Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

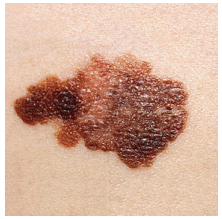
Mark Gerstein
Yale

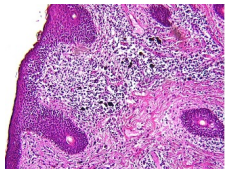
Slides freely downloadable from
Lectures.GersteinLab.org
& “tweetable”
(via @markgerstein).

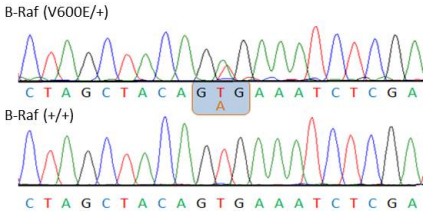
See last slide for more info.

Molecular pathology extends the diagnostic precision gains of surgical pathology by probing even more fundamental elements of biology

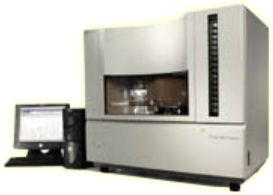
Biological Causation







http://wrightstatephysicians.org/whatsnew/melanoma.html
http://pathology.osu.edu/residents/InternalGate/Area51/ResidentSlideCollection/images/A100.jpg
https://rikengenesis.jp/ori/50279/etc_img/BRAFV600E.jpg



Diagnostic Precision

https://upload.wikimedia.org/wikipedia/commons/d/d2/Stethoscope-2.png
http://www.microscope.com/student-microscopes/university-student-microscopes/omano-om139-infinity-corrected-plan-optics.html#gref
http://sequetech.com/

Next generation sequencing is an exciting addition to the molecular pathology suite

Miniaturization and massive parallelization has sparked a genomics revolution

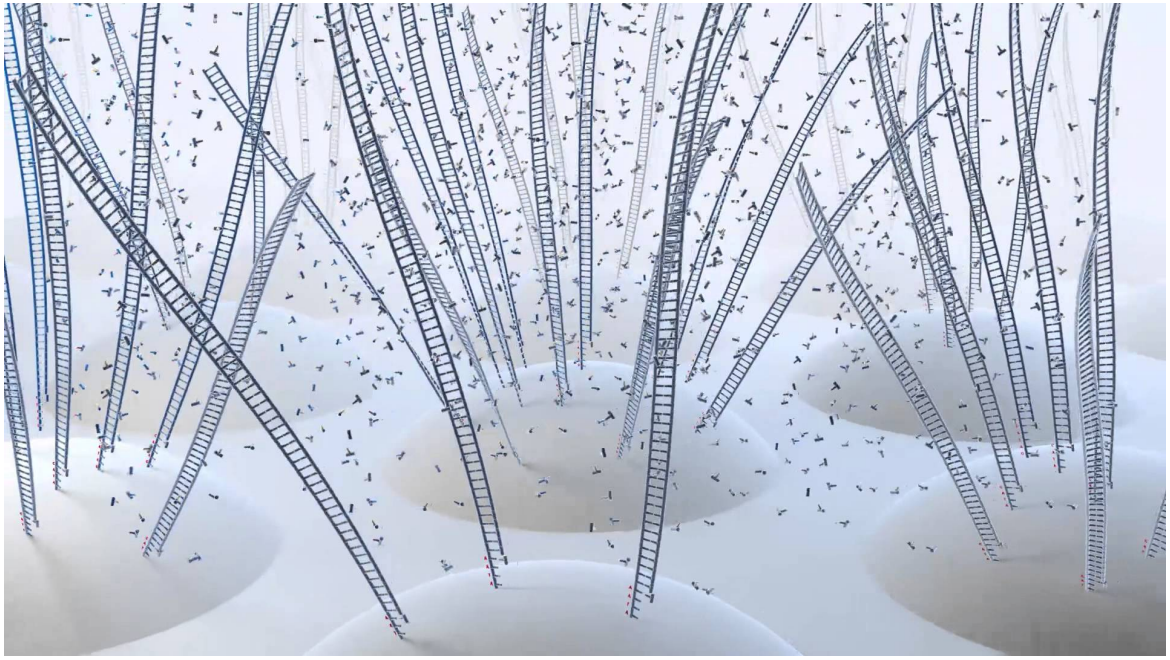


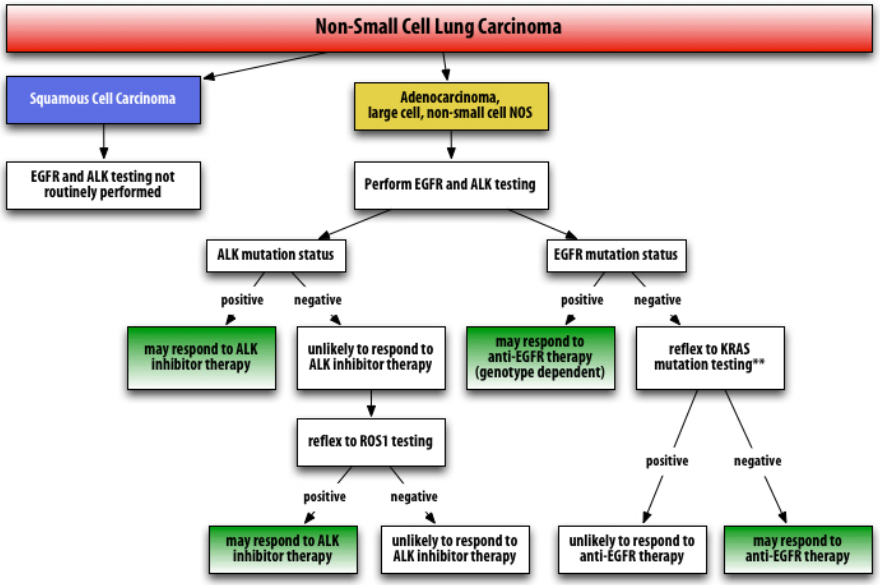
Image credit: http://perception7.com/

In the future, all stages of clinical care will depend on bioinformatics and genomics

- Prevention – molecular well-visits for early cancer screening
- Risk-prediction – large genomic and transcriptomic data-sets
- Diagnosis – identify the molecular subtype of a patient's condition
- Personalized treatments
  - Targeted therapy – treat a patient's underlying molecular pathology
  - Smarter experiential learning – treat patients based on what worked for patients who were most molecularly similar

Genomic technologies have begun to enter the molecular pathology suite

- Precision Oncology



http://www.apmggroup.net/innovation/molecular_testing/Lung_Pathways/lung.html

- Neonatal screening for Mendelian disease

http://www.ngsleaders.org/blogs.aspx

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

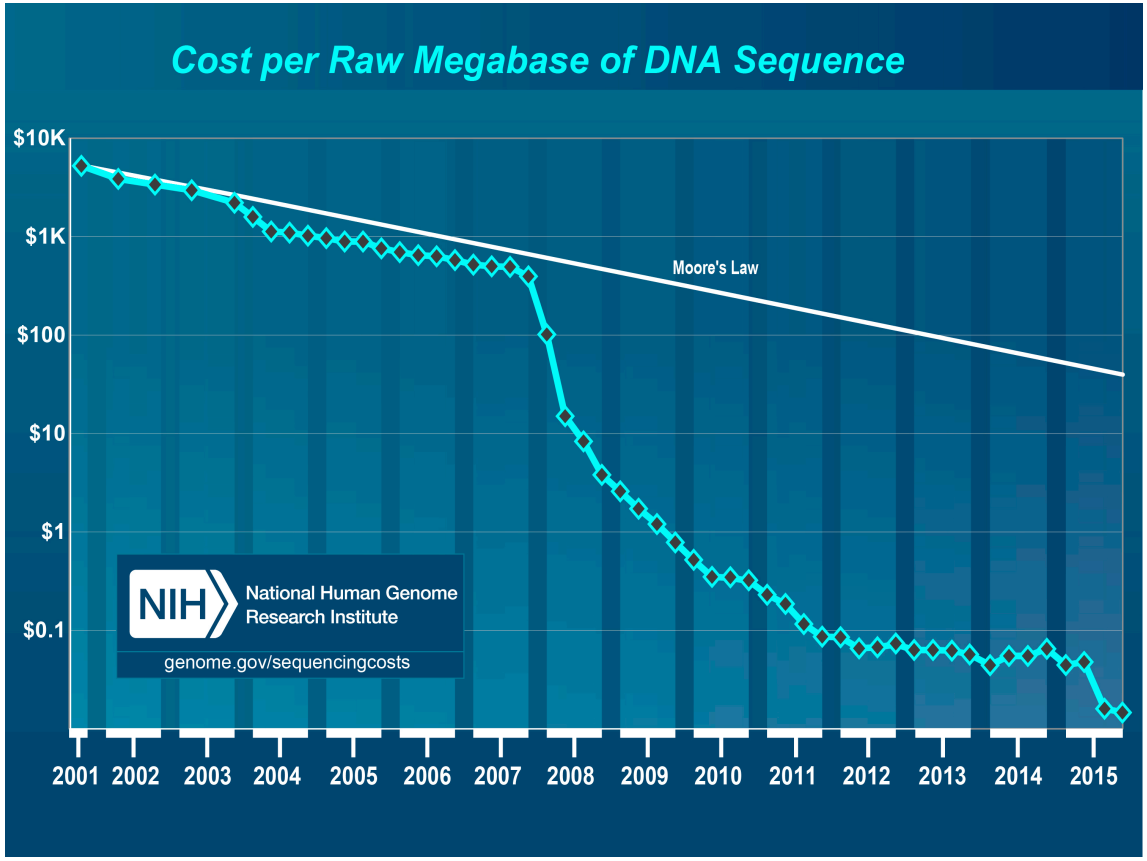
- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

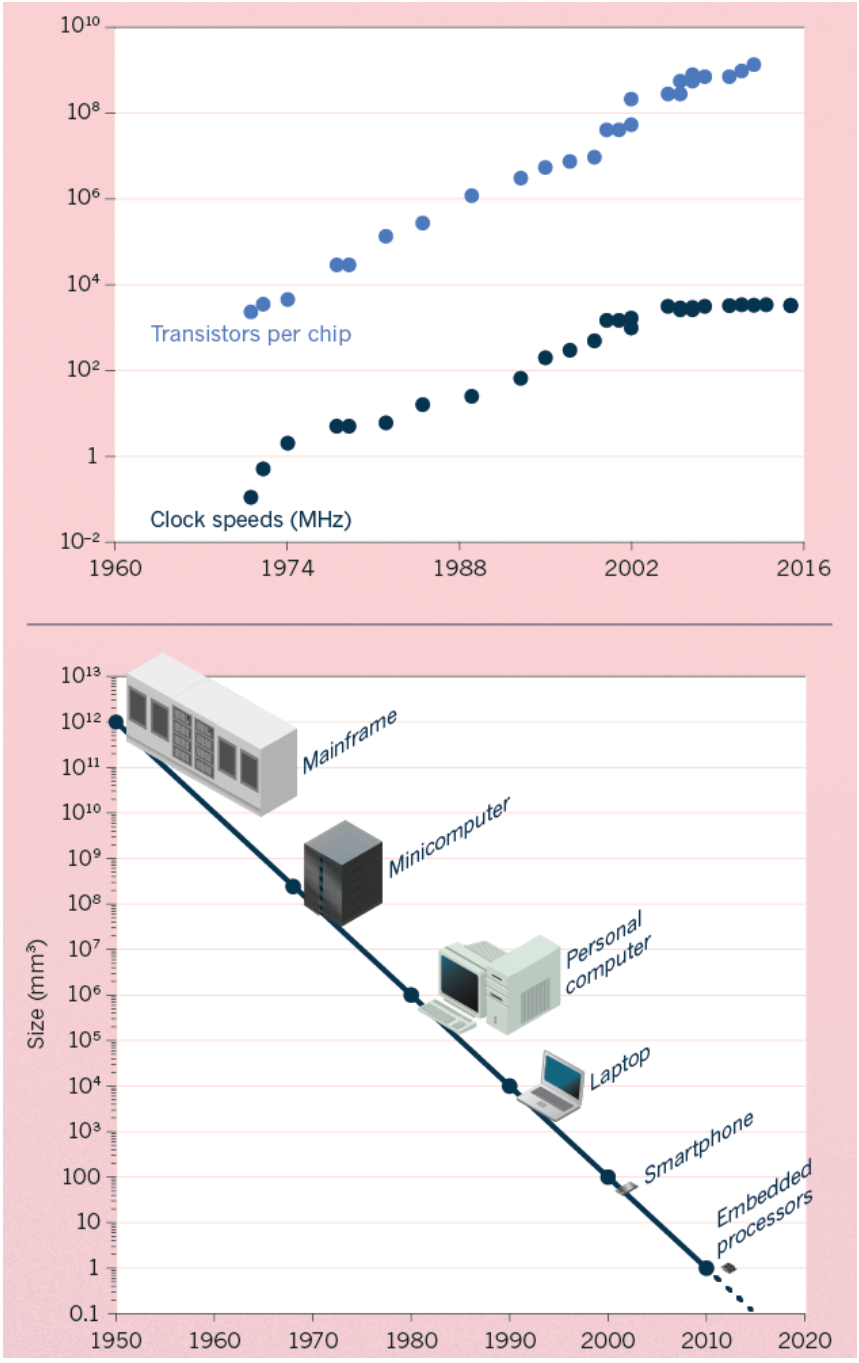
Sequencing Data Explosion: Faster than Moore's Law for a Time

- DNA sequencing has gone through technological S-curves
  - The advent of NGS was a shift to a new technology with dramatic decrease in cost).



Moore's Law: Exponential Scaling of Computer Technology

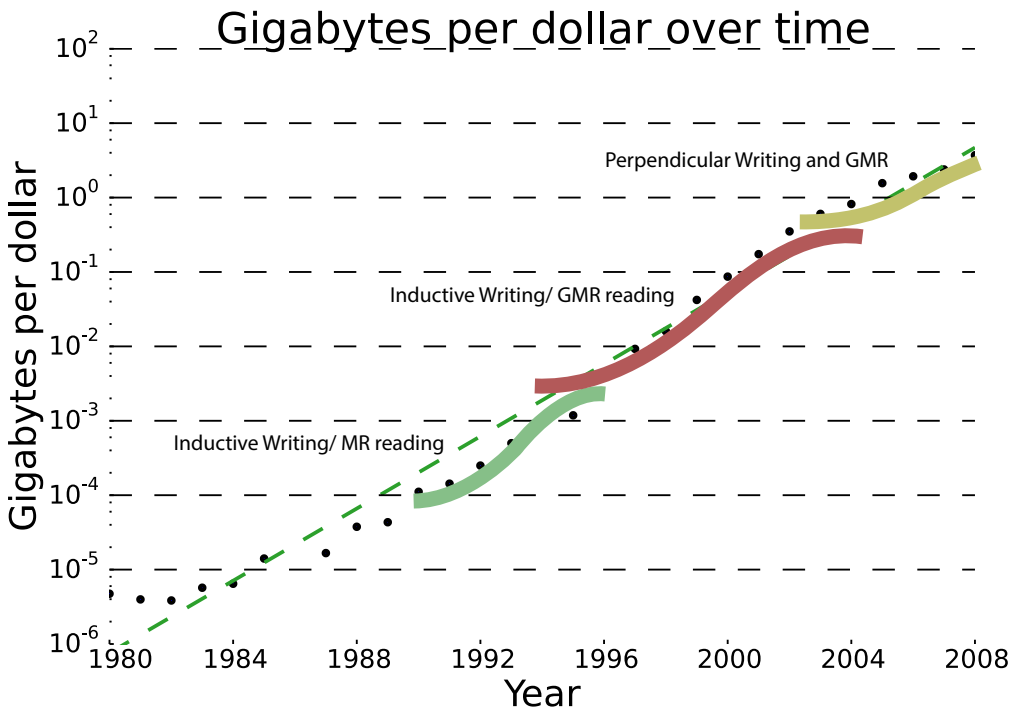
- Exponential increase in the number of transistors per chip.
- Led to improvements in speed and miniaturization.
- Drove widespread adoption and novel applications of computer technology.

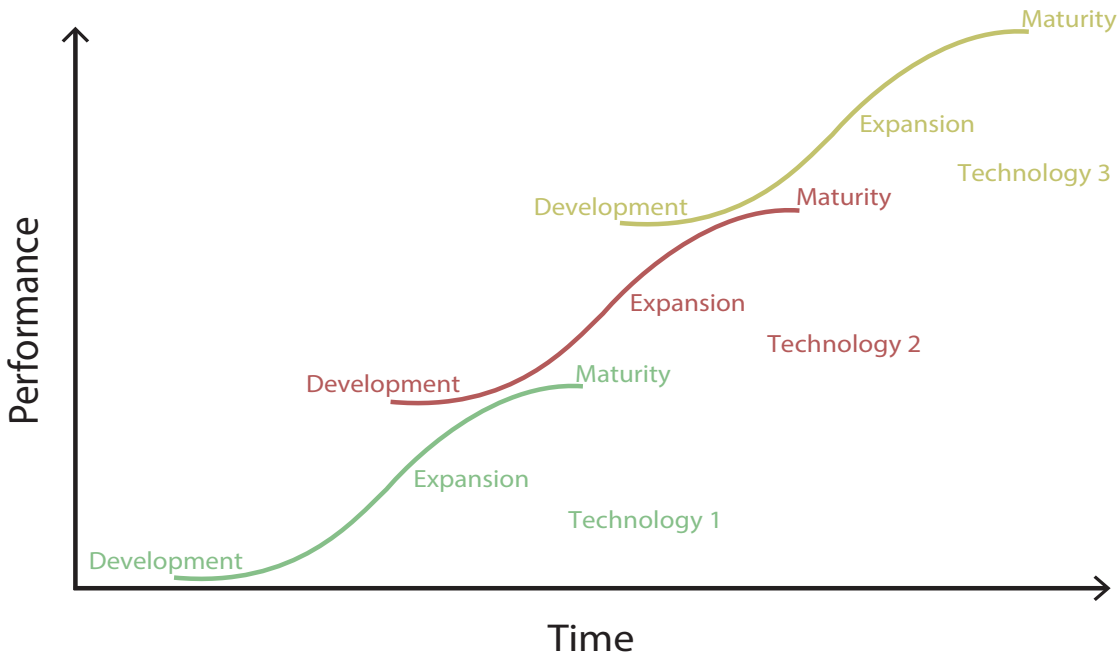


[Waldrop ('15) Nature]

Kryder's Law and S-curves underlying exponential growth

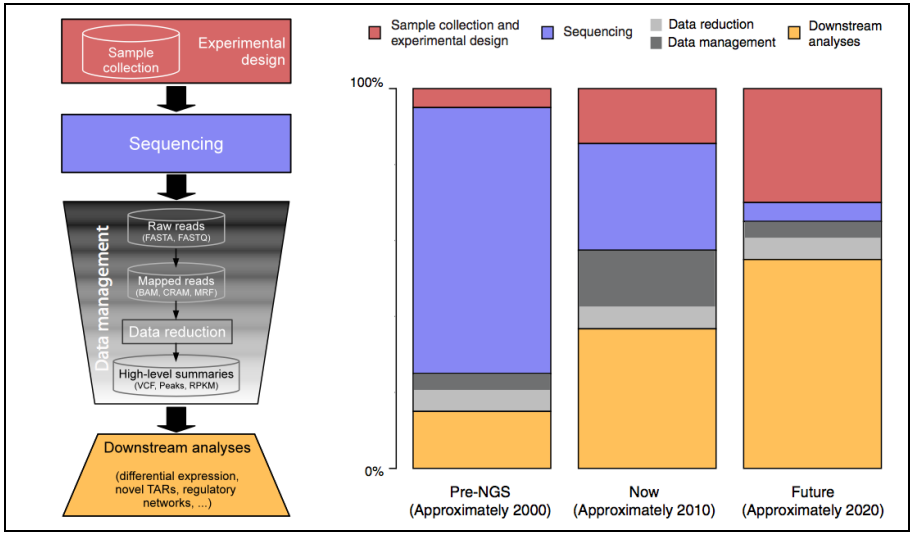
- Moore's & Kryder's Laws
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies





[Muir et al. ('15) GenomeBiol.]

