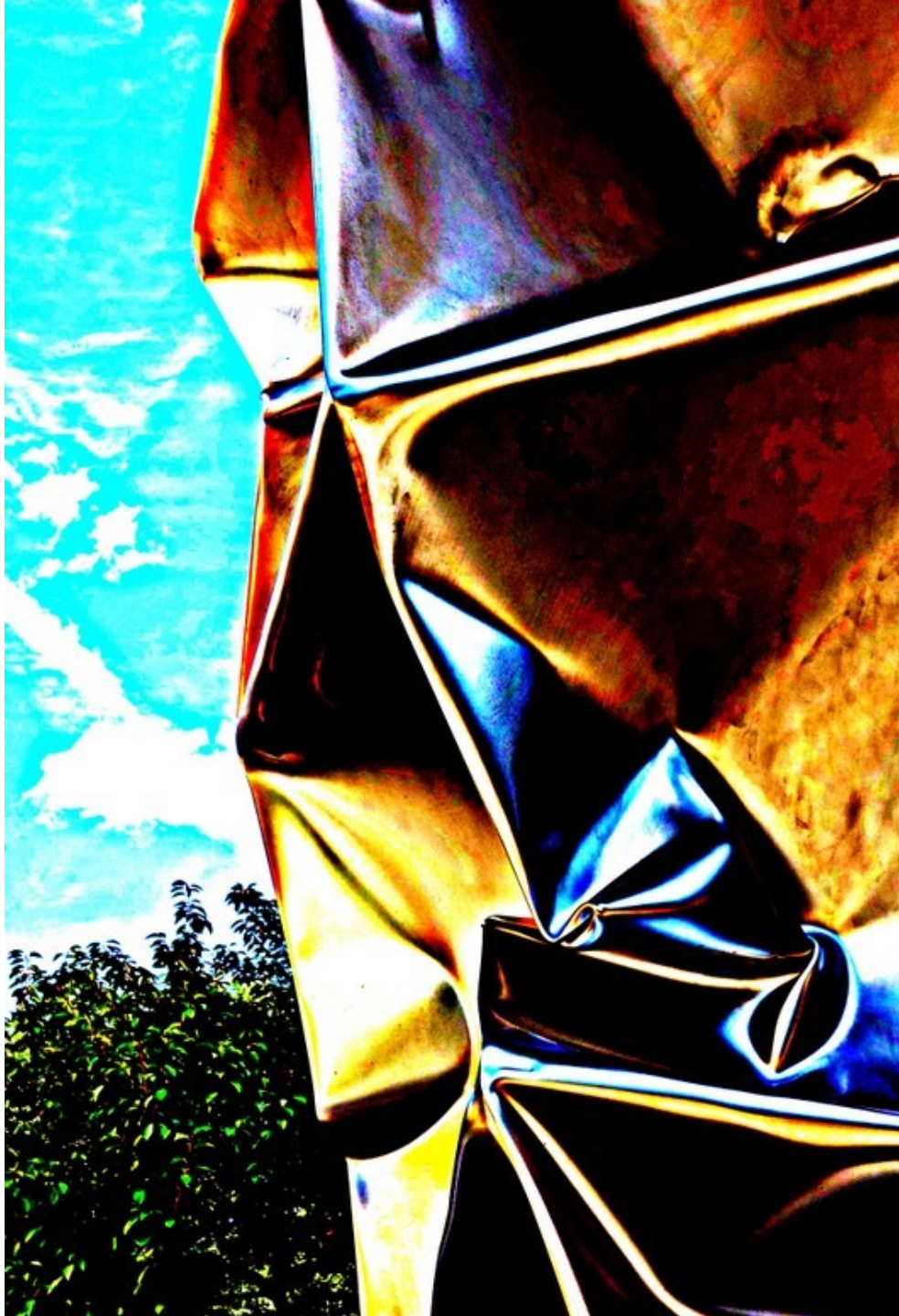
Personal Genomics:

Handling Exponential Data Scaling through Prioritizing High-impact Variants

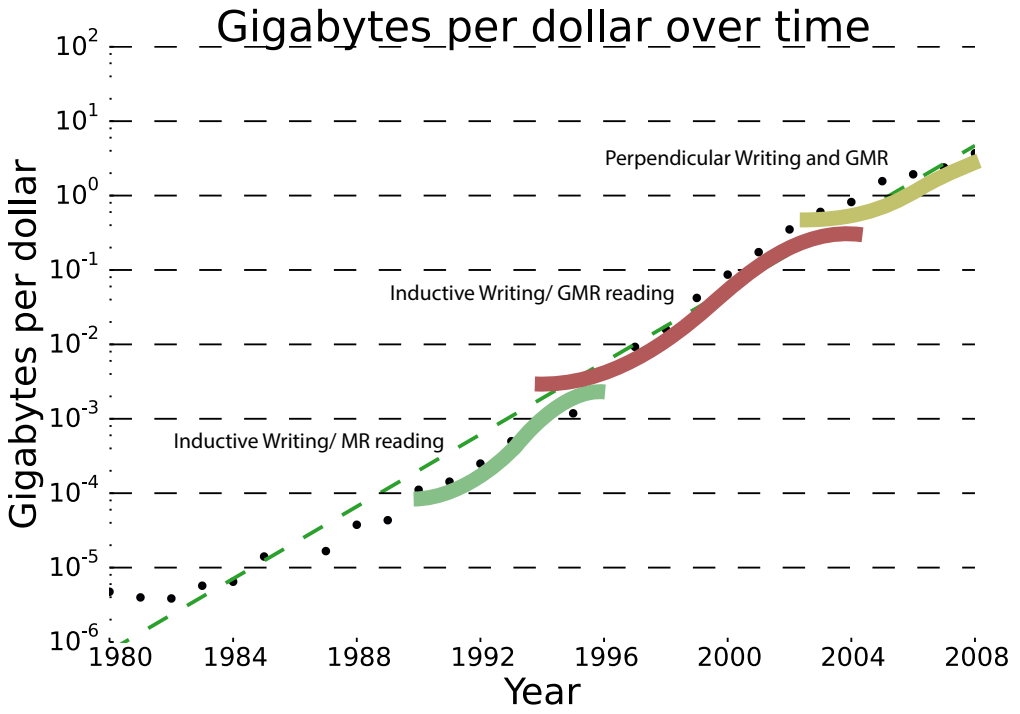
Mark Gerstein
Yale

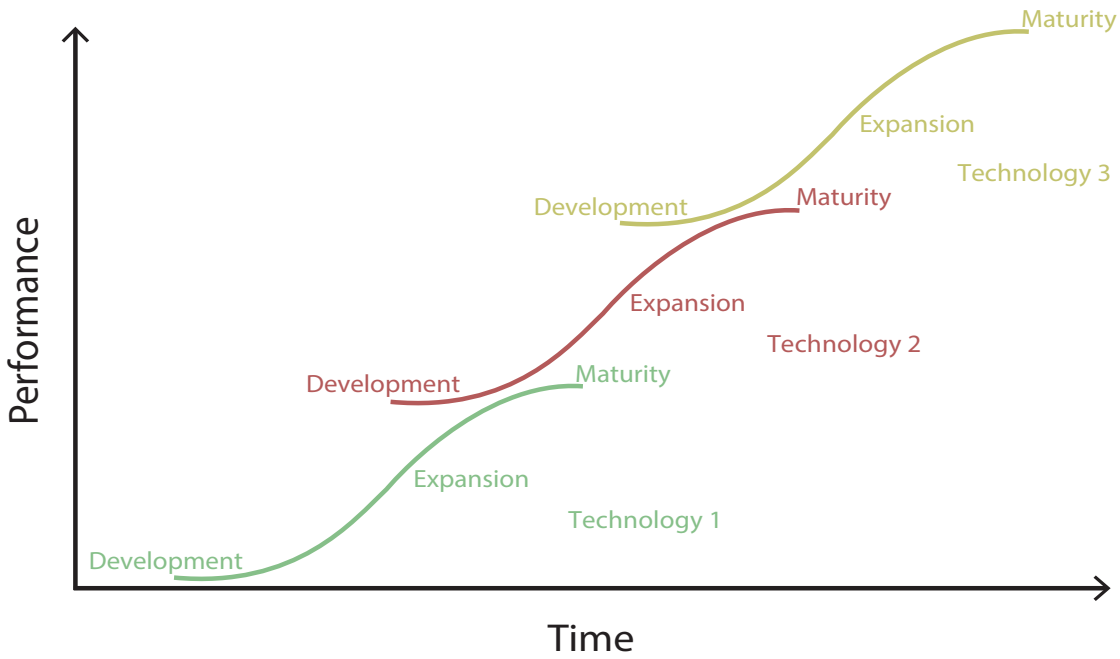
Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein).
See last slide for more info.



Moore's Law: Exponential Scaling of Computer Technology

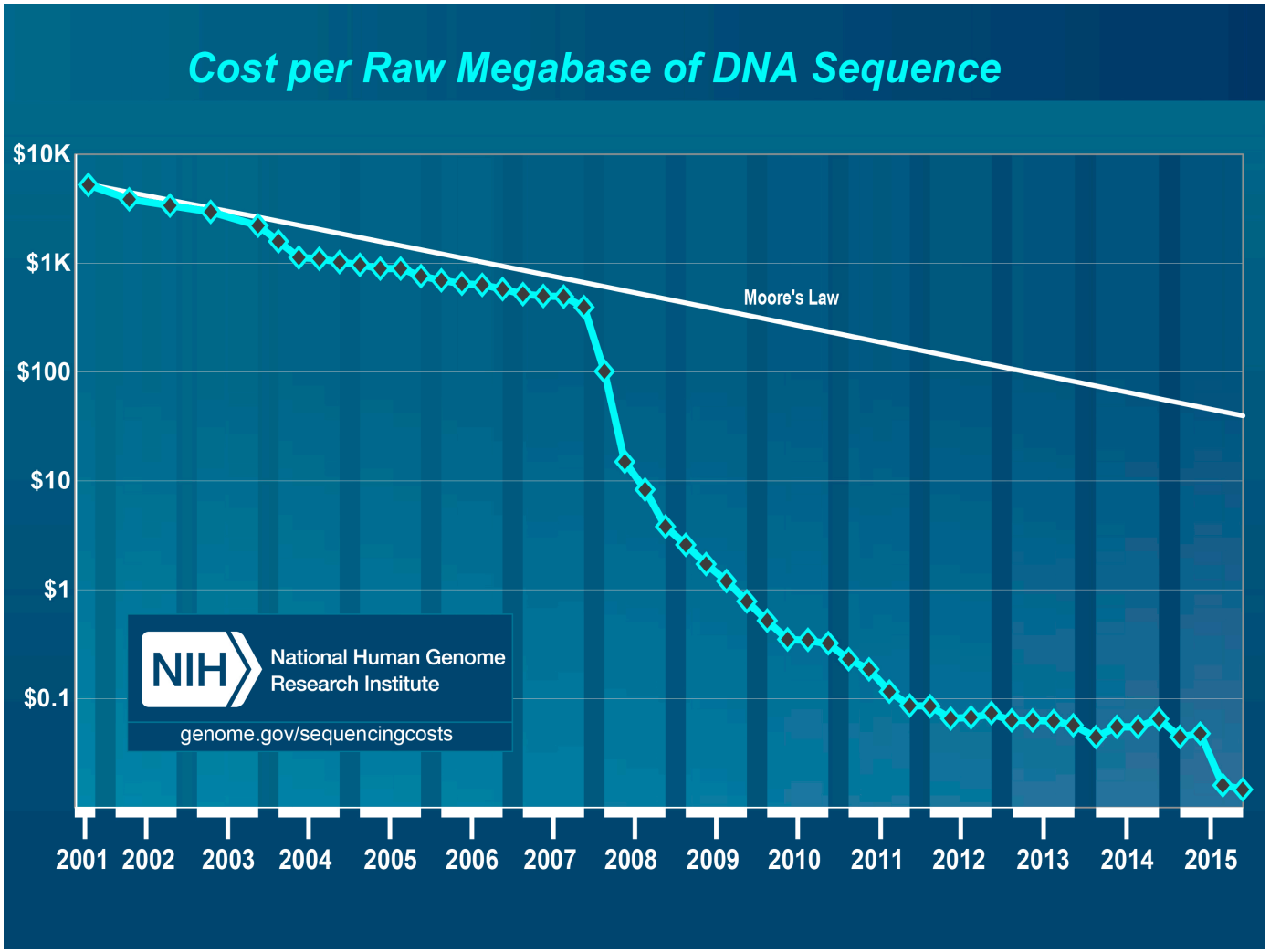
- Moore's & Kryder's Laws
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Moore's law is a superposition of s-curves for different technologies





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

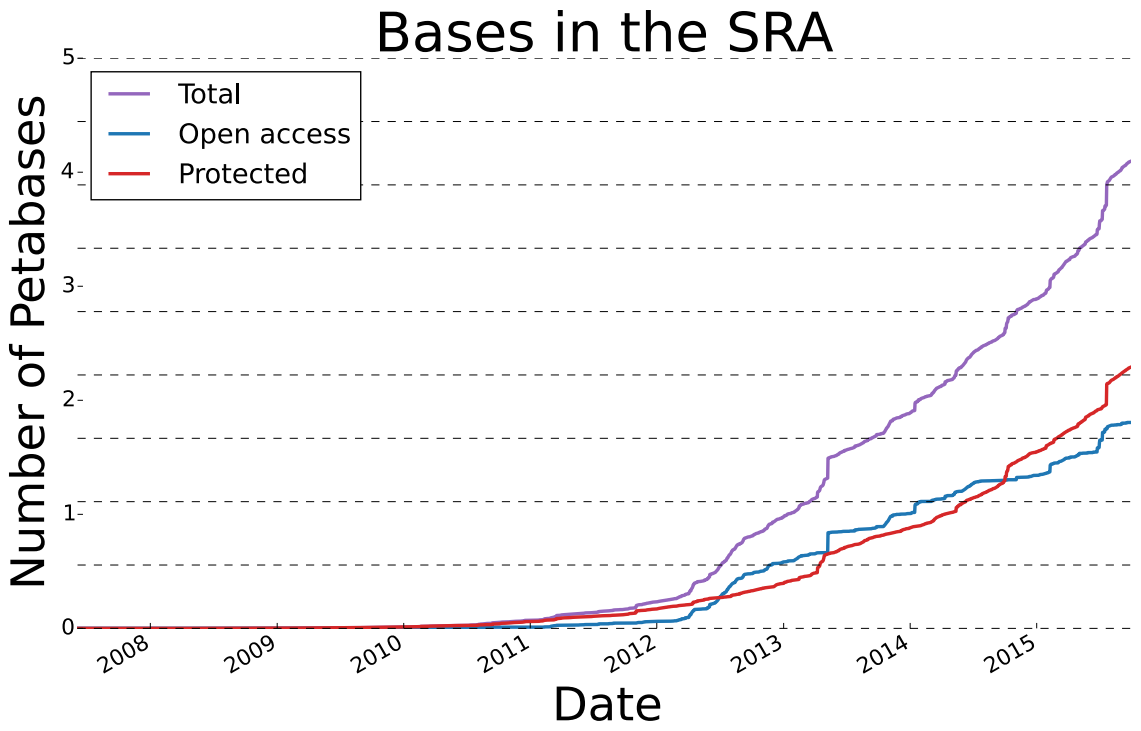
Sequencing Data Explosion: Faster than Moore's Law for a Time (or a S-curve)

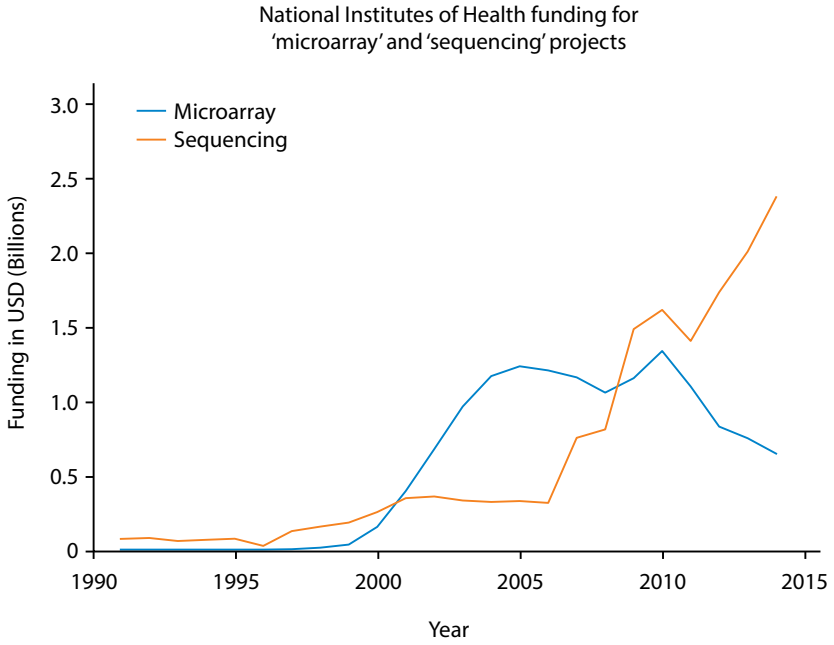


Sequencing cost reductions have resulted in an explosion of data

- 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.





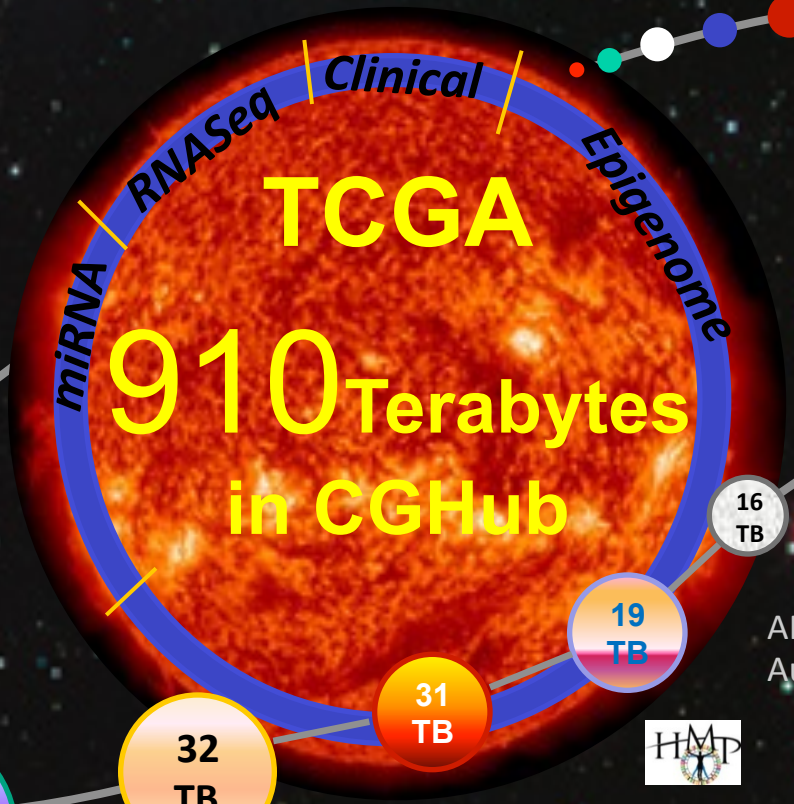
[Muir et al. ('15) GenomeBiol.]

