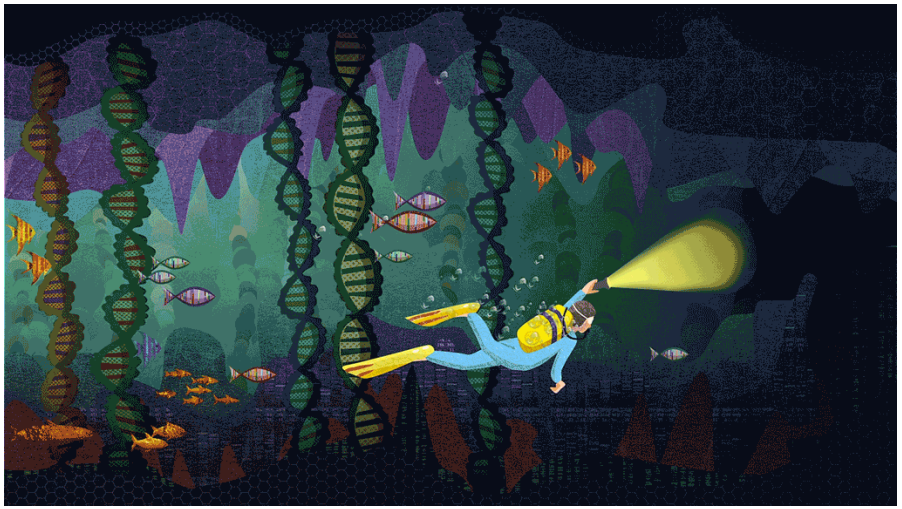
- What else are in the genome?
- Pseudogenes
- About 14000 pseudogenes carried in our genome



Gerstein and Zheng SiAm 2006

Looking beyond the genome

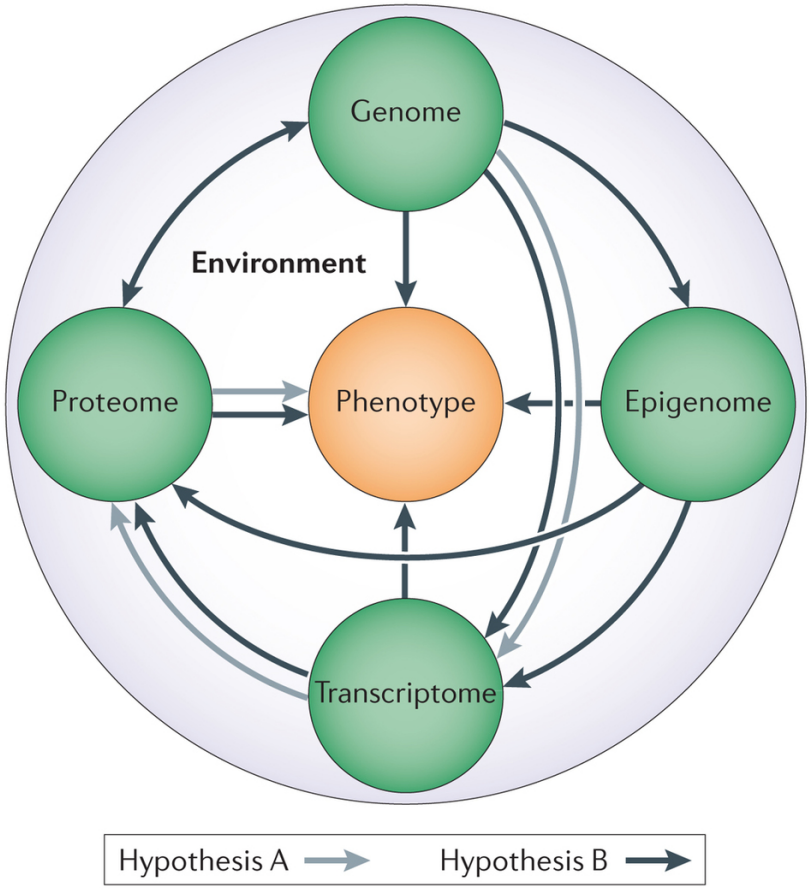
- In the Game of Genomes Carl Zimmer explored his genomic sequence.
- The genome provides a mostly static view.
- Misses the active regulation, transcription, and translation



https://www.statnews.com/feature/game-of-genomes/

Feedback is important in biological systems

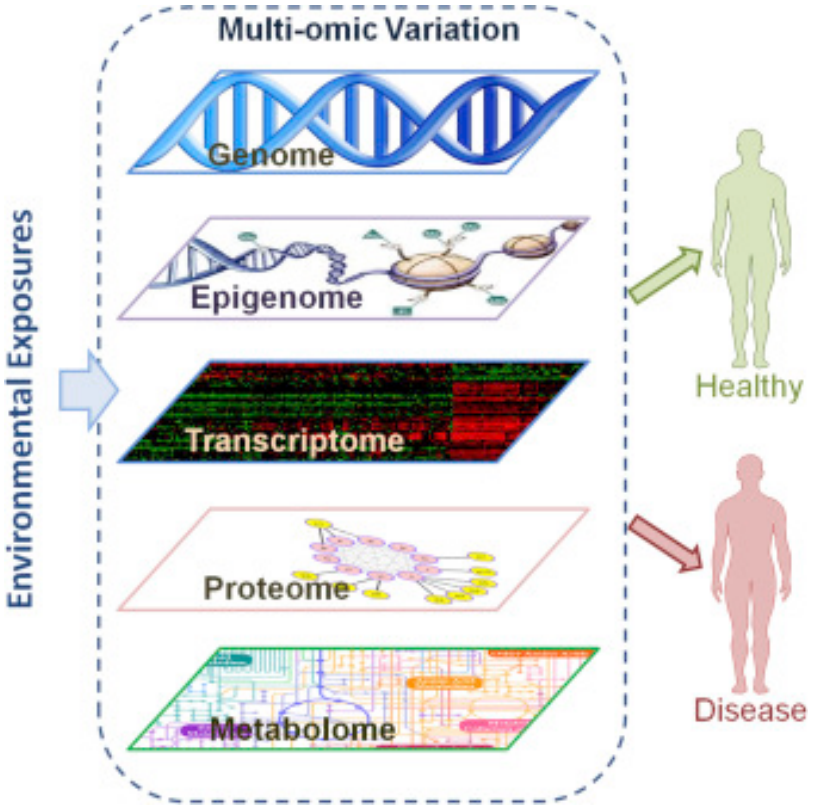
- The flow of genetic information isn't linear.
- Constant feedback between the genome, transcriptome, and proteome and environment.
- Can't always accurately predict phenotype from any single 'ome'.



Ritchie et al. Nat. Rev. Genetics 2015

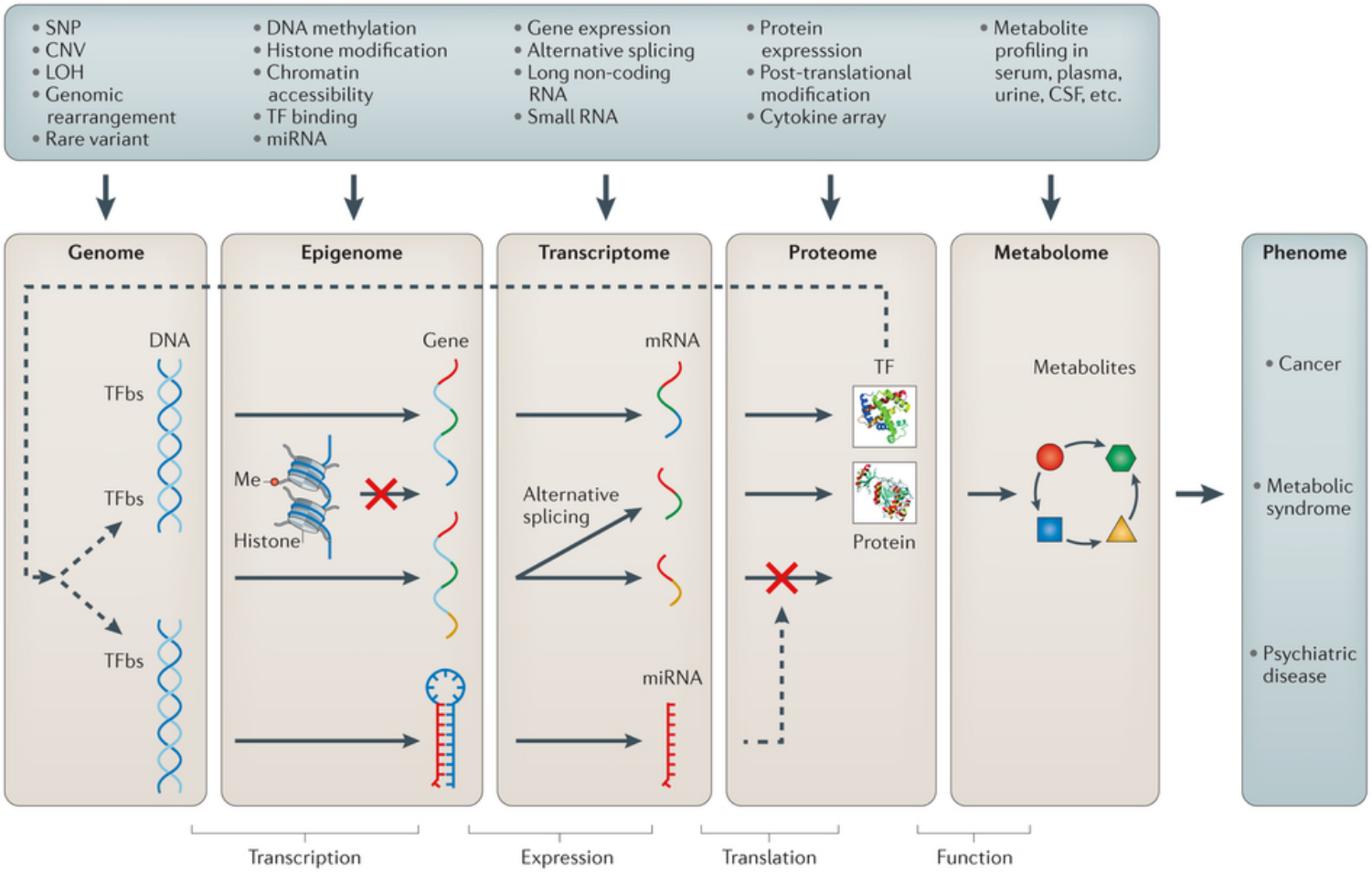
Integrating environmental factors, genetic background, and large scale datasets

- Difference between health and disease depends on many factors.
- Environment, genome, cellular contents, etc. all play a a role.
- Important to integrate information from multiple large-scale datasets.