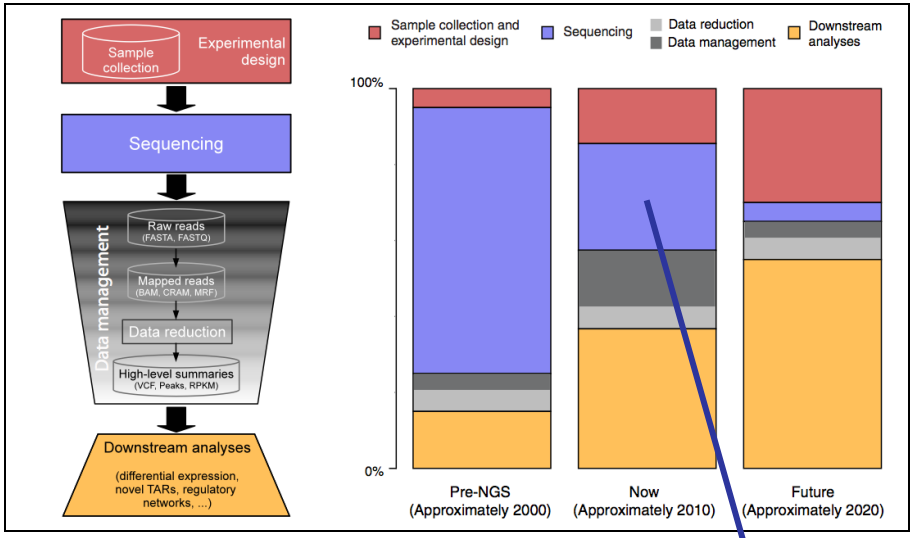
The changing costs of a sequencing pipeline



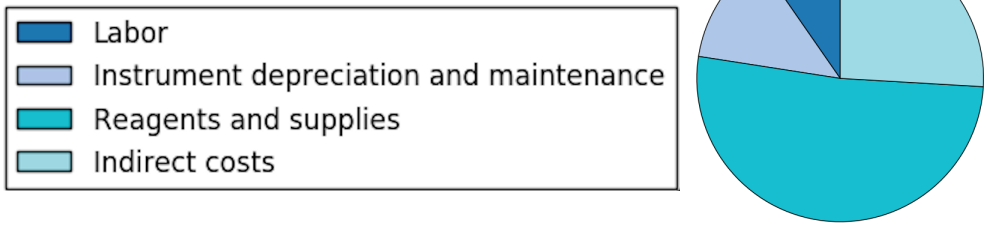
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline

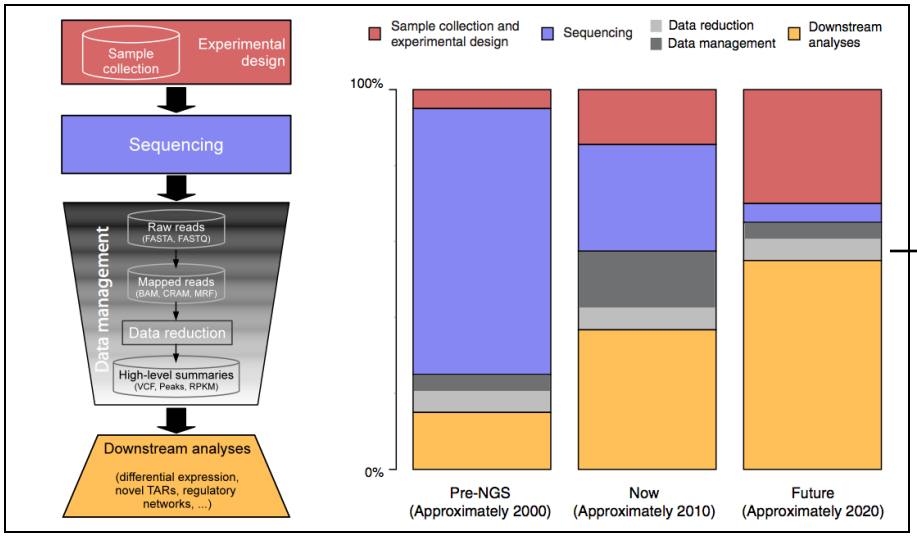


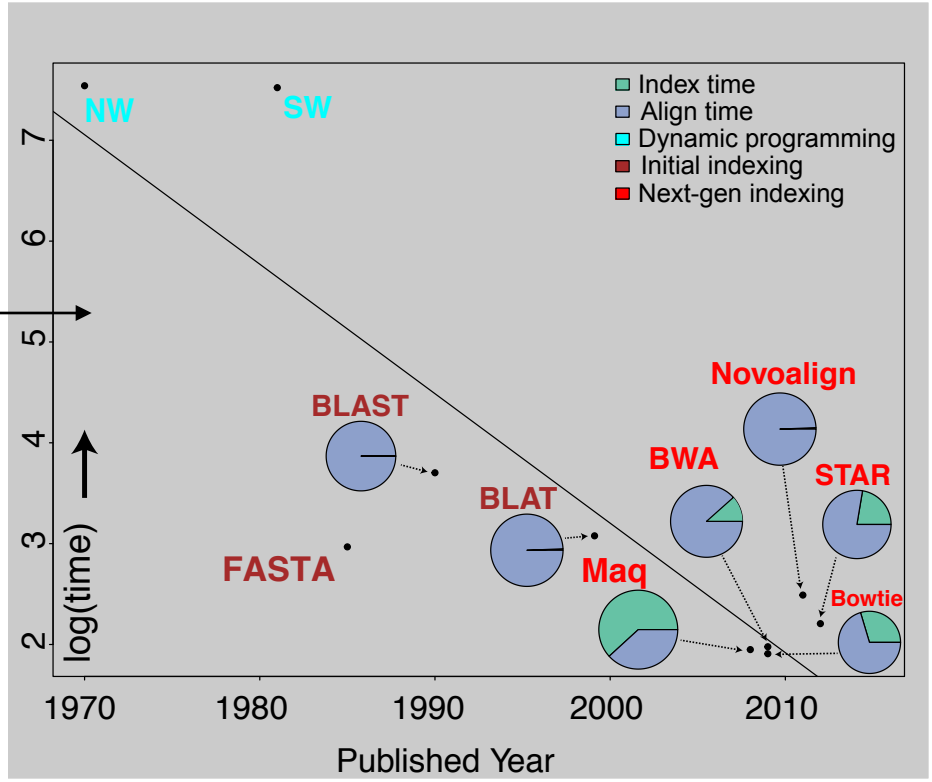
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline



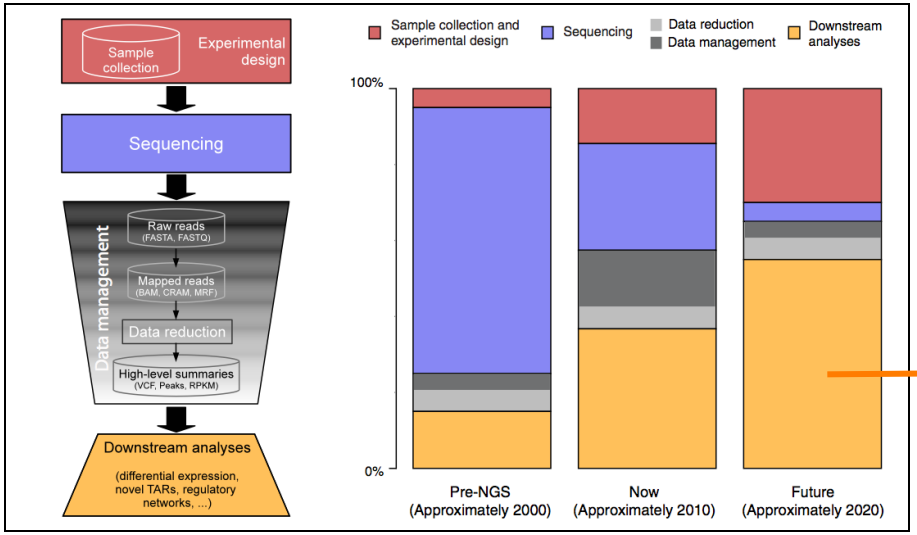


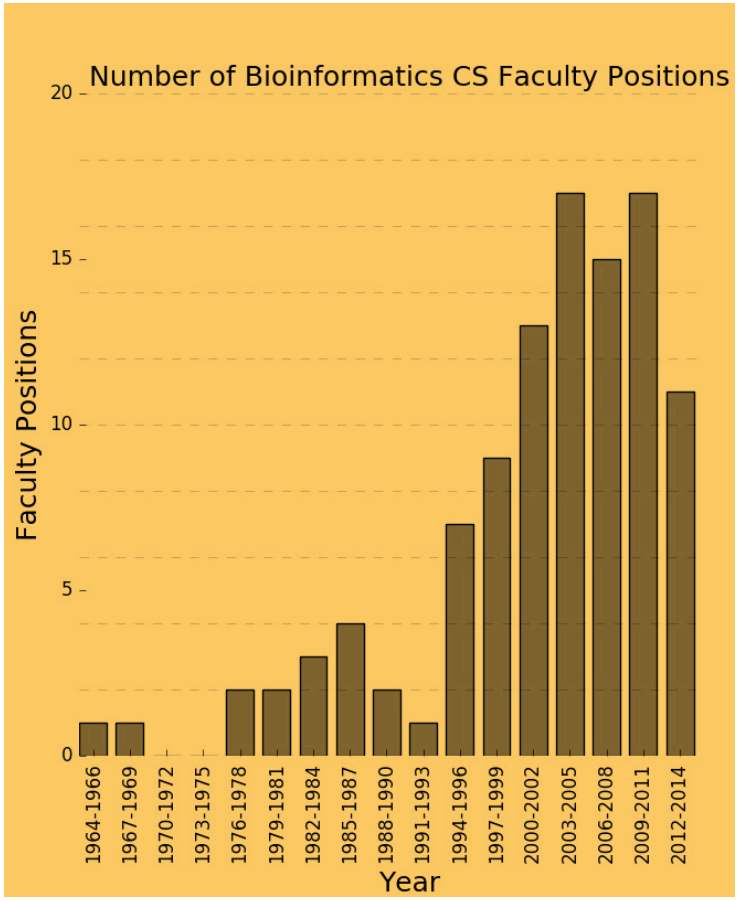
From '00 to ~' 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

Alignment algorithms scaling to keep pace with data generation

[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

The changing costs of a sequencing pipeline





From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

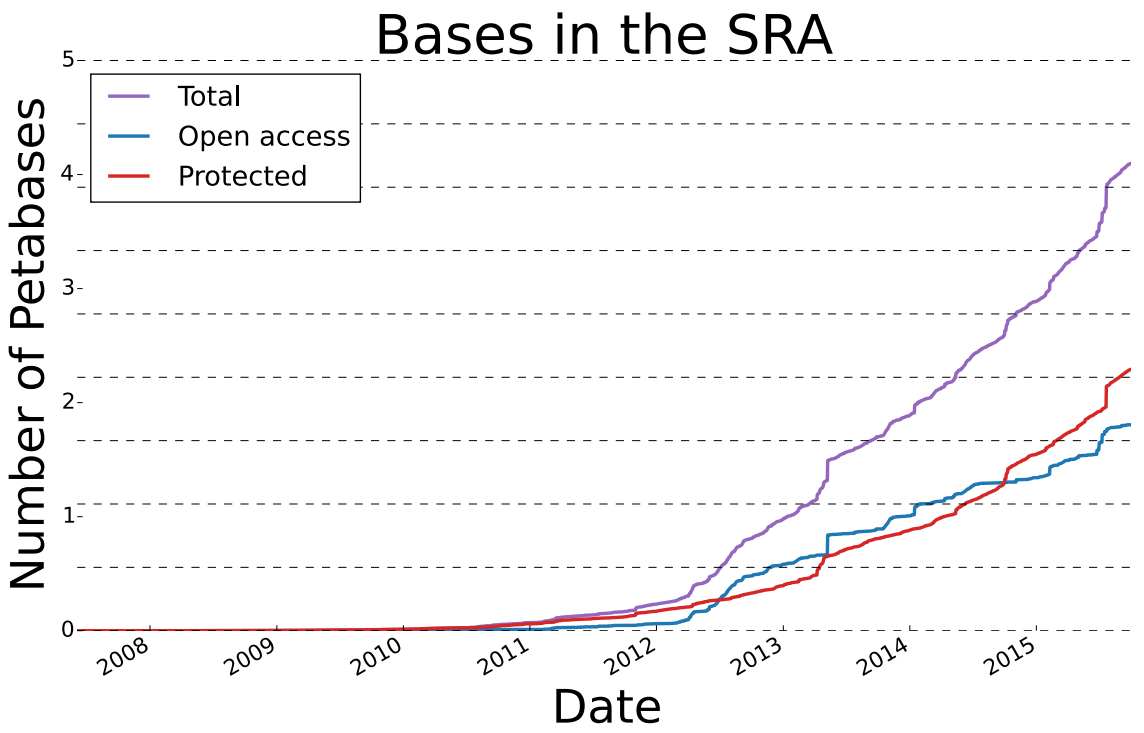
[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

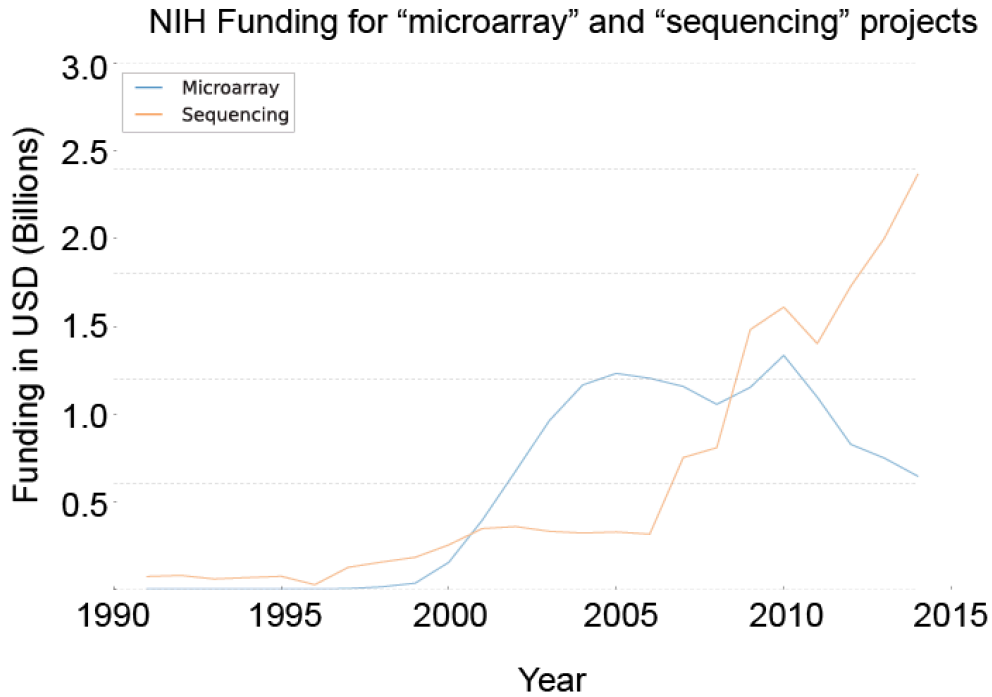
Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

Sequencing cost reductions have resulted in an explosion of data

- The type of sequence data deposited has changed as well.
  - Protected data represents an increasing fraction of all submitted sequences.
  - Data from techniques utilizing NGS machines has replaced that generated via microarray.

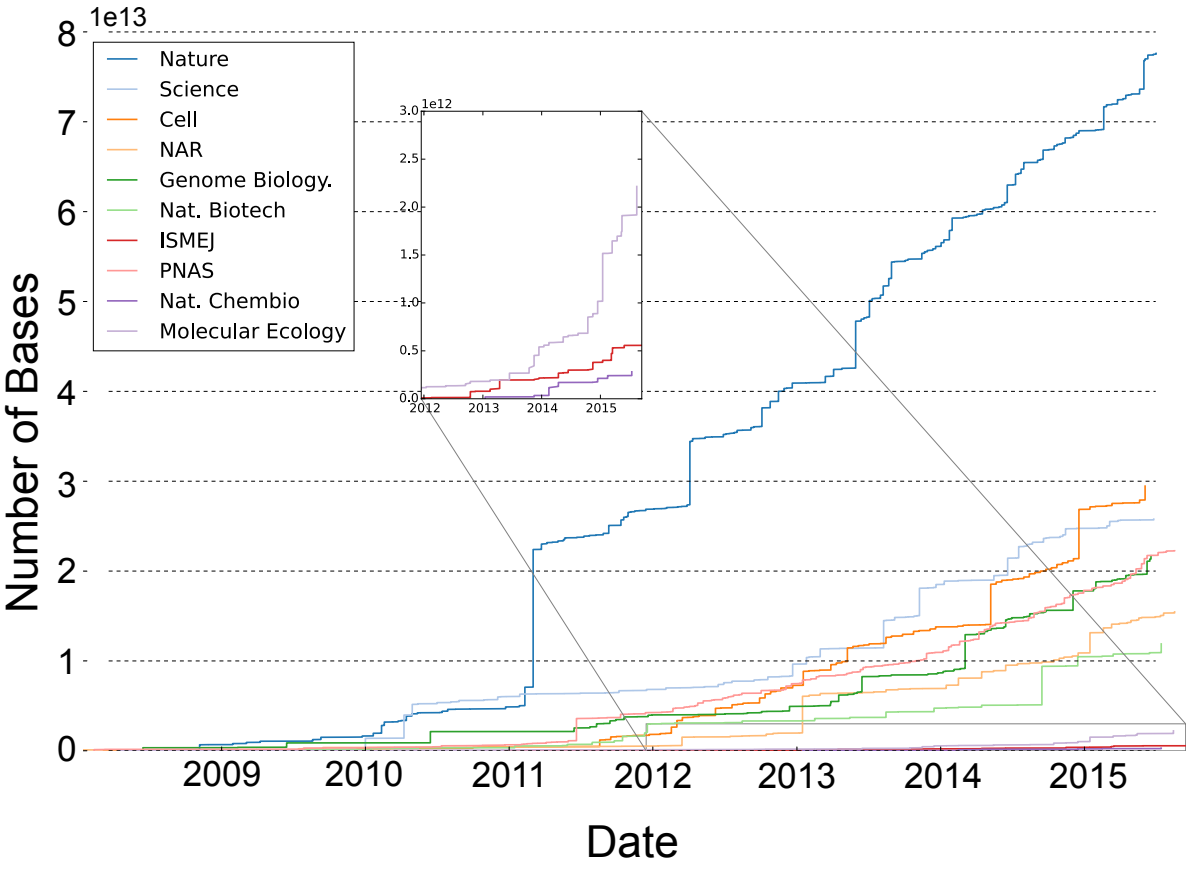


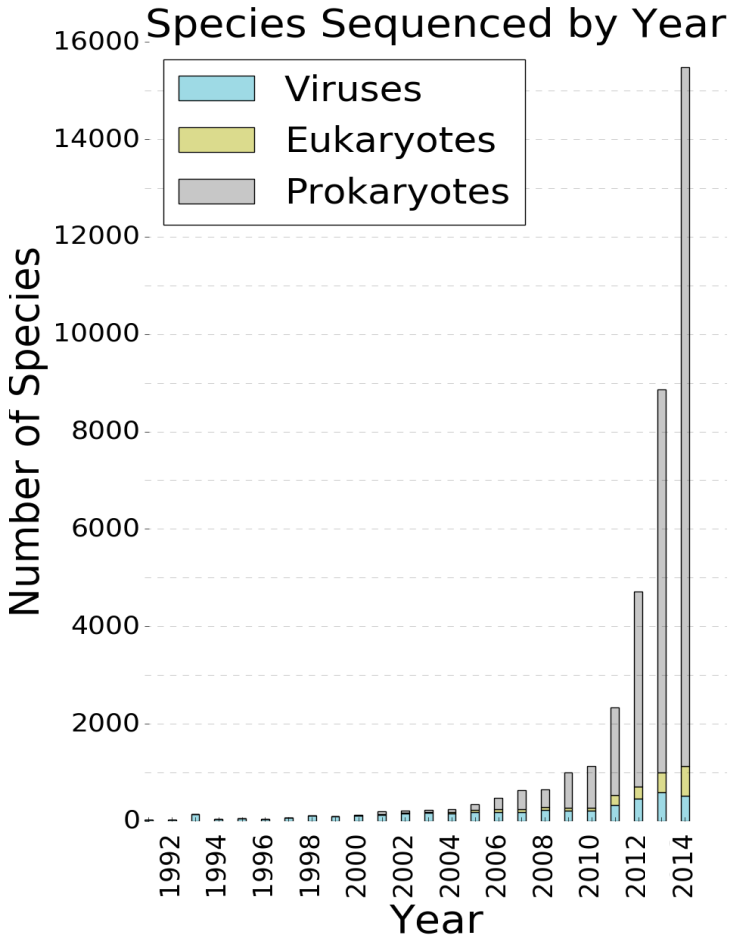


[Muir et al. (‘15) GenomeBiol.]