Seq Universe

[from Heidi Sofia, NHGRI, '14]

SRA >1 petabyte

TCGA endpoint: ~2.5 Petabytes
- **~1.5 PB exome**
- **~1 PB whole genome**



220 TB

46 TB

32 TB

31 TB

19 TB

16 TB

9 TB

ADSP

NHLBI ESP

NHGRI LSSP

ARRA Autism

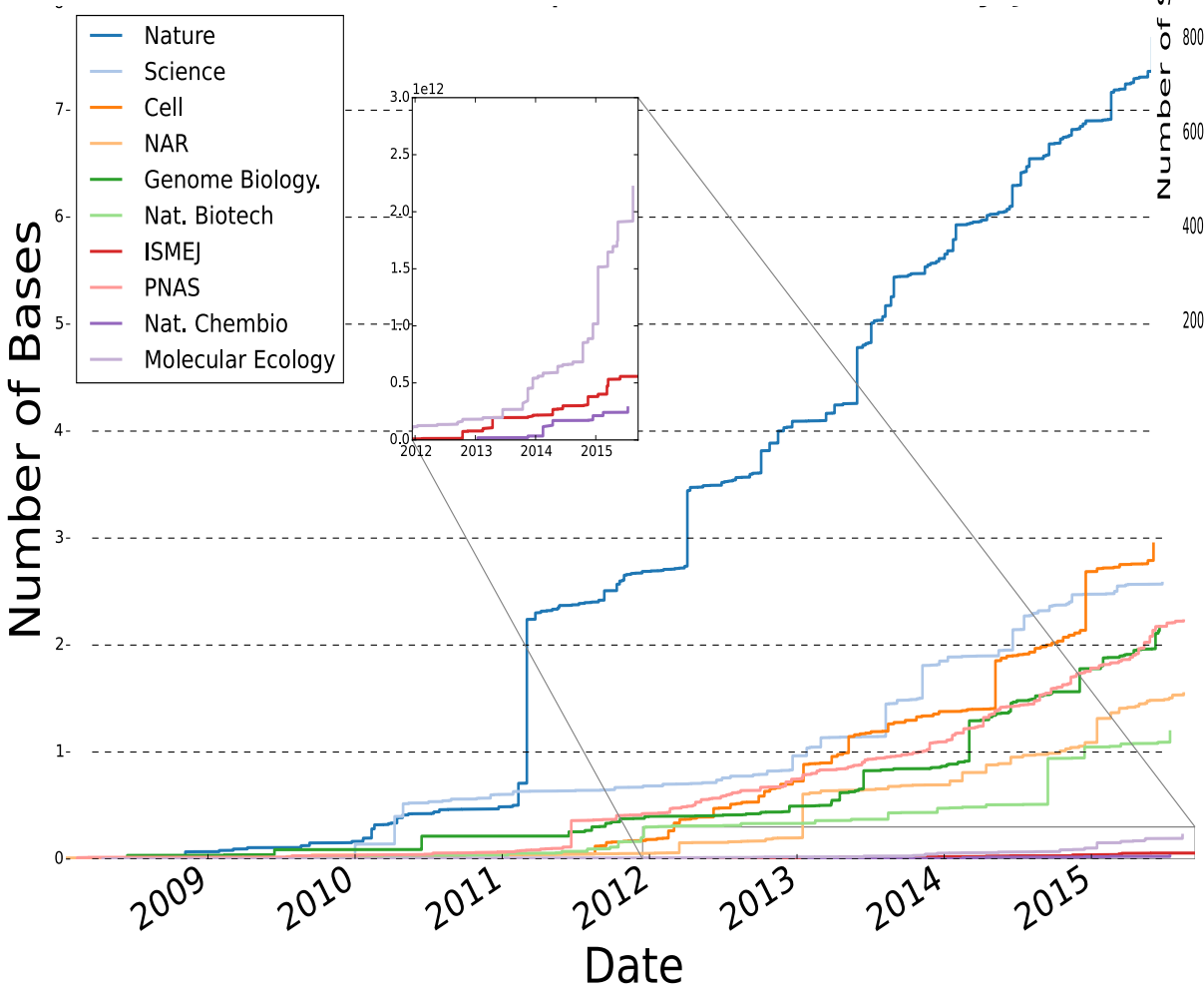
National Human Genome Research Institute

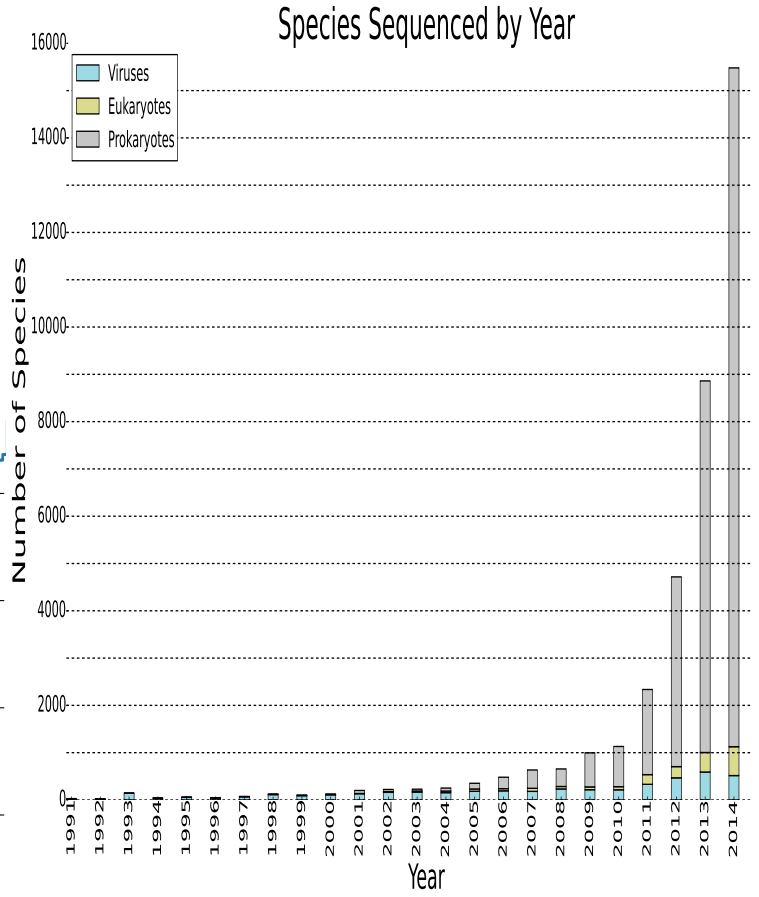
2016

TCGA	2300 TB
1000 Genomes	222 TB
ADSP	68 TB
NHGRI LSSP	40 TB
GTeX	34 TB
NHLBI ESP	32 TB
HMP	29 TB
ARRA Autism	24 TB
ENCODE	9 TB

Sofia, 2-28-14

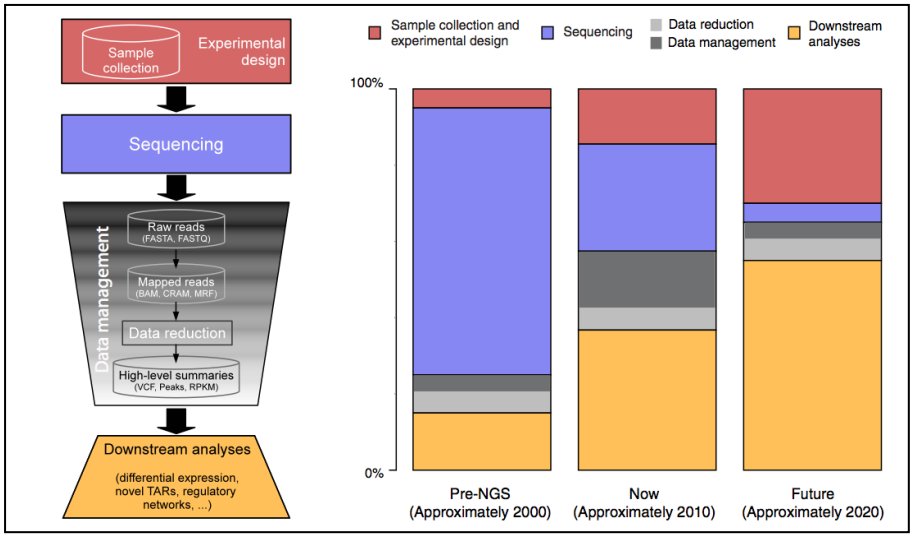
Increasing diversity in sequence data sources

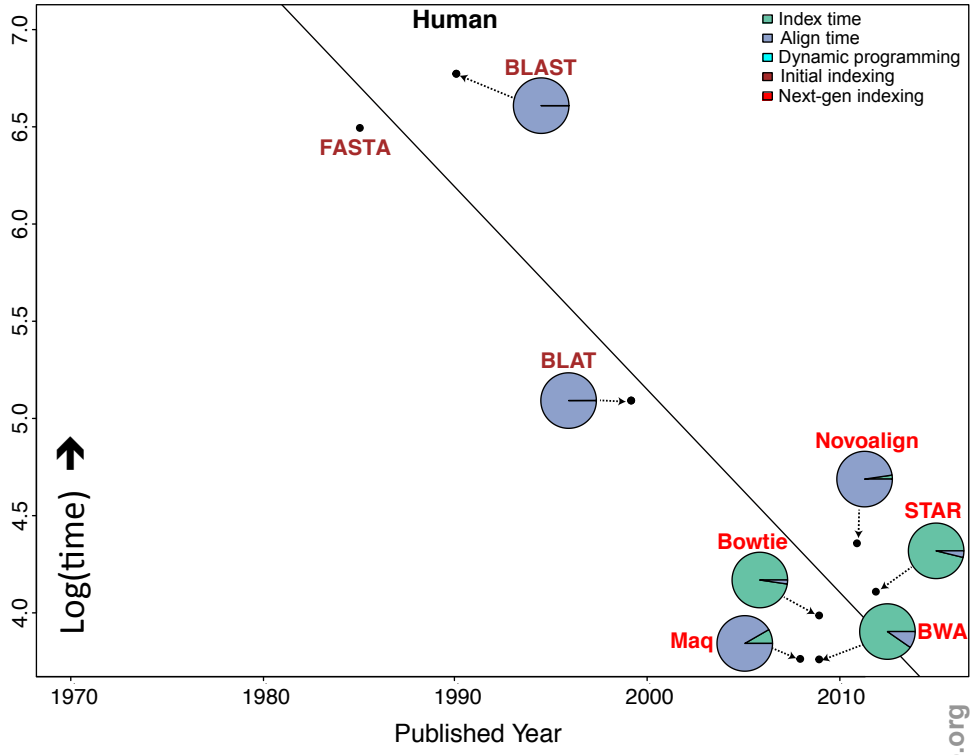




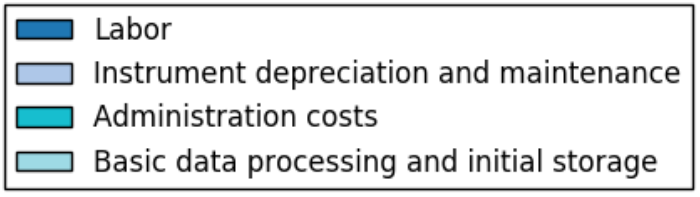
[Muir et al. ('15) GenomeBiol.]

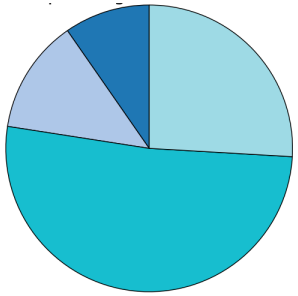
The changing costs of a sequencing pipeline





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

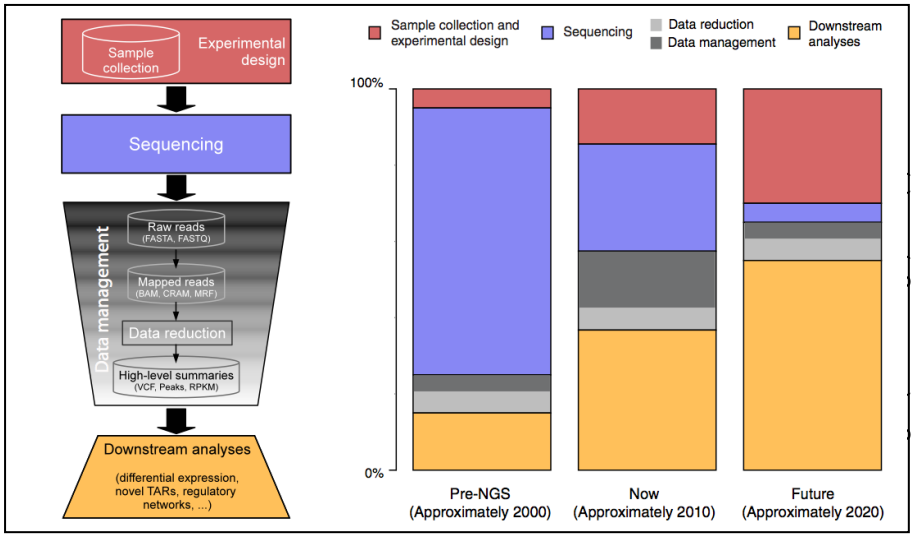


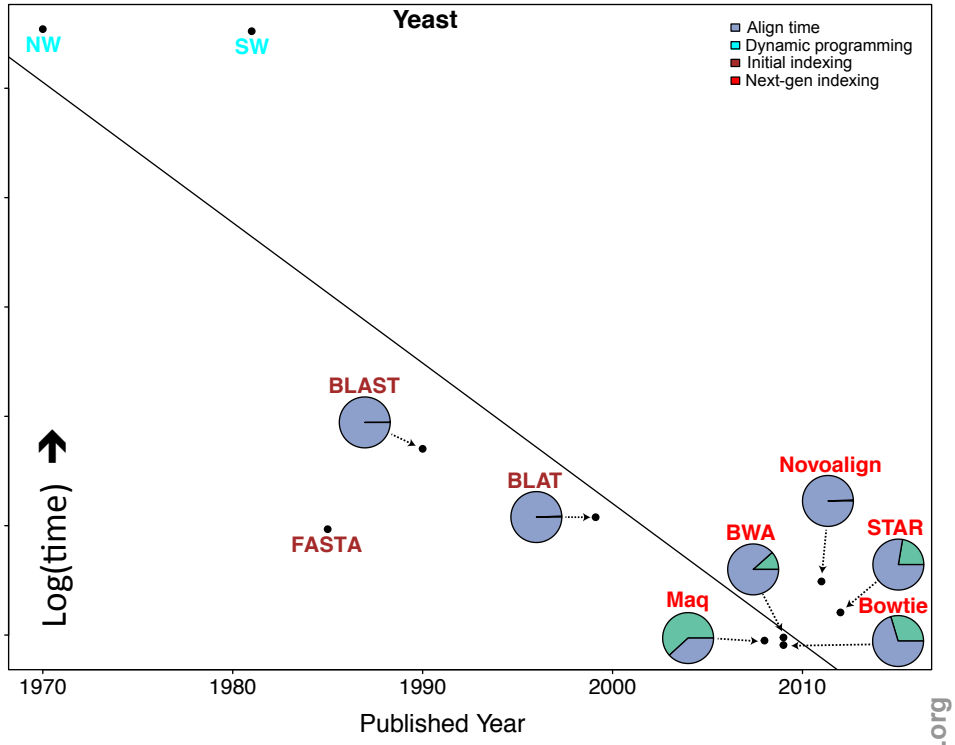


Alignment algorithms scaling to keep pace with data generation

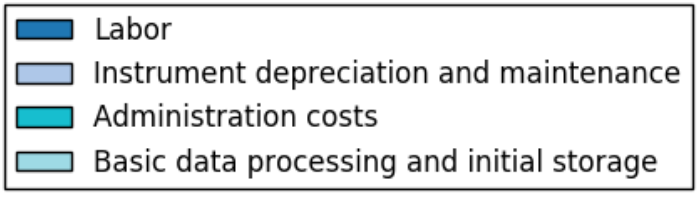
[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

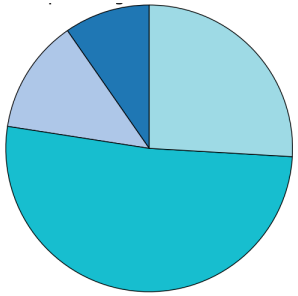
The changing costs of a sequencing pipeline





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

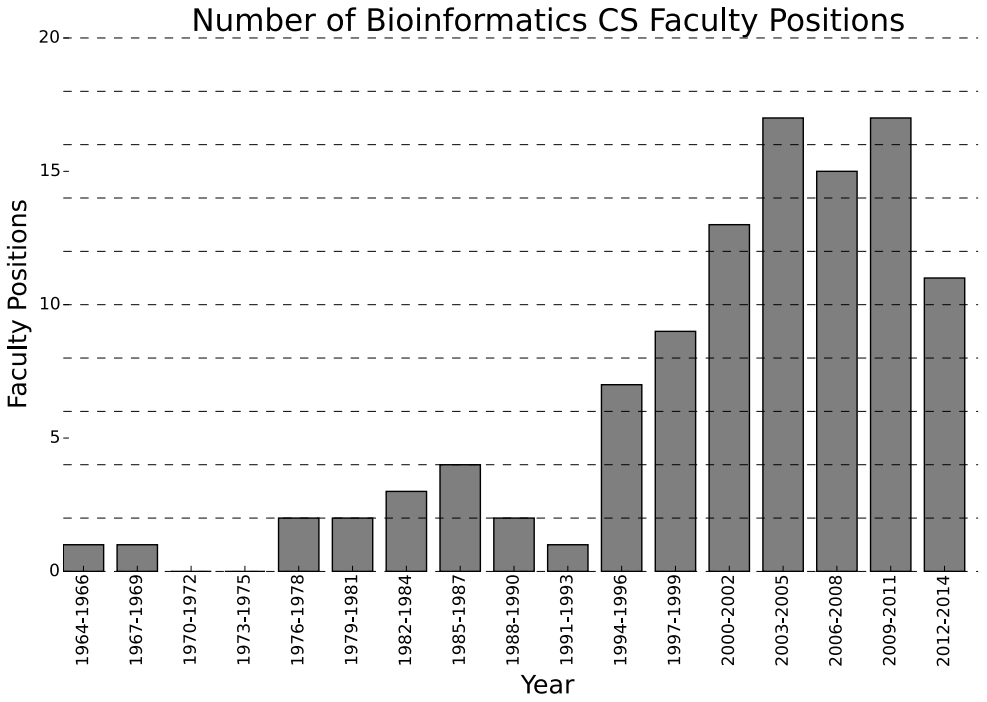




Alignment algorithms scaling to keep pace with data generation

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

Jobs: Bioinformatics is born!