Sun et al. Advances in Genetics 2016

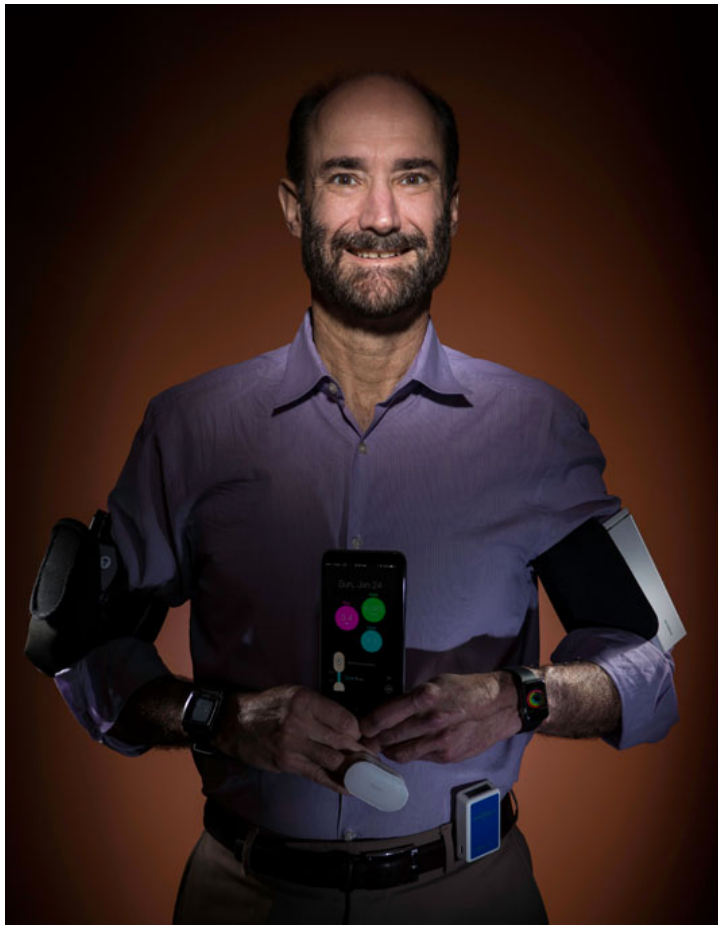
Different large-scale assays provide information on many types of biological regulation



Ritchie et al. Nat. Rev. Genetics 2015

Expanding personalized medicine beyond the genome.

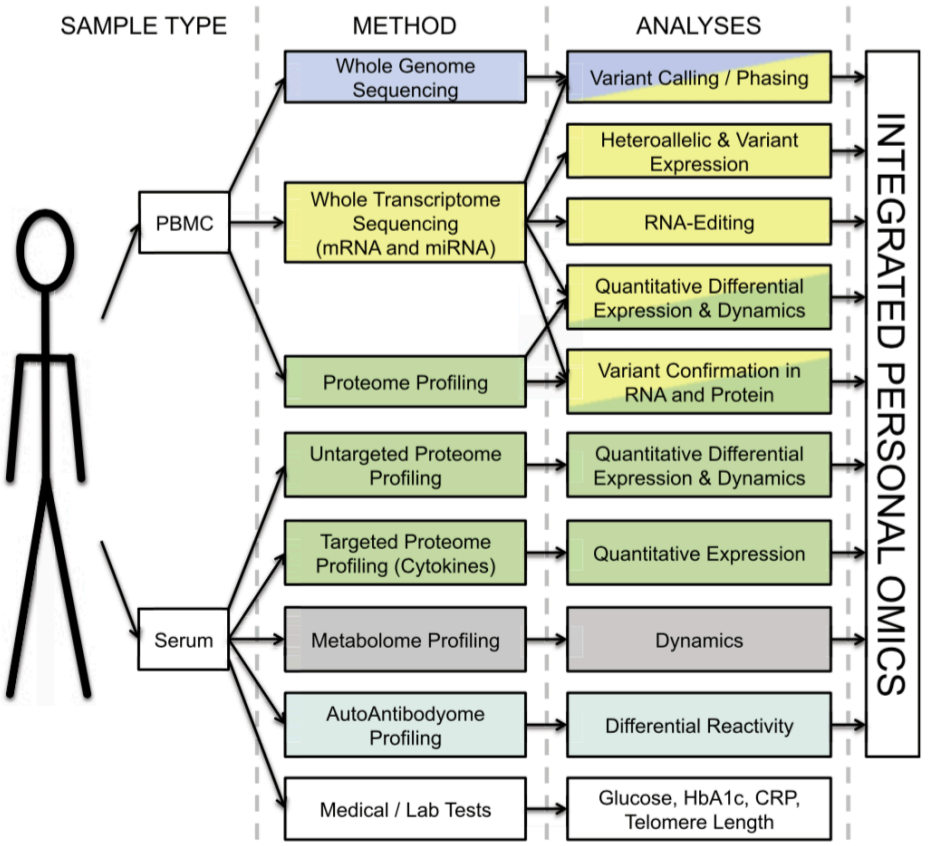
- An integrated personal omics profile (iPOP) is an example of a more comprehensive version of personalized medicine.
- Michael Snyder had his genome sequenced and collected many other large scale datasets over an extended period of time.



https://med.stanford.edu/news/all-news/2017/01/wearable-sensors-can-tell-when-you-are-getting-sick.html

Integrated personal omics profile (iPOP)

- Numerous types of data were collected, primarily from blood samples. The datasets include:
  - Transcriptomic
  - Proteomic
  - Metabolomic
  - Cytokine profiling
  - Autoantibody profiling
  - Medical exams



Chen et al. Cell 2012

Michael Snyder's personal genome: a starting point

Table 1. Summary and Breakdown of DNA Variants

Type	Total Variants	Total High Confidence	Heterozygous High Confidence	Homozygous High Confidence
Total SNVs	3,739,701	3,301,521	1,971,629	1,329,892
Total gene-associated SNVs	1,312,780	1,183,847	717,485	466,362
Total coding/UTR	49,017	44,542	27,383	17,159
Missense	10,592	9,683	5,944	3,739
Nonsense	83	73	49	24
Synonymous	11,459	10,864	6,747	4,117
5′UTR	4,085	2,978	1,802	1,176
3′UTR	22,798	20,944	12,841	8,103
Intron	1,263,763	1,139,305	690,102	449,203
Ts/Tv	—	2.14	—	—
dbSNP	3,493,748	3,167,180	—	—
Candidate private SNV	245,953	134,341	—	—
Indels (−107∼ +36 bp)	1,022,901	216,776	—	—
Coding	3,263	302	—	—
Structural variants (>50 bp)	44,781	2,566	—	—
In 1000G project[a]	4,434	1,967	—	—

Chen et al. Cell 2012

Prioritizing variants by leveraging mutation databases

- Using existing databases of population level genetic variation, rare and disease associated variants could be identified.
- Helped prioritize medical conditions for monitoring (e.g. glucose for diabetes)

High Interest Disease-Associated Rare Variants.

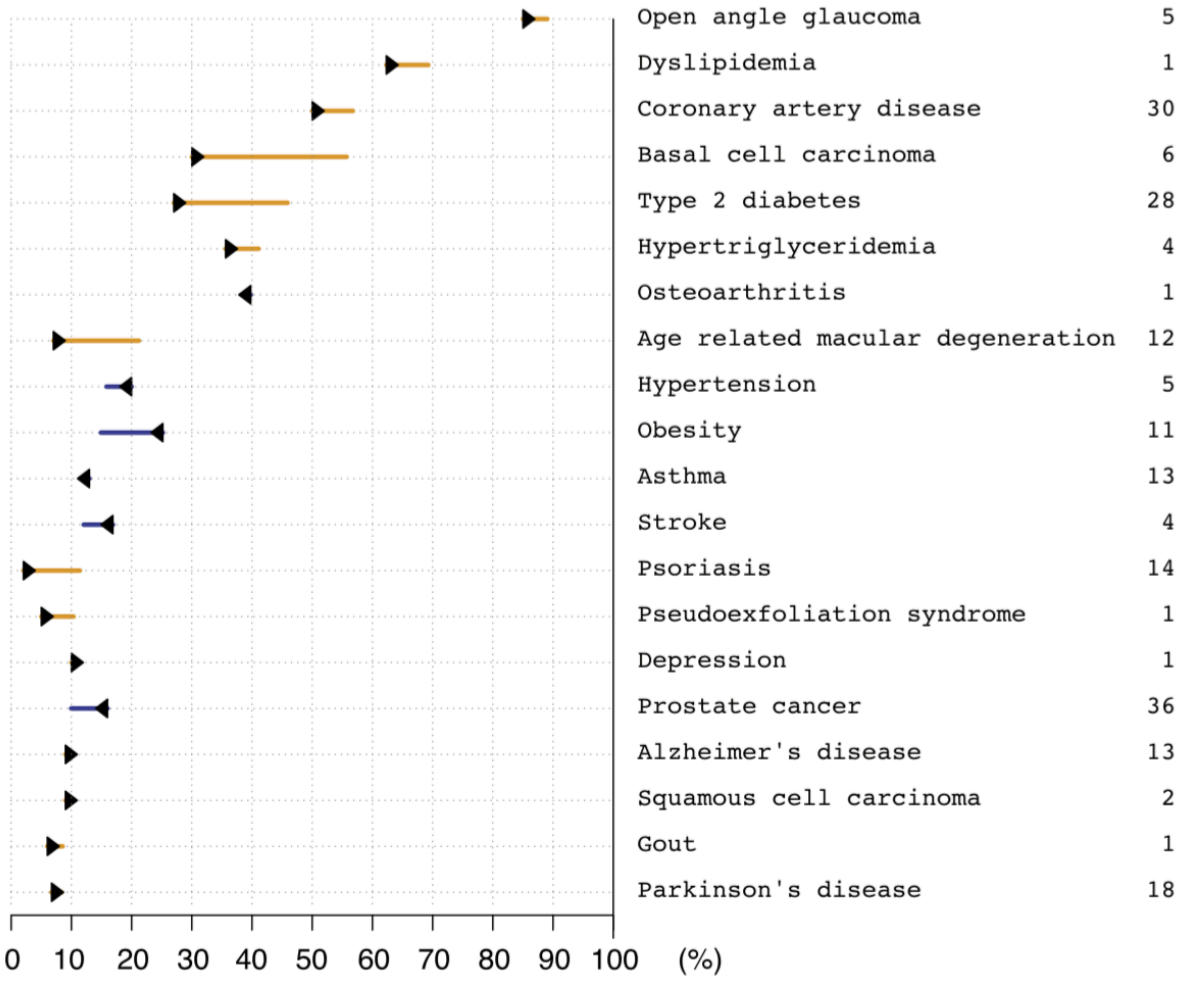
Gene	Position	Genotype	OMIM
SERPINA1	14:94844947	C/T	Emphysema due to AAT deficiency
TERT	5:1294397	C/T	Aplastic anemia
KCNJ11	11:17409571	T/T	Type 2 diabetes
GCKR	2:27730939	T/T	Hypertriglyceridemia
NUP54	4:77055431	G/A	Nuclear Pore Complex Protein

High Interest Drug-Related Variants.