Increasing diversity in sequence data sources

[Muir et al. ('15) GenomeBiol.]



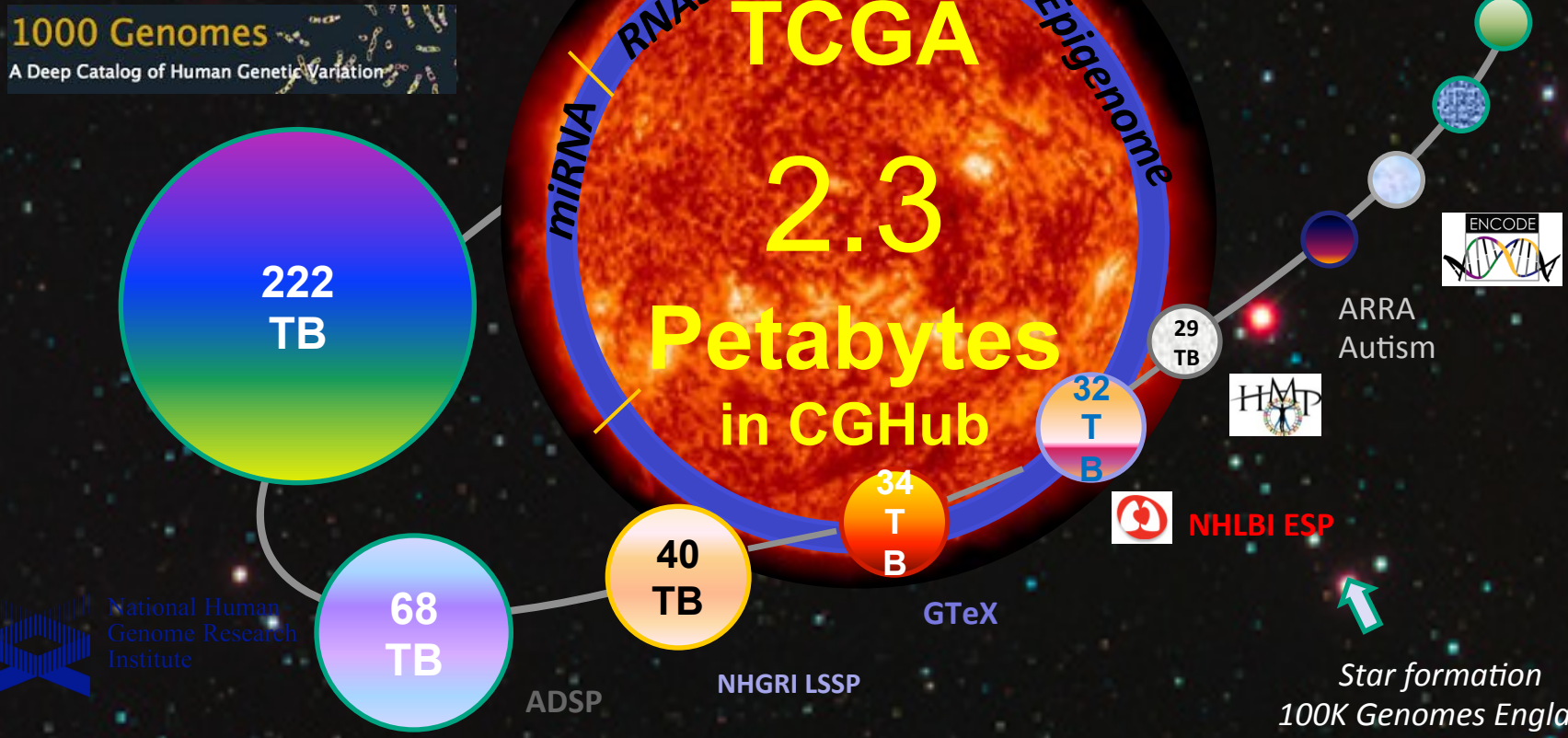


Sequence Universe

SRA ~1 petabyte

TCGA endpoint: ~2.5 Petabytes

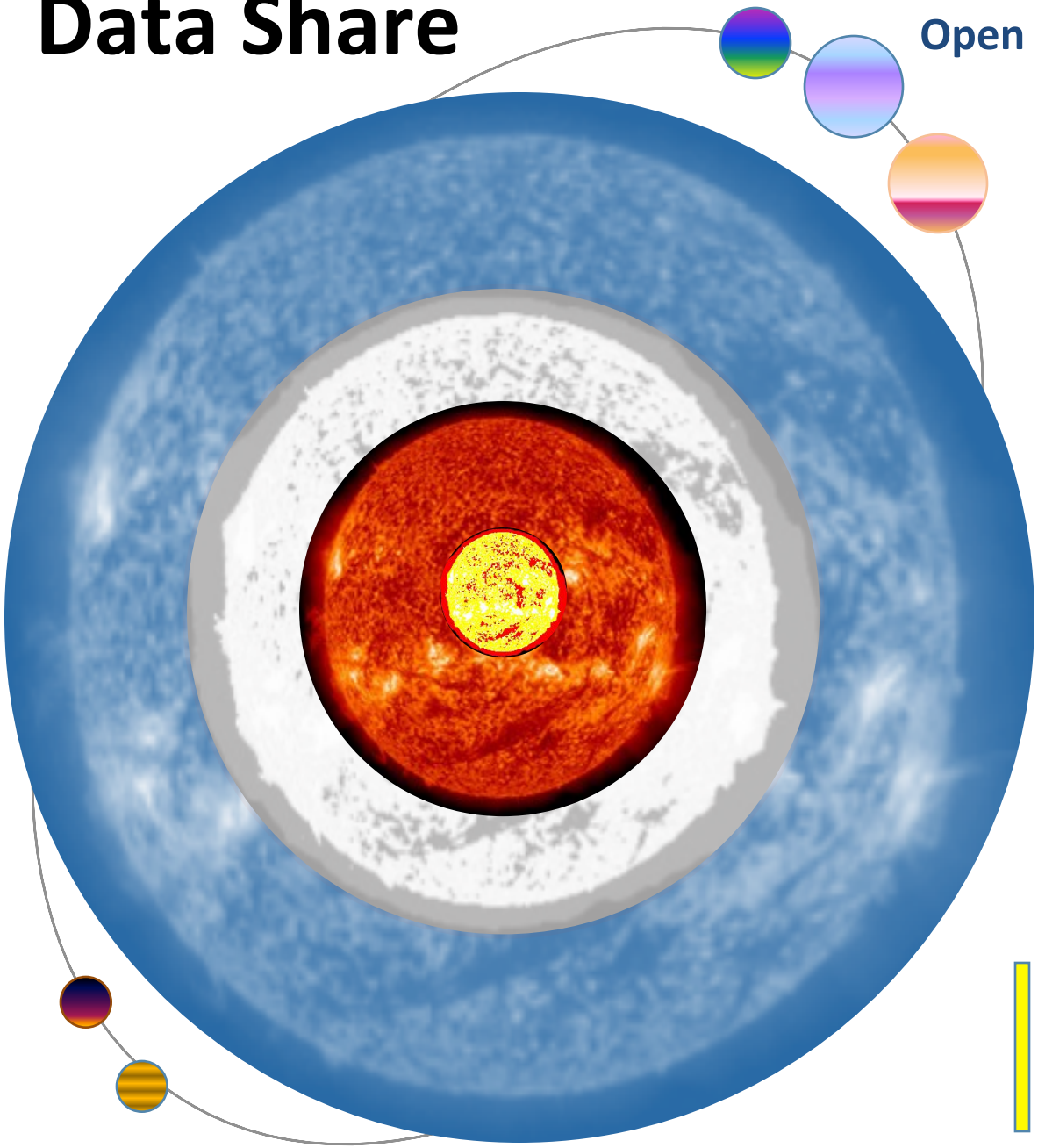
- **~1.5 PB exome**
- **~1 PB whole genome**



Heidi Sofia, 7-16-15

JESS3

Data Share



Open resources interface with API

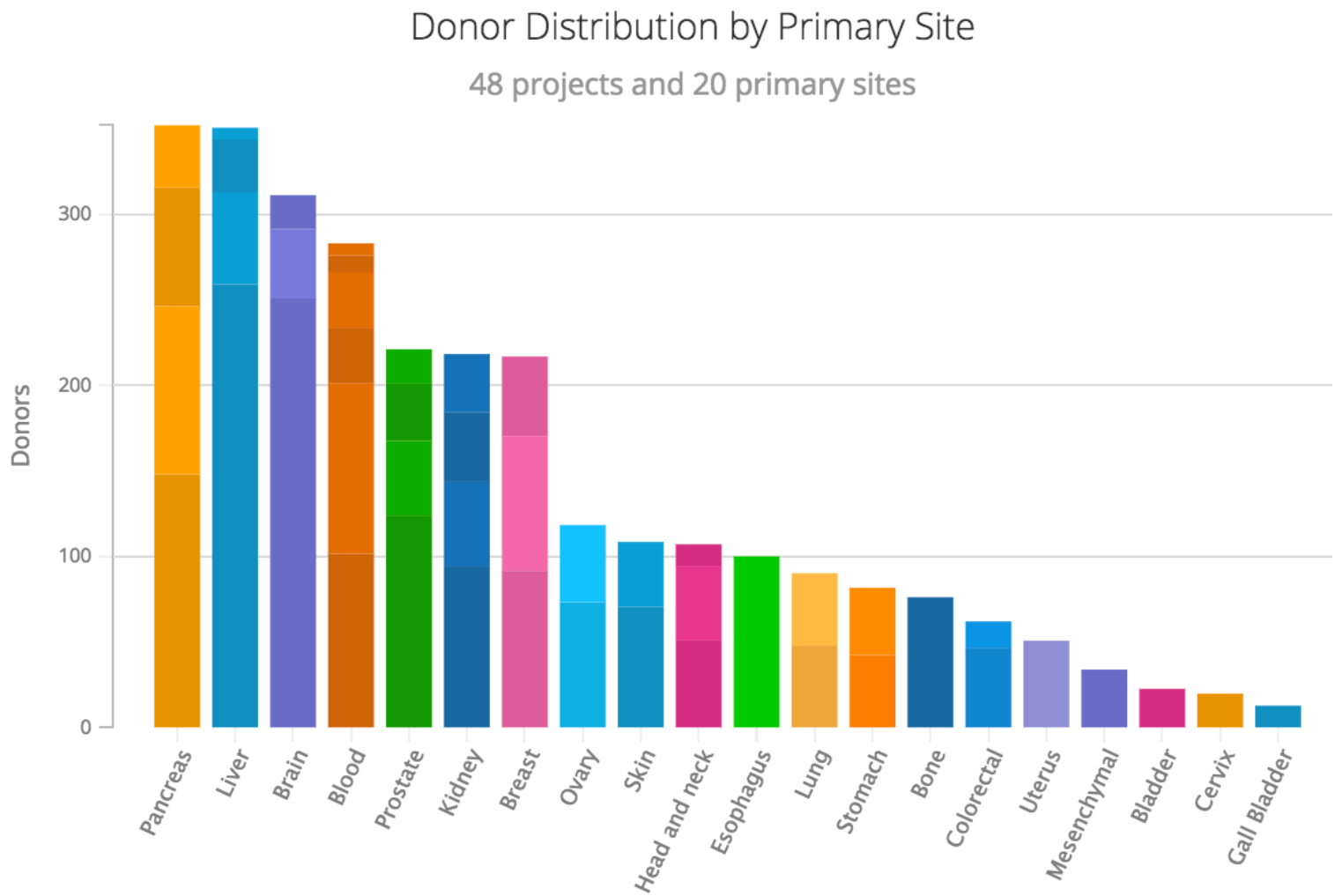
Privacy Belt
Cutting-edge cryptographic technology to ensure privacy for results returned outside of dbGaP authorization

Secure Resource
Must use internal tools
Requires user registration

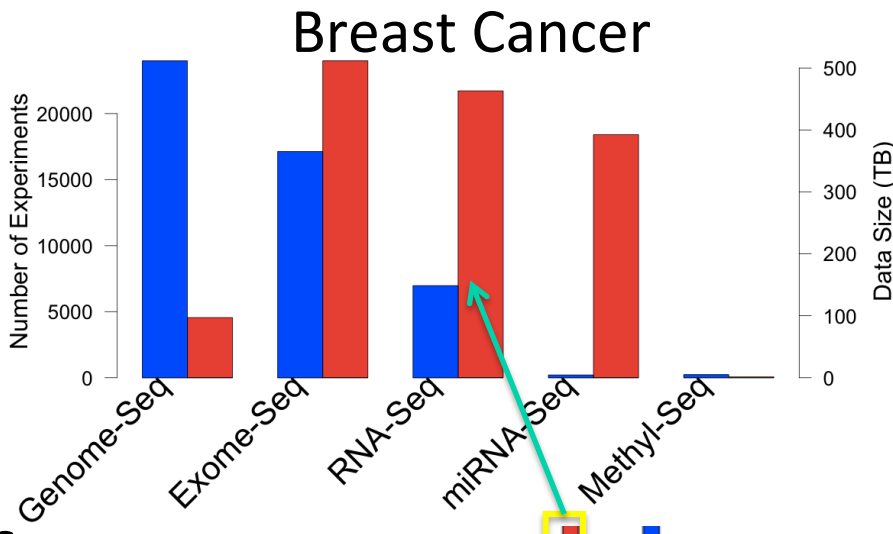
Limited Partner Grant
Bring outside tools to data
Download results only
Requires dbGaP authorization

Trusted Partner Contract
Allows data download
Requires dbGaP authorization

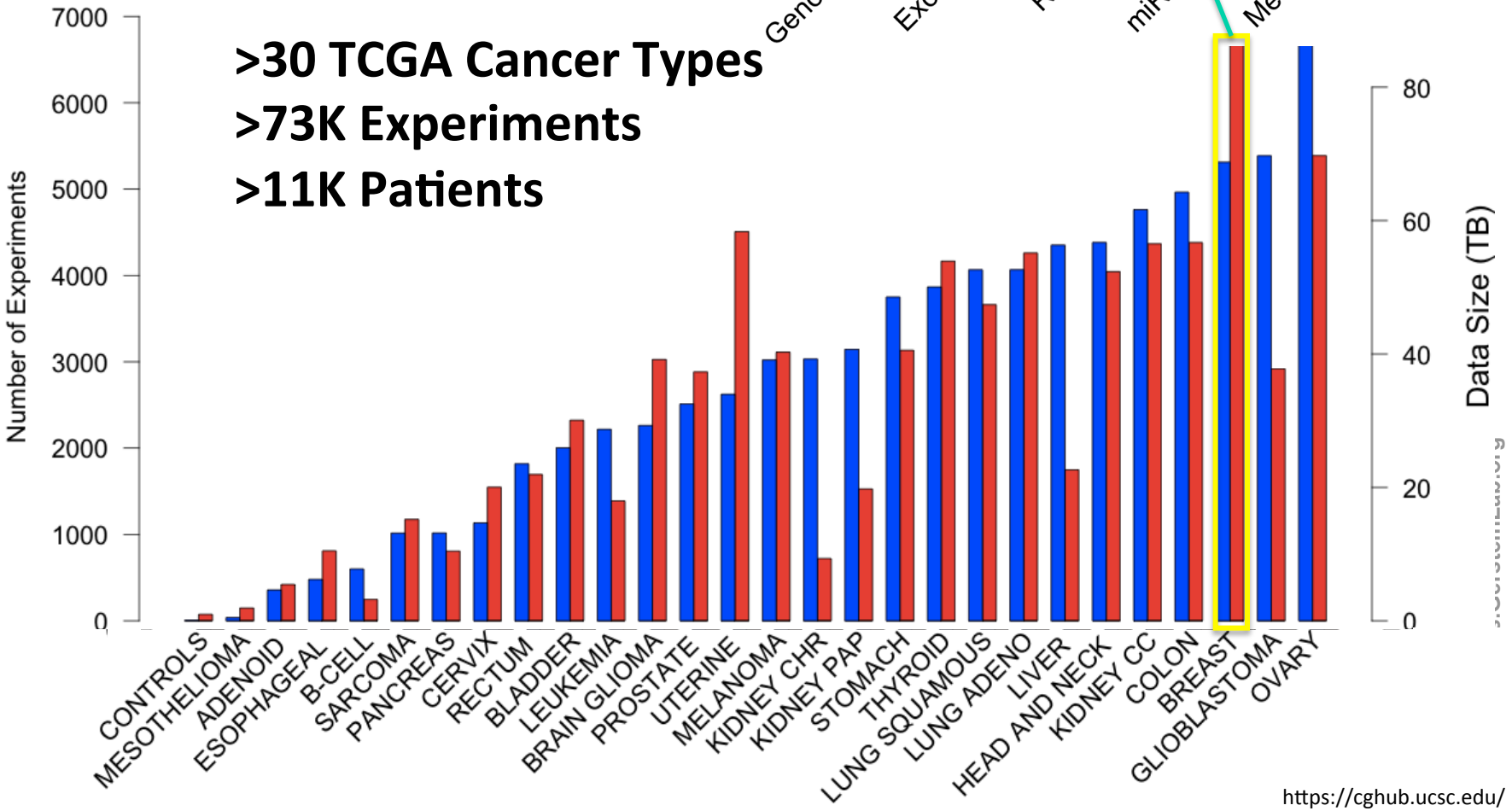
PCAWG: PANCANCER ANALYSIS OF WHOLE GENOMES



TCGA: What's in a petabyte?