Hires at 51 US universities in 3-year bins

http://jeffhuang.com/computer_science_professors.html

Personal Genomics: Handling Exponential Data Scaling through Prioritizing High-impact Variants

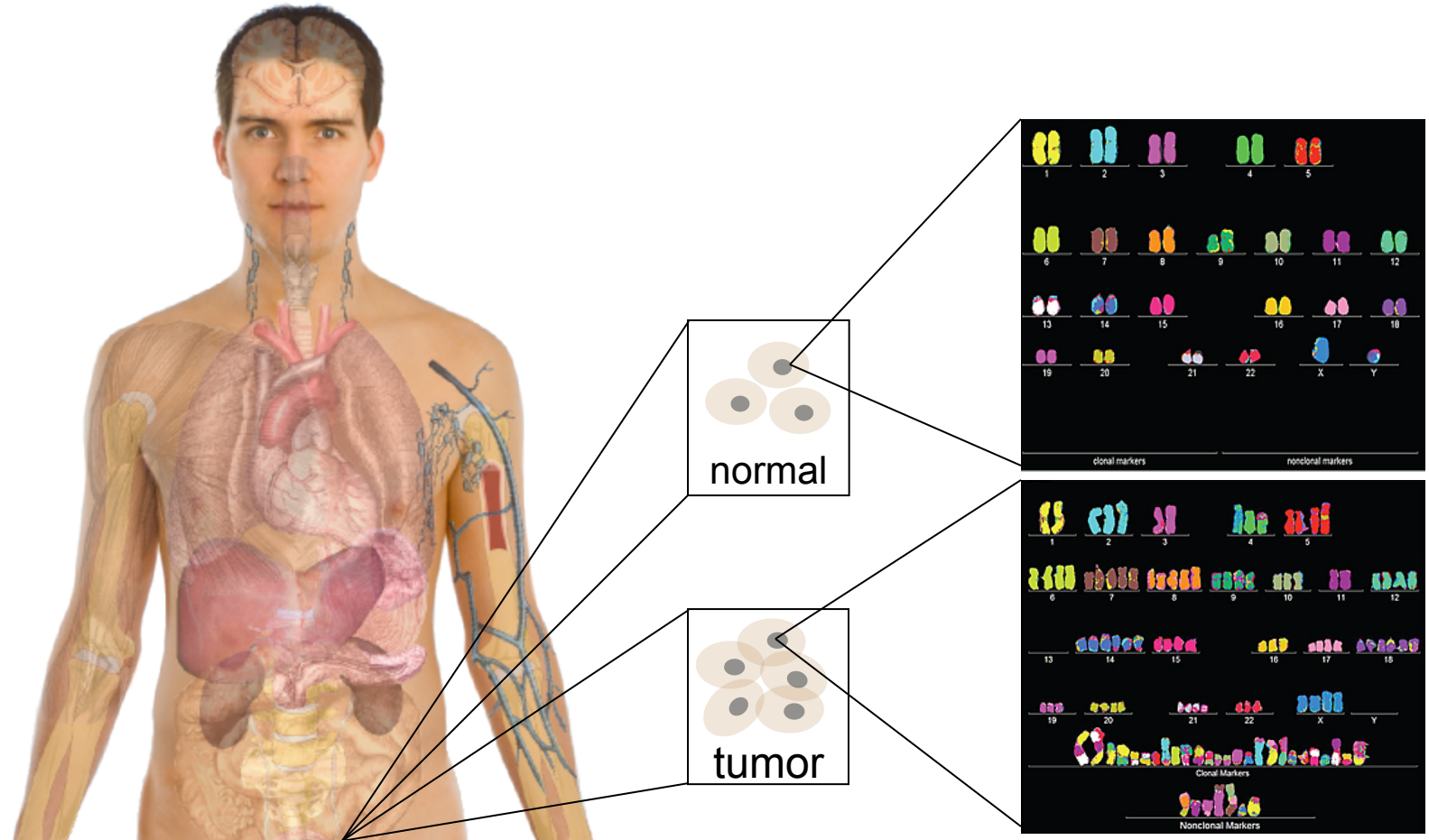
- **Introduction**
  - **The exponential scaling of data generation & data processing**
  - **The landscape of variants in personal genomes**
- **Characterizing Rare Variants in Coding Regions**
  - **Identifying with STRESS cryptic allosteric sites**
  - On surface & in interior bottlenecks
- **Non-coding Variants : Prioritizing using AlleleDB in terms of allelic elements**
  - Having observed difference in molecular activity in many contexts
- **Putting it together in workflows: Integrating evidence on non-coding variants with FunSeq**
  - Systematically weighting all the features
  - suggesting non-coding drivers
  - Prioritzing rare germline variants

Personal Genomics: Handling Exponential Data Scaling through Prioritizing High-impact Variants

- **Introduction**
  - The **exponential scaling** of data generation & data processing
  - The **landscape of variants** in personal genomes
- **Characterizing Rare Variants in Coding Regions**
  - **Identifying with STRESS cryptic allosteric sites**
  - On surface & in interior bottlenecks
- **Non-coding Variants : Prioritizing using AlleleDB in terms of allelic elements**
  - Having observed difference in molecular activity in many contexts
- **Putting it together in workflows: Integrating evidence on non-coding variants with FunSeq**
  - Systematically weighting all the features
  - suggesting non-coding drivers
  - Prioritzing rare germline variants

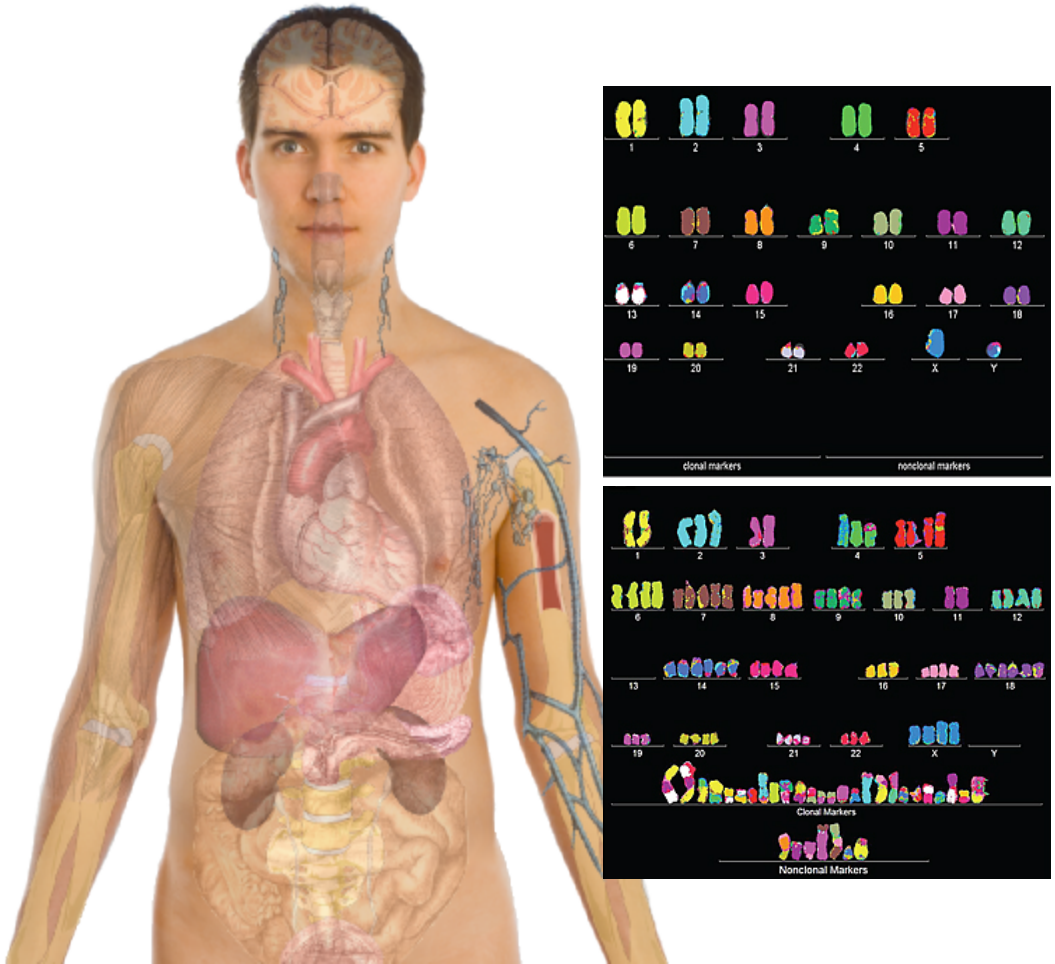
Personal Genomics as a Gateway into Biology

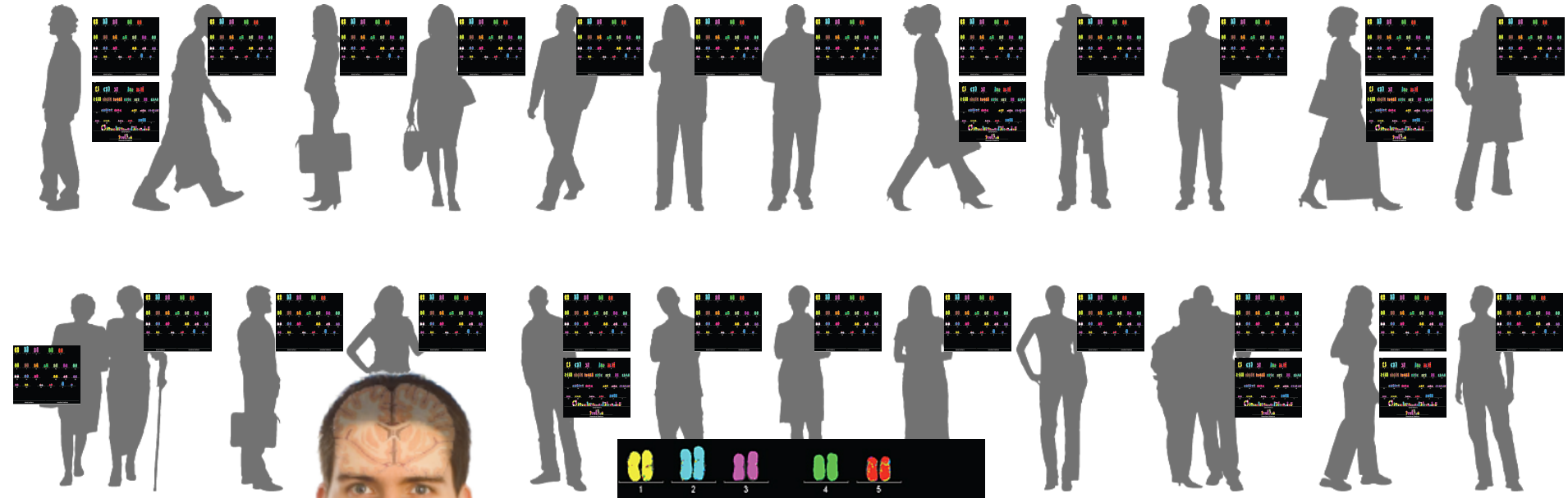
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

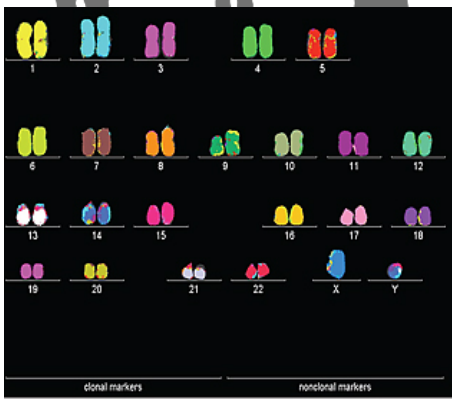


Personal Genomics as a Gateway into Biology

Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

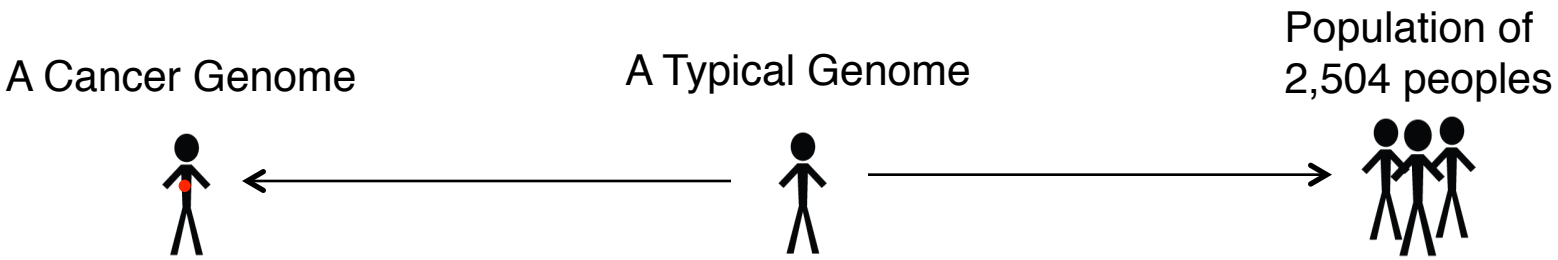





clonal markers
nonclonal markers

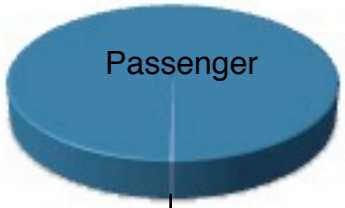

Clonal Markers
Nonclonal Markers

Human Genetic Variation



Origin of Variants

	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

Prevalence of Variants

SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M

* 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

Finding Key Variants

Germline

CAN YOU FIND THE PANDA?

- **Common variants**
  - Can be associated with phenotype (ie disease) via a Genome-wide Association Study (GWAS), which tests whether the frequency of alleles differs between cases & controls.
  - Usually their functional effect is weaker.
  - Many are non-coding
  - Issue of LD in identifying the actual causal variant.
- **Rare variants**
  - Associations are usually underpowered due to low frequencies.
  - They often have larger functional impact
  - Can be collapsed in the same element to gain statistical power (burden tests).
  - In some cases, causal variants can be identified through tracing inheritance of Mendelian subtypes of diseases in large families.

McCarthy, M. et al. Nat. Rev. Genet. 2008. 9, 356-369, Zuk, O. et al. PNSA. 2014. Vol. 11, no. 4, MacArthur DG et al. Nature 2014. 508:469-476

Finding Key Variants

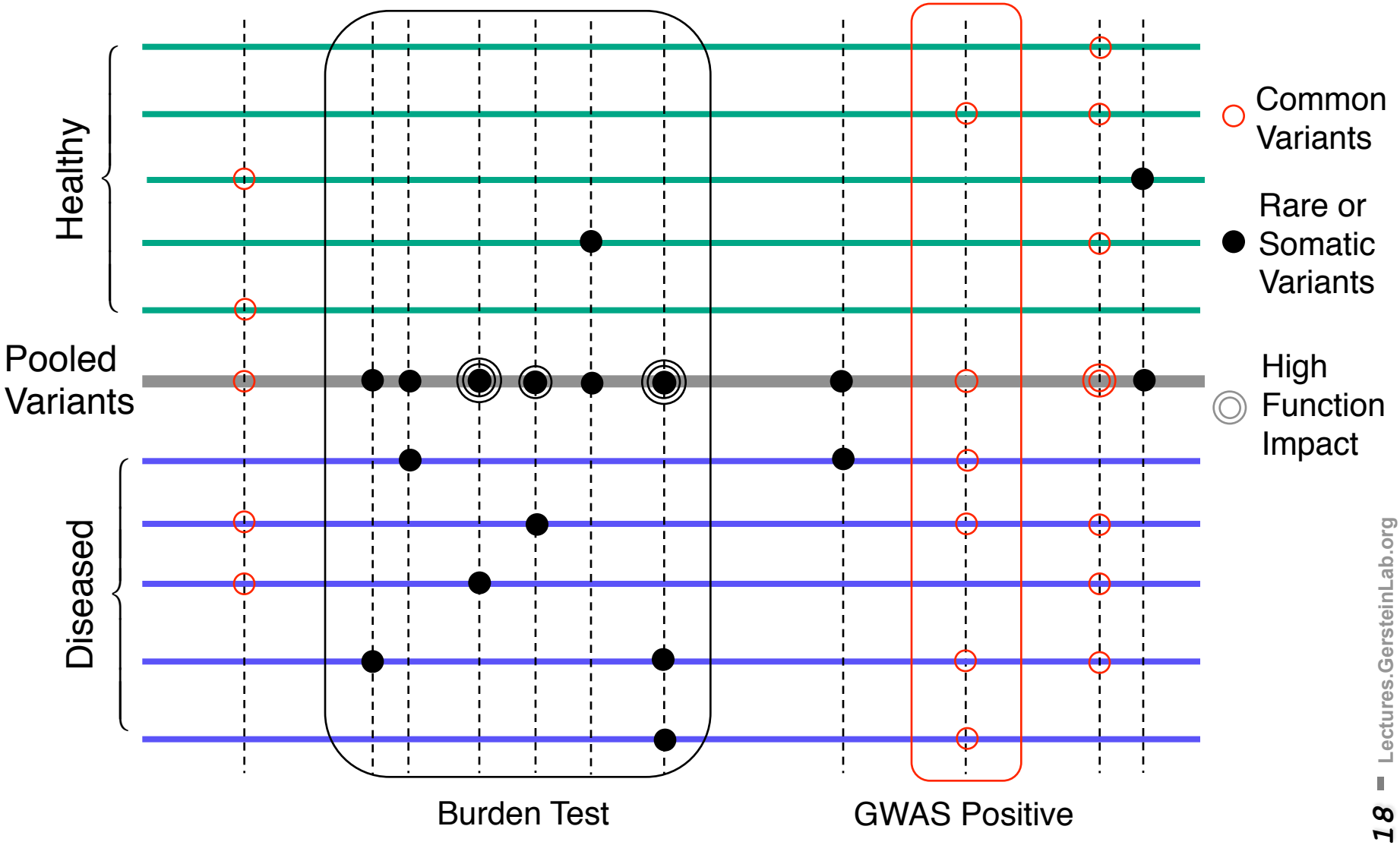
Somatic

CAN YOU FIND THE PANDA?