Gene	rsID	Genotype	Drug Response Affected
	rs10811661	C/T	Troglitazone (Increased Beta-Cell Function)
CYP2C19	rs12248560	C/T	Clopidogrel (Increased Activation)
LPIN1	rs10192566	G/G	Rosiglitazone (Increased Effect)
SLC22A1	rs622342	A/A	Metformin (Increased Effect)
VKORC1	rs9923231	C/T	Warfarin (Lower Dose Required)

Chen et al. Cell 2012

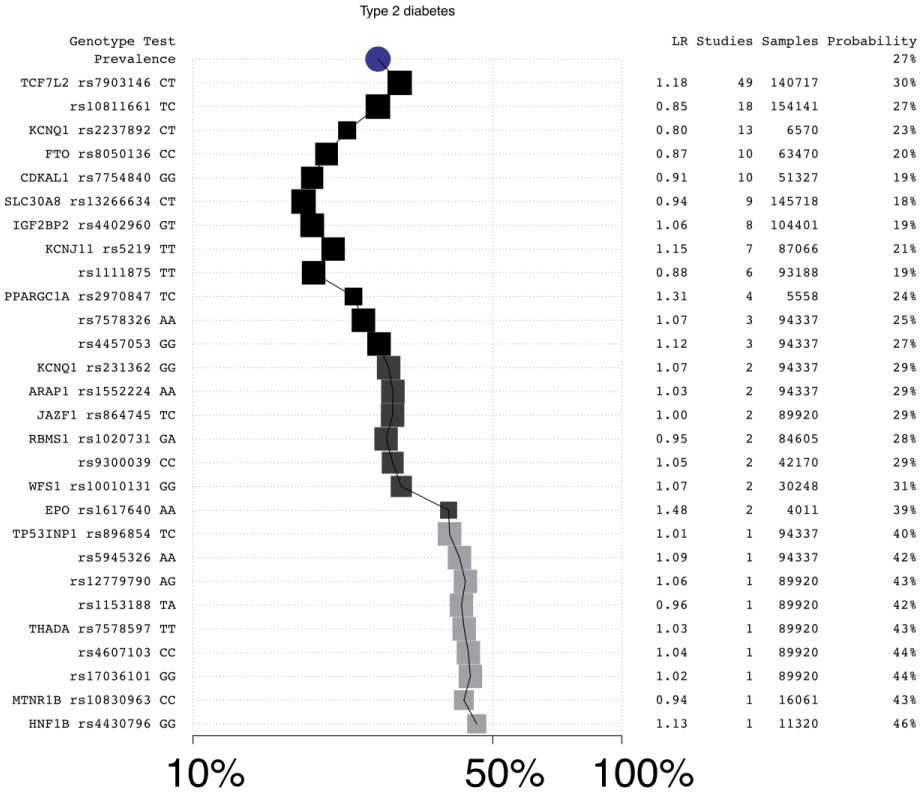
Genomic information helps refine disease risk estimates



Chen et al. Cell 2012

Prioritizing variants by leveraging mutation databases

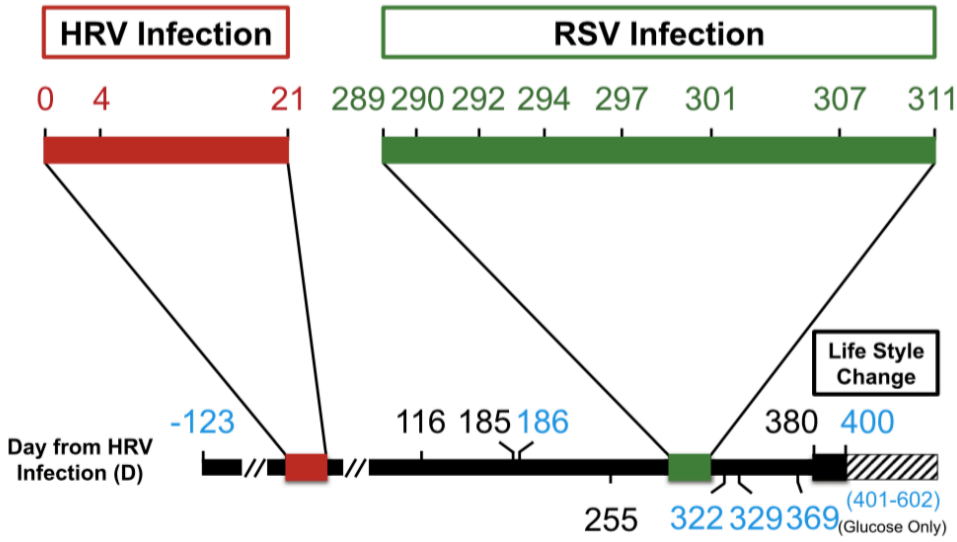
- Variants at multiple loci are taken into account to produce an risk estimate.
- The cumulative impact of both protective and deleterious mutations is evaluated.



Chen et al. Cell 2012

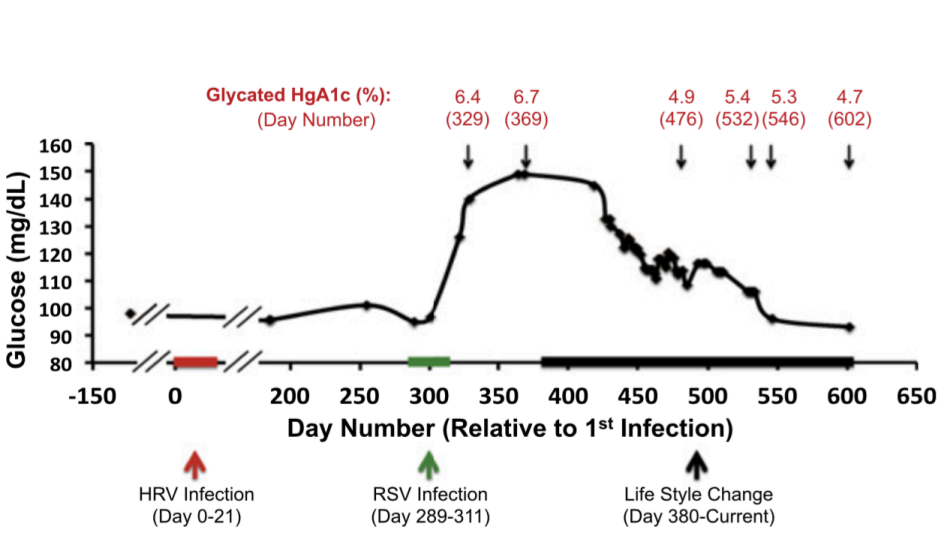
Personal omics profiling time course

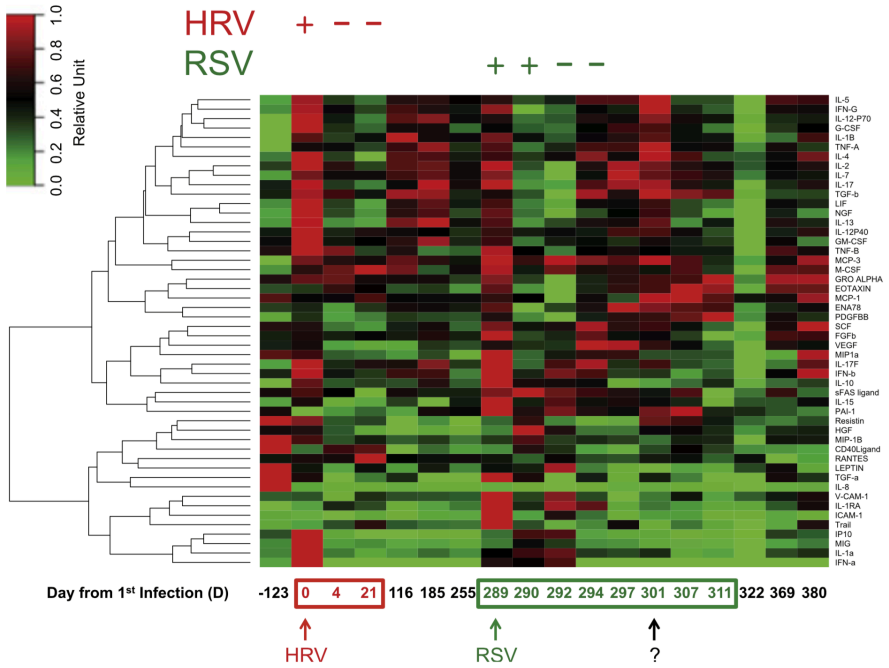
- Changing cellular state and environmental perturbations impact the genome.
- Longitudinal data collection tracks the dynamic regulation of the genome.



Chen et al. Cell 2012

Longitudinal medical data

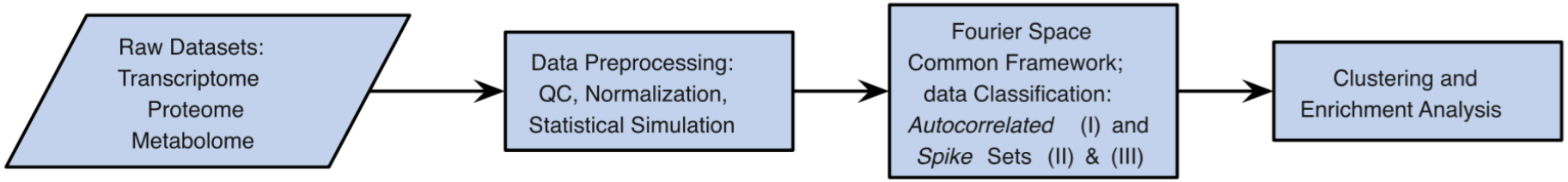




- Tracking relevant medical (e.g. blood glucose) data over time helps link phenotypic changes with changes at the molecular level.

Chen et al. Cell 2012

Multi omic time course data collection

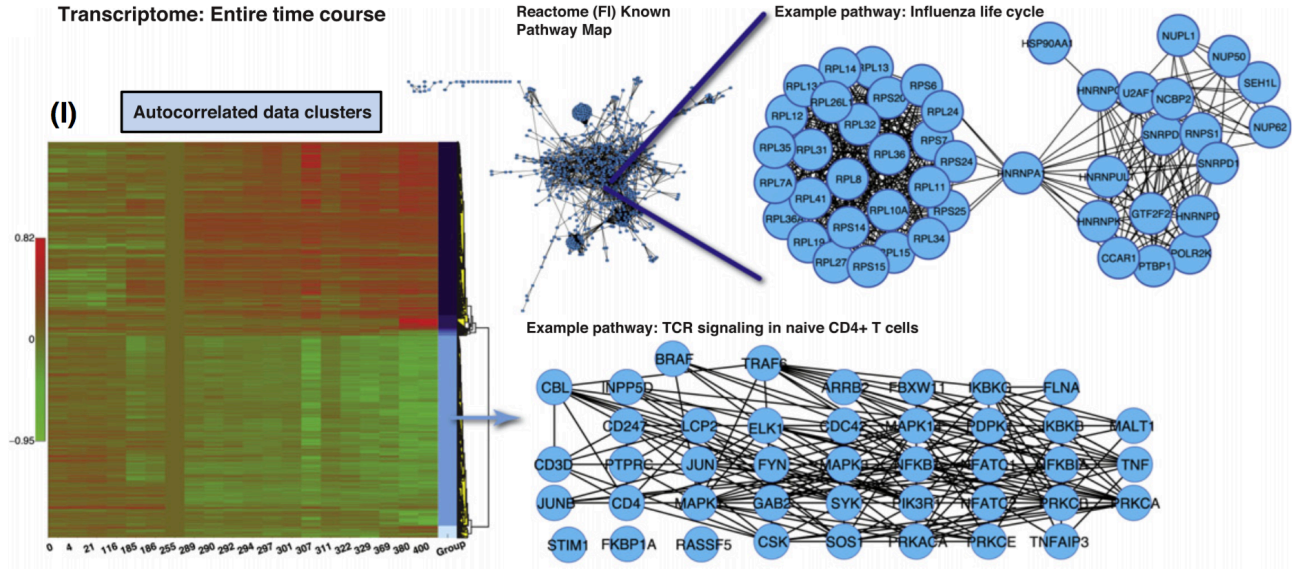


- Time course data enabled studying expression dynamics before, during, and after viral infection.