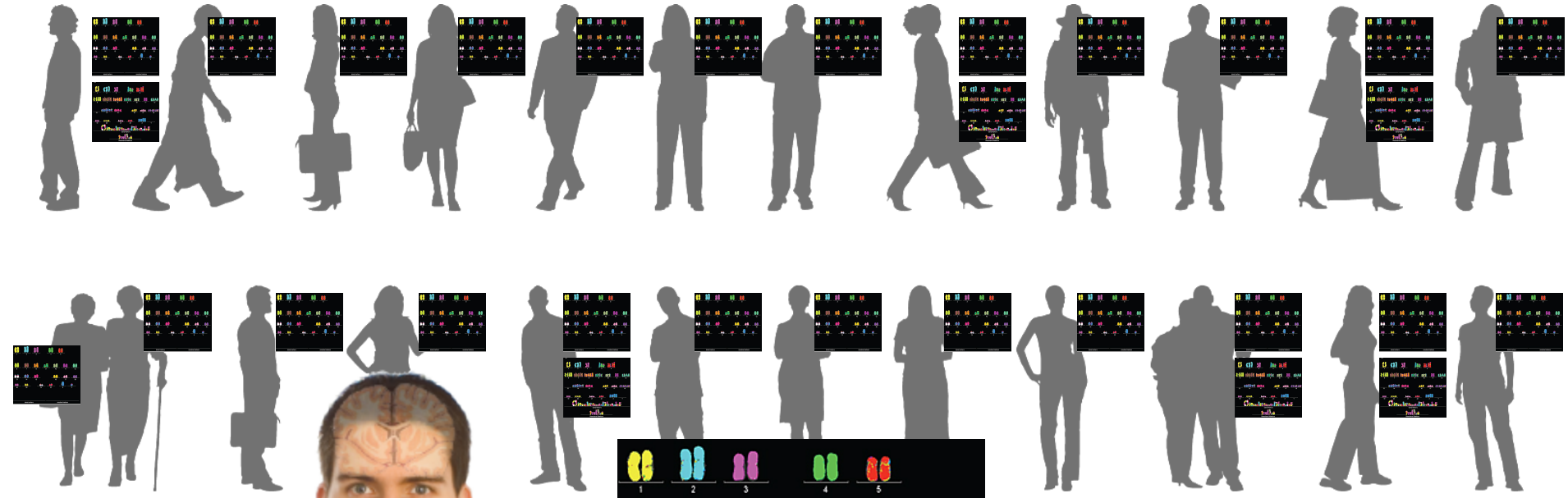
>30 TCGA Cancer Types
>73K Experiments
>11K Patients

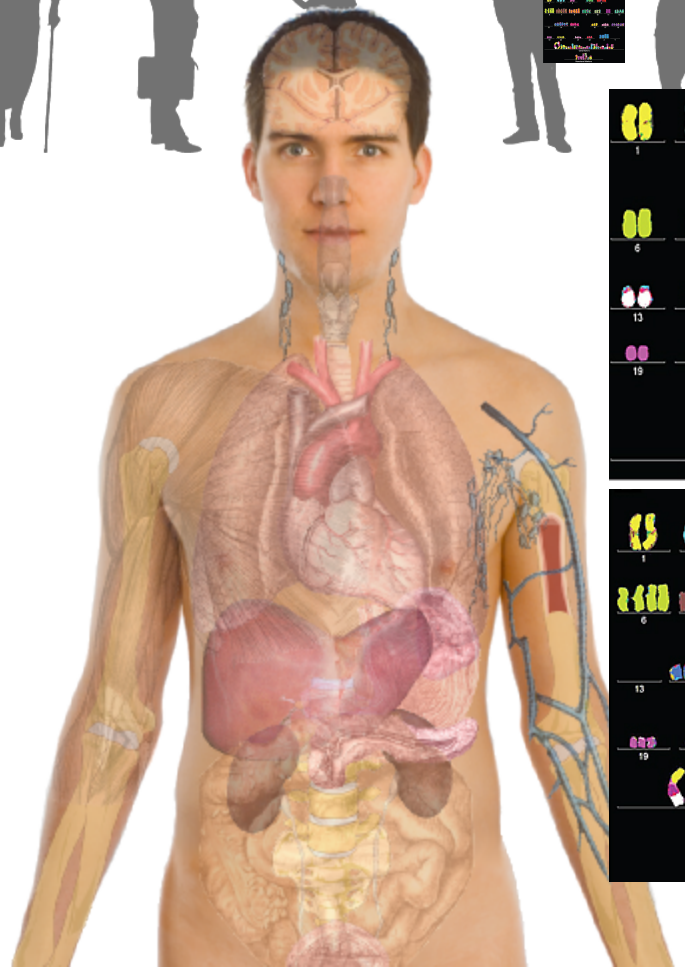


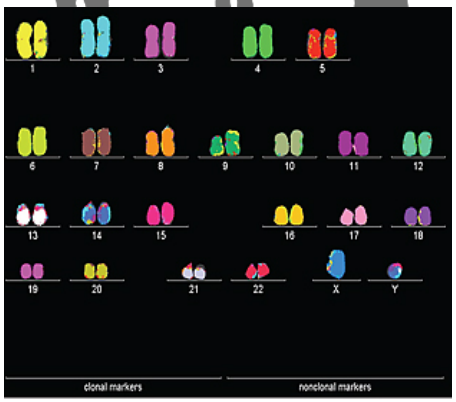
https://cghub.ucsc.edu/

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

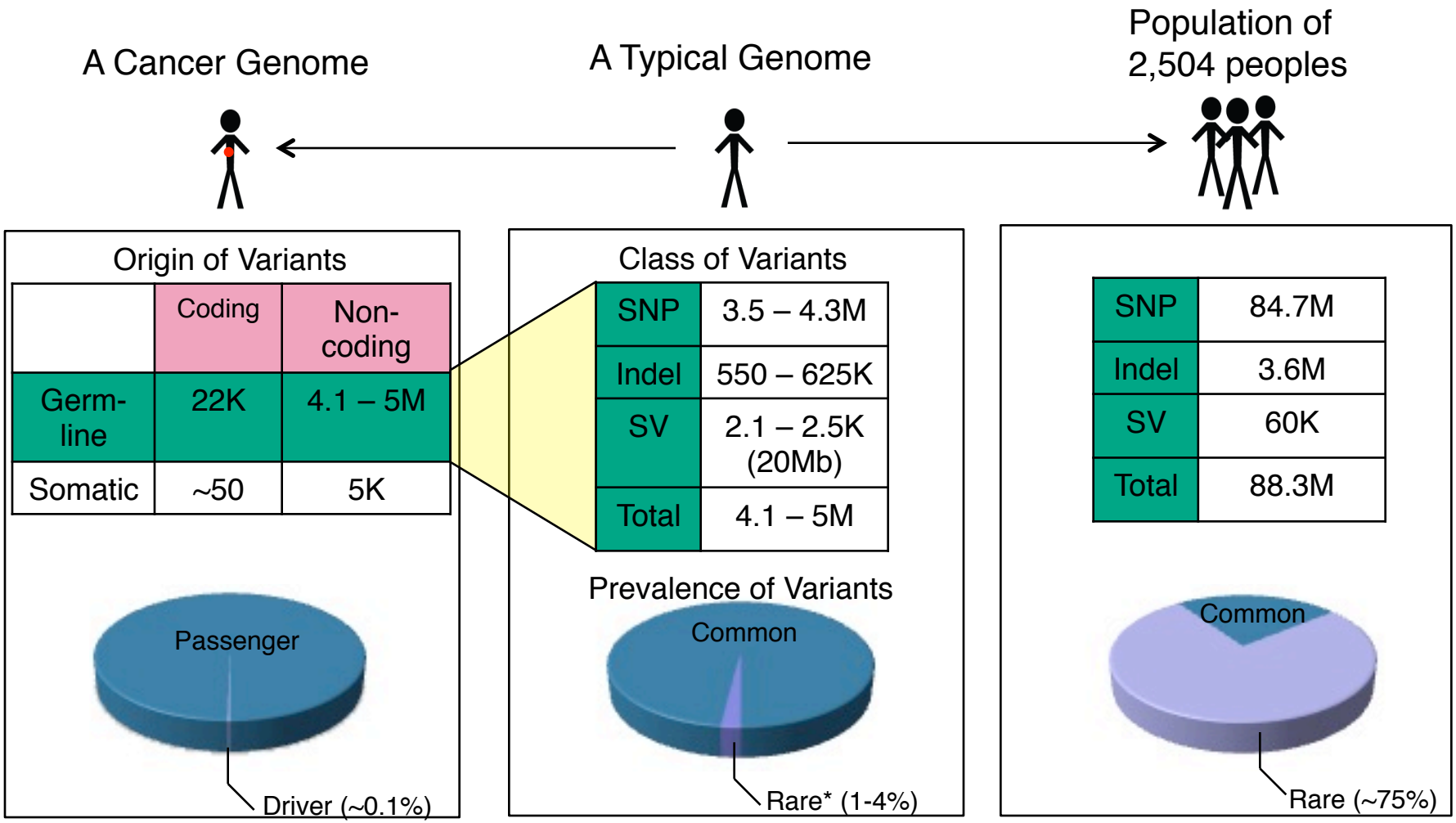






Placing the individual into the context of the population
&
using the population to build a interpretative model

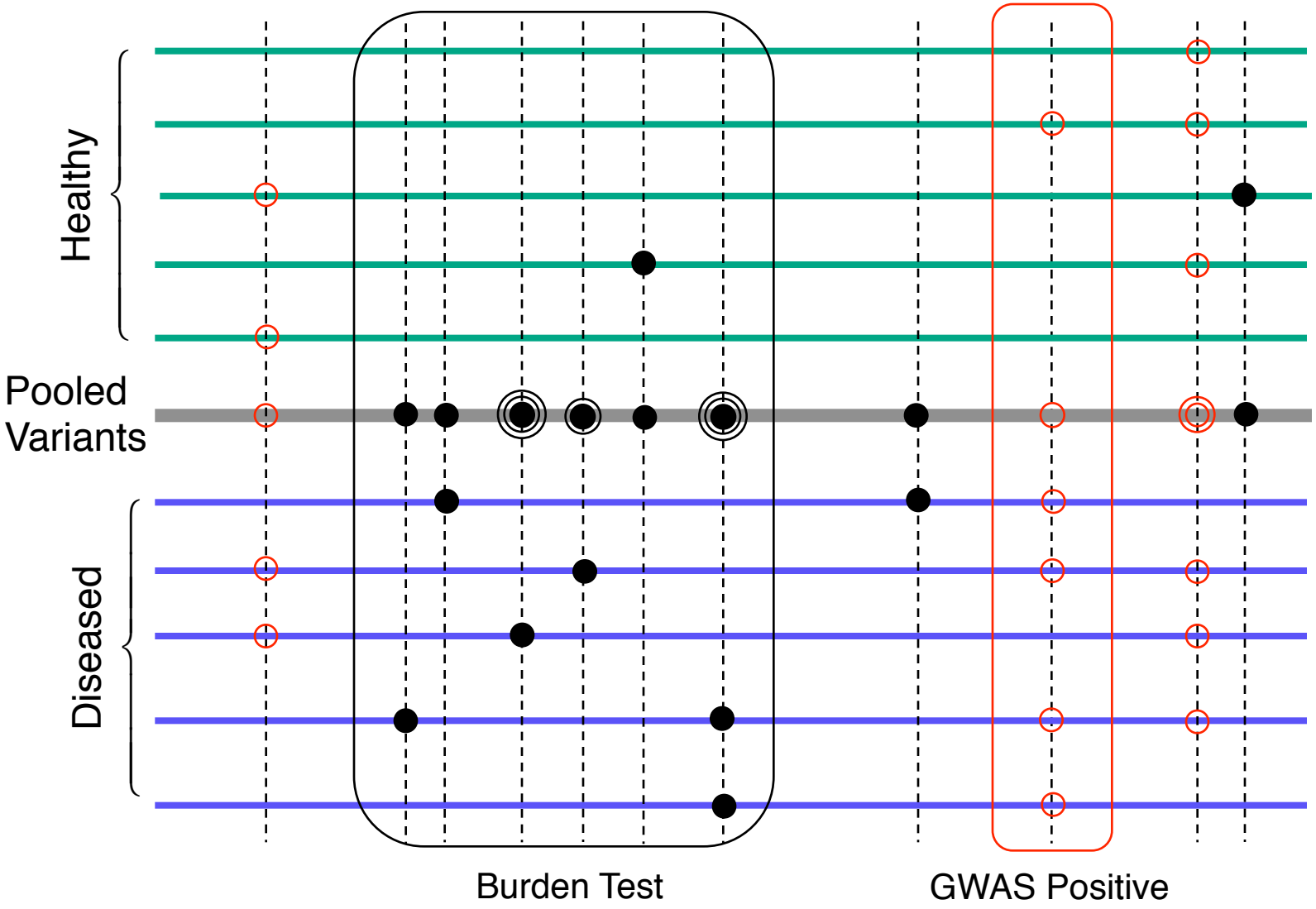
Human Genetic Variation



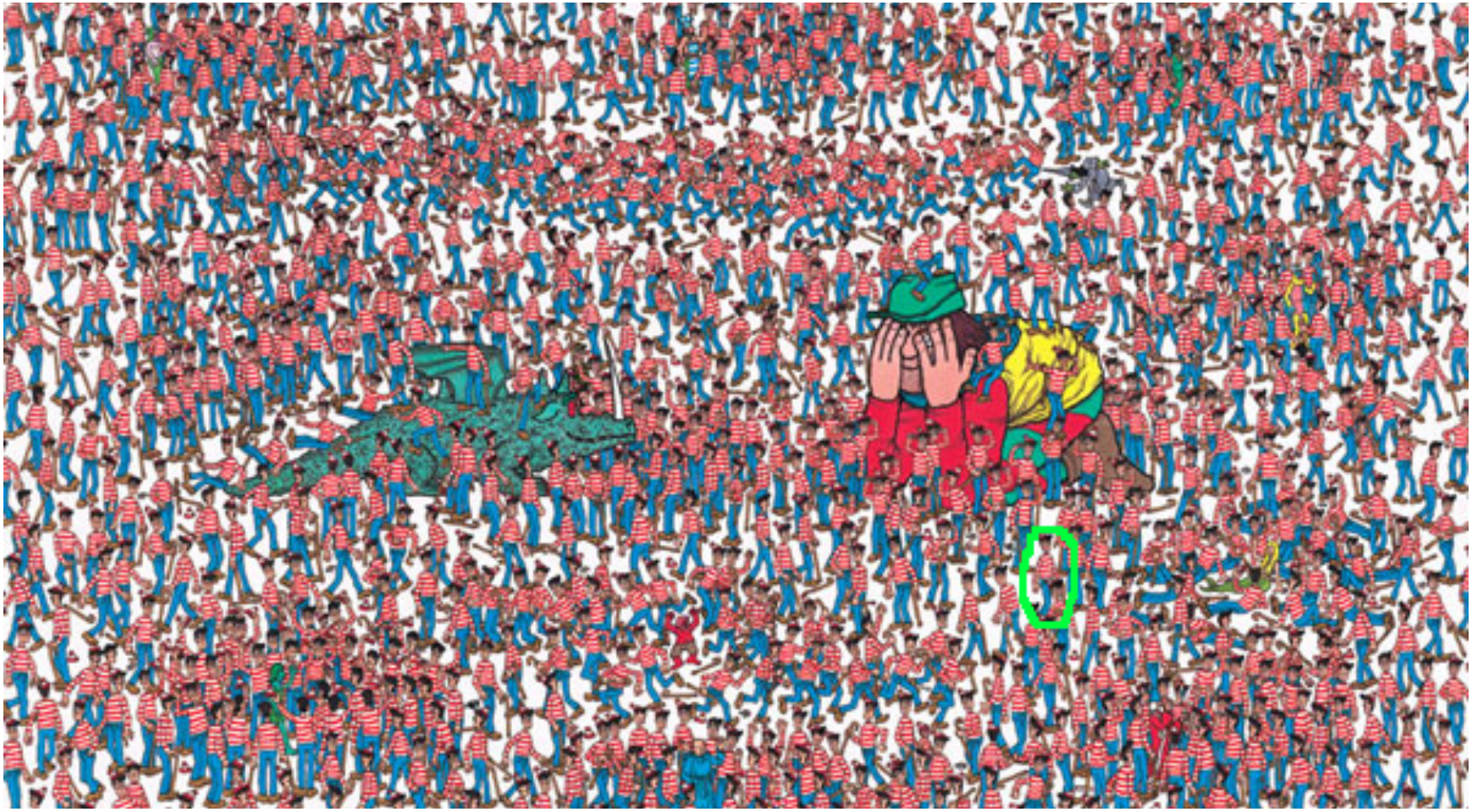
* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

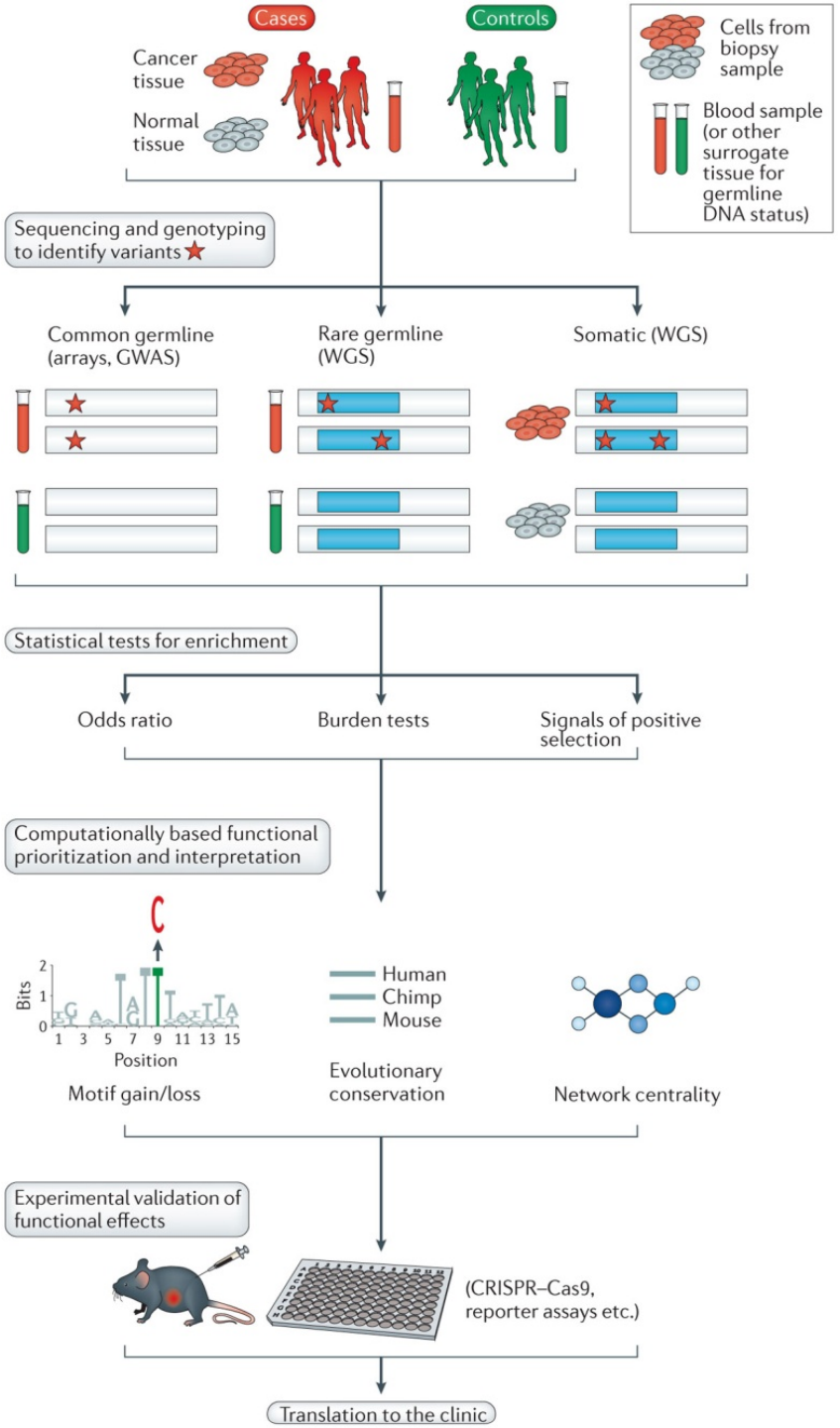
Association of Variants with Diseases



Where is Waldo?
(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)



Combined workflows for finding key variants

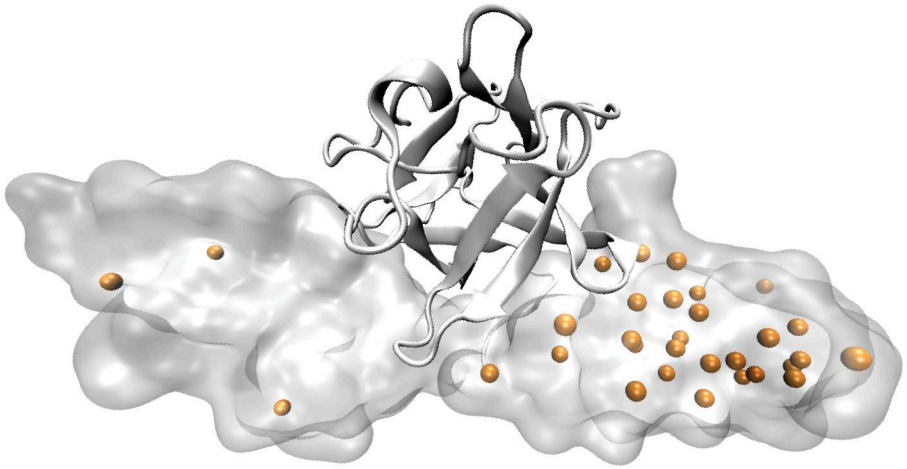


Khurana, Ekta, et al. Nature Reviews Genetics (2016)