- **Overall**
  - Often these can be conceptualized as very rare variants
  - A challenge to identify somatic mutations contributing to cancer is to find driver mutations & distinguish them from passengers.
- **Drivers**
  - Driver mutation is a mutation that directly or indirectly confers a selective growth advantage to the cell in which it occurs.
  - A typical tumor contains 2-8 drivers; the remaining mutations are passengers.
- **Passengers**
  - Conceptually, a passenger mutation has no direct or indirect effect on the selective growth advantage of the cell in which it occurred.

Vogelstein B. Science 2013. 339(6127):1546-1558

Association of Variants with Diseases

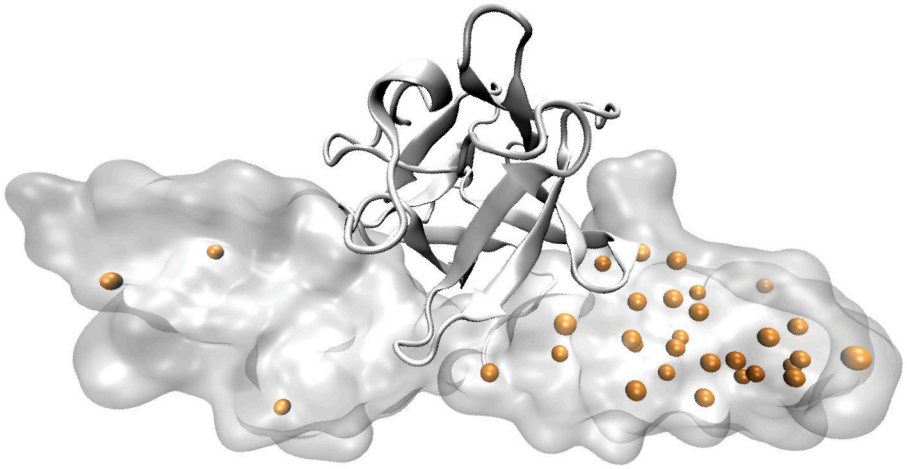


Personal Genomics: Handling Exponential Data Scaling through Prioritizing High-impact Variants

- Introduction
  - The exponential scaling of data generation & data processing
  - The landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
  - On surface & in interior bottlenecks
- Non-coding Variants : Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
- Putting it together in workflows: Integrating evidence on non-coding variants with FunSeq
  - Systematically weighting all the features
  - suggesting non-coding drivers
  - Prioritzing rare germline variants

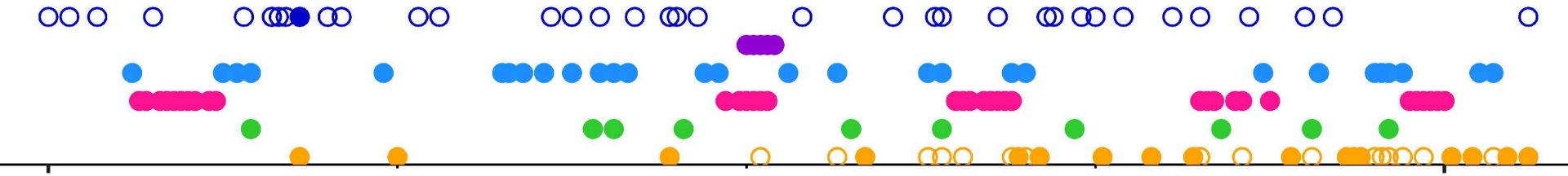
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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)

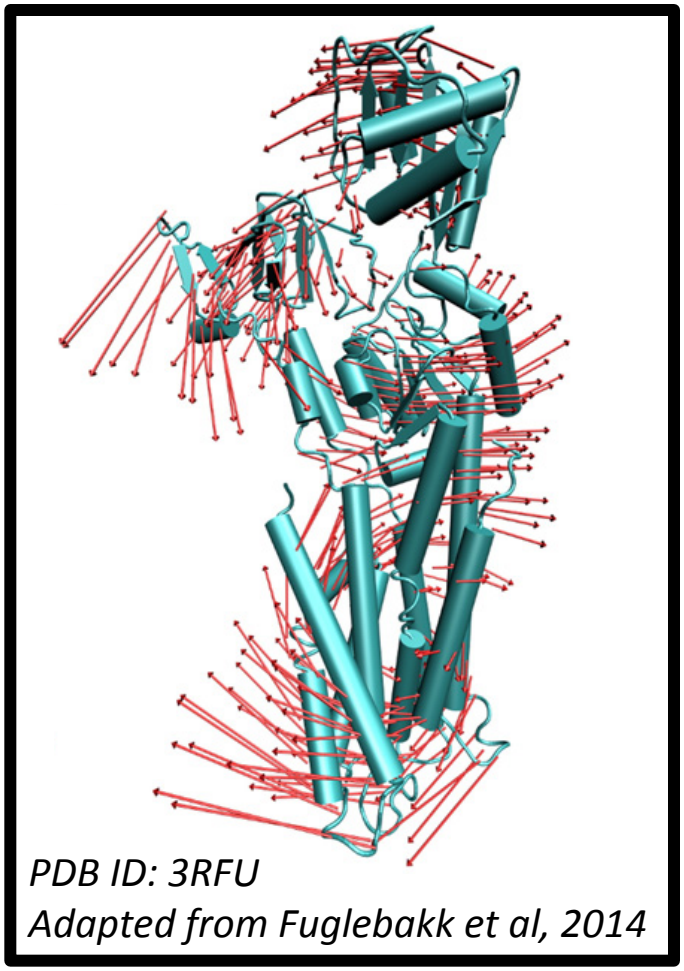
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



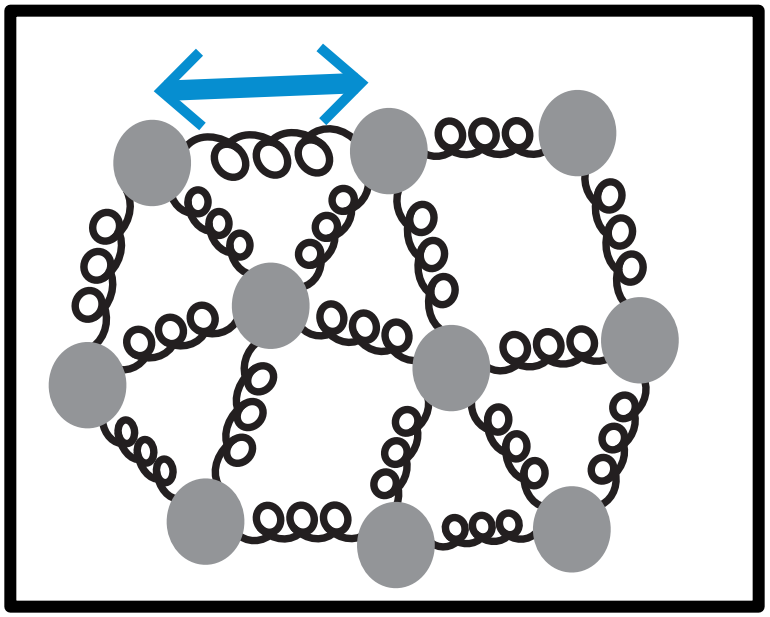
[Sethi et al. COSB ('15)]

Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)



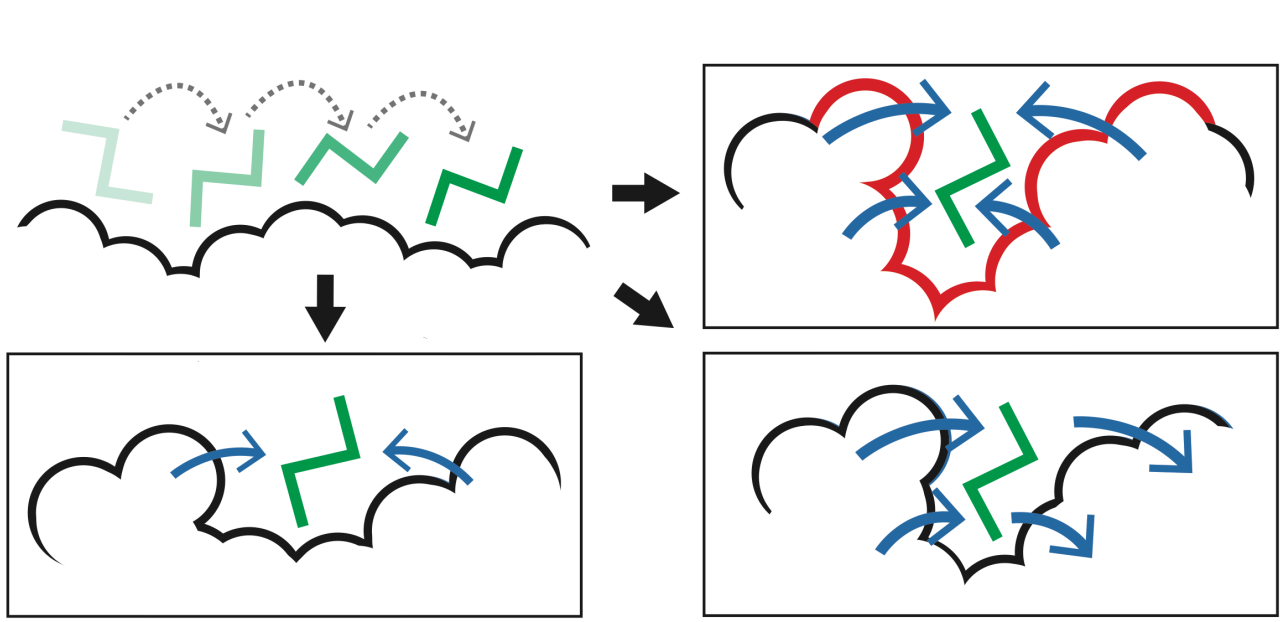
PDB ID: 3RFU
Adapted from Fuglebakk et al, 2014



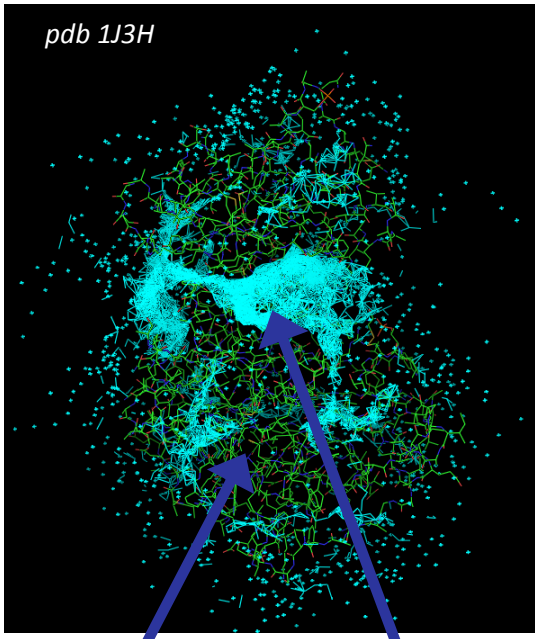
Characterizing uncharacterized variants
<= Finding Allosteric sites
<= Modeling motion

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

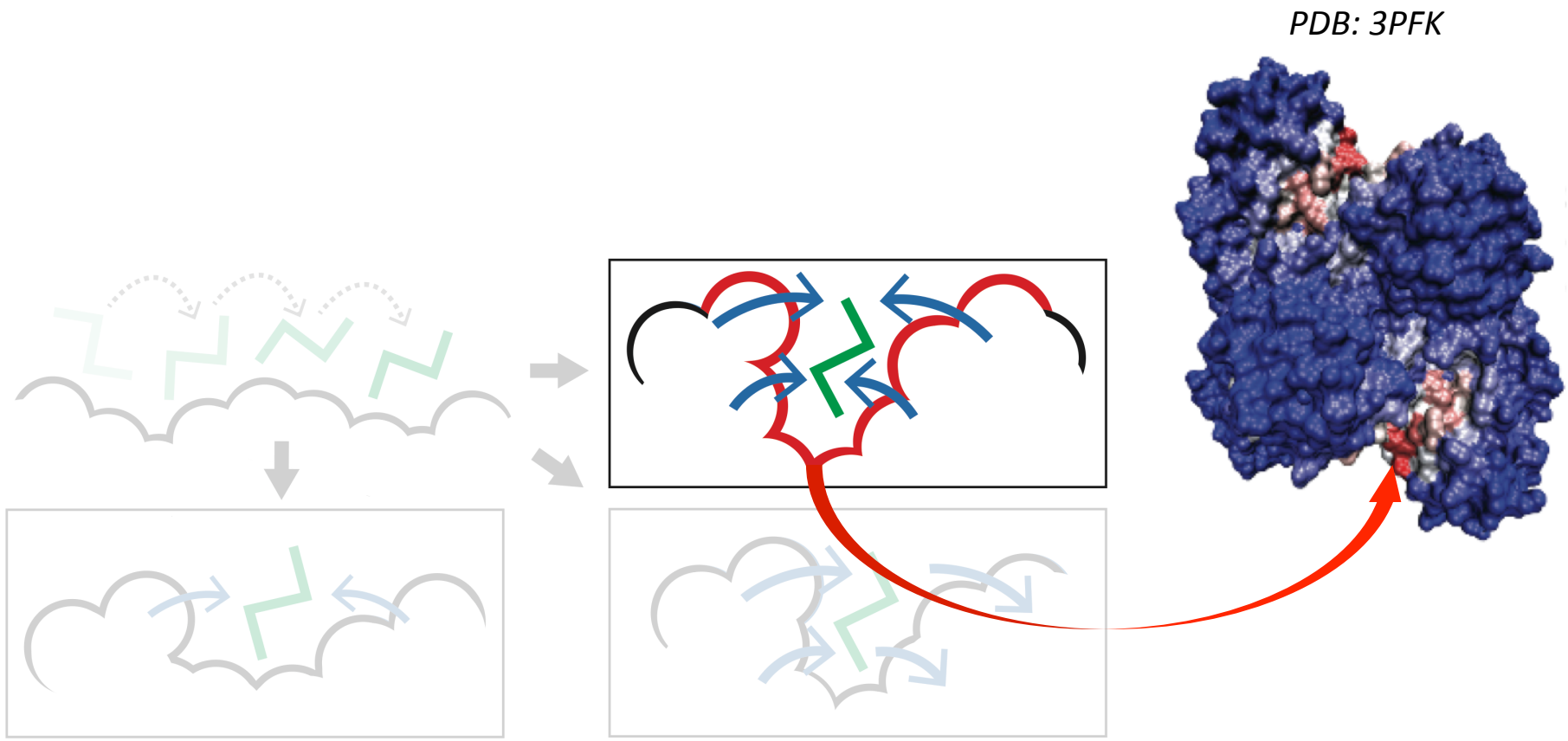


$$\textit{binding leverage} = \sum_{m=1}^{10}\left(\sum_{i}\sum_{j}\Delta d_{ij(m)}^{2}\right)$$



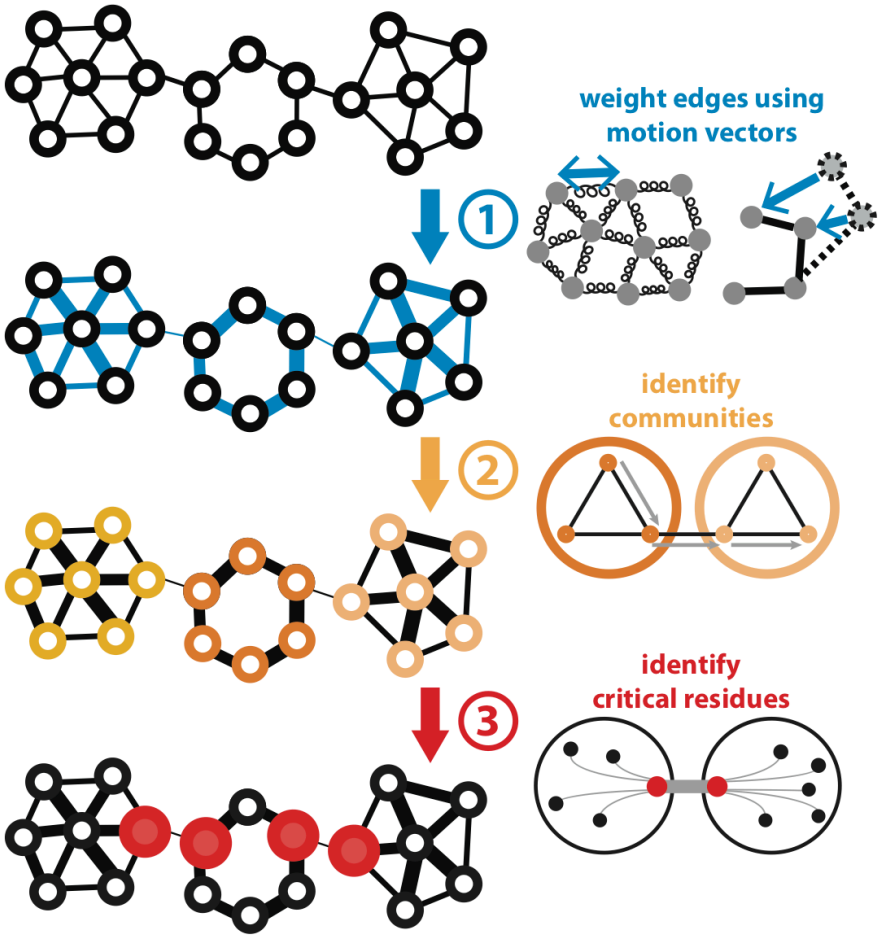
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surface