- Clustering, pathway analysis, and Gene Ontology term enrichment help identify key cellular processes undergoing change.

Chen et al. Cell 2012

Transcriptomic time course



Gene Ontology analysis

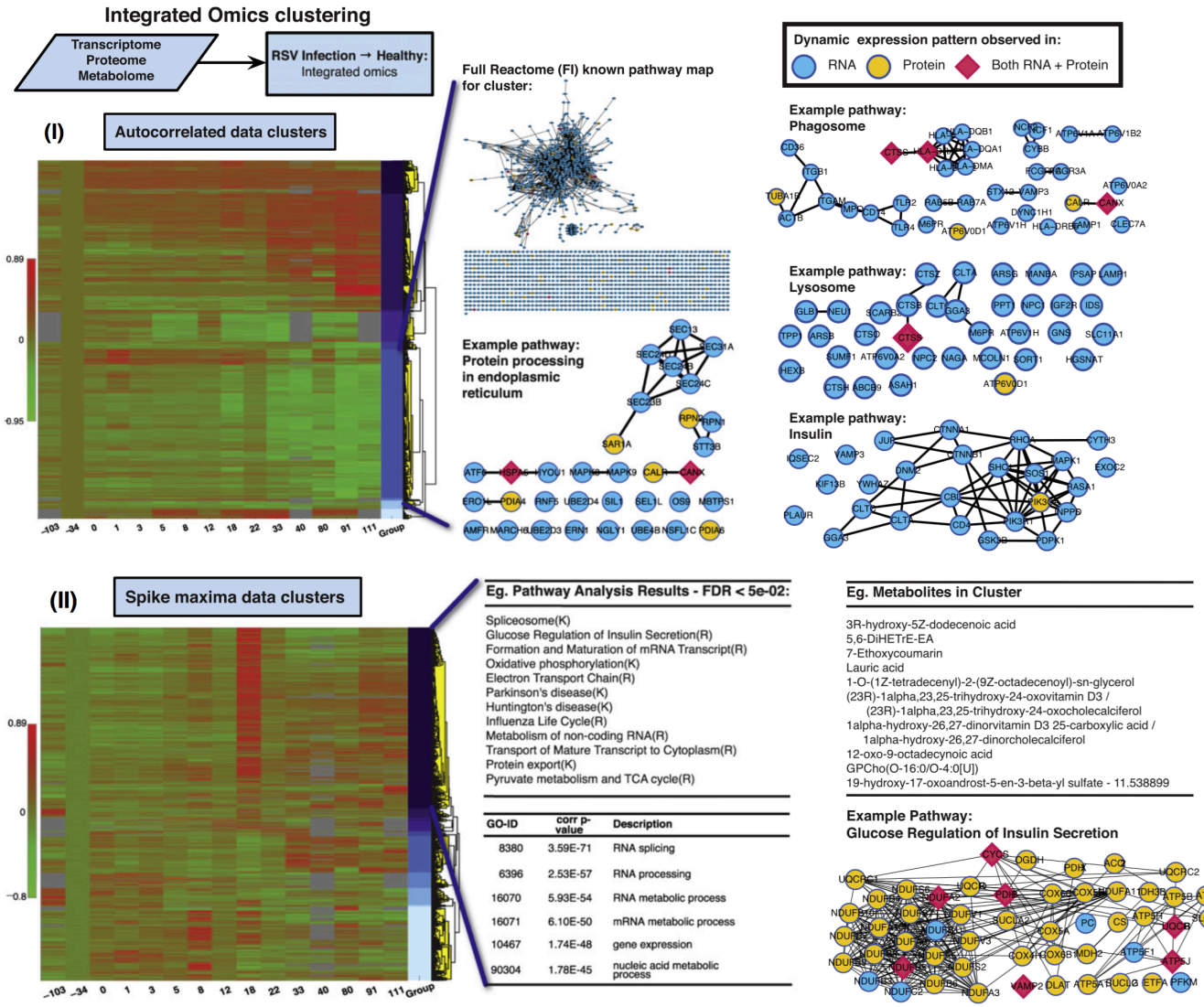
GO-ID	corr p-value	Description
19882	5.33E-12	antigen processing and presentation
2376	1.06E-08	immune system process
6955	1.24E-08	immune response
2504	1.69E-07	antigen processing & presentation of peptide or polysaccharide antigen via MHC class II
6952	1.59E-04	defense response
6954	7.39E-03	inflammatory response

Eg. Pathway analysis results - FDR < 1e-03:

Allograft rejection(K)
Intestinal immune network for IgA production(K)
Graft-versus-host disease(K)
Type I diabetes mellitus(K)
Phagosome(K)
Asthma(K)
Autoimmune thyroid disease(K)
Viral myocarditis(K)
Endocytosis(K)
Fc gamma R-mediated phagocytosis(K)

Chen et al. Cell 2012

Integration of multiple omics datasets

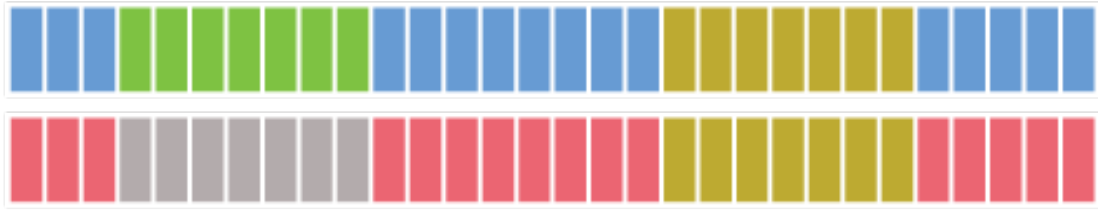


Chen et al. Cell 2012

Constructing a diploid personal genome

Haploid Reference genome

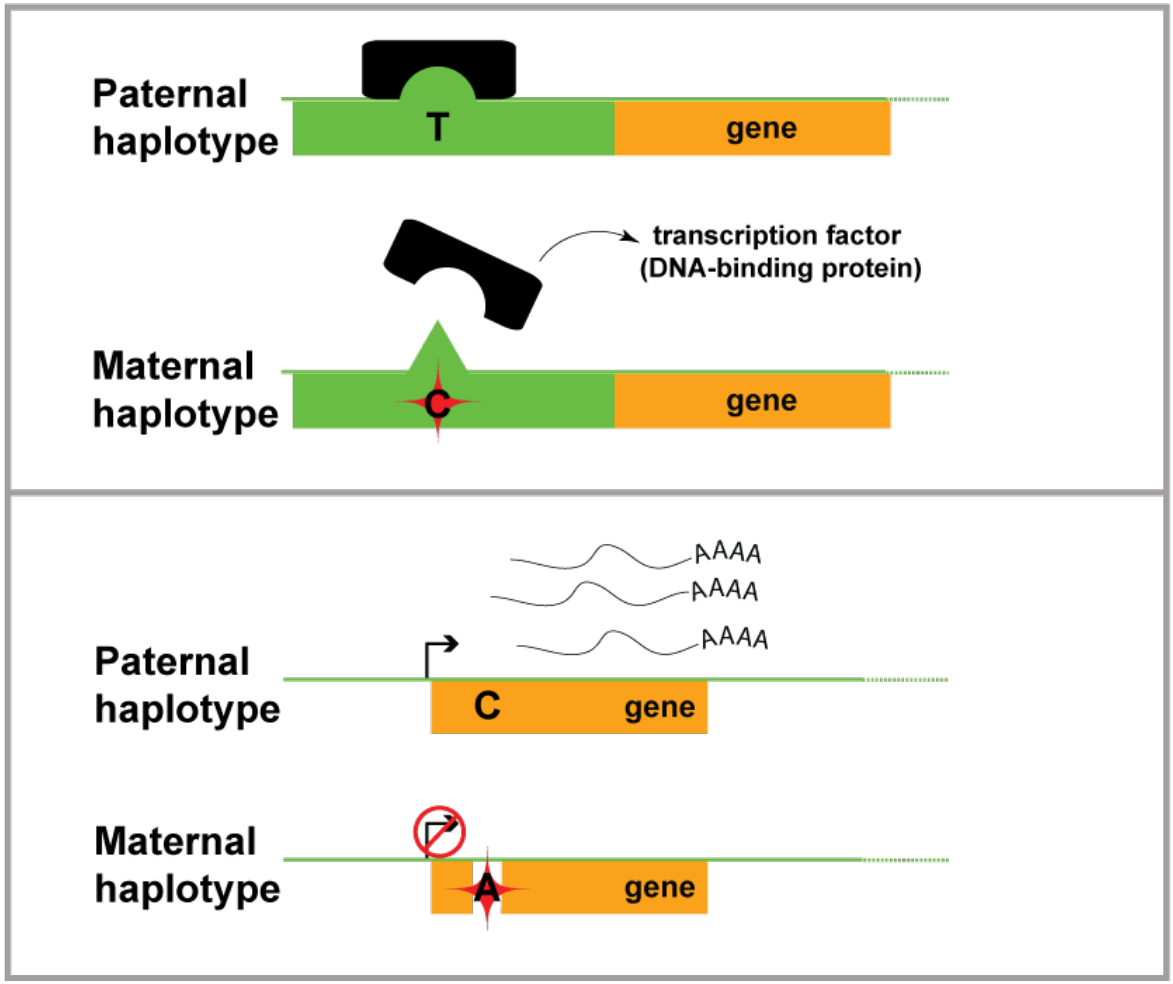
Diploid Personal genome



- Sequencing both Mike Snyder and his mother enabled creation of a high confidence diploid genome.