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

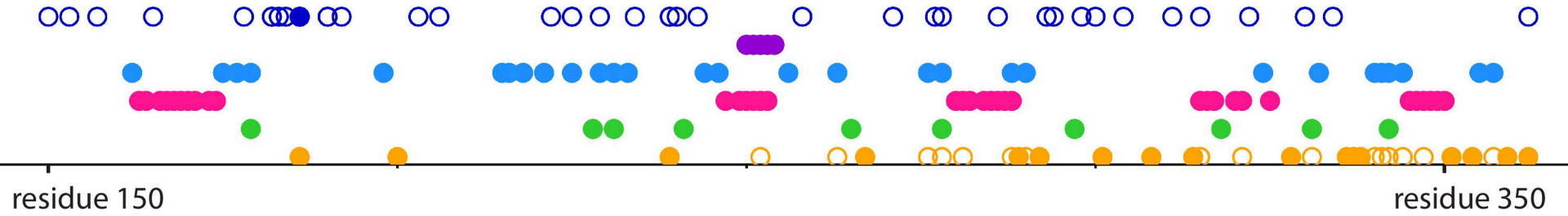
- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated



Fibroblast growth factor receptor 2 (pdb: 1IIL)

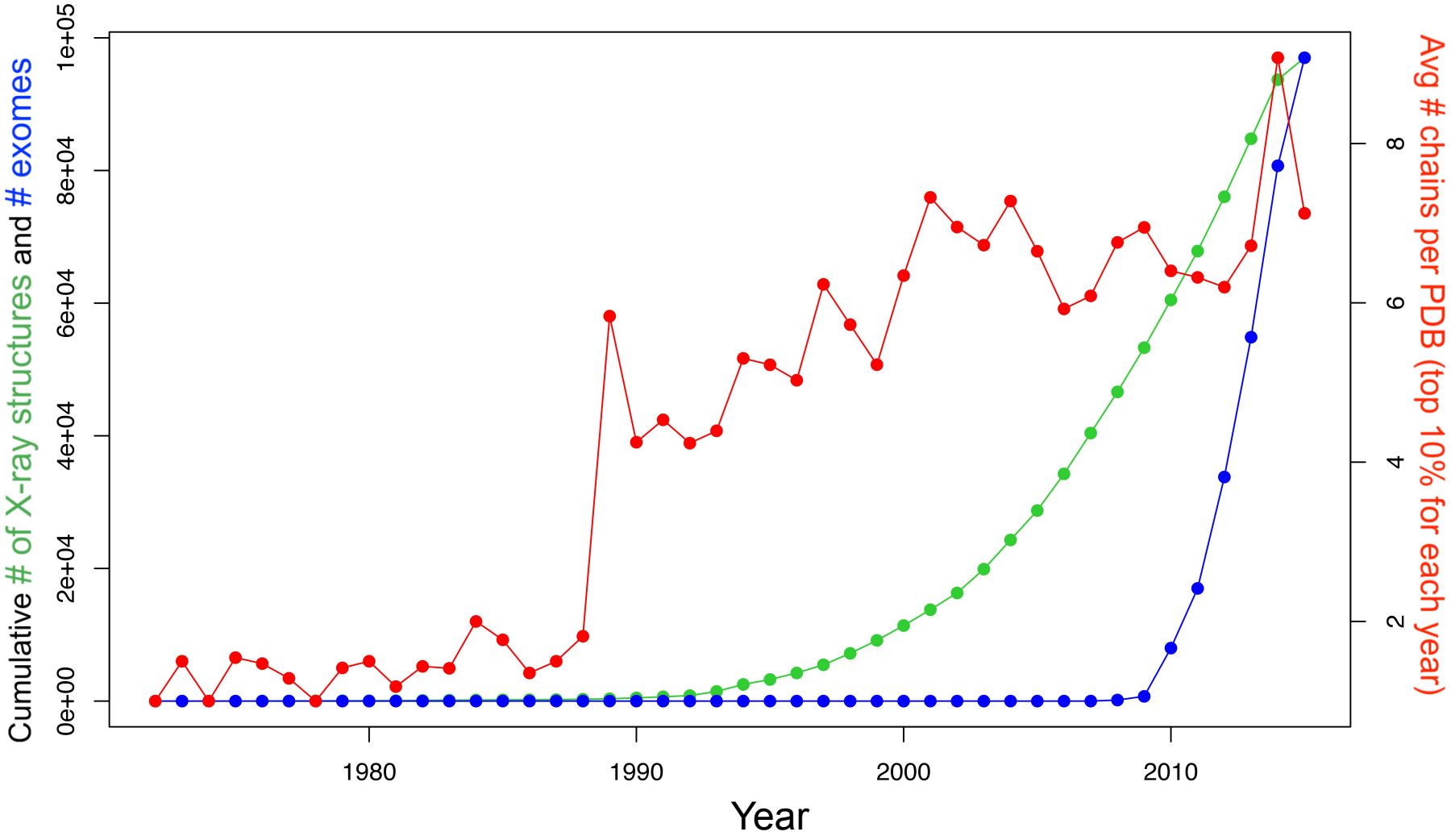
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa)

The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature.



Exome data hosted on NCBI Sequence Read Archive (SRA)

[Sethi et al. COSB ('15)]

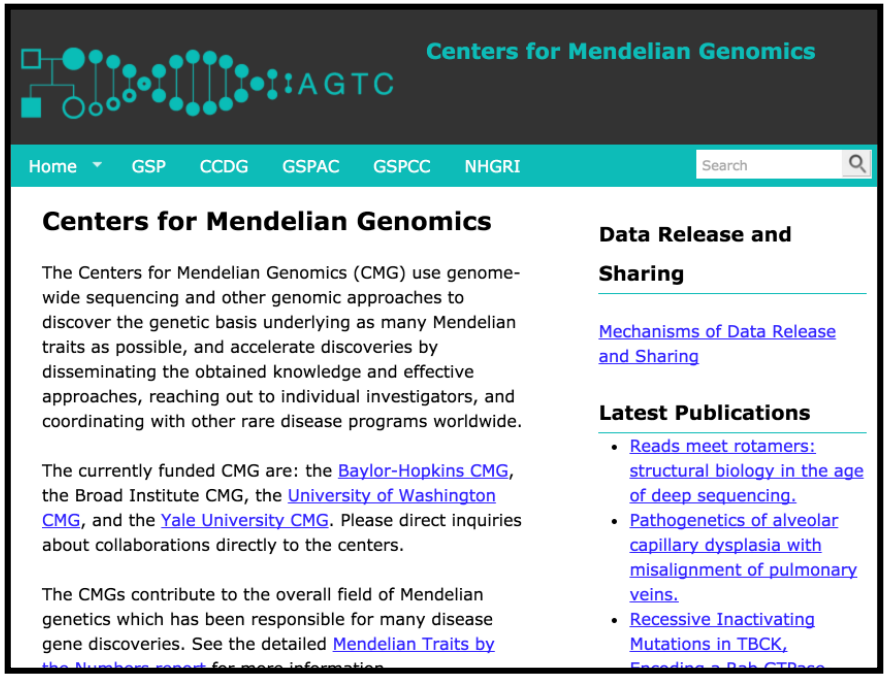
Growing sequence redundancy in the PDB (as evidenced by a reduced pace of novel fold discovery) offers a more comprehensive view of how such sequences occupy conformational landscapes

[Sethi et al. COSB ('15)]

PDB: Berman HM, et al. NAR. (2000)
CATH: Sillitoe I, et al. NAR. (2015)
SCOP: Fox NK et al. NAR. (2014)

Rare variant analysis particularly applicable at the moment to Exomes

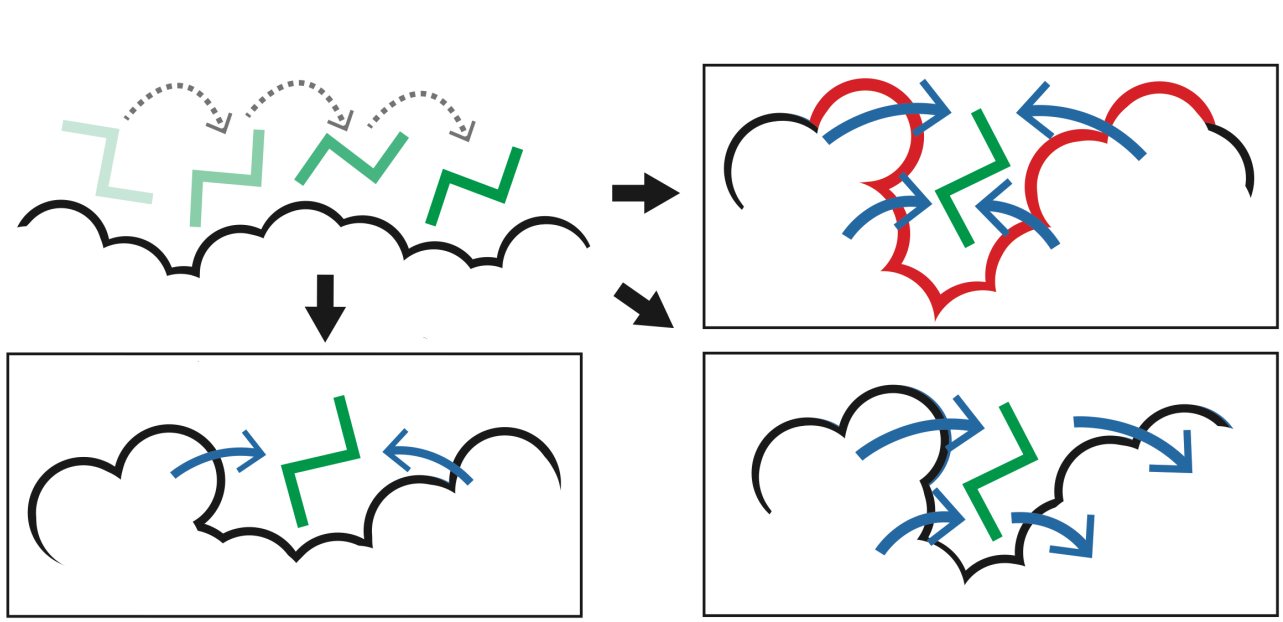
- CMG rare disease variants & TCGA somatic variants
  - Main NIH disease genomic project
  - Both of these focus on ”rare” variant for which GWAS is not meaningful
  - Larger numbers of individual exomes more important than WGS



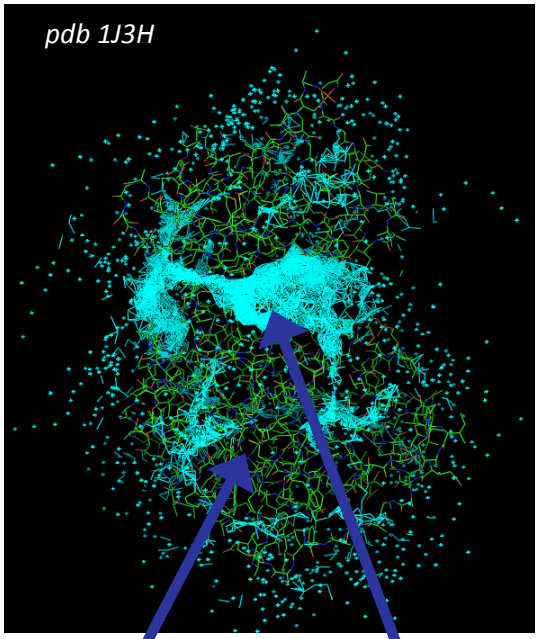
- Exomes have the current potential for great scale with the better impact interpretability of coding variants, often in a region of known protein structure
  - Scale of EXAC, >60K exomes [Lek et al. ‘16]

Predicting Allosterically-Important Residues at the Surface

1. MC simulations generate a large number of candidate sites
2. Score each candidate site by the degree to which it perturbs large-scale motions
3. Prioritize & threshold the list to identify the set of high confidence-sites



$$\textit{binding leverage} = \sum_{m=1}^{10}\left(\sum_{i}\sum_{j}\Delta d_{ij(m)}^{2}\right)$$

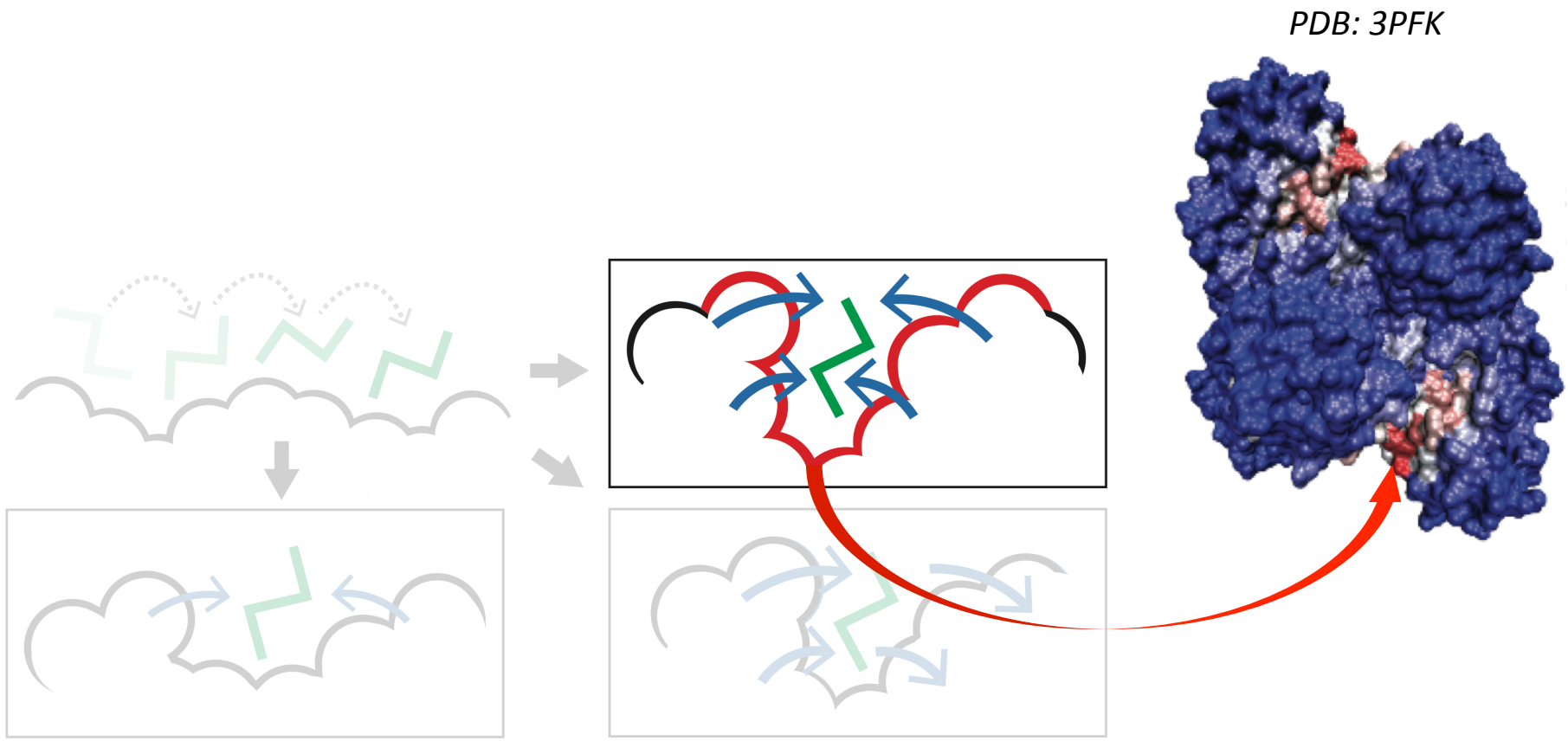


Surface region with high density of candidate sites

Surface region with low density of candidate sites

Adapted from Clarke*, Sethi*, et al *(in press)*

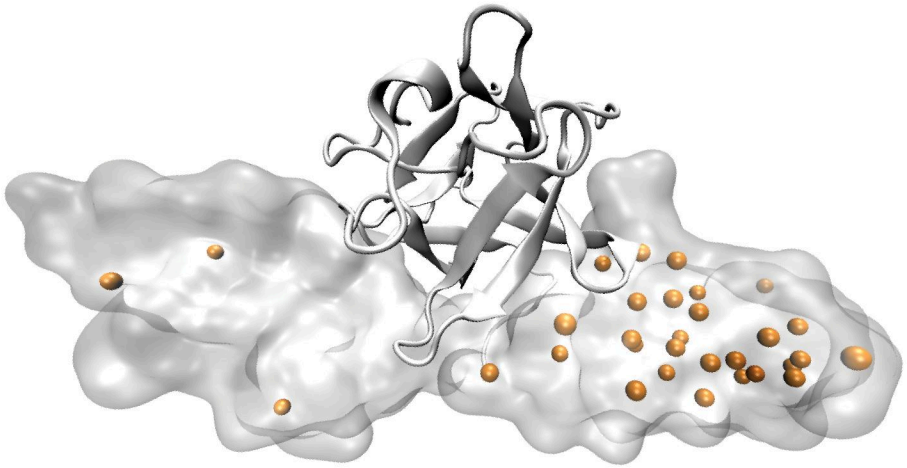
Predicting Allosterically-Important Residues at the Surface



Adapted from Clarke*, Sethi*, et al *(in press)*

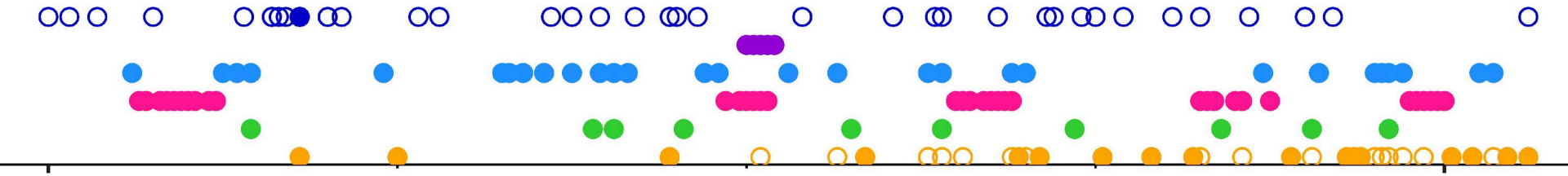
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated



Fibroblast growth factor receptor 2 (pdb: 1IIL)

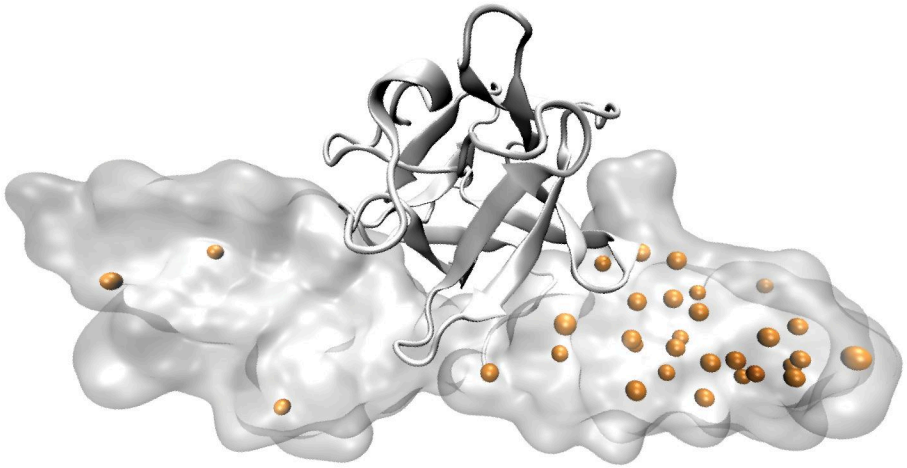
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