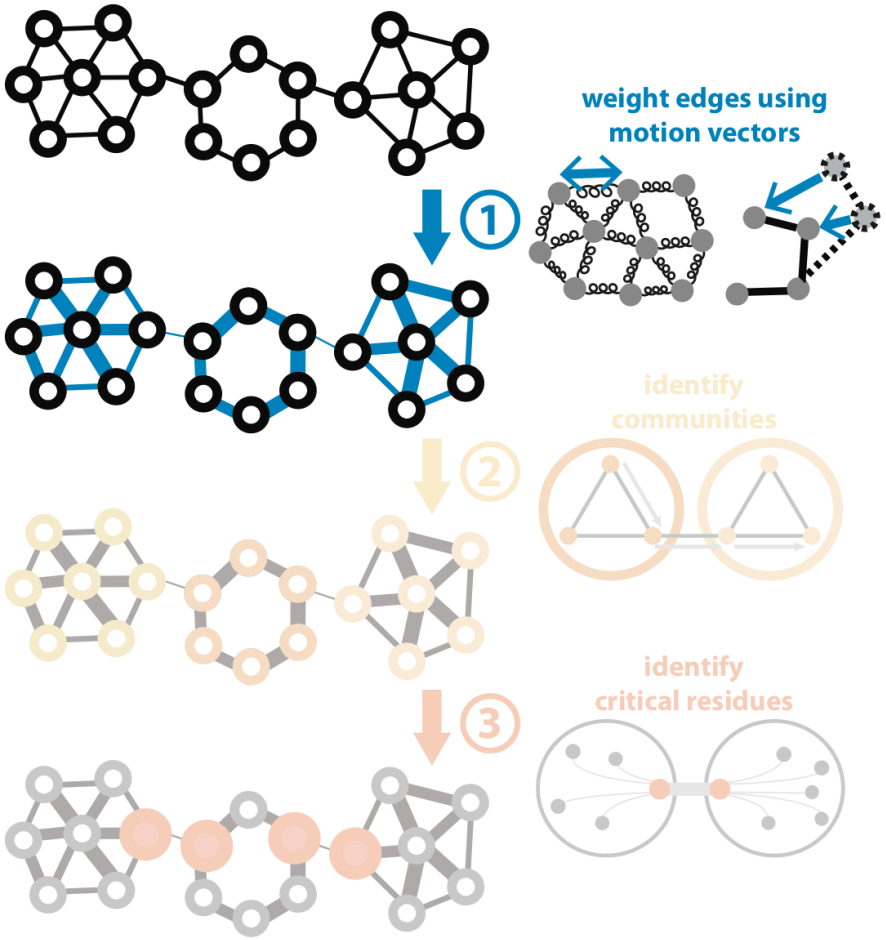
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues within the Interior



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

Predicting Allosterically-Important Residues within the Interior



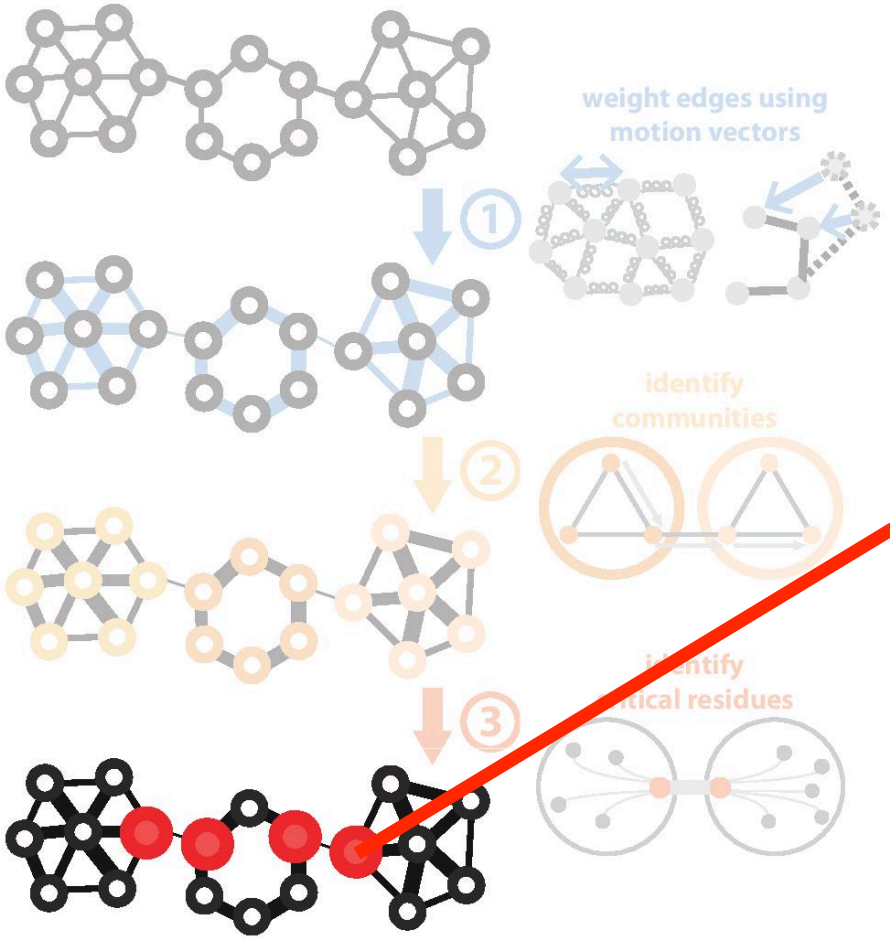
$$Cov_{ij} = \langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$$

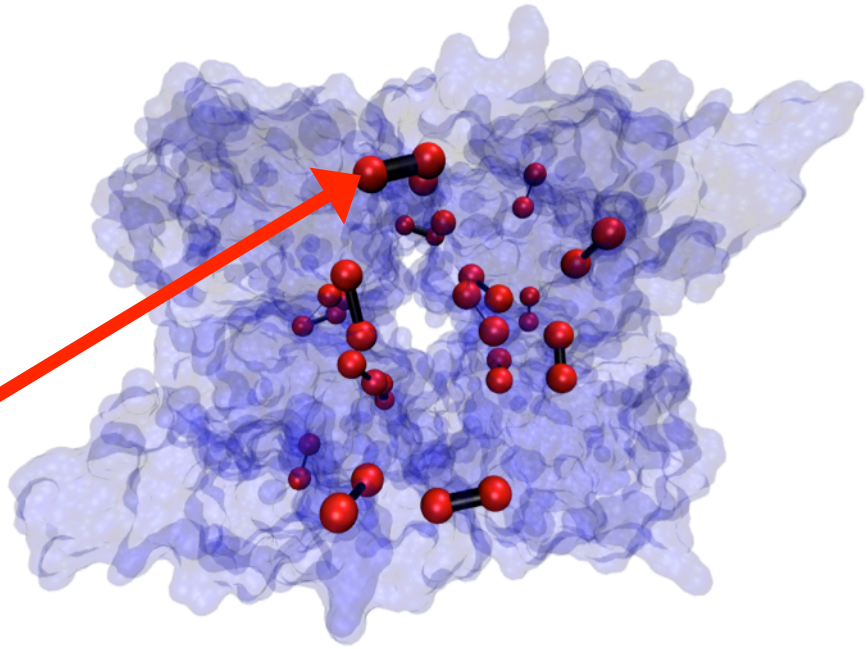
$$C_{ij} = Cov_{ij} \; / \; \sqrt{(\langle \mathbf{r}_i^2 \rangle \langle \mathbf{r}_j^2 \rangle)}$$

$$D_{ij} = -\log(|C_{ij}|)$$

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

Predicting Allosterically-Important Residues within the Interior



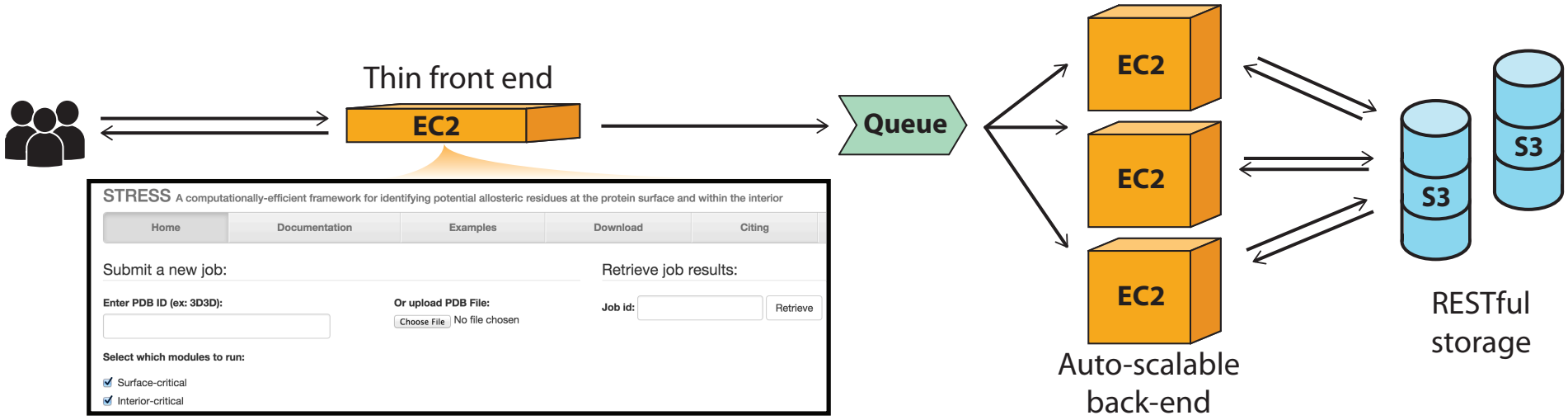


PDB: 1XTT

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

STRESS Server Architecture: Highlights

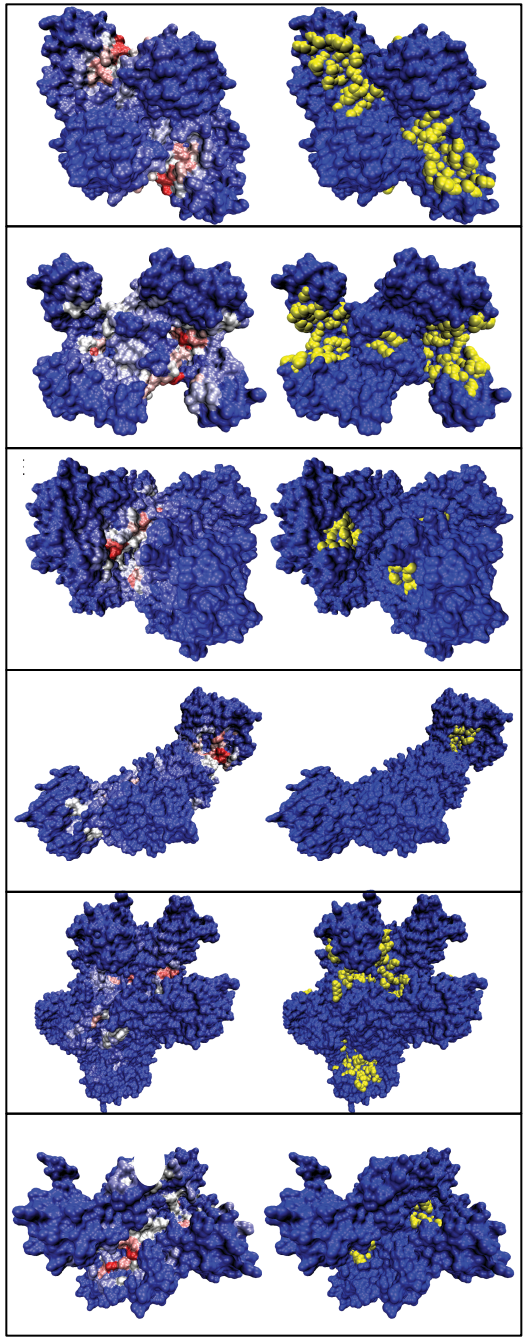
stress.molmovdb.org



- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are stored in S3 buckets.

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

Predicting Allosteric Sites at the Surface: Predicted & Known



PFK (100)

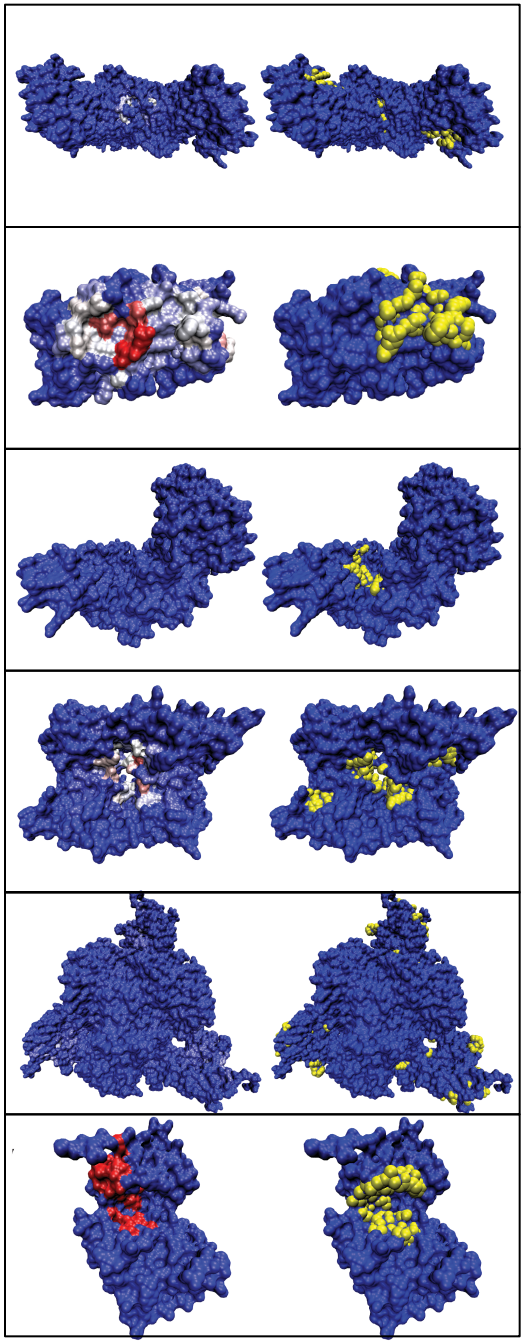
ADK (100)

G6P deamin. (50)

Trp Synth. (25)

Glu dehyd. (25)

Thr Synth. (67)



Malic enz. (0)

Tyr P-ase (100)

Arg. Kin. (0)

UPRTase (100)

ATCase (25)

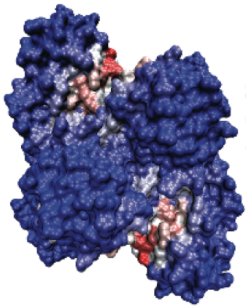
PKA (100)

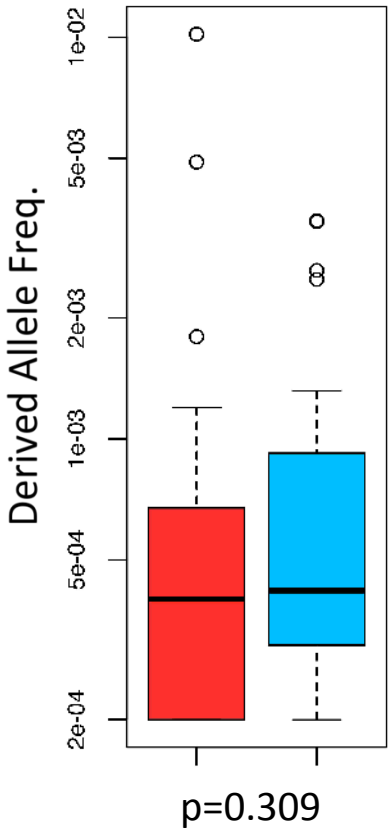
[Clarke et al. ('16) *Structure*]

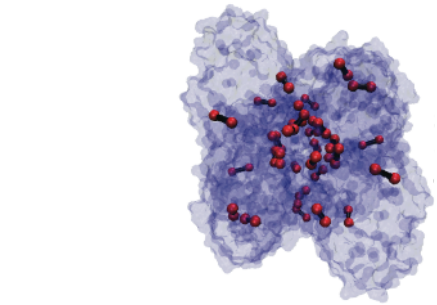
Intra-species conservation of predicted allosteric residues

1000 Genomes

Surface



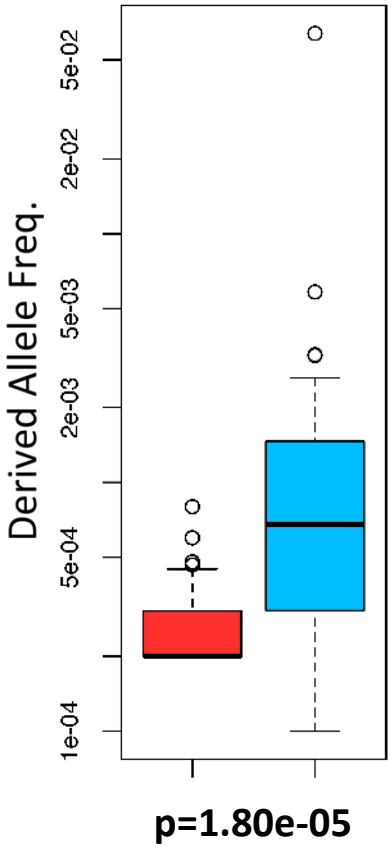




critical

non-critical

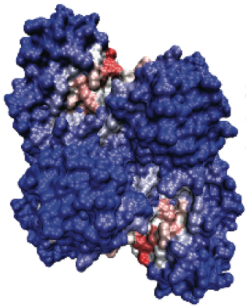
Interior



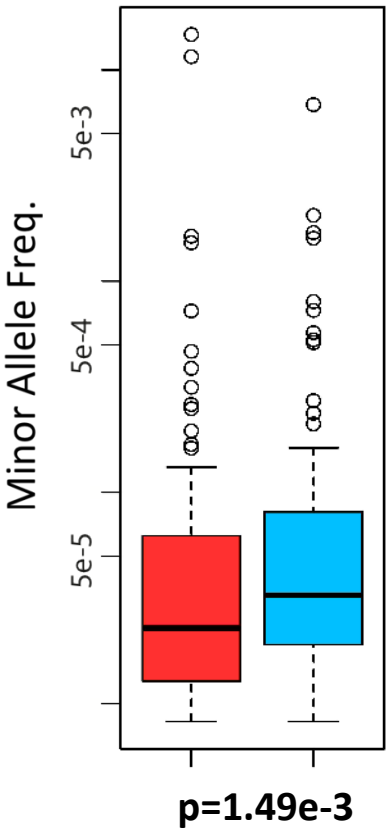
Adapted from Clarke*, Sethi*, et al *(in press)*