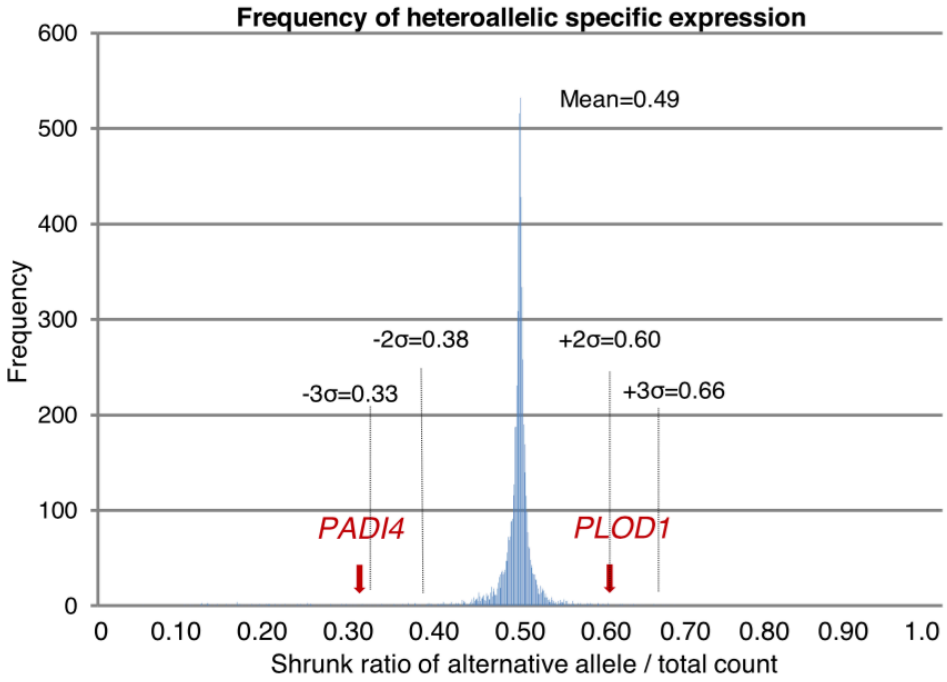
Courtesy of Joel Rozowsky

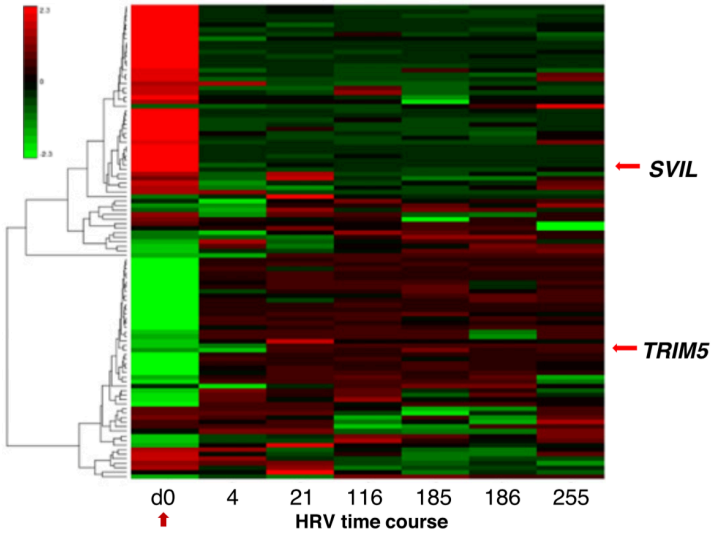
Allele Specific behavior



Courtesy of Joel Rozowsky

Allele specific expression



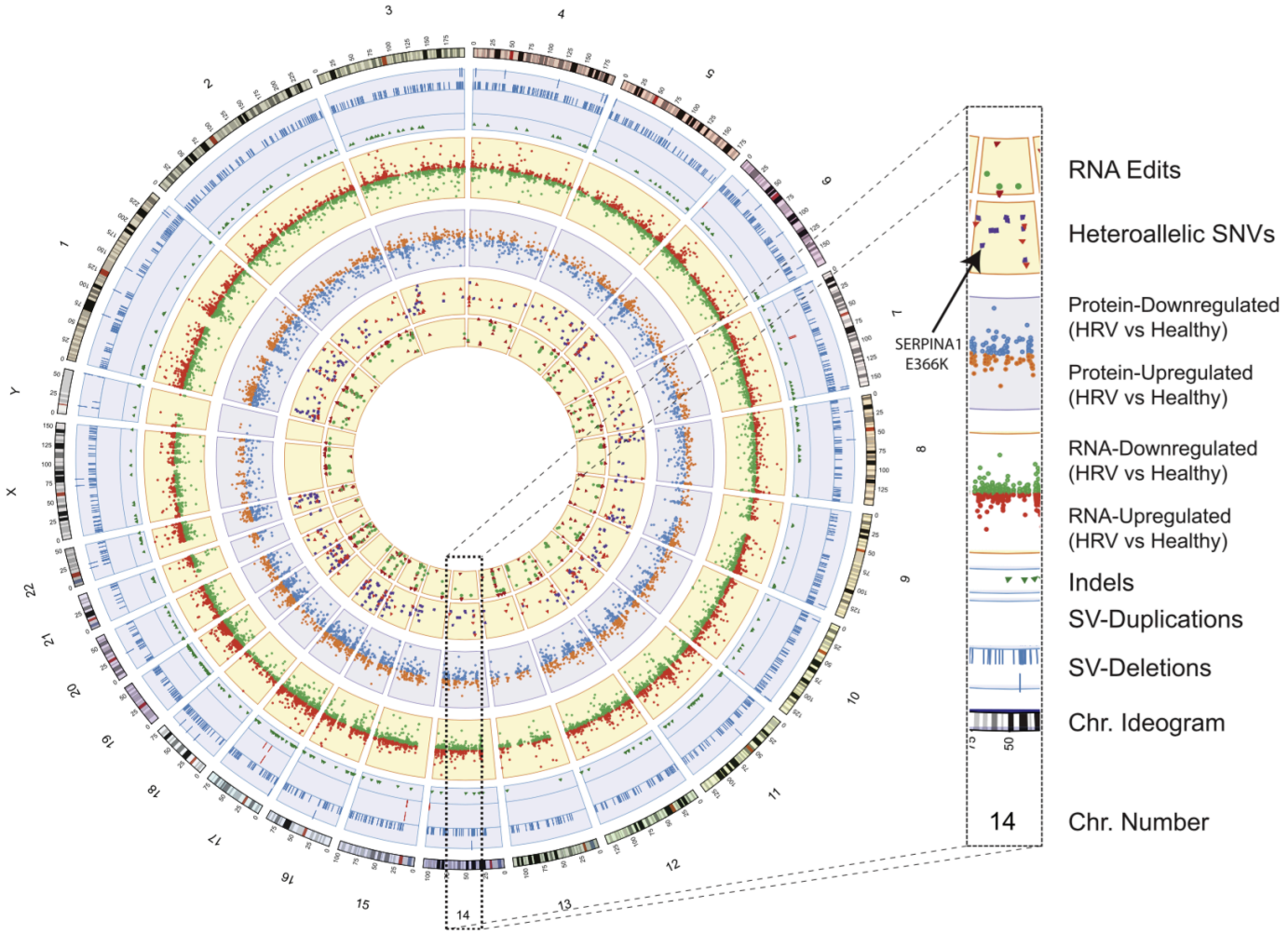


KEGG Pathway	Genes	p-value	Benjamini p-value
SNARE interactions in vesicular transport	*VAMP5, SEC22B, SNAP23, TSNARE1*	4.02E-4	2.27E-2

- The combination of transcriptomic data and a personal diploid genome enabled the identification of allele specific expression patterns.

Chen et al. Cell 2012

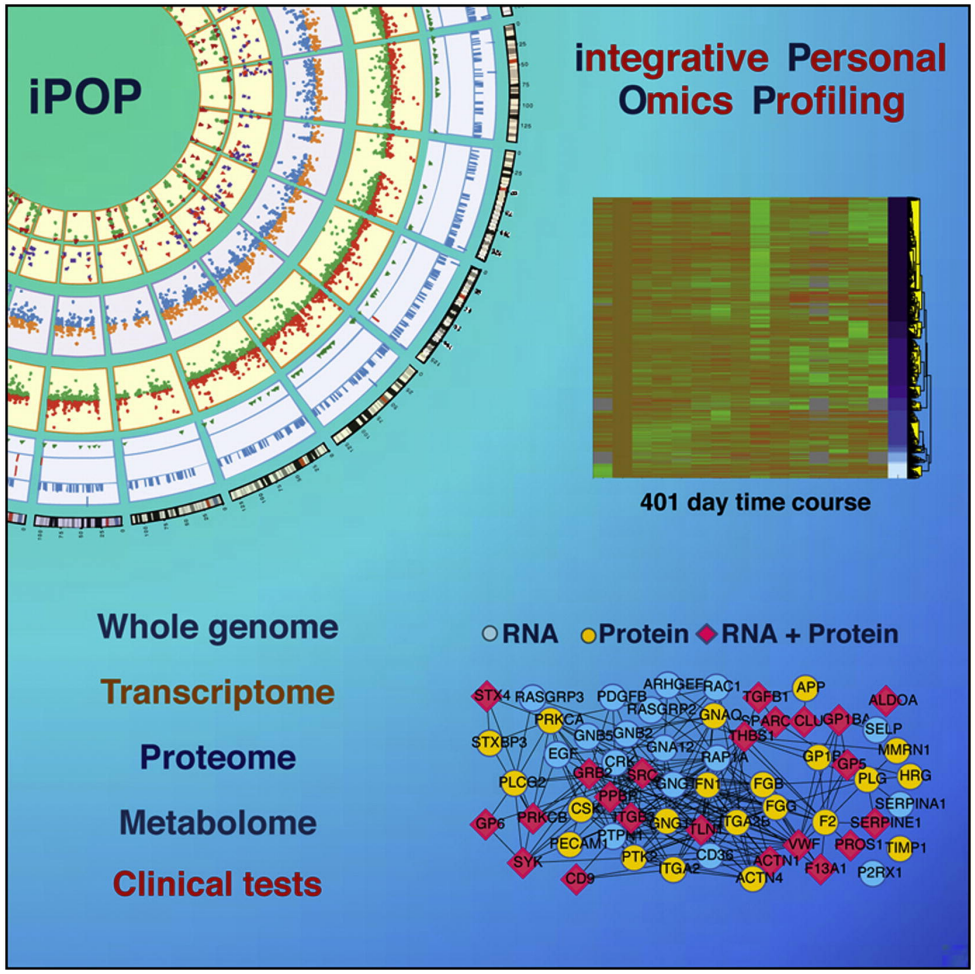
Mapping data back to the genome illustrates connections between genotype and phenotype



Chen et al. Cell 2012

Integrated personal omics profile

- iPOP: Longitudinal study integrating multiple large-scale datasets.
- Recording medical and molecular data helps reveal molecular underpinnings of health and disease.



Chen et al. Cell 2012

Precision medicine in the clinic

- Increasingly genomic information is playing a role in the clinic.
  - Targeted therapeutics
  - Pharmacogenomics
    - informing treatment options based on patient drug sensitivity

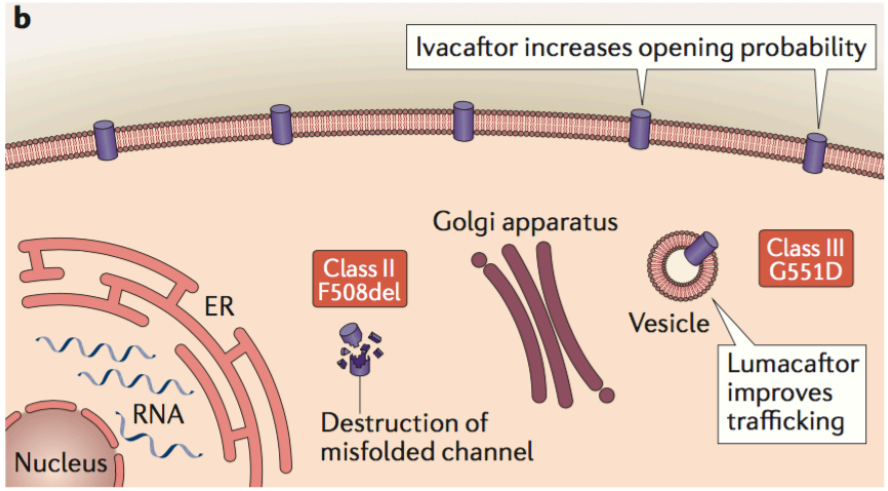
Table 1 | **Examples of precision medicine**

Condition	Gene	Action
Mendelian disease		
Cystic fibrosis	*CFTR*	Specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor
Long QT syndrome	*KCNQ1*, *KCNH2* and *SCN5A*	Specific therapy for patients with *SCN5A* mutations
Duchenne muscular dystrophy	*DMD*	Ongoing phase III clinical trials of exon-skipping therapies
Malignant hyperthermia susceptibility	*RYR1*	Avoid volatile anaesthetic agents; avoid extremes of heat
Familial hypercholesterolaemia (FH)	*PCSK9*, *APOB* and *LDLR*	• Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs • Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen
Dopa-responsive dystonia	*SPR*	Therapy with dopamine precursor L-dopa and the serotonin precursor 5-hydroxytryptophan
Thoracic aortic aneurysm	*SMAD3*, *ACTA2*, *TGFBR1*, *TGFBR2* and *FBN1*	Customization of surgical thresholds based on patient genotype
Left ventricular hypertrophy	*MYH7*, *MYBPC3*, *GLA* and *TTR*	Sarcomeric cardiomyopathy, Fabry disease and transthyretin cardiac amyloid disease have specific therapies
Precision oncology		
Lung adenocarcinoma	*EGFR* and *ALK*	Targeted kinase inhibitors, such as gefitinib and crizotinib
Breast cancer	*HER2*	HER2 (also known as ERBB2)-targeted treatment, such as trastuzumab and pertuzumab
Gastrointestinal stromal tumour	*KIT*	Targeted KIT kinase activity inhibitors, such as imatinib
Melanoma	*BRAF*	BRAF inhibitors, such as vemurafenib and dabrafenib
Pharmacogenomics		
Warfarin sensitivity	*CYP2C9* and *VKORC1*	Adjust dosage of warfarin or consider alternative anticoagulant
Clopidogrel sensitivity, post-stent procedure	*CYP2C19*	Consider alternative antiplatelet therapy (for example, prasugrel or ticagrelor)
Thiopurine sensitivity	*TPMT*	Reduce thiopurine dosage or consider alternative agent
Codeine sensitivity	*CYP2D6*	Avoid use of codeine; consider alternatives such as morphine and non-opioid analgesics
Simvastatin sensitivity	*SLCO1B1*	Reduce dose of simvastatin or consider an alternative statin; consider routine creatine kinase surveillance

Ashley. Nat. Rev. Genetics 2016

Example: cystic fibrosis

- The type of mutation in the CFTR gene informs treatment options.
- Patients with a G551D mutation receive Ivacaftor.
- Patients with F508del might take both Lumacaftor and Ivacaftor.

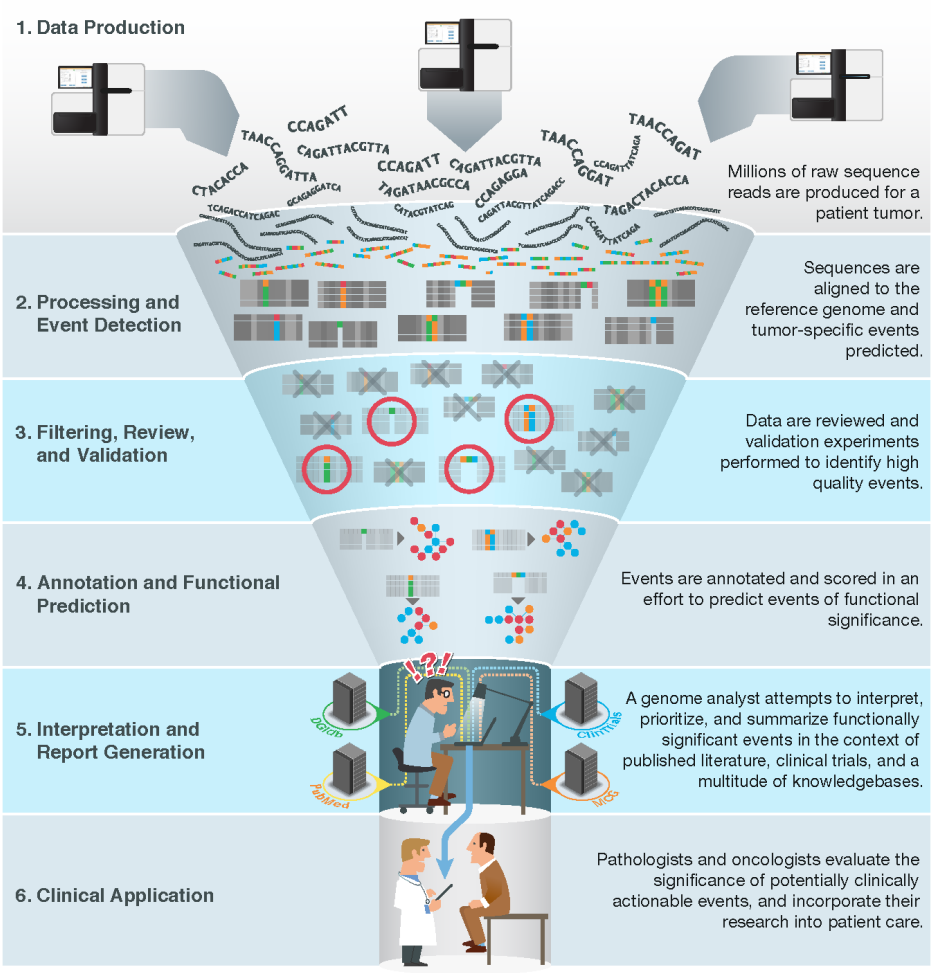


Subcategory	Functional effect	Example variant
I	No functional protein	G542x
II	Trafficking defect	F508del
III	Defective regulation	G551D
IV	Decreased conductance	R117H
V	Reduced synthesis	3120+1G>A
VI	Reduced stability	Q1412x

Ashley. Nat. Rev. Genetics 2016

Clinical considerations for genomic medicine

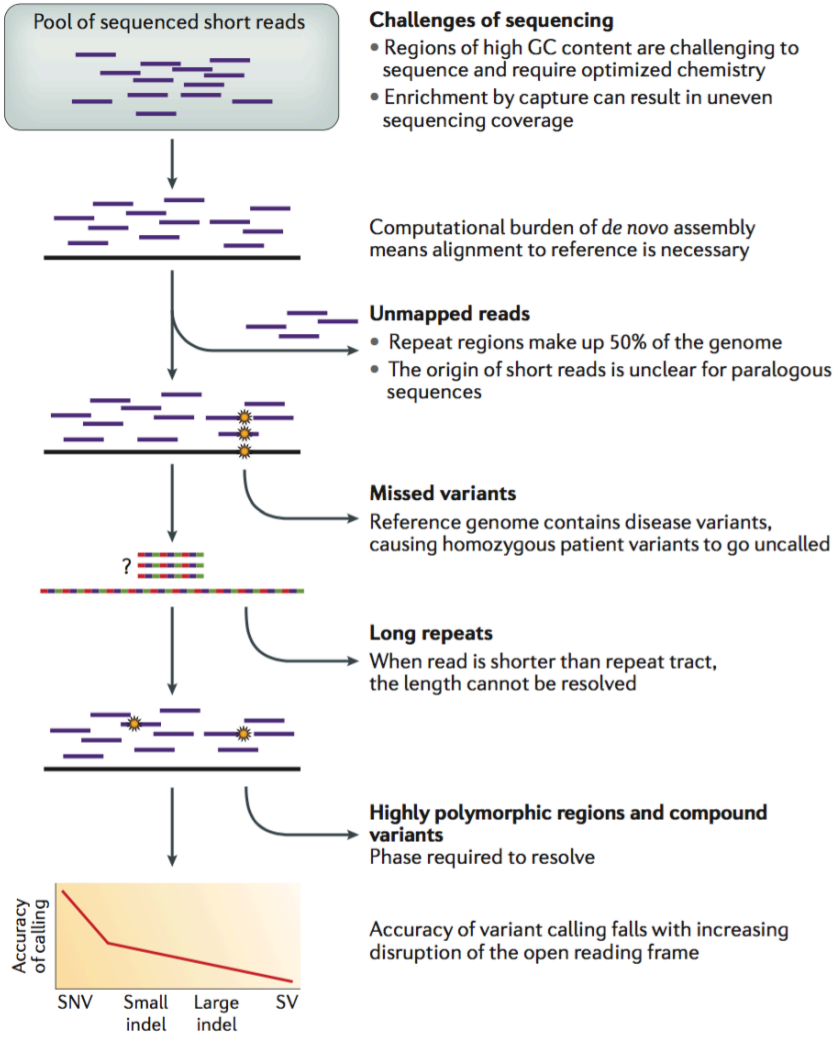
- Analysis of the Zimmer and Snyder genomes were done for journalistic and research purposes
- High quality genome coverage is of paramount importance in clinical genomics.
  - The risk of false positive or false negative genomic variants.



http://genome.wustl.edu/projects/detail/civic-clinical-interpretations-of-variants-in-cancer/

Challenges in assembly of clinical-grade genome sequences

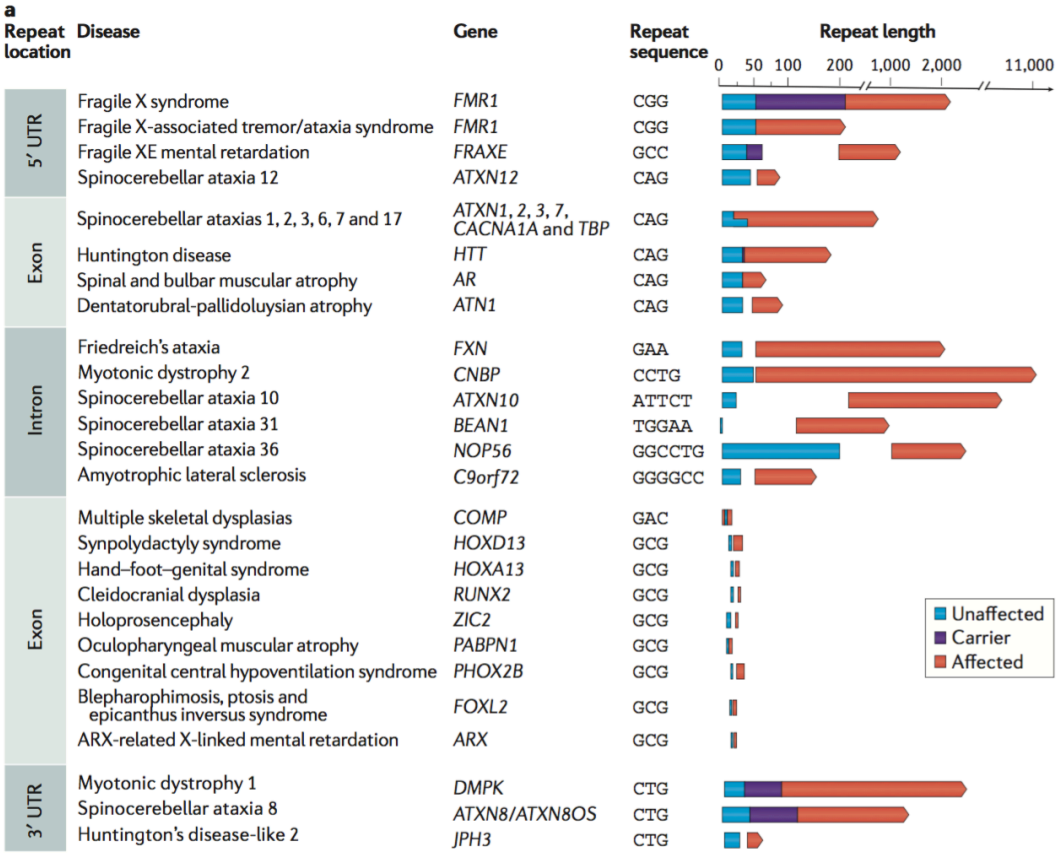
- Difficult to sequence or map regions of the genome lead to approximately 5% of reads not mapping to the genome. These regions include:
  - High GC content
  - Repeats
  - Paralogous regions (gene families, pseudogenes)



Ashley. Nat. Rev. Genetics 2016

Difficulty with short repeat regions

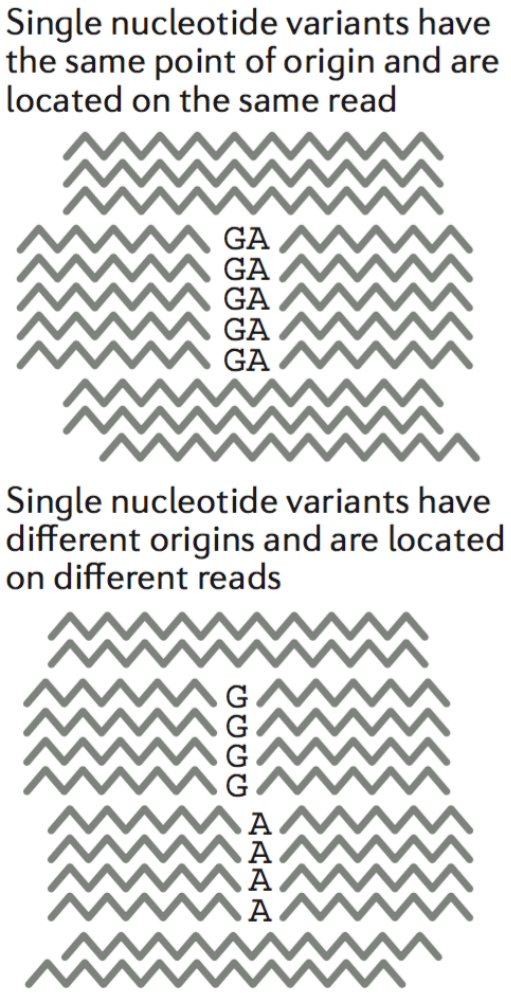
- Repeat regions are both difficult to sequence and medically relevant.
- Longer stretches of repeats underlie a number of diseases
  - e.g. Huntington's disease



Ashley. Nat. Rev. Genetics 2016

Diploid genomes and compound variant phasing

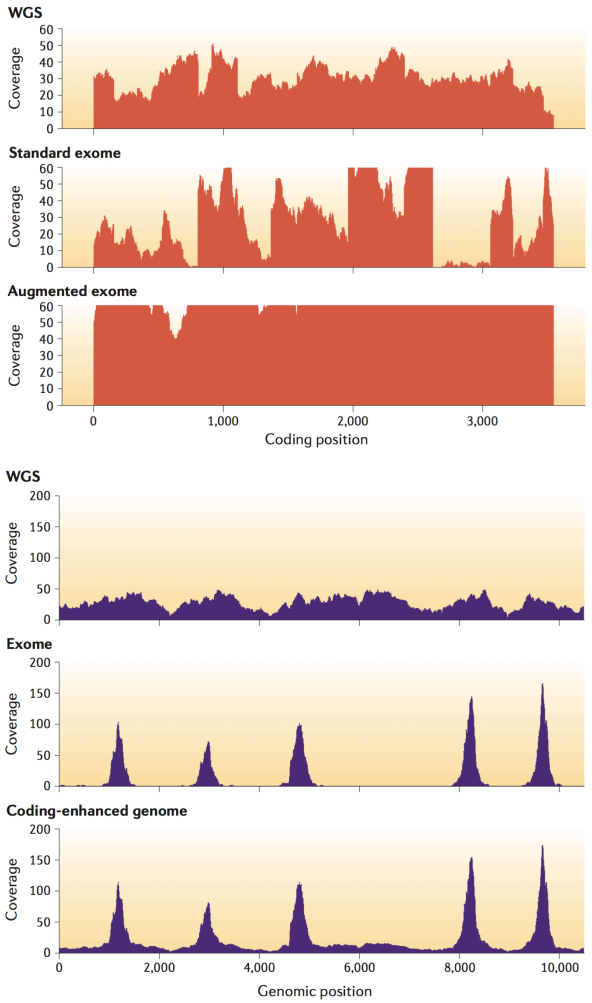
- What if two mutations occur in a single gene?
  - If they are spanned by a single read its easy to determine if they belong to one allele or two.
  - More difficult to determine if they are too far apart



Ashley Nat. Rev. Genetics 2016

Exome sequencing vs. whole genome sequencing

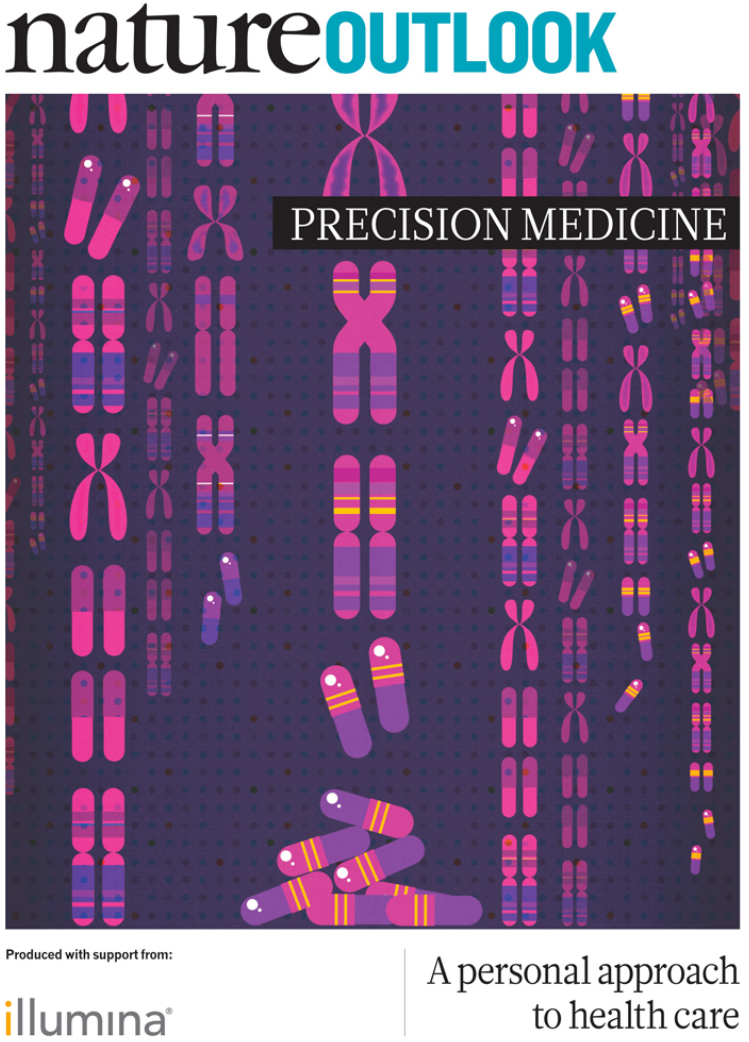
- Tradeoff between depth and breadth of sequencing at a given cost constraint.
  - Augmented exome sequencing enables higher depth sequencing of coding regions than WGS.
  - WGS provides more uniform coverage and also includes regulatory regions or unknown coding regions of relevance.



Ashley. Nat. Rev. Genetics 2016

Precision medicine in the clinic

- Precision medicine is leading to better defining and treating disease at the molecular level.
- It is both changing the prescription of existing medications and inspiring new targeted therapies.
- Precision medicine requires high quality patient genome sequences be obtained at reasonable cost.



Additional Game of Genomes Slides

From Mark's lecture:

Drivers-for-Making-Personal-Genomics-into-Tool-20161114-i0siem

Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

3,559,137
Germline SNVs

Rare variants: MAF<0.05%

3,239,212
Common SNVs

319,925
Rare SNVs

1,800
Coding

318,125
Non-coding

1,018
Non-synonymous

733
Synonymous

15,007
Enhancer

3,043
Promoter

15
PreMatureStop

42,389
Other annotations (TFP, DHS, ncRNA, pseudogene)

Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

Overview & Coding Variants

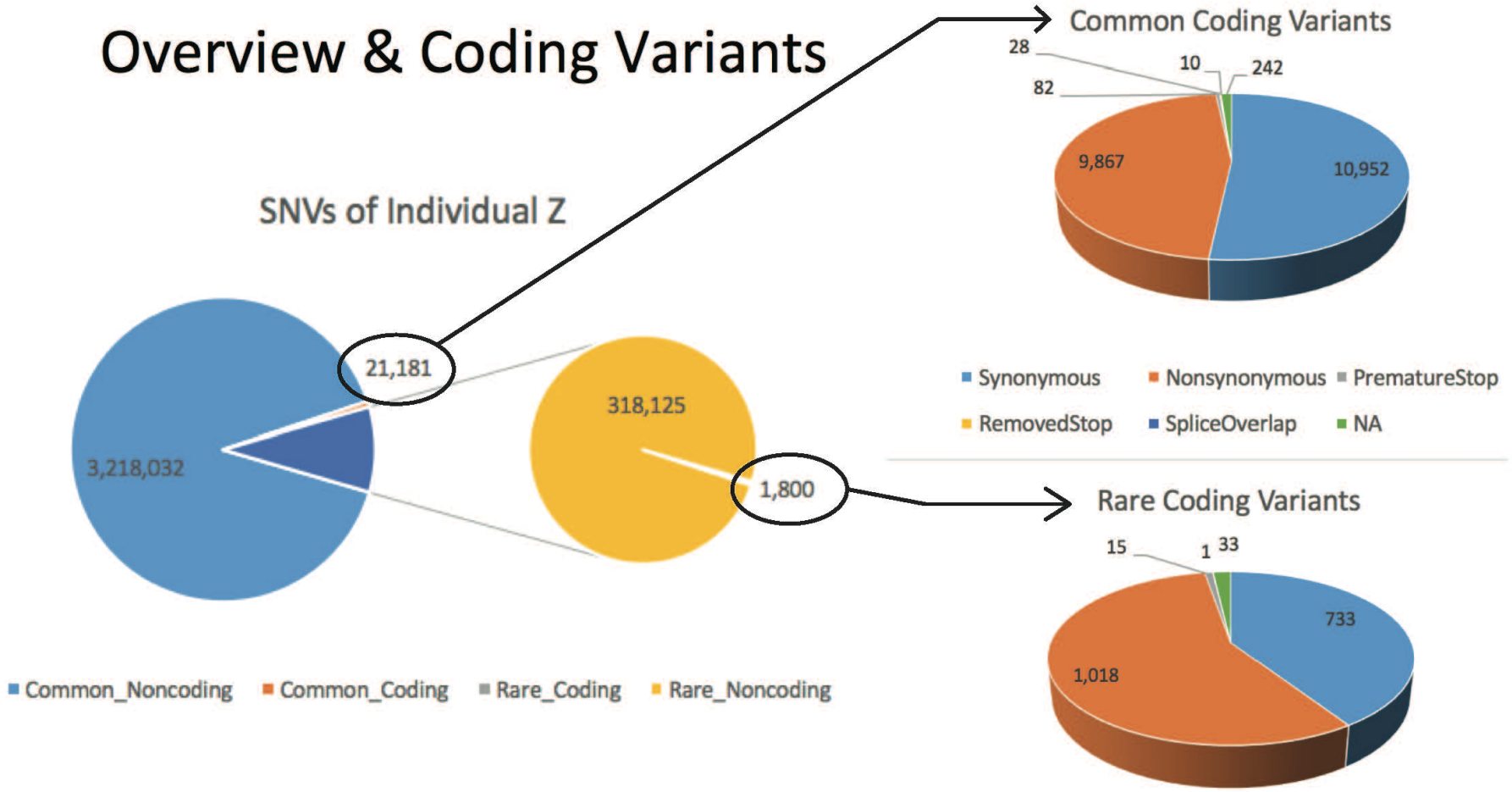


Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

Rare Non-synonymous Coding Variants

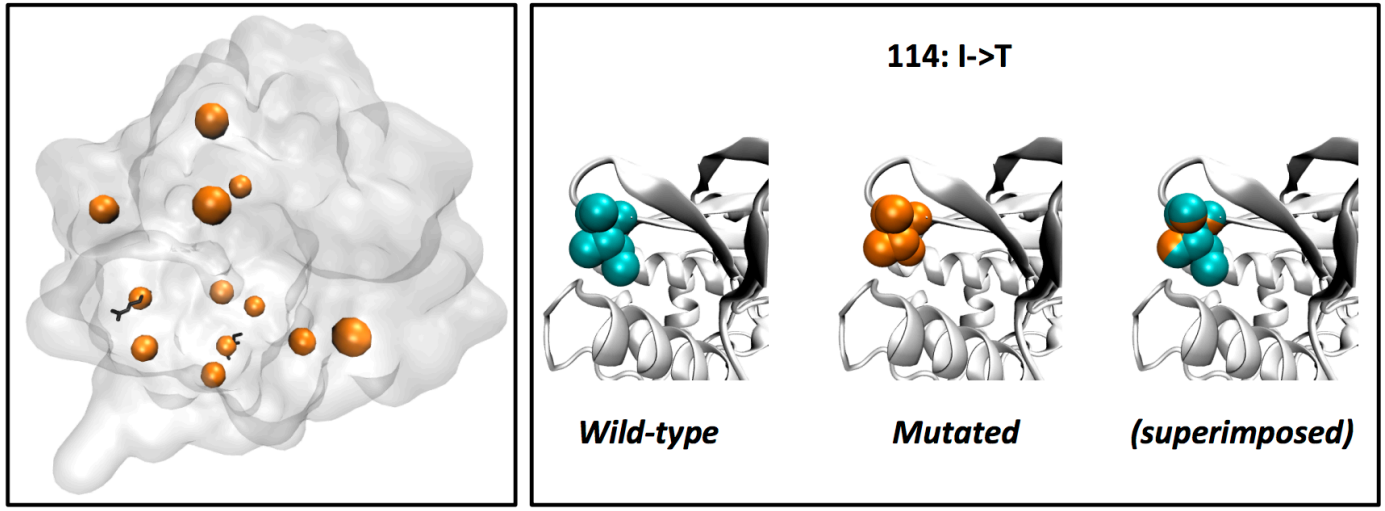
- 1018 SNVs -> 824 target genes

Gene Annotation	Gene Name
Cancer-related	NOTCH2; PDE4DIP; TPR; CRTC3; CDH11; MLLT6; ASXL1; HMGA1; KDM6A
DNA repair	RECQL; RAD51; PPM1D; XRCC1; AP1B1; FANCI; PTPRH; RBBP7; SLX4; POLR2A; DCLRE1C; ANKLE1
Cancer & DNA repair	ATM; PMS2; ERCC5
Actionable Gene	ATM; KDM6A; INSR; FOXP4

- ATM: Serine/Threonine Kinase; Regulator of p53 and BRCA1; leukemia; ataxia-telangiectasia; breast cancer
- PMS2: Direct p53 effectors; mismatch repair cancer syndrome; colorectal cancer; hereditary nonpolyposis
- ERCC5: Chks in Checkpoint Regulation; DNA Repair; xeroderma pigmentosum
- KDM6A: Transcriptional misregulation in cancer
- INSR: Insulin Receptor; PI3K-Akt signaling pathway; GPCR Pathway; Diabetes mellitus
- FOXP4: Transcriptional repressor that represses lung-specific expression

Example of Molecular Effect of Impactful Coding Variant

Arylamine N-acetyltransferase (PDB: 2PFR_A ; gene: NAT2)



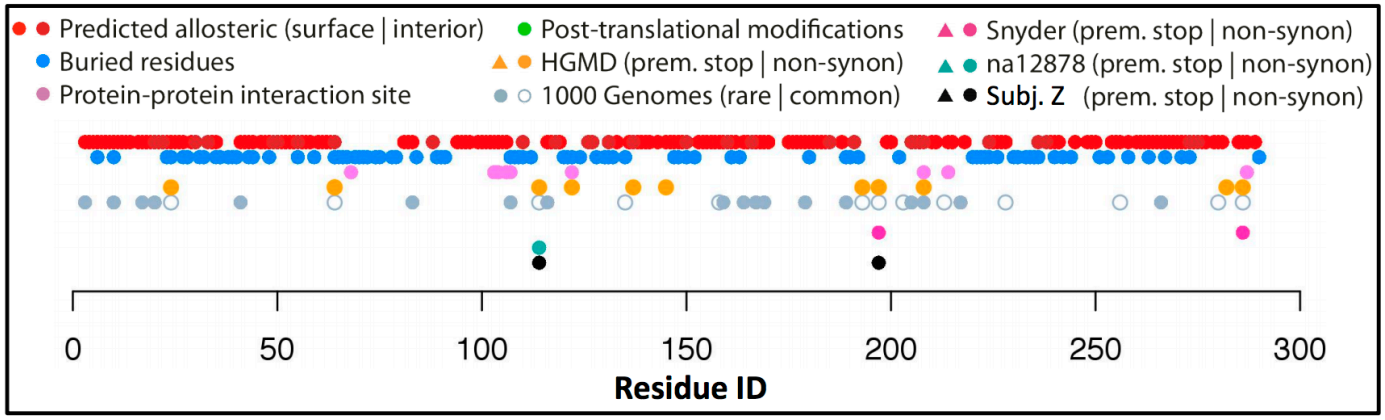
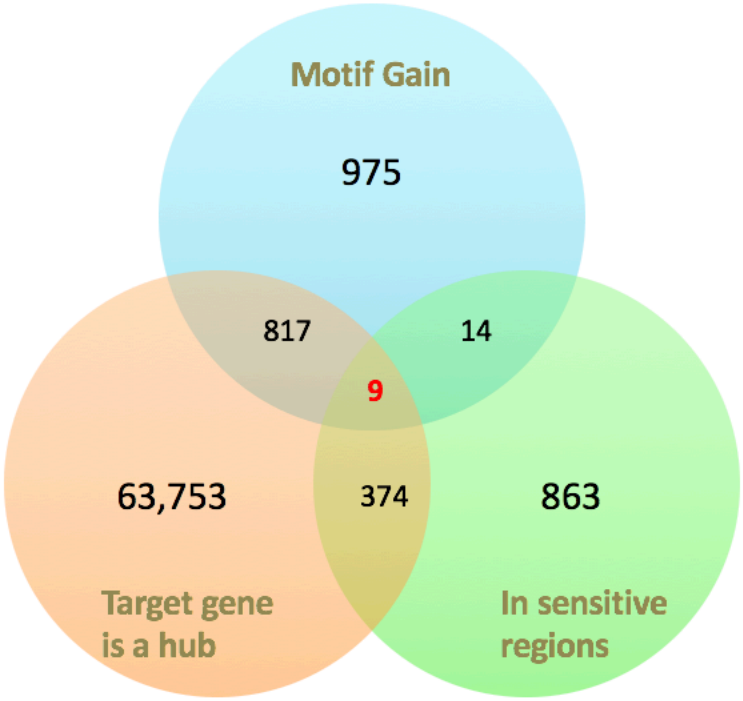


Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

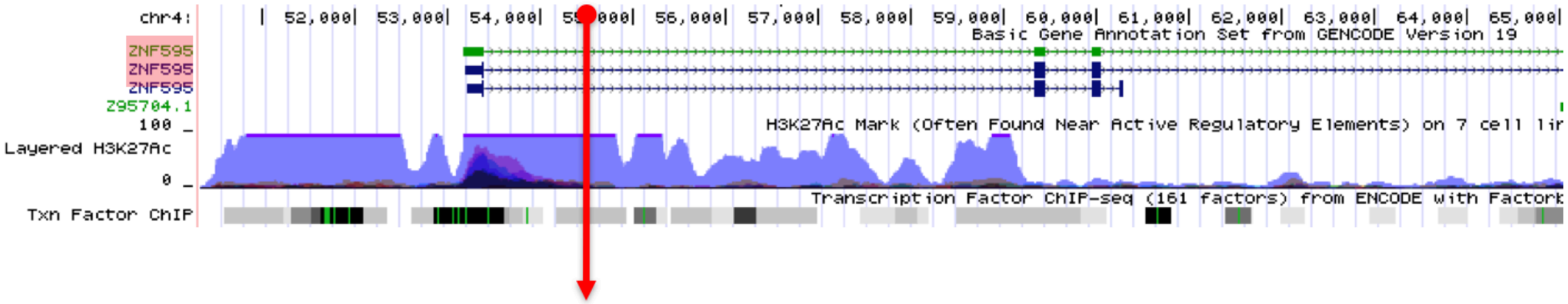
Annotation of Rare Noncoding Variants



- 9 variants -> 11 target genes

Gene Name	Variant Location	Function Annotation
RPL10	(Promoter&UTR)	[cancer]
PDE4DIP	(Distal&Intron)	[cancer]
ZNF595	(Intron&Promoter)	
GADD45G	(Promoter)	[DNA_repair]
CCND2	(Distal)	[actionable][cancer]
ACAP3	(Intron)	
VANGL2	(Promoter)	
SEC22B	(Distal)	
RNU1-9	(Distal)	
PARP11	(Distal)	
PUSL1	(Promoter)	

Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

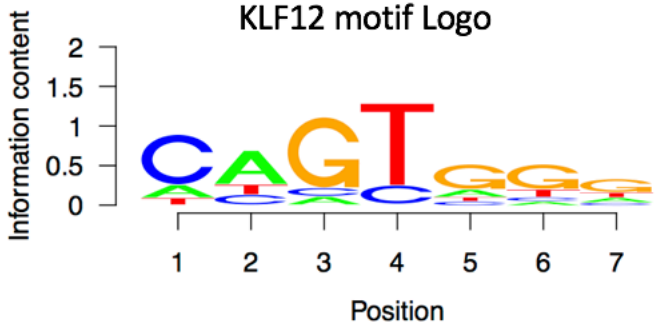


Rare noncoding SNV

- Chr4: 54475
- C => T
- Target gene: Intron of ZNF595

Motif Gain: **KLF12 (AP-2)**

- Chr4:54469-54476
- Minus strand



A

P-value < 4e-8