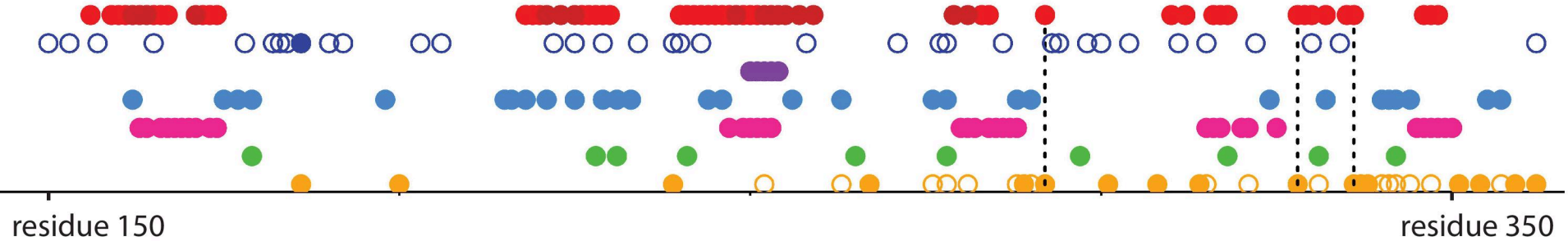
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

Rationalizing disease variants in the context of allosteric behavior with allostery as an added annotation



Fibroblast growth factor receptor 2 (pdb: 1IIL)

- Predicted allosteric (surface | interior)
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

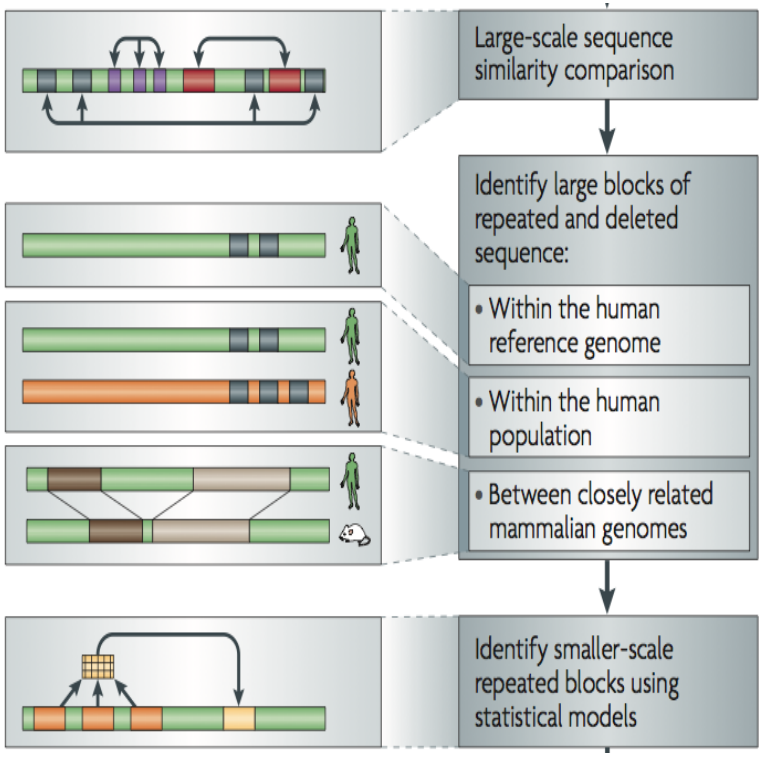
Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

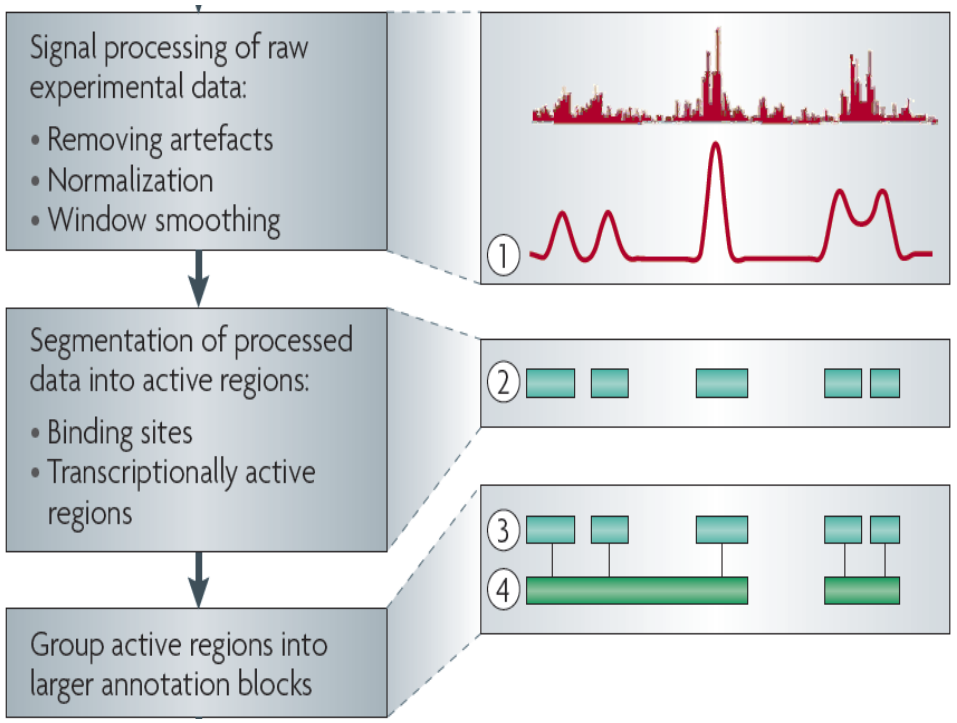
- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

Non-coding Annotations: Overview

Sequence features, incl. **Conservation**

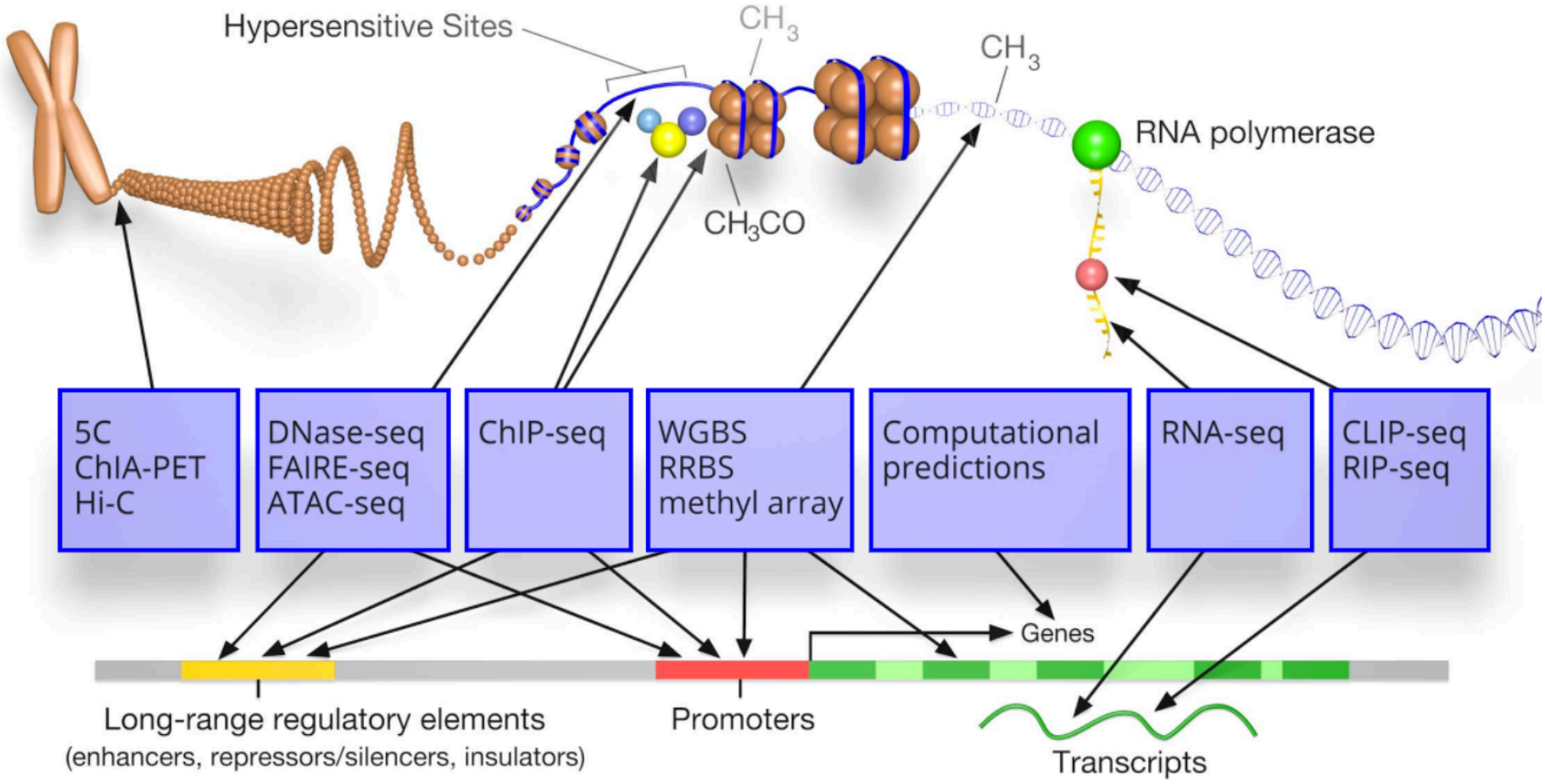
Functional Genomics
Chip-seq (Epigenome & seq. specific TF)
and ncRNA & un-annotated transcription





[Alexander et al., *Nat. Rev. Genet.* ('10)]

ENCODE: Encyclopedia of DNA Elements

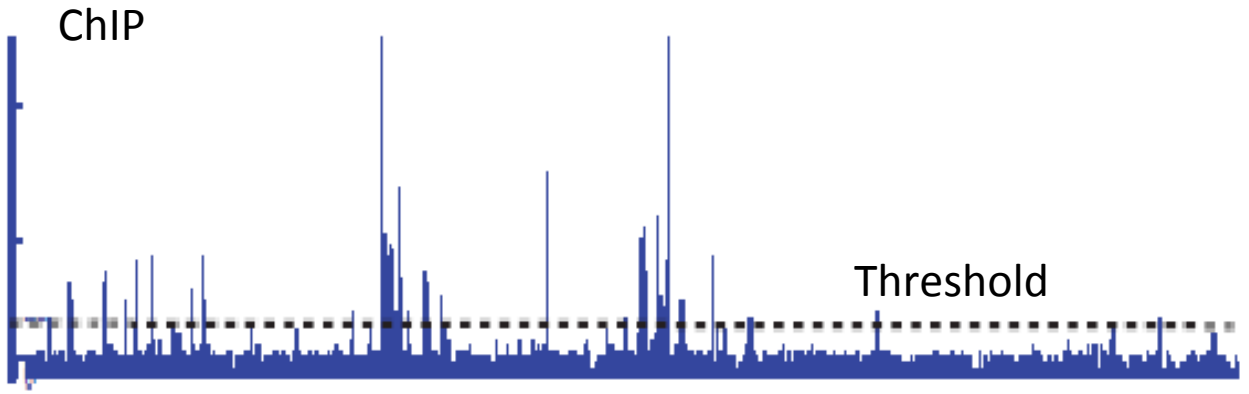


ENCODE

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)



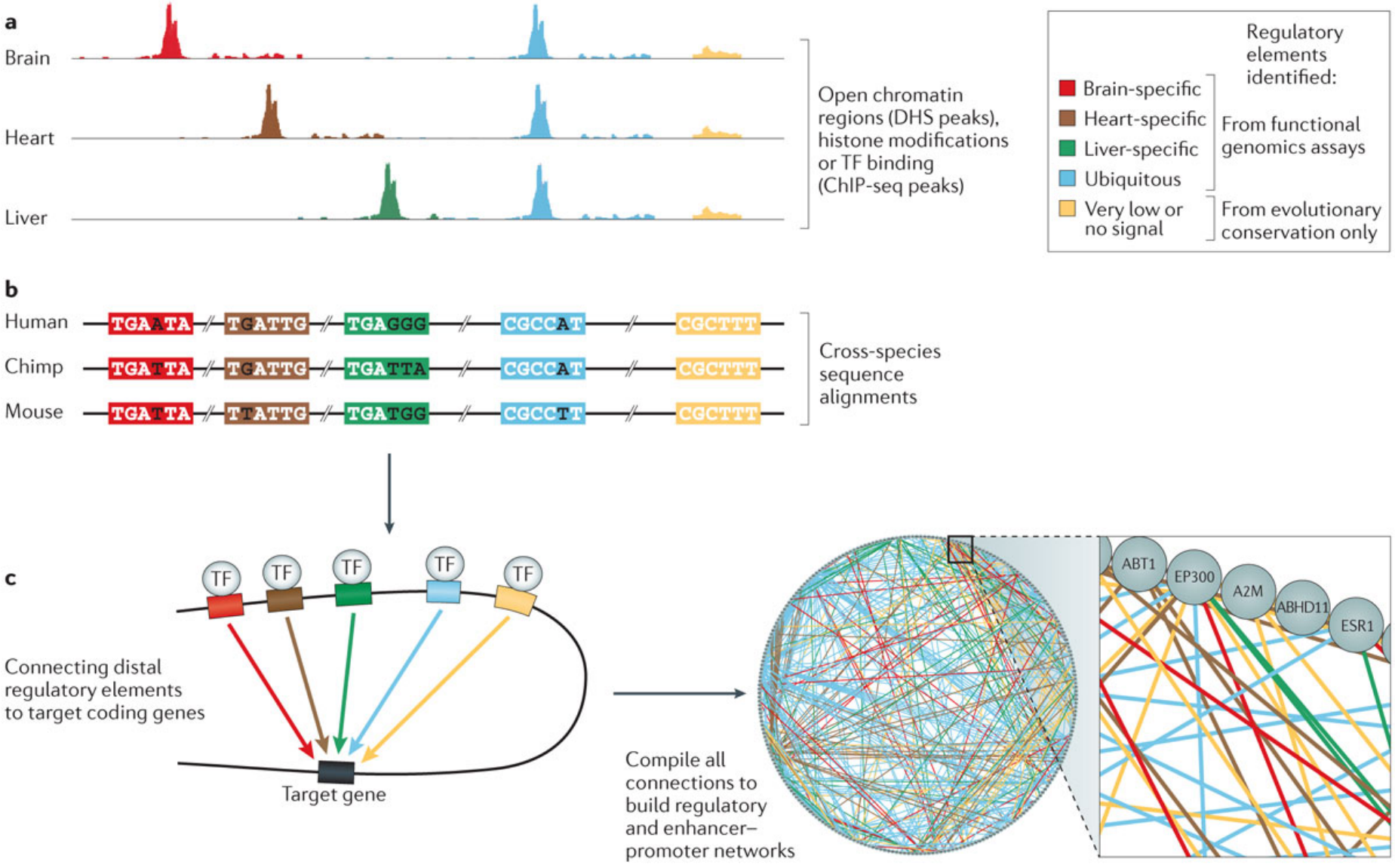
Potential Targets →

- Score against the control

Significantly Enriched targets →

[Rozowsky et al. ('09) *Nat Biotech*]

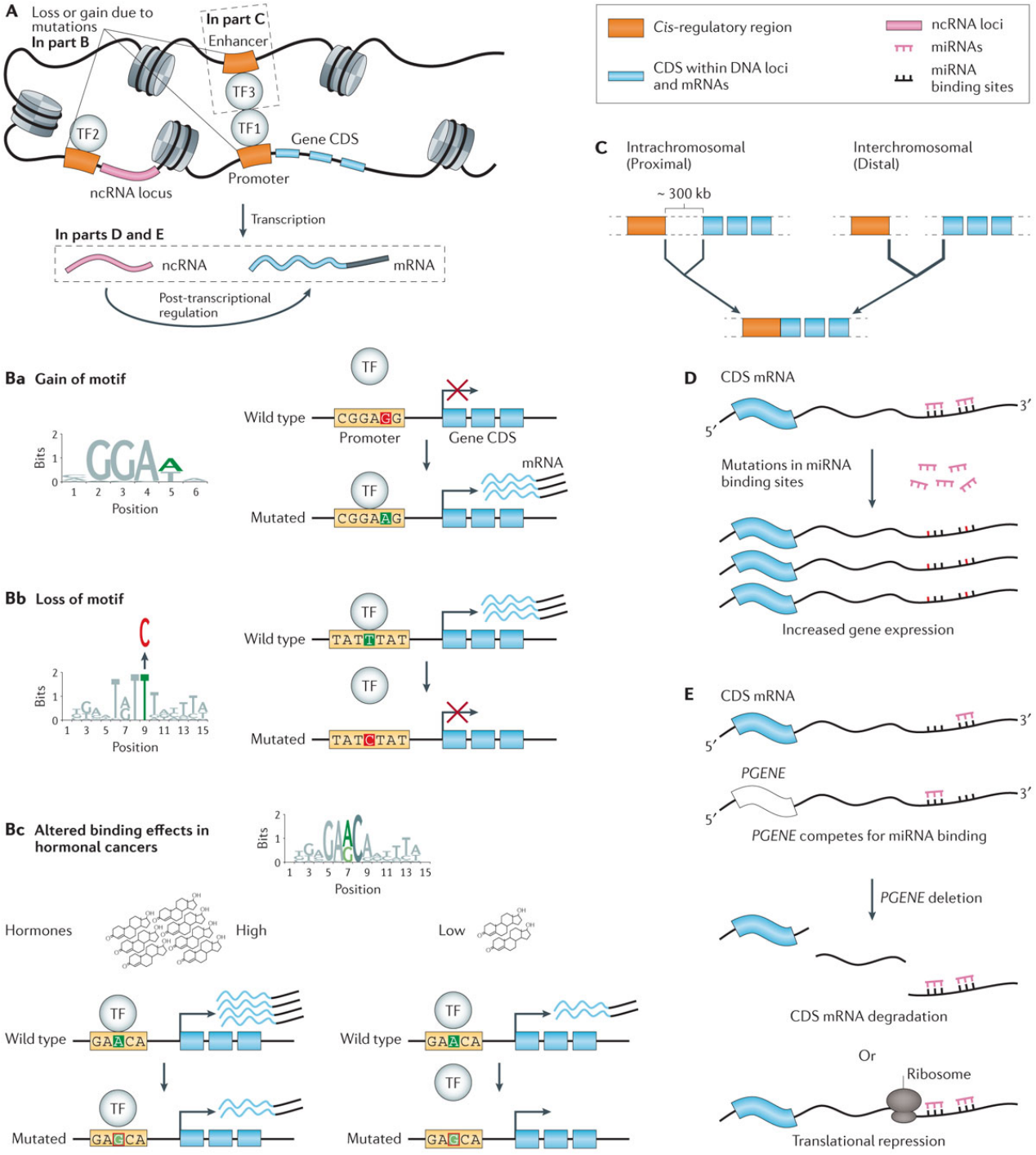
Different Active Enhancers in Different Epigenetic Contexts (ie tissues); Linking these enhancers to their target gene



Khurana, Ekta, et al. Nature Reviews Genetics (2016)