Intra-species conservation of predicted allosteric residues

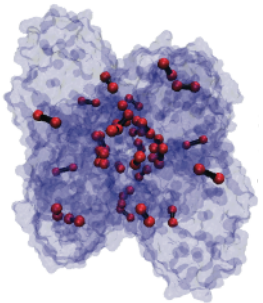
ExAC



Surface



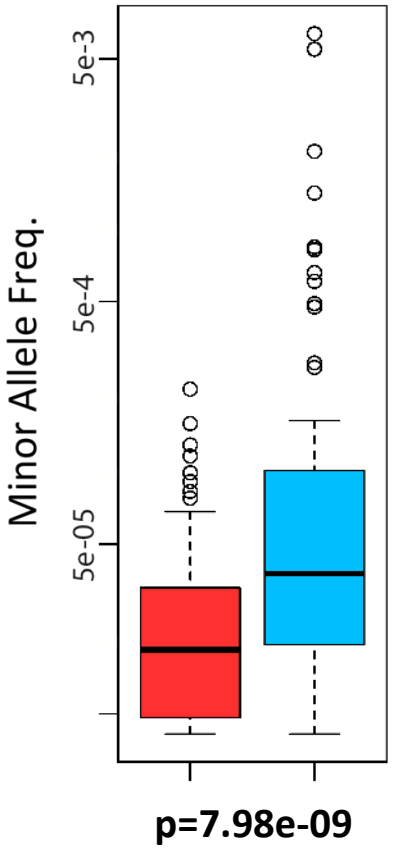
p=1.49e-3



critical

non-critical

Interior

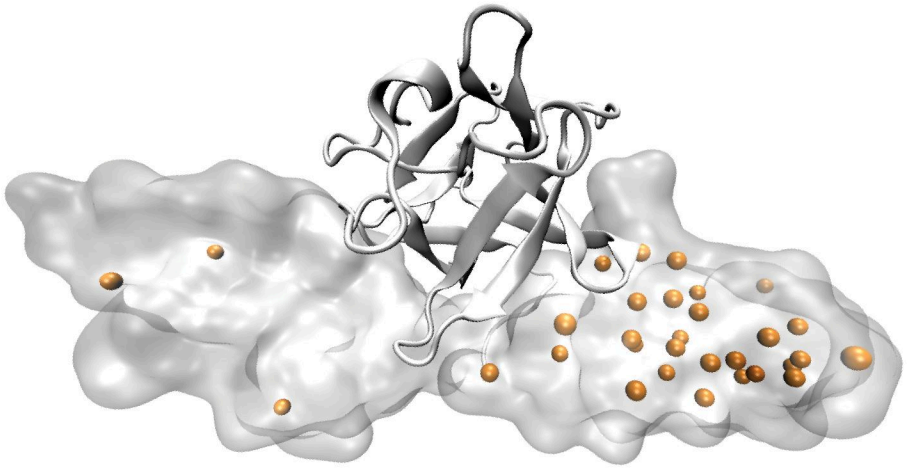


p=7.98e-09

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

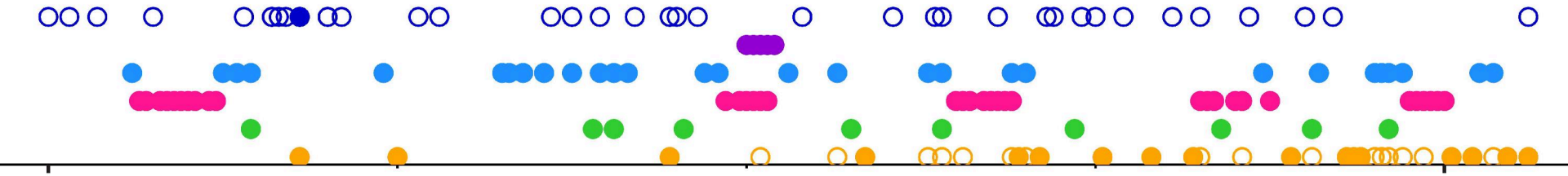
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited

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)

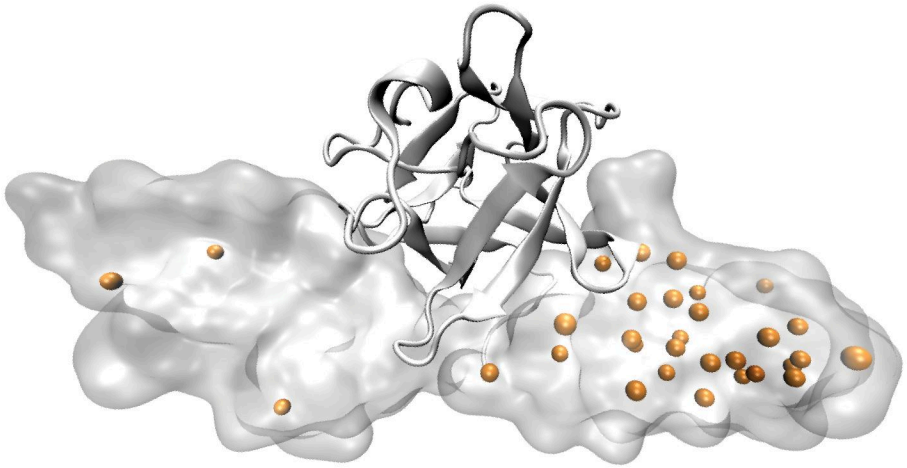
- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)



[Sethi et al. COSB ('15)]

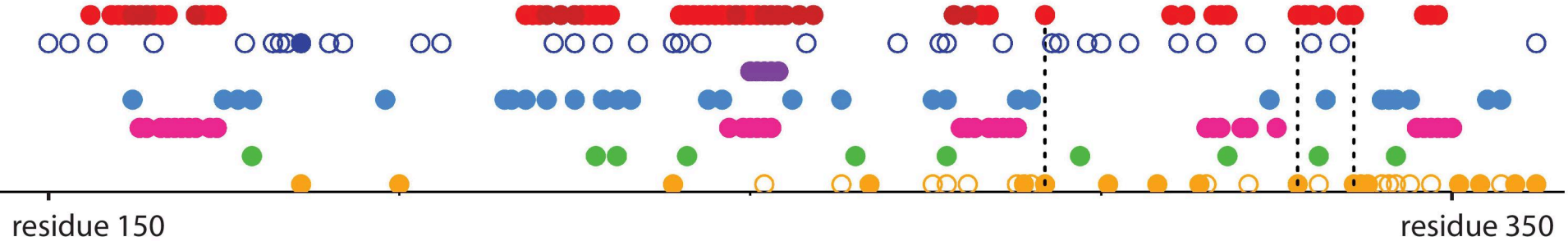
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated

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)

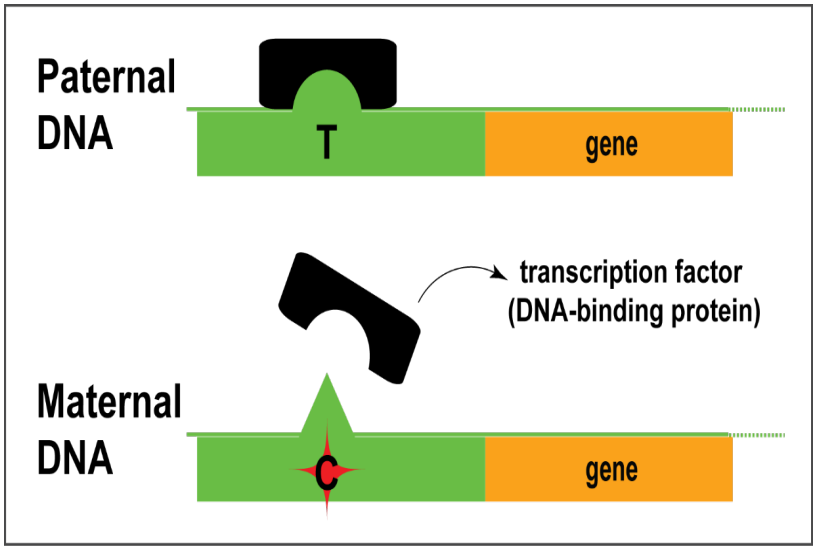


[Sethi et al. COSB ('15)]

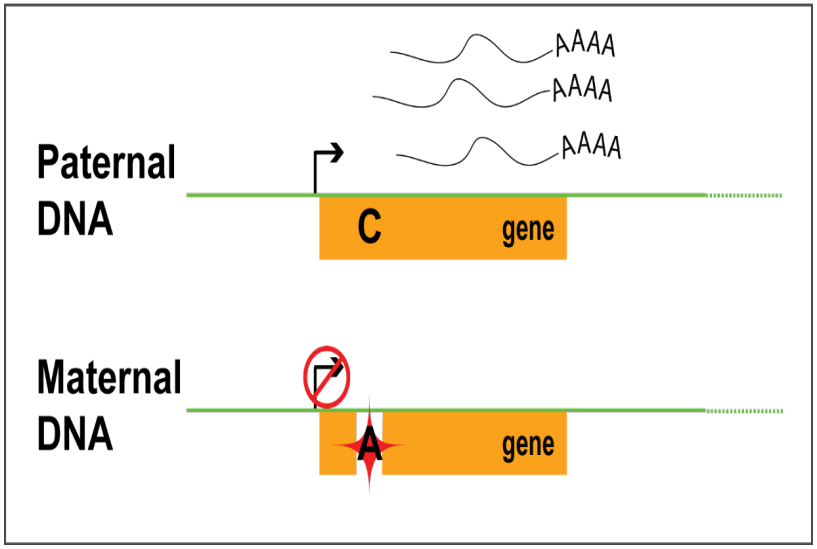
Personal Genomics: Handling Exponential Data Scaling through Prioritizing High-impact Variants

- Introduction
  - The exponential scaling of data generation & data processing
  - The landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
  - On surface & in interior bottlenecks
- Non-coding Variants : Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
- Putting it together in workflows: Integrating evidence on non-coding variants with FunSeq
  - Systematically weighting all the features
  - suggesting non-coding drivers
  - Prioritzing rare germline variants

Allele-specific binding and expression



Genomic variants affecting allele-specific behavior e.g. allele-specific binding (ASB)



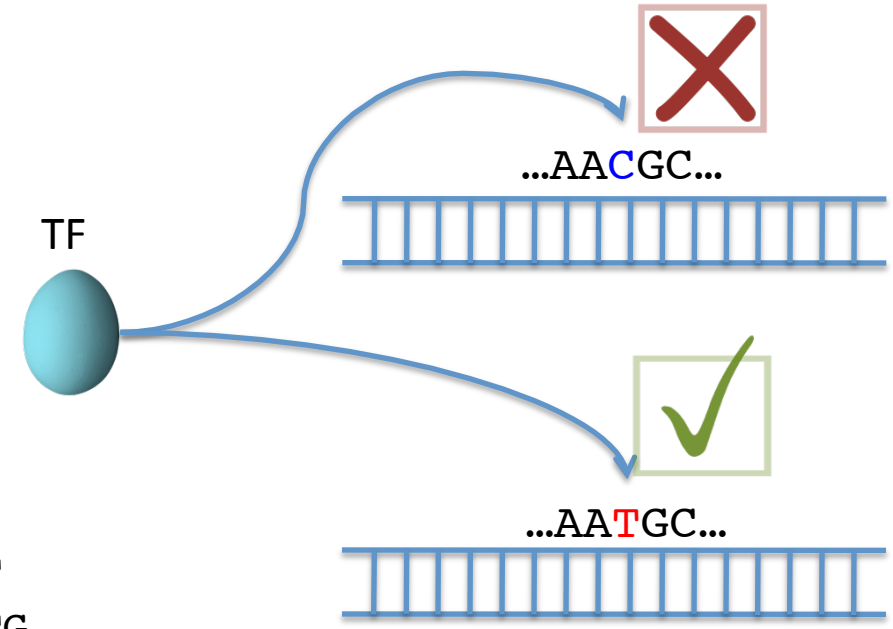
e.g. allele-specific expression (ASE)

Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
              AATGCACGTCGGGAGTTG
                 TGCACGTTGGGAGTTGGC
```

```
10 x T
 2 x C
```

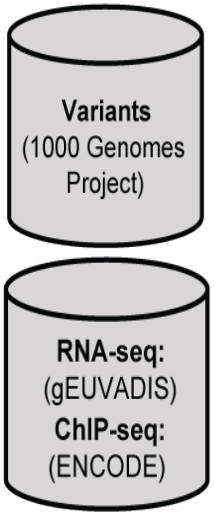


Haplotypes with a Heterozygous Polymorphism

Interplay of the annotation and individual sequence variants

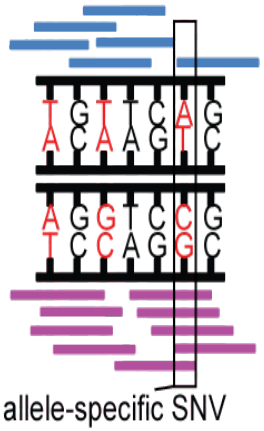
AlleleDB: Building 382 personal genomes to detect allele-specific variants on a large-scale

1. Build personal genomes



2. Align ChIP-seq & RNA-seq reads

3. Detect allele-specific variants via a series of filters and tests

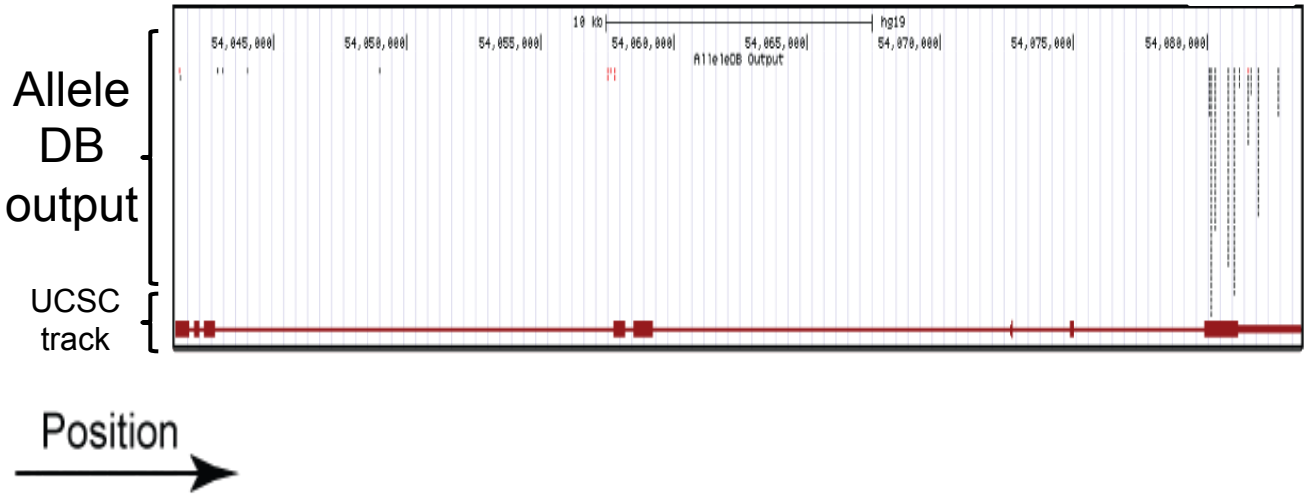


Many Technical Issues: Reference bias, Ambiguous mapping bias, Over-dispersed (non binomial null)

alleledb.gersteinlab.org

Chen J. *et al.* (*Nature Commun*, in press)

AlleleDB: Annotating rare & common allele-specific variants over a population

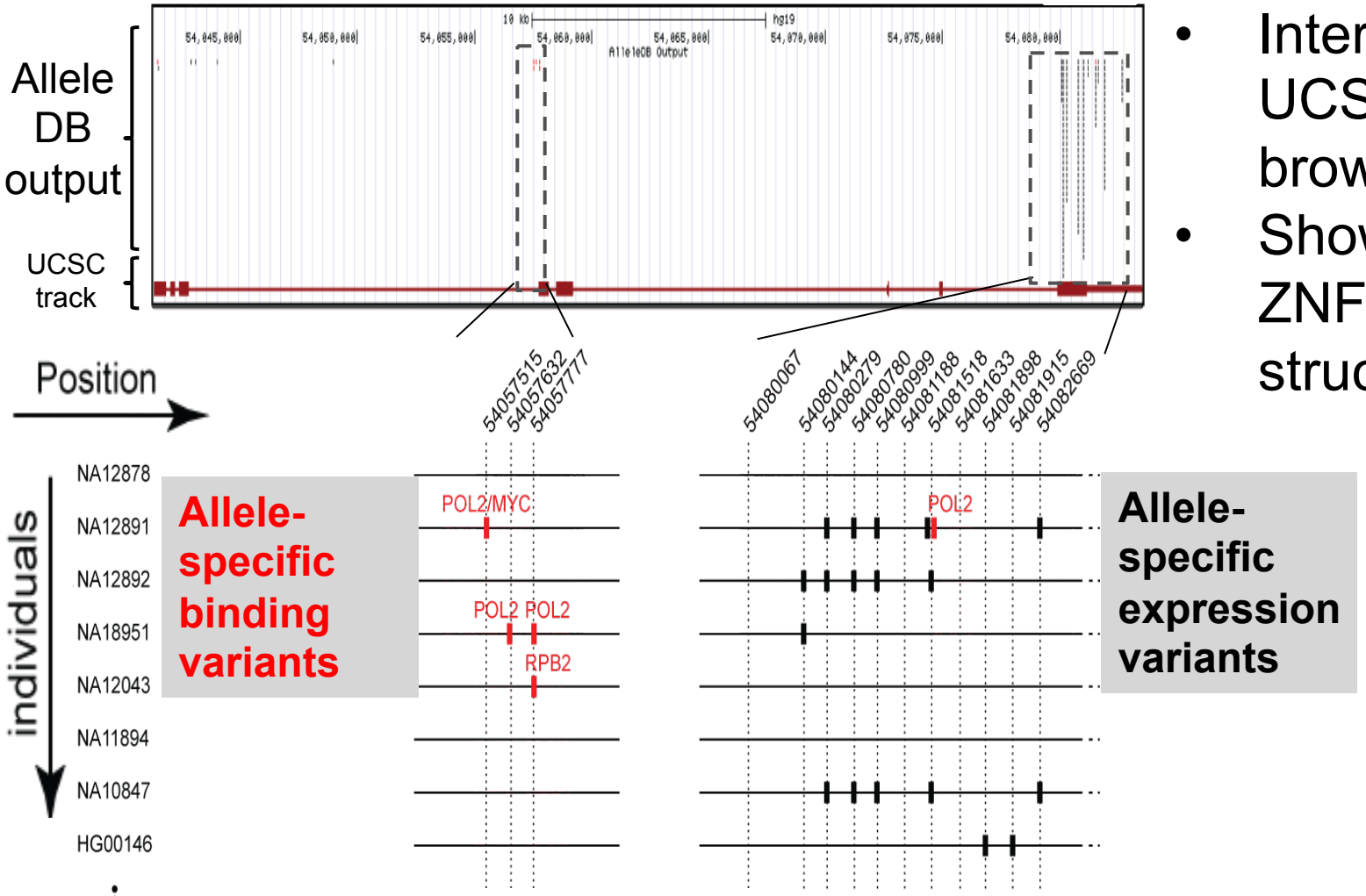


- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

alleledb.gersteinlab.org

Chen J. *et al.* (*Nature Commun*, in press)

AlleleDB: Annotating rare & common allele-specific variants over a population

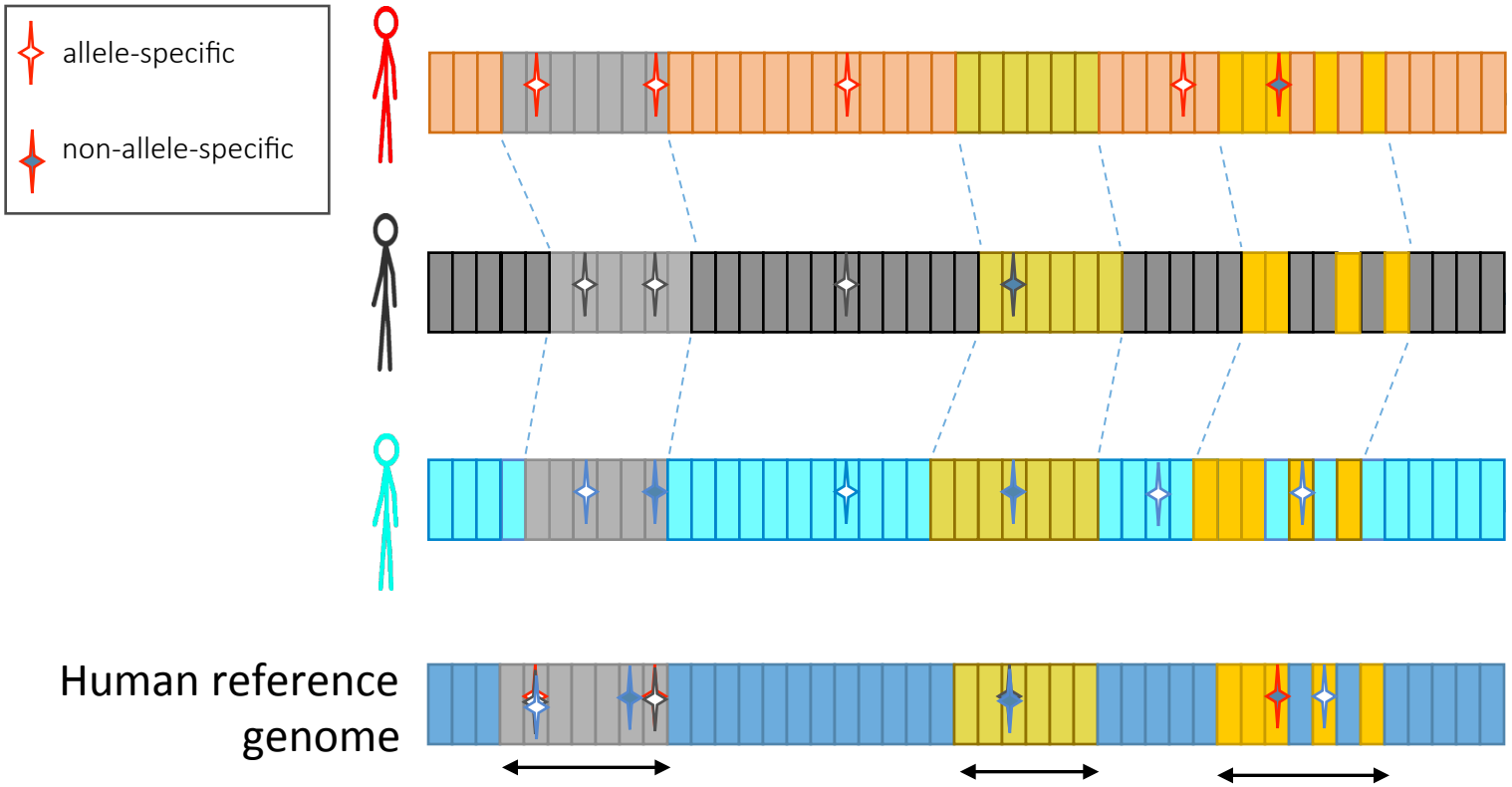


- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

Chen J. *et al.* (*Nature Commun*, in press)

Collecting ASE/ASB variants into allele-specific genomic regions

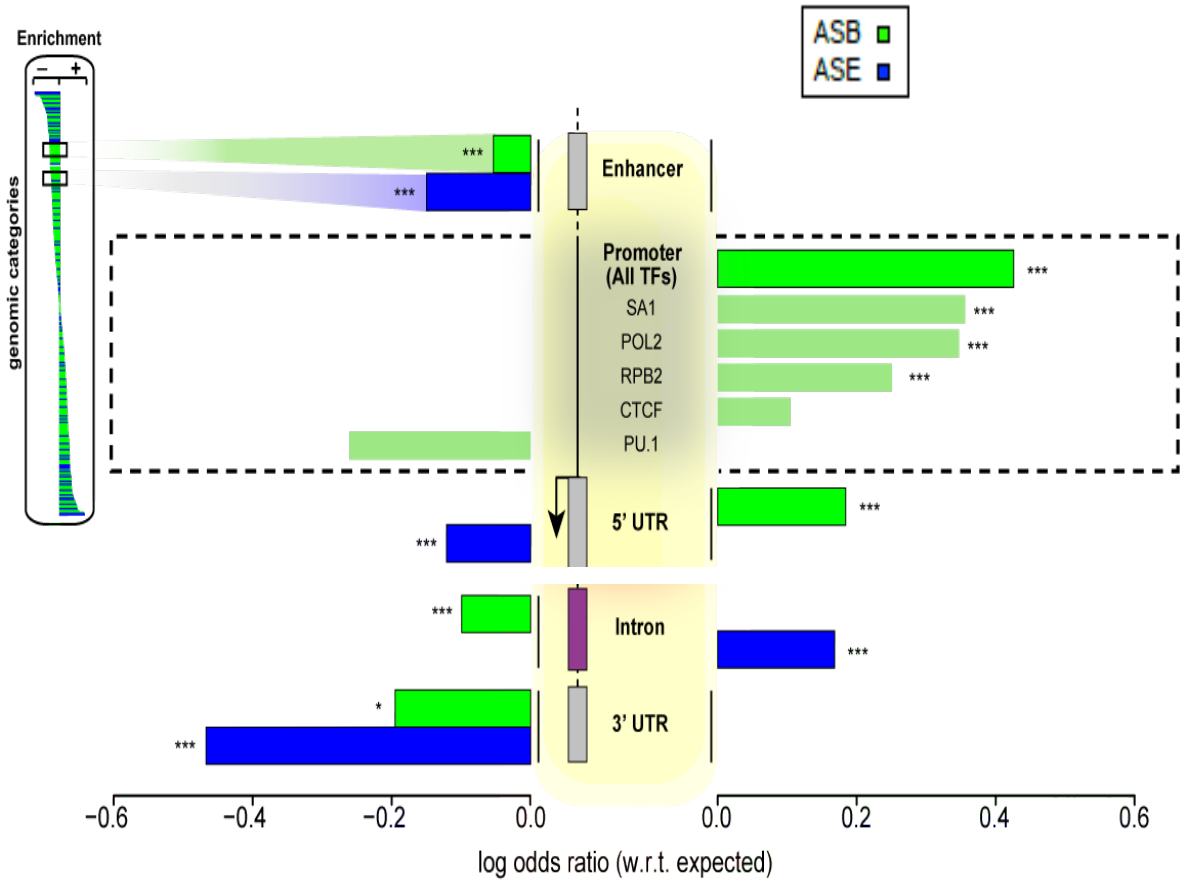
Does a particular genomic element have a higher tendency to be allele-specific? Fisher's exact test, for the **enrichment** of allele-specific variants in the element (with respect to non-allele-specific variants that could potentially be called as allelic)



alleledb.gersteinlab.org

Chen J. *et al.* (*Nature Commun*, in press)

Groups of elements that are enriched or depleted in allelic activity

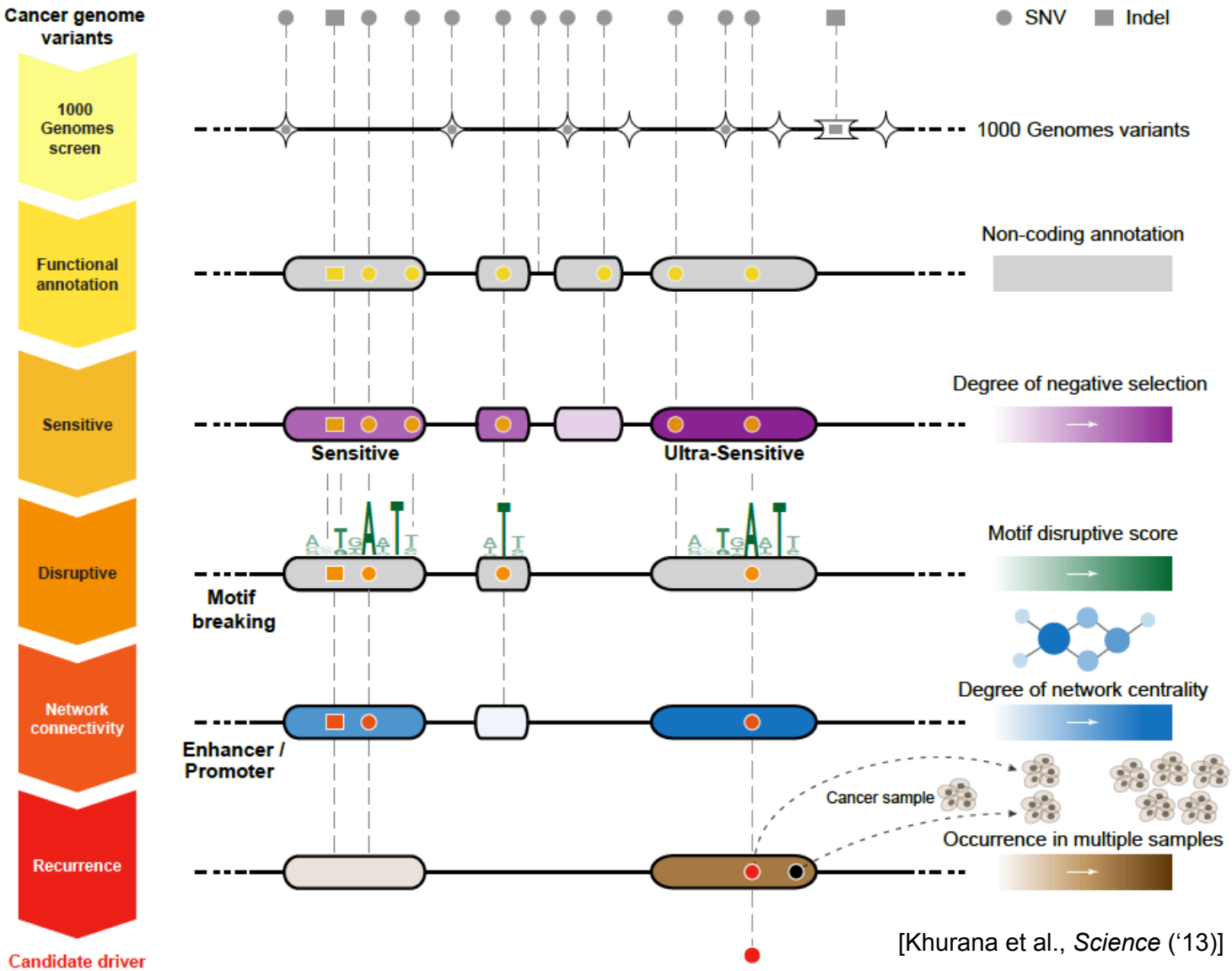


Chen J. *et al.* (*Nature Commun*, in press)

Personal Genomics: Handling Exponential Data Scaling through Prioritizing High-impact Variants

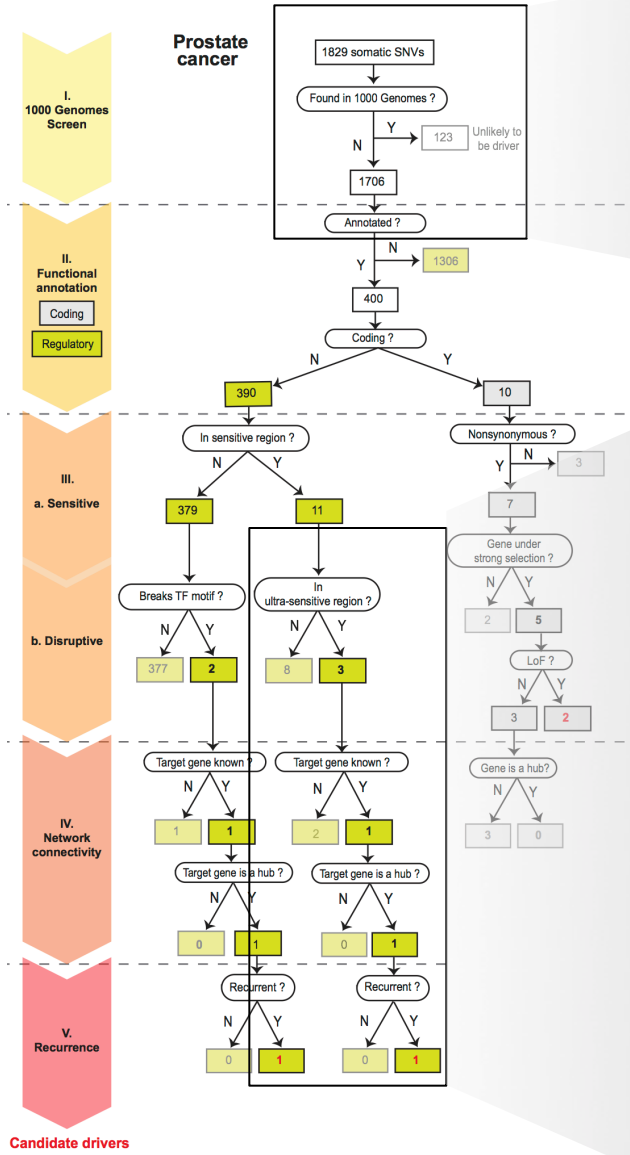
- **Introduction**
  - The **exponential scaling** of data generation & data processing
  - The **landscape of variants** in personal genomes
- **Characterizing Rare Variants in Coding Regions**
  - **Identifying with STRESS cryptic allosteric sites**
  - On surface & in interior bottlenecks
- **Non-coding Variants : Prioritizing using AlleleDB in terms of allelic elements**
  - Having observed difference in molecular activity in many contexts
- **Putting it together in workflows: Integrating evidence on non-coding variants with FunSeq**
  - Systematically weighting all the features
  - suggesting non-coding drivers
  - Prioritzing rare germline variants

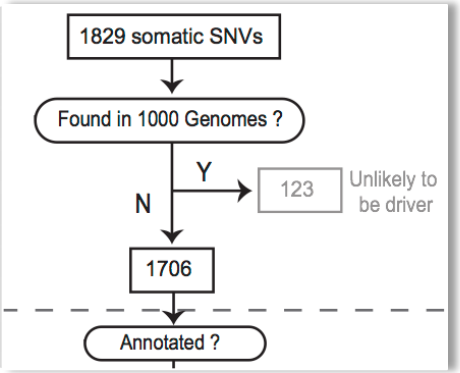
Identification of non-coding candidate drivers amongst somatic variants: Scheme

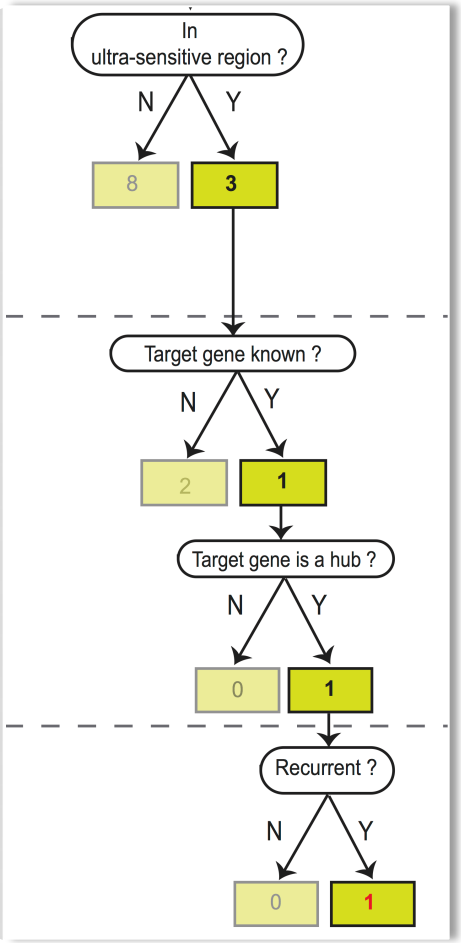


[Khurana et al., *Science* ('13)]

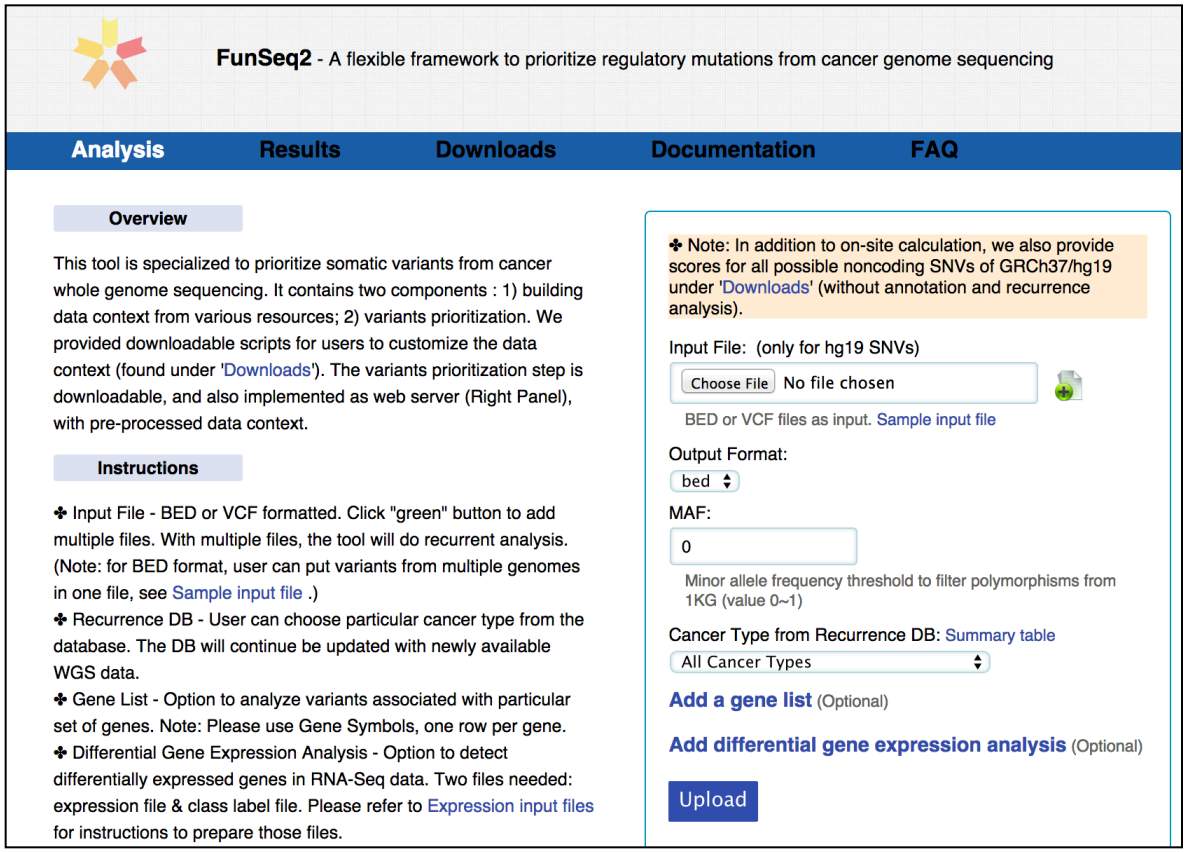
Flowchart for 1 Prostate Cancer Genome (from Berger et al. '11)





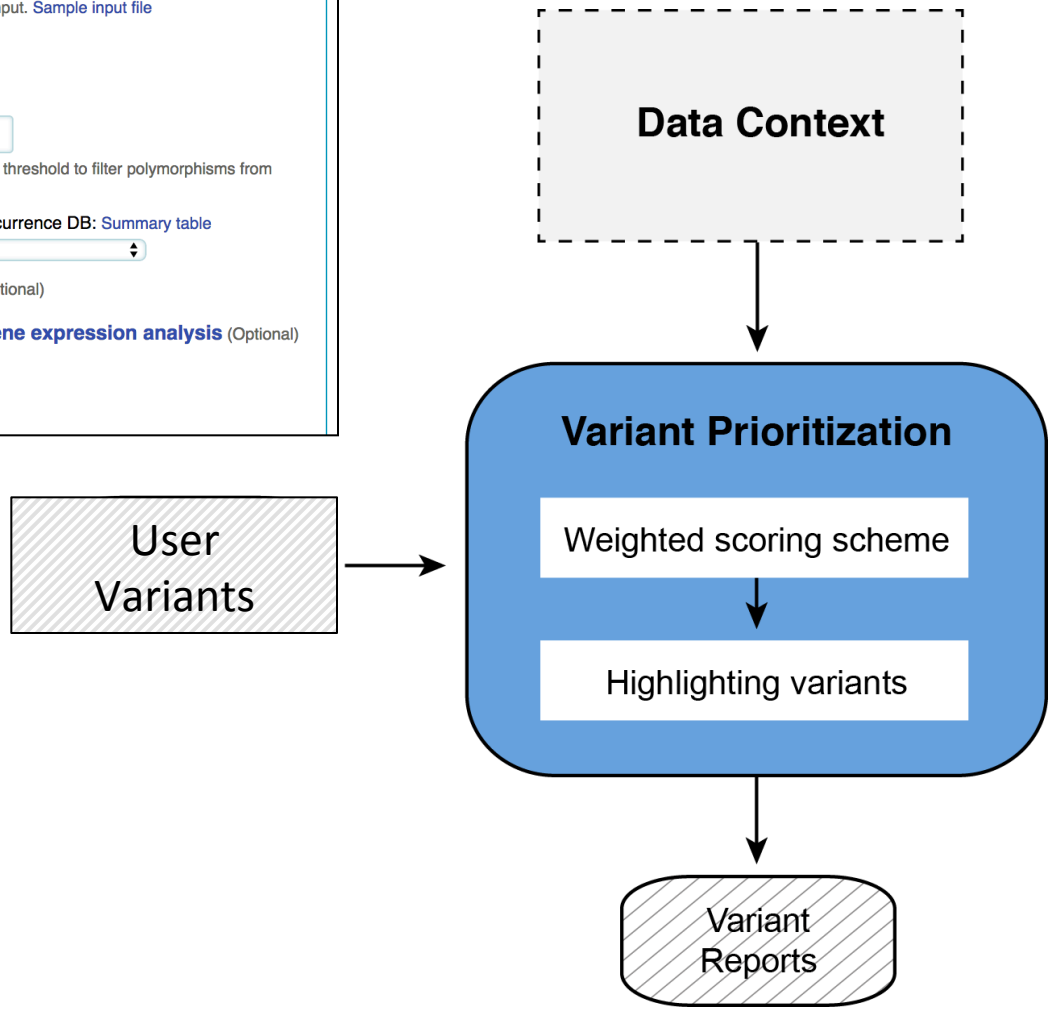


[Khurana et al., *Science* ('13)]



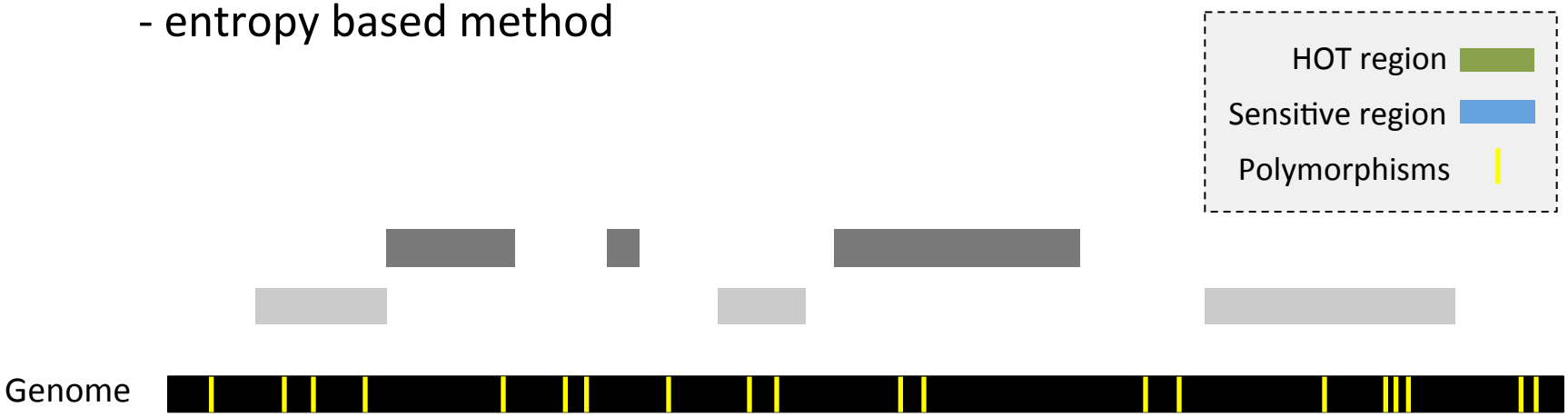
Site integrates user variants with large-scale context

FunSeq.gersteinlab.org



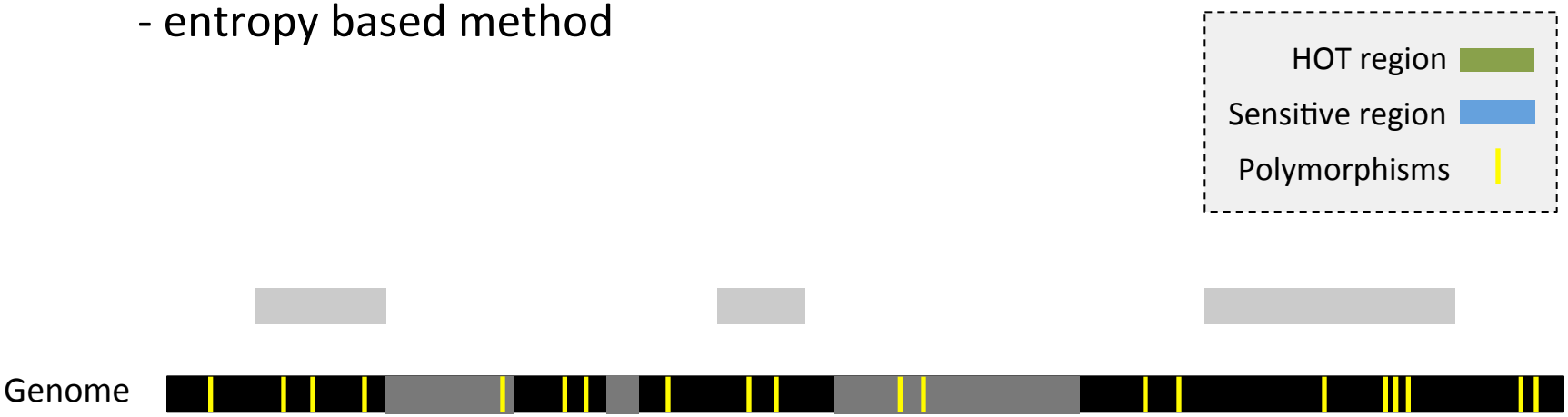
[Fu et al., GenomeBiology ('14)]

- Feature weight
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method



[Fu et al., GenomeBiology ('14)]

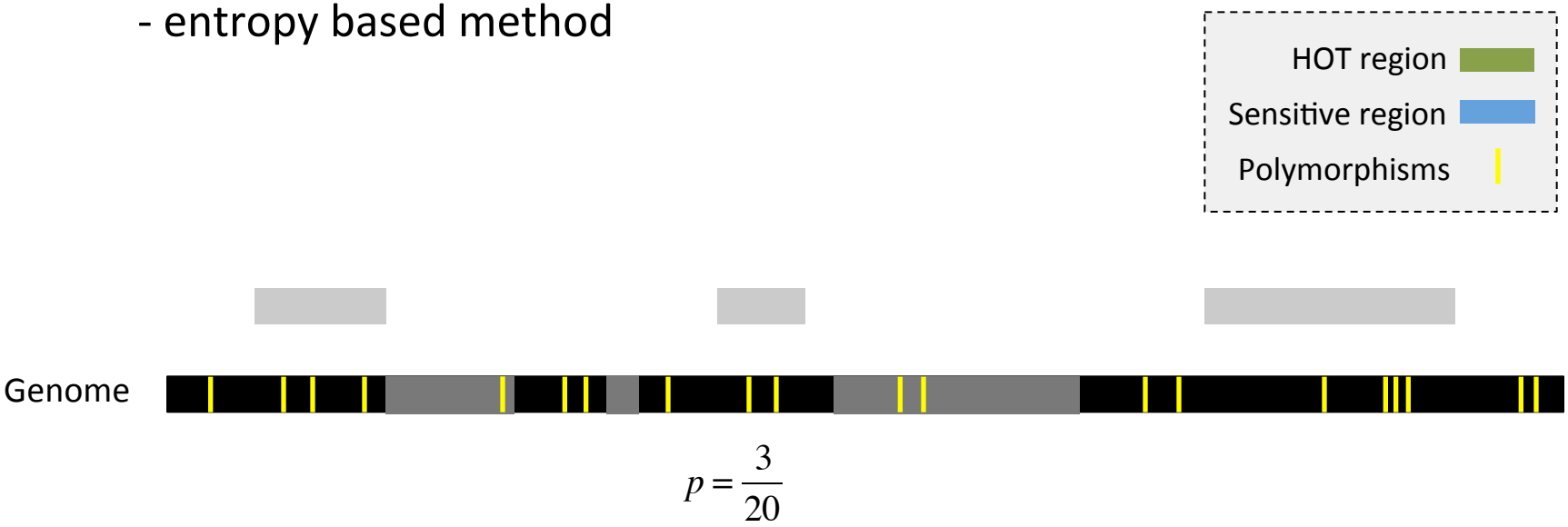
- Feature weight
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method



$$p = \frac{3}{20}$$

[Fu et al., GenomeBiology ('14)]

- Feature weight
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method



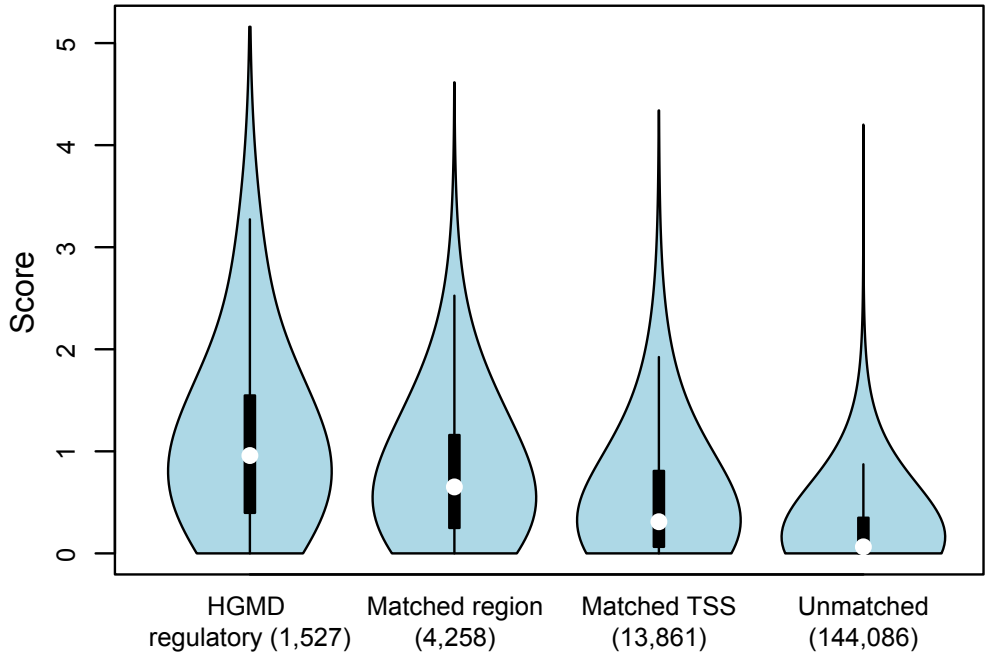
Feature weight: $w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$

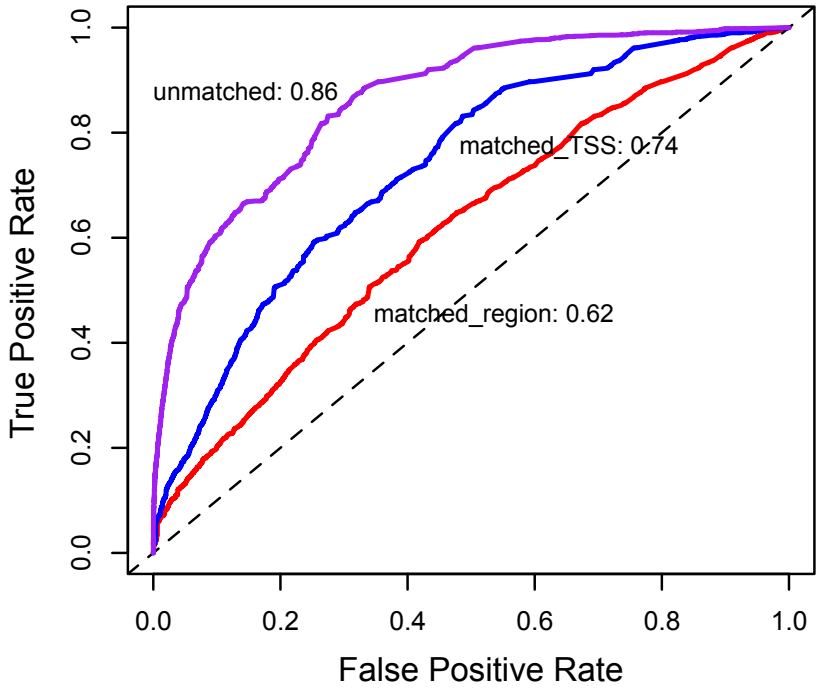
$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

For a variant: $Score = \sum w_d \quad of\ observed\ features$

[Fu et al., GenomeBiology ('14)]

Germline pathogenic variants show higher core scores than controls





3 controls with natural polymorphisms (allele frequency >= 1%)

1. Matched region: 1kb around HGMD variants
2. Matched TSS: matched for distance to TSS
3. Unmatched: randomly selected

Ritchie et al., Nature Methods, 2014

[Fu et al., GenomeBiology ('14, in revision)]

Personal Genomics: Handling Exponential Data Scaling through Prioritizing High-impact Variants

- Introduction
  - The exponential scaling of data generation & data processing
  - The landscape of variants in personal genomes
- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
  - On surface & in interior bottlenecks
- Non-coding Variants : Prioritizing using AlleleDB in terms of allelic elements
  - Having observed difference in molecular activity in many contexts
- Putting it together in workflows: Integrating evidence on non-coding variants with FunSeq
  - Systematically weighting all the features
  - suggesting non-coding drivers
  - Prioritzing rare germline variants

Personal Genomics: Handling Exponential Data Scaling through Prioritizing High-impact Variants

- **Introduction**
  - **The exponential scaling of data generation & data processing**
  - **The landscape of variants in personal genomes**
- **Characterizing Rare Variants in Coding Regions**
  - **Identifying with STRESS cryptic allosteric sites**
  - On surface & in interior bottlenecks
- **Non-coding Variants : Prioritizing using AlleleDB in terms of allelic elements**
  - Having observed difference in molecular activity in many contexts
- **Putting it together in workflows: Integrating evidence on non-coding variants with FunSeq**
  - Systematically weighting all the features
  - suggesting non-coding drivers
  - Prioritzing rare germline variants

Acknowledgments

AlleleDB.gersteinlab.org
J **Chen**,
J **Rozowsky**,
TR **Galeev**, A Harmanci, R Kitchen, J Bedford, A Abyzov, Y Kong, L Regan

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

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

Hiring Postdocs. See gersteinlab.org/jobs

FunSeq.gersteinlab.org & **FunSeq2**.gersteinlab.org
Y **Fu**, E **Khurana**, Z Liu, S Lou, J Bedford, XJ Mu, KY Yip, V Colonna, XJ Mu, … ,
1000 Genomes Project Consortium, et al

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2015.
  - 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