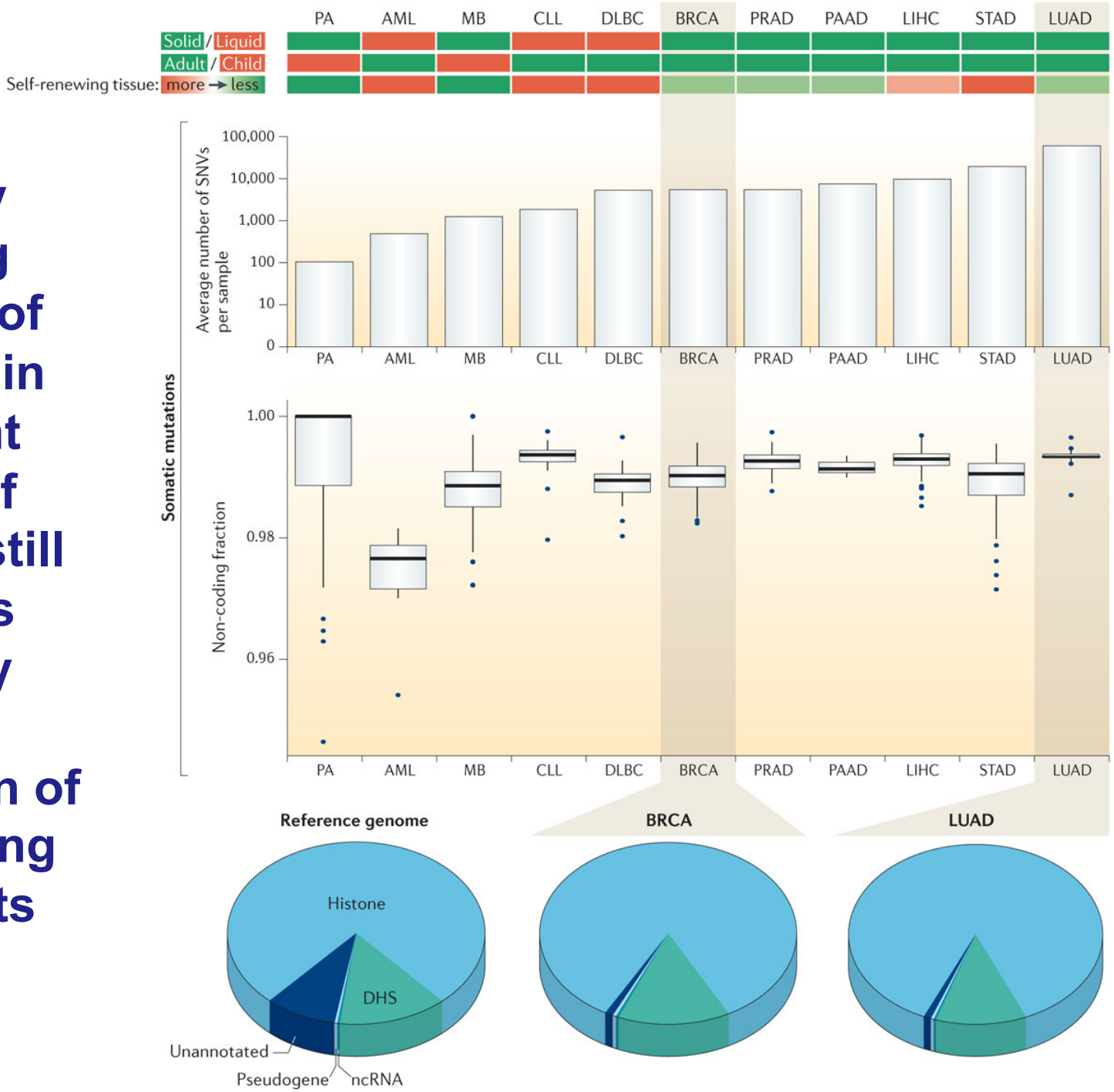
Many different ways that variants can impact non-coding elements

~80% of disease-associated GWAS variants in noncoding regions (Hindorff et al. 2009 PNAS)



Khurana, Ekta, et al. Nature Reviews Genetics (2016)

Greatly varying number of variants in different types of cancers still impacts roughly same proportion of non-coding elements



Khurana, Ekta, et al. Nature Reviews Genetics (2016)

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

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

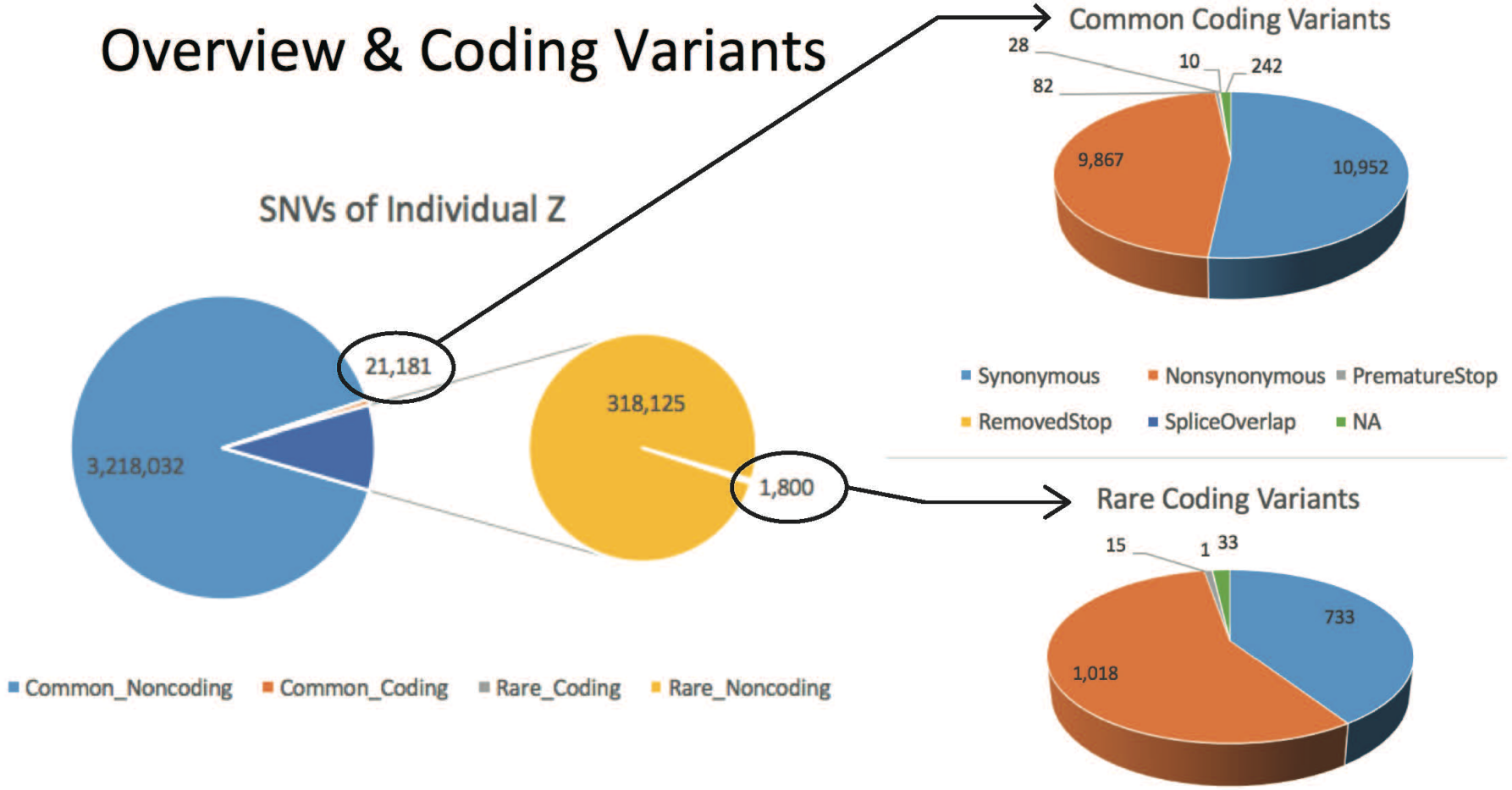


Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

Rare Non-synonymous Coding Variants

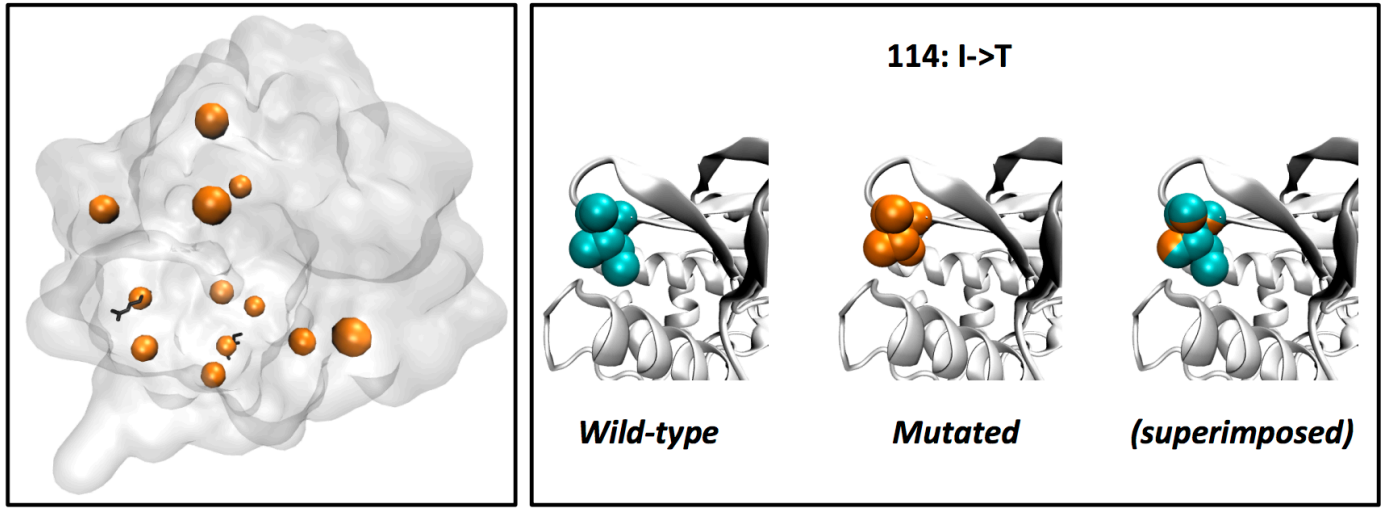
- 1018 SNVs -> 824 target genes

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)



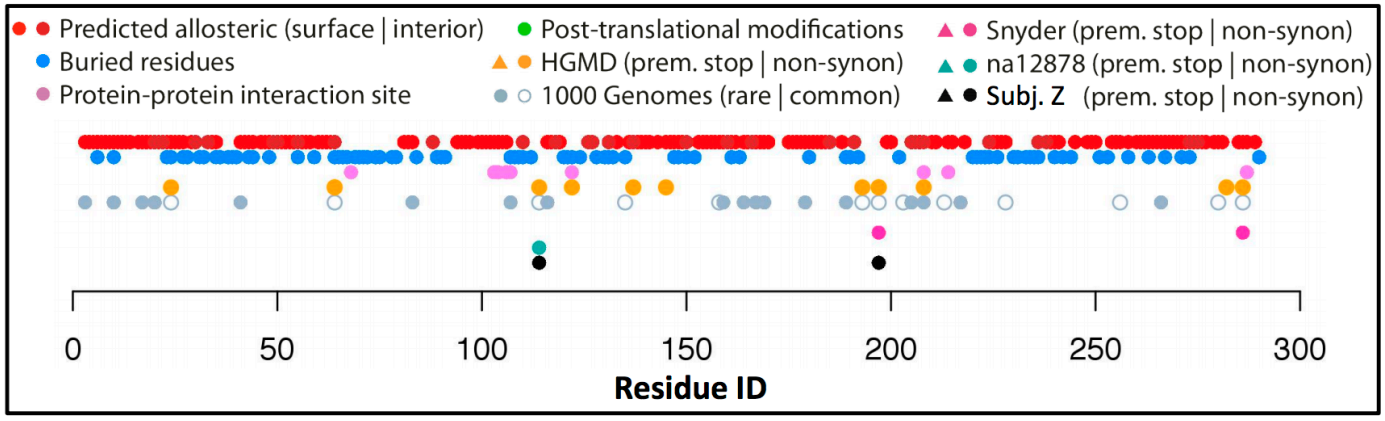
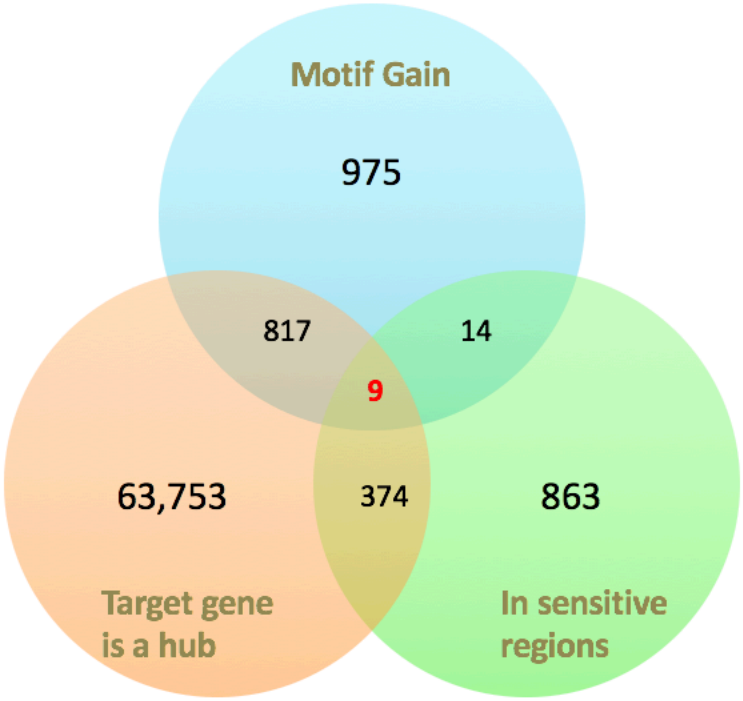


Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

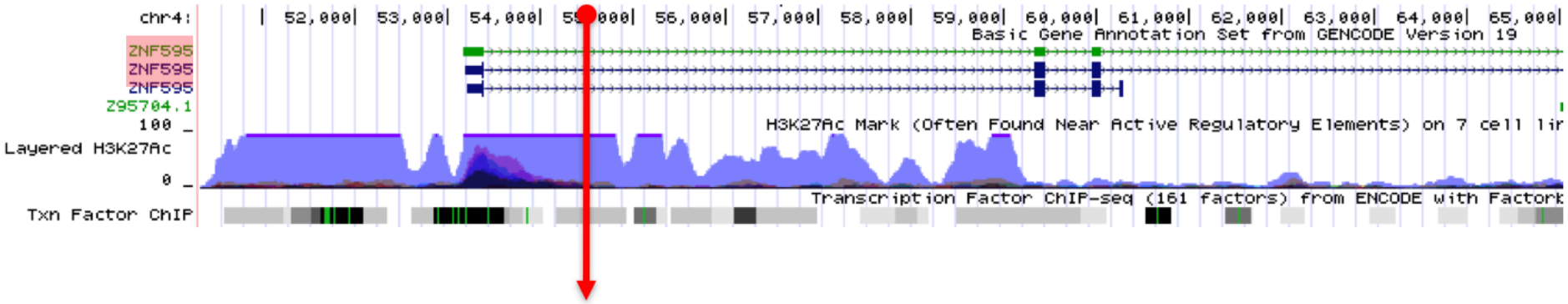
Annotation of Rare Noncoding Variants



- 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

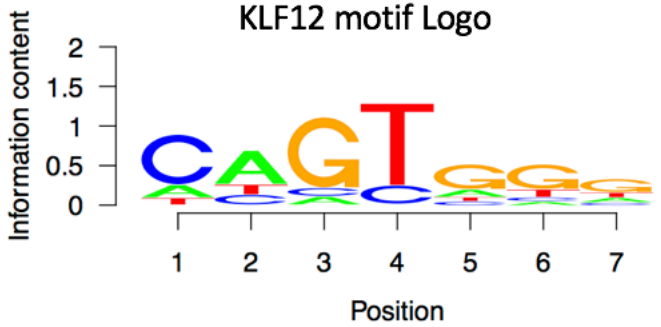


Rare noncoding SNV

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

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

- Chr4:54469-54476
- Minus strand



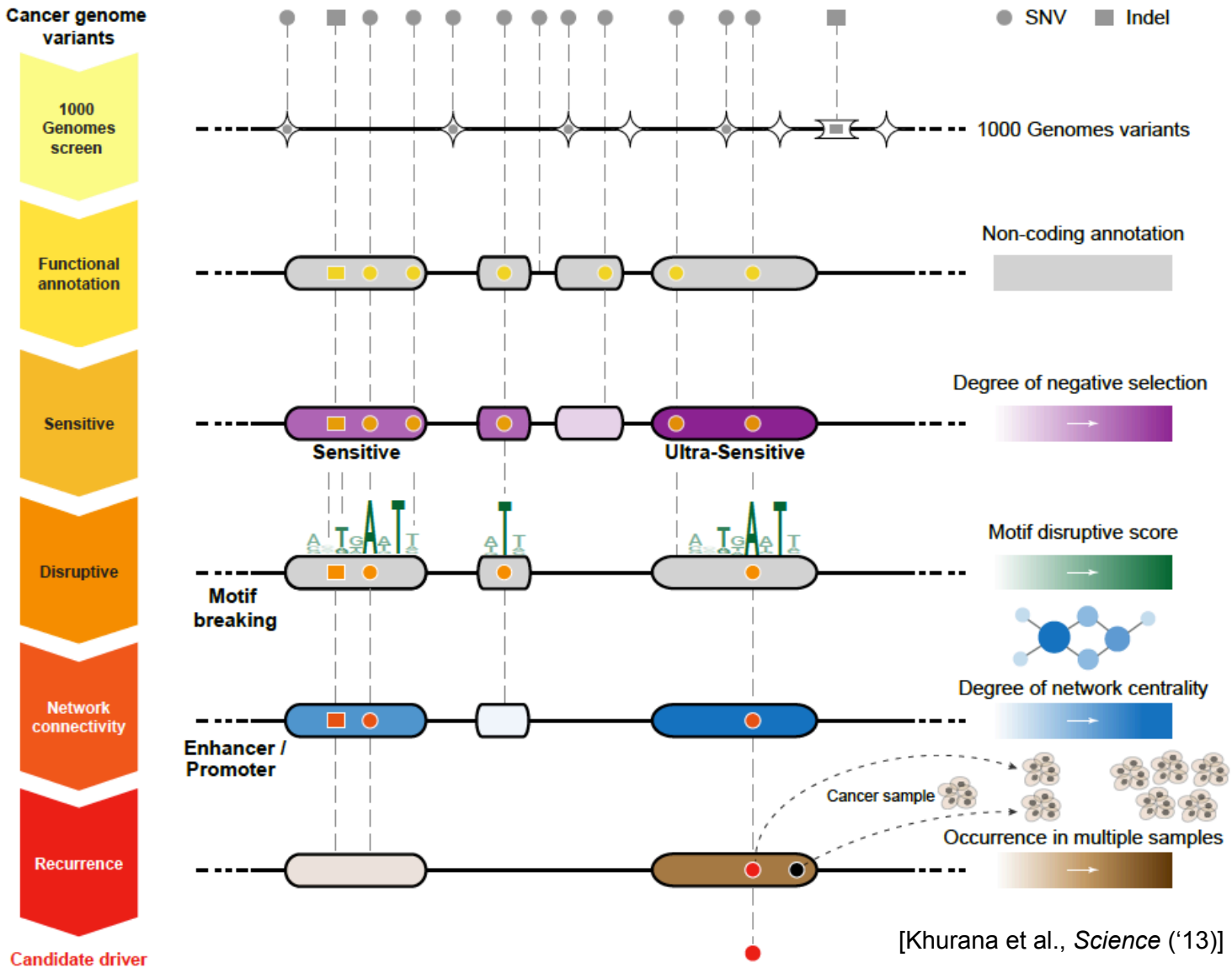
A

P-value < 4e-8

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

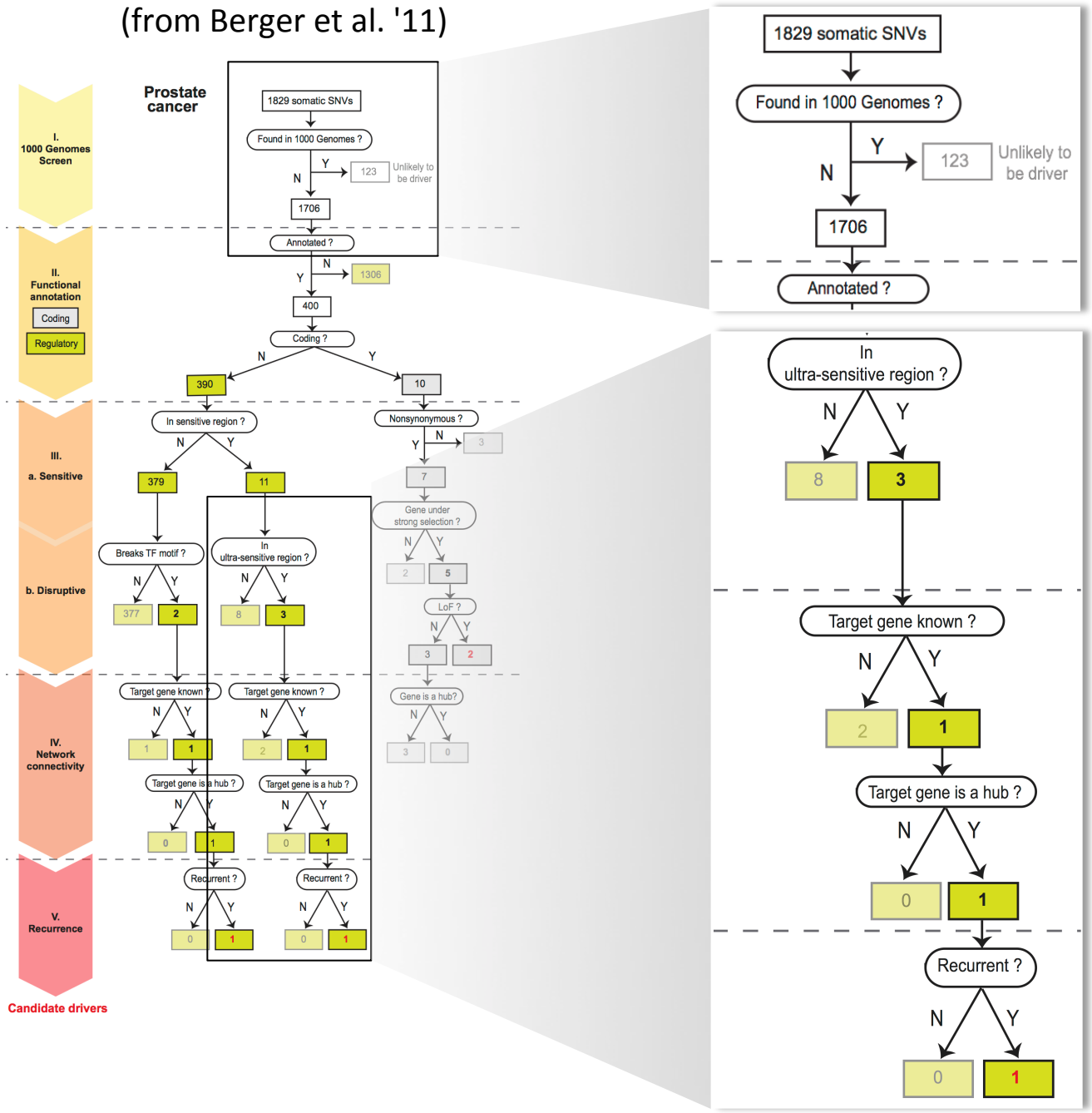
- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

Identification of non-coding candidate drivers amongst somatic variants: Scheme



[Khurana et al., *Science* ('13)]

Flowchart for 1 Prostate Cancer Genome (from Berger et al. '11)



[Khurana et al., *Science* ('13)]

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

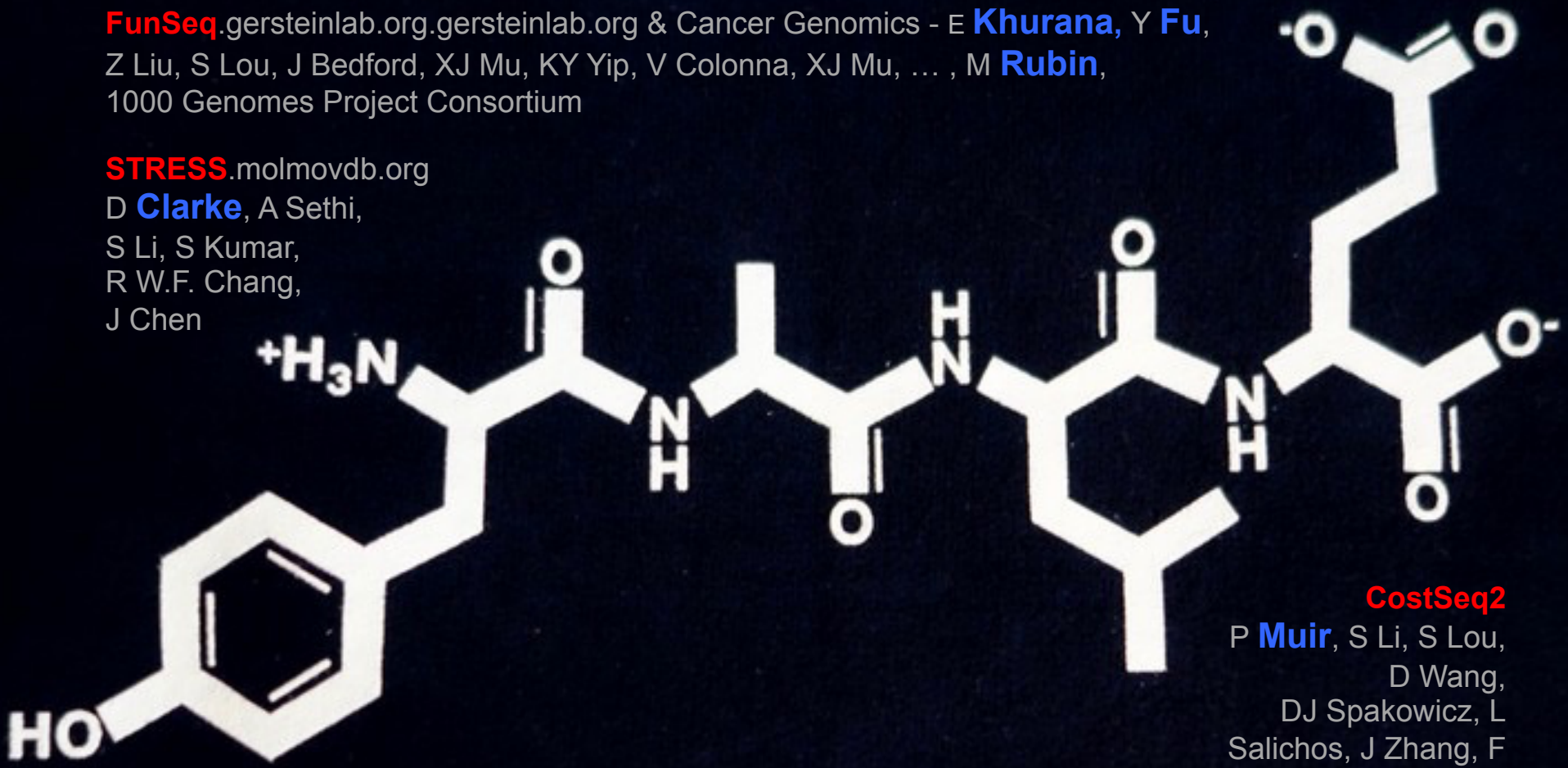
- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

Key Drivers for Making Personal Genomic Sequencing into a Useful Tool

- Falling costs of sequencing & computing
  - How this exponential scaling will change personal genomics
- One result of the falling costs: a huge & diverse DB
  - Particularly with large amounts of cancer genomics data
- Using the database to help prioritize variants
  - Rare & common variants
  - Prioritizing rare coding variants with molecular structures
    - Ex of STRESS
  - Interpreting non-coding variants with non-coding annotation
- Putting it together in Workflows
  - Game of Genomes
    - Illustrating the prioritization on an identifiable individual
    - For variants, going from millions to tens
  - Using FunSeq to integrate evidence on noncoding variants
    - Suggesting a non-coding drivers

FunSeq.gersteinlab.org.gersteinlab.org & Cancer Genomics - E **Khurana,** Y **Fu**, Z Liu, S Lou, J Bedford, XJ Mu, KY Yip, V Colonna, XJ Mu, … , M **Rubin**, 1000 Genomes Project Consortium

STRESS.molmovdb.org
D **Clarke**, A Sethi, S Li, S Kumar, R W.F. Chang, J Chen



CostSeq2
P **Muir**, S Li, S Lou, D Wang, DJ Spakowicz, L Salichos, J Zhang, F Isaacs, J Rozowsky

statnews.com/feature/**game-of-genomes** + **Zimmerome**.gersteinlab.org
C **Zimmer**, S **Kumar**, J **Rozowsky**, W **Meyerson,** D Clarke, X Li, F Navarro

Y - A - L - E

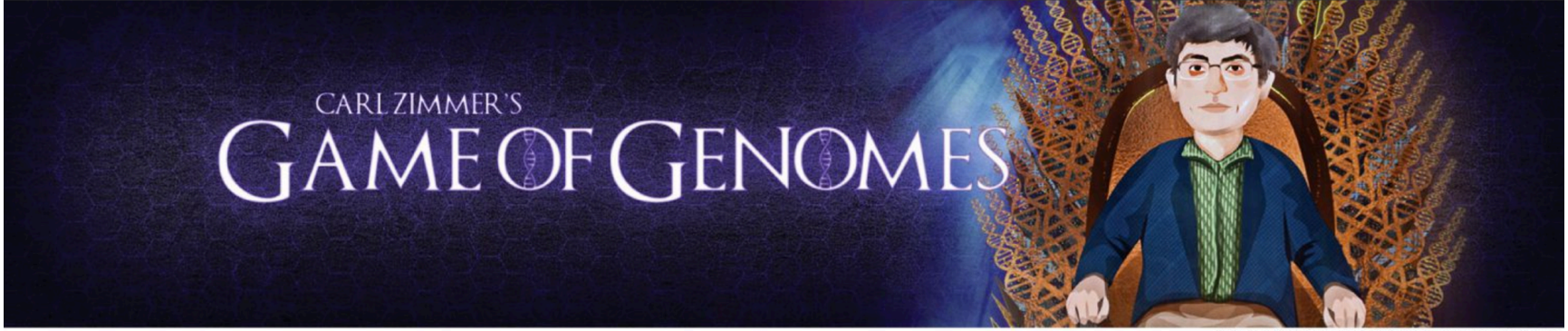


Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

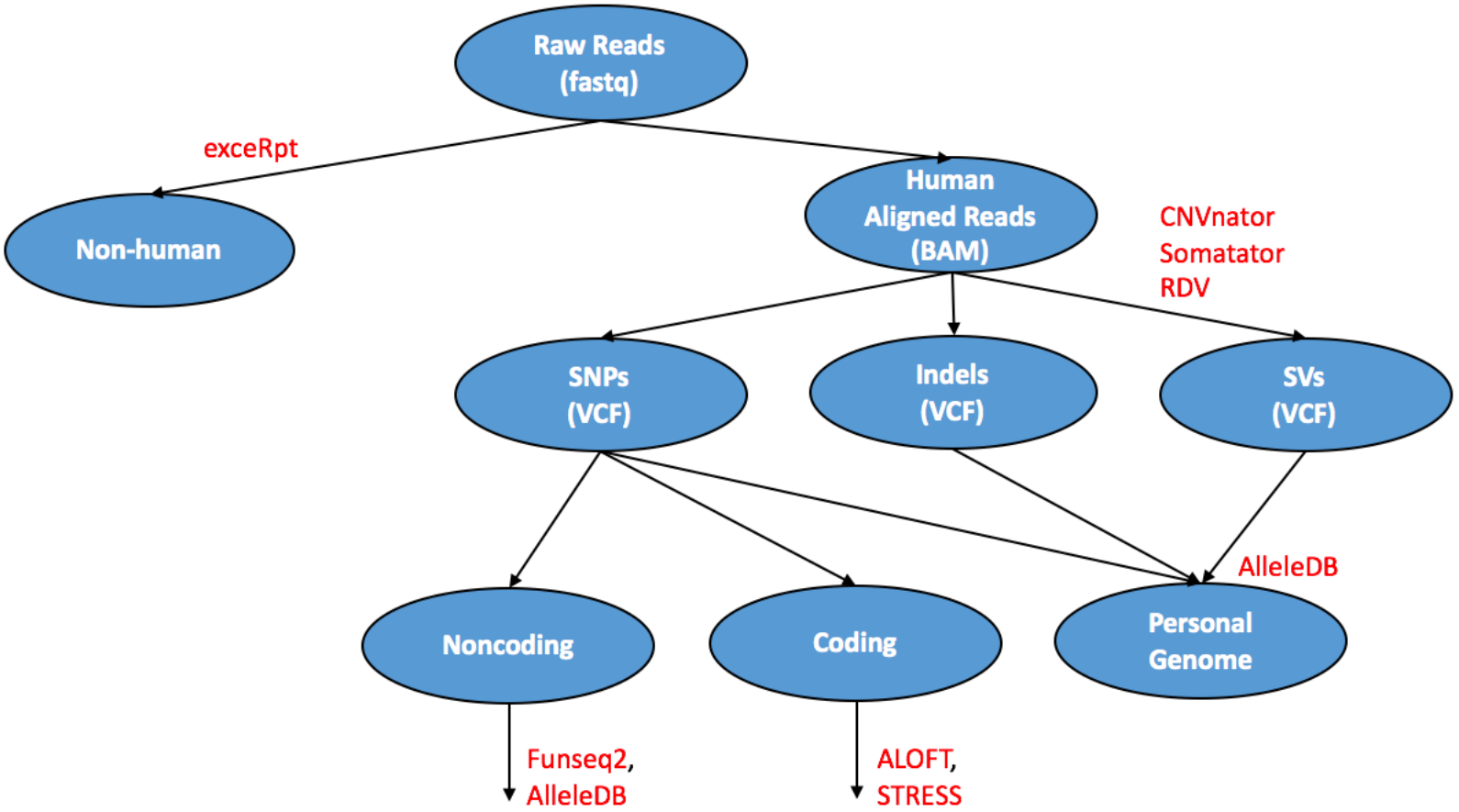
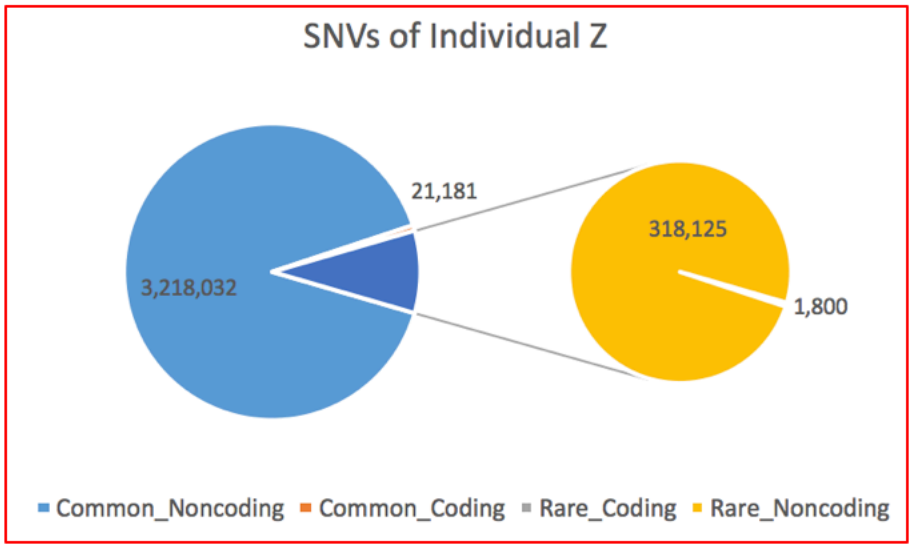
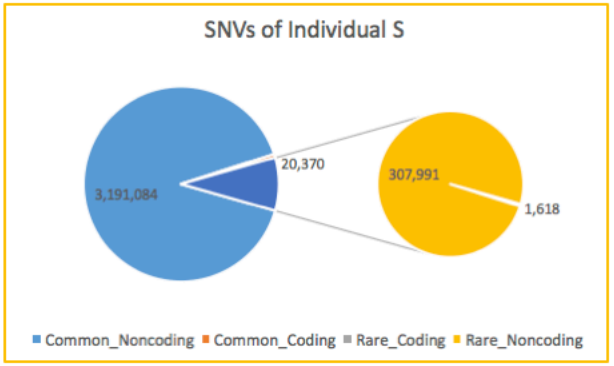
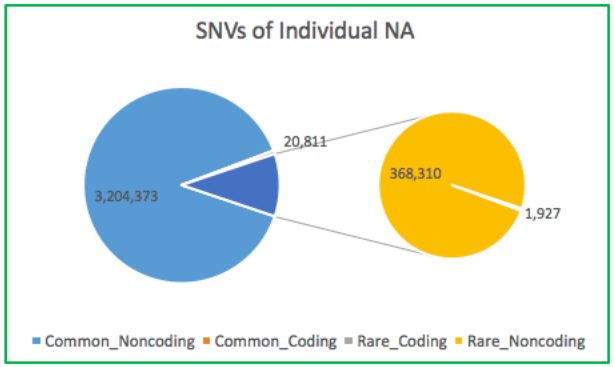


Illustration: Molly Ferguson for STAT; Animation: Dom Smith/STAT

Overview





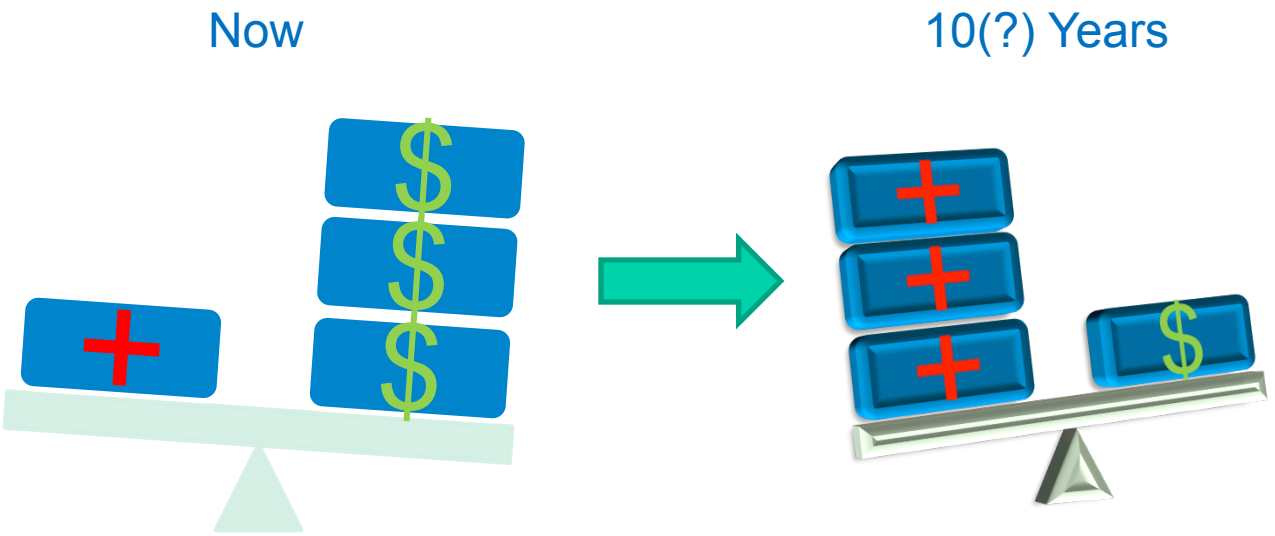


Annotations for rare noncoding vairants

	Subject Z	Subject S	Subject N
Promoter	3,043	2,473	3,038
Enhancer	15,007	14,186	15,190
Other noncoding annotation*	42,389	40,048	44,510

* TFP, DHS, ncRNA, pseudogene

Genomic technologies will find widespread clinical adoption when their clinical utility justifies their cost across disease domains



Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt