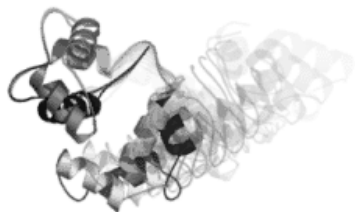
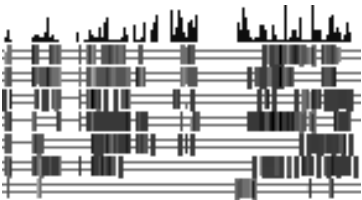
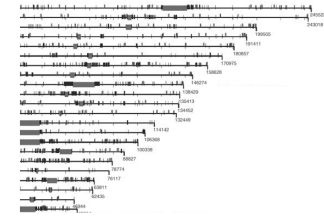
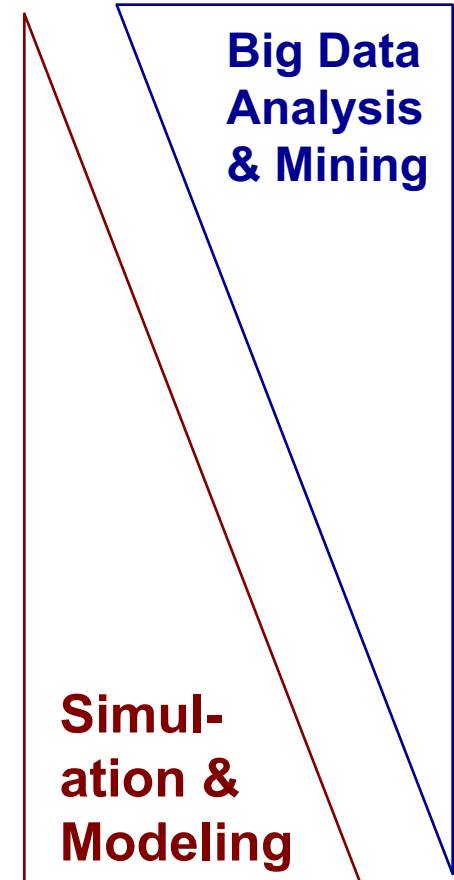


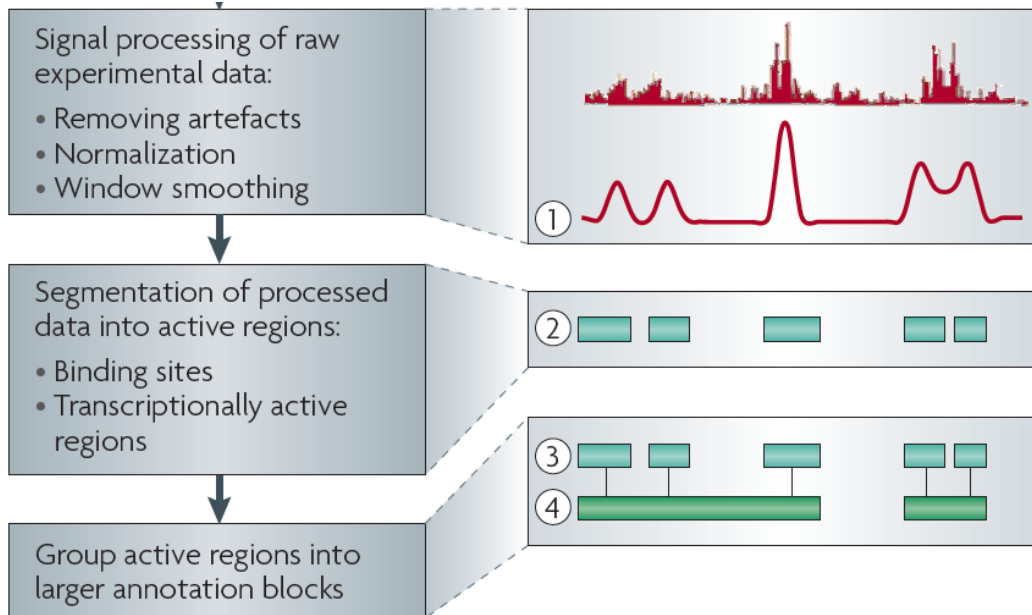
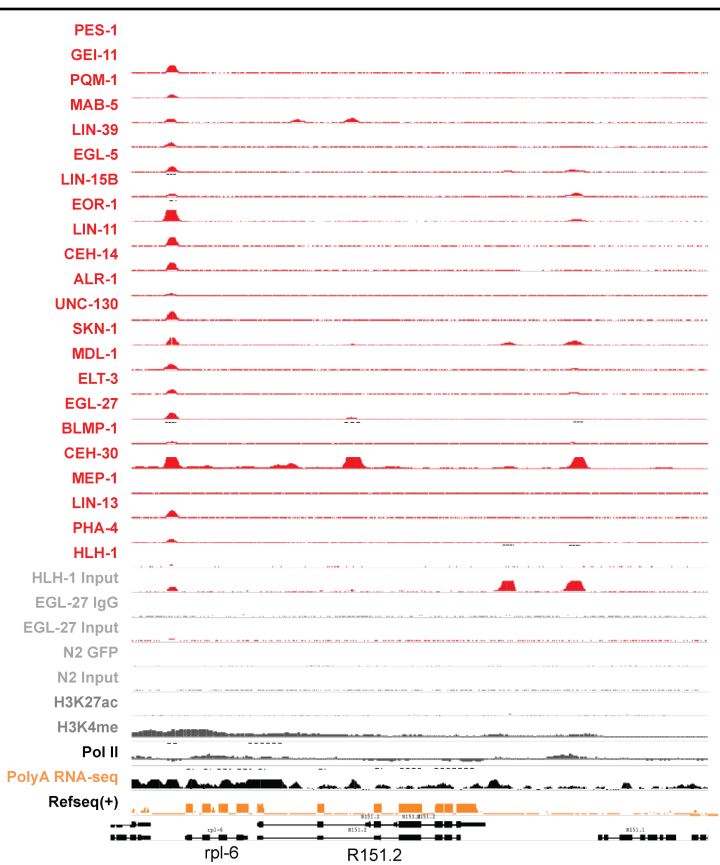
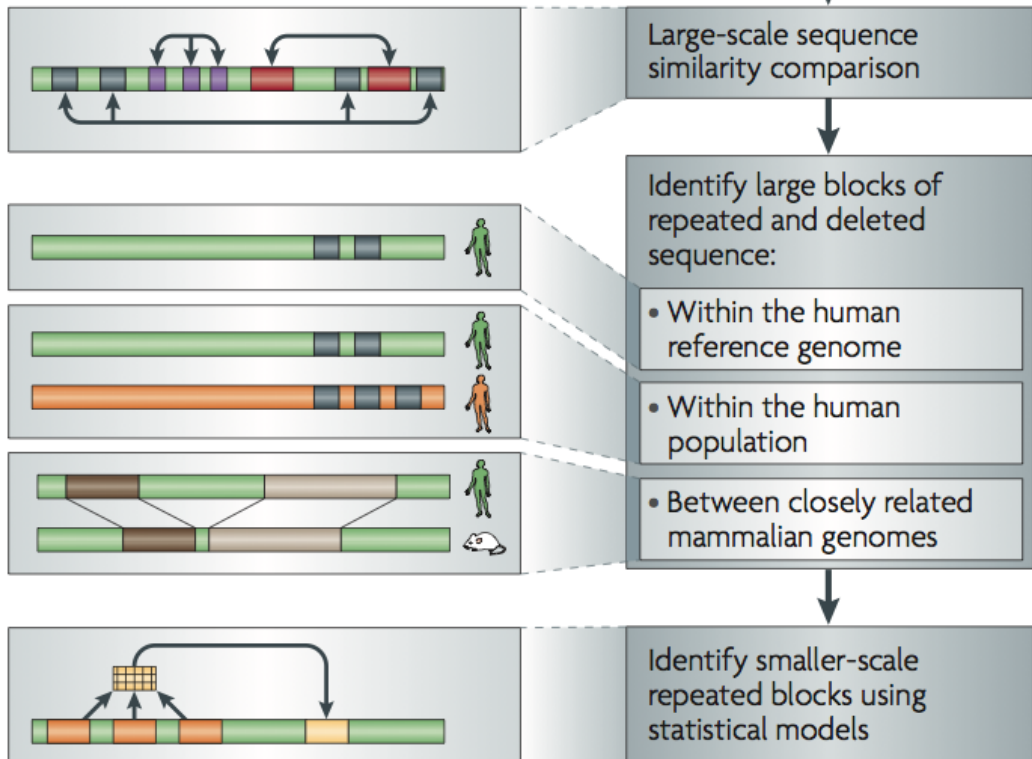
GersteinLab.org Overview



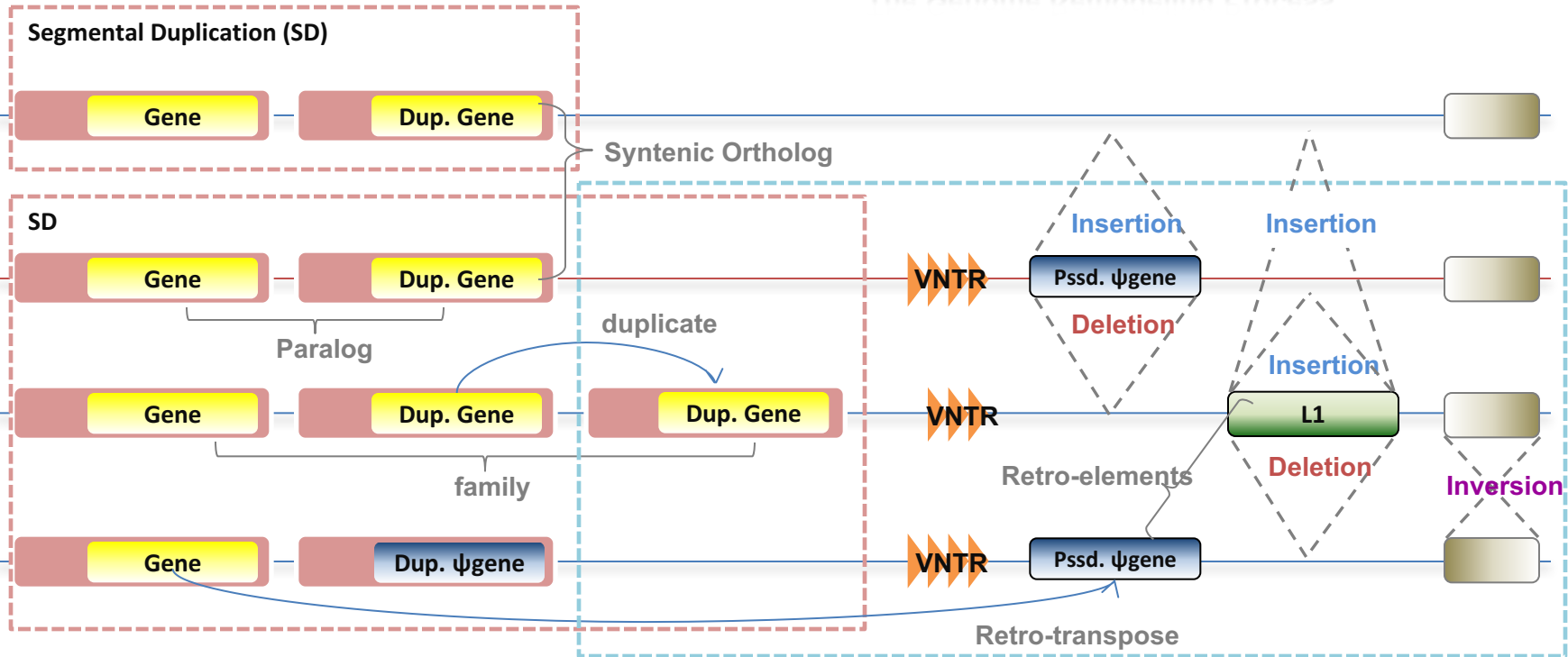
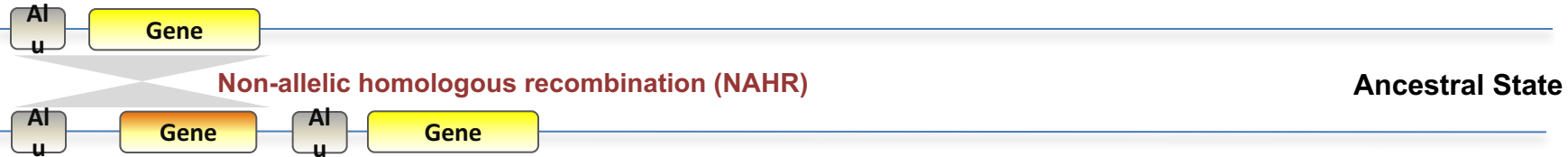
- Biological Knowledge Representation, Literature Mining, **Genomic Privacy**
- Genome Structural Variation & **Personal Genomics**
- **Human Genome Annotation & Disease Genomics**
- **Networks of Genes** & Protein Pathways
- Macromolecular Structures & Motions

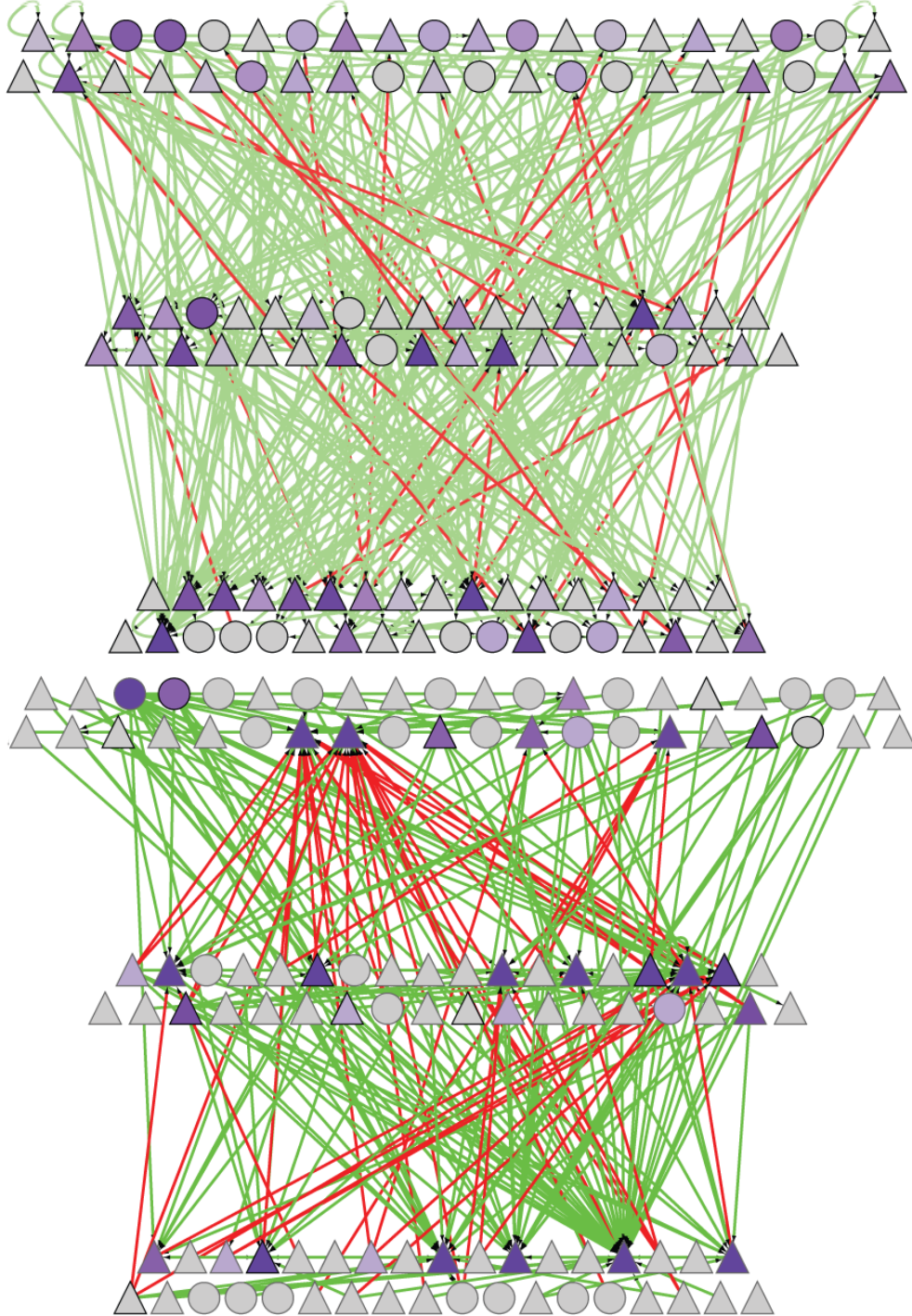


Annotating the Human Genome: Comparative & Functional



Comparative Genomics: Pseudogenes & Genome Remodeling





Recasting Functional Genome Annotation as Networks:

Comparing Proximal & Distal Networks

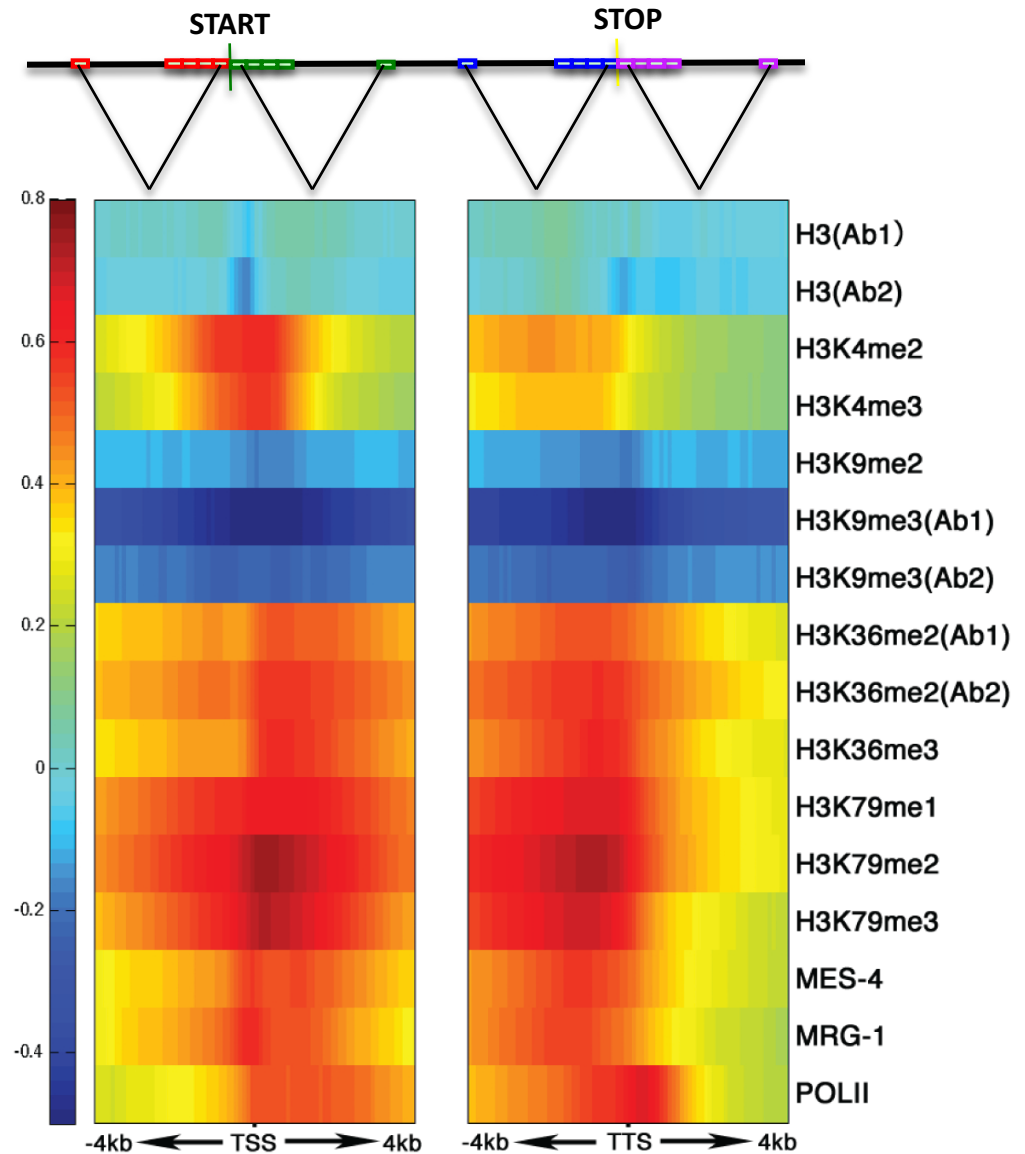
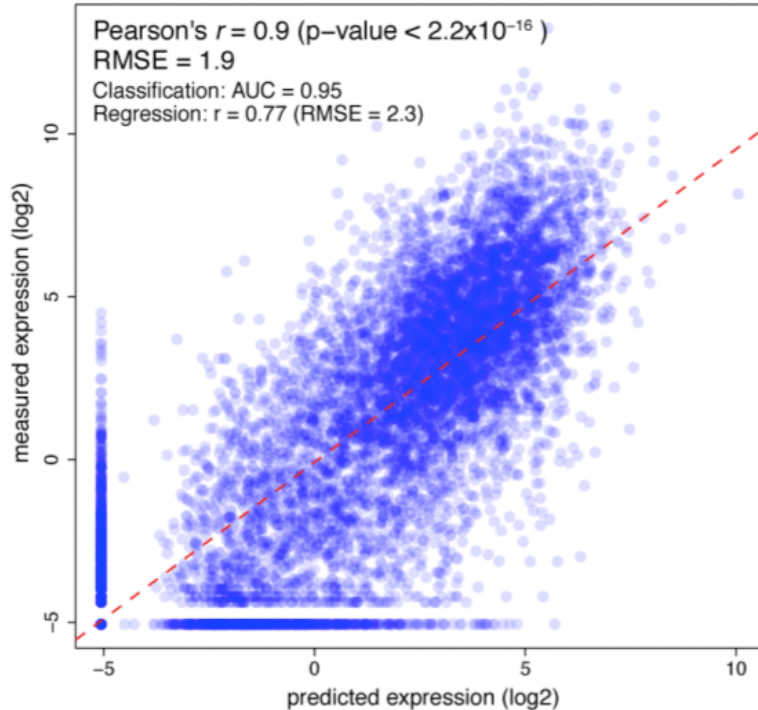
[Gerstein et al. Nature (in press, '12)]

Machine Learning Models from Genome Annotation

His. mods around TSS & TTS are clearly related to level of gene expression

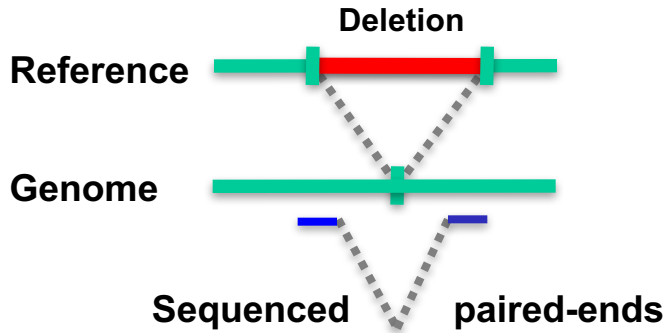
Early work in '09/'10

Science 330:6012

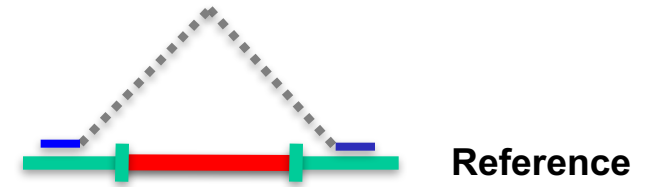


Finding Variants (SVs) in Personal Genomes

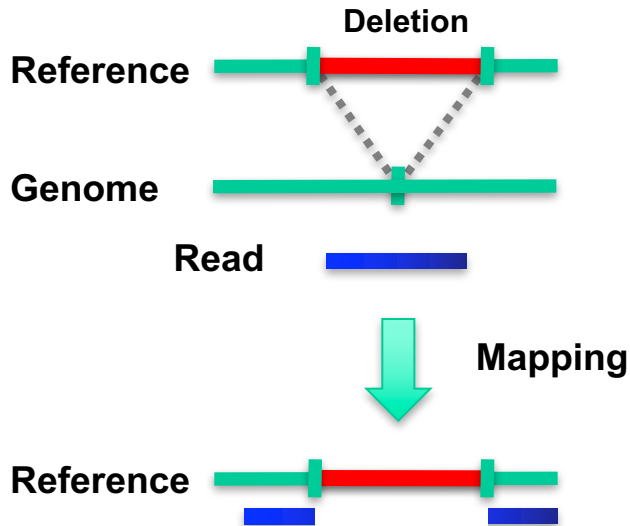
1. Paired ends



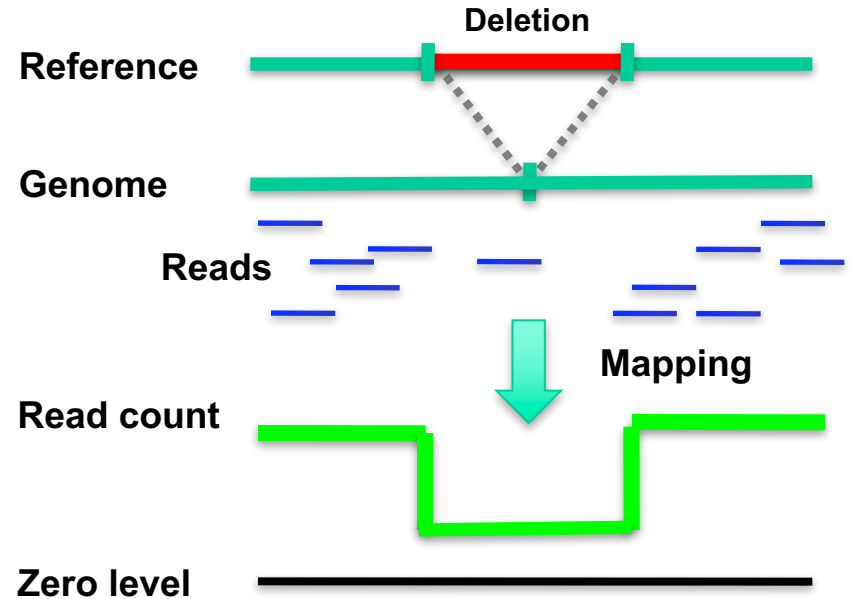
Mapping
→



2. Split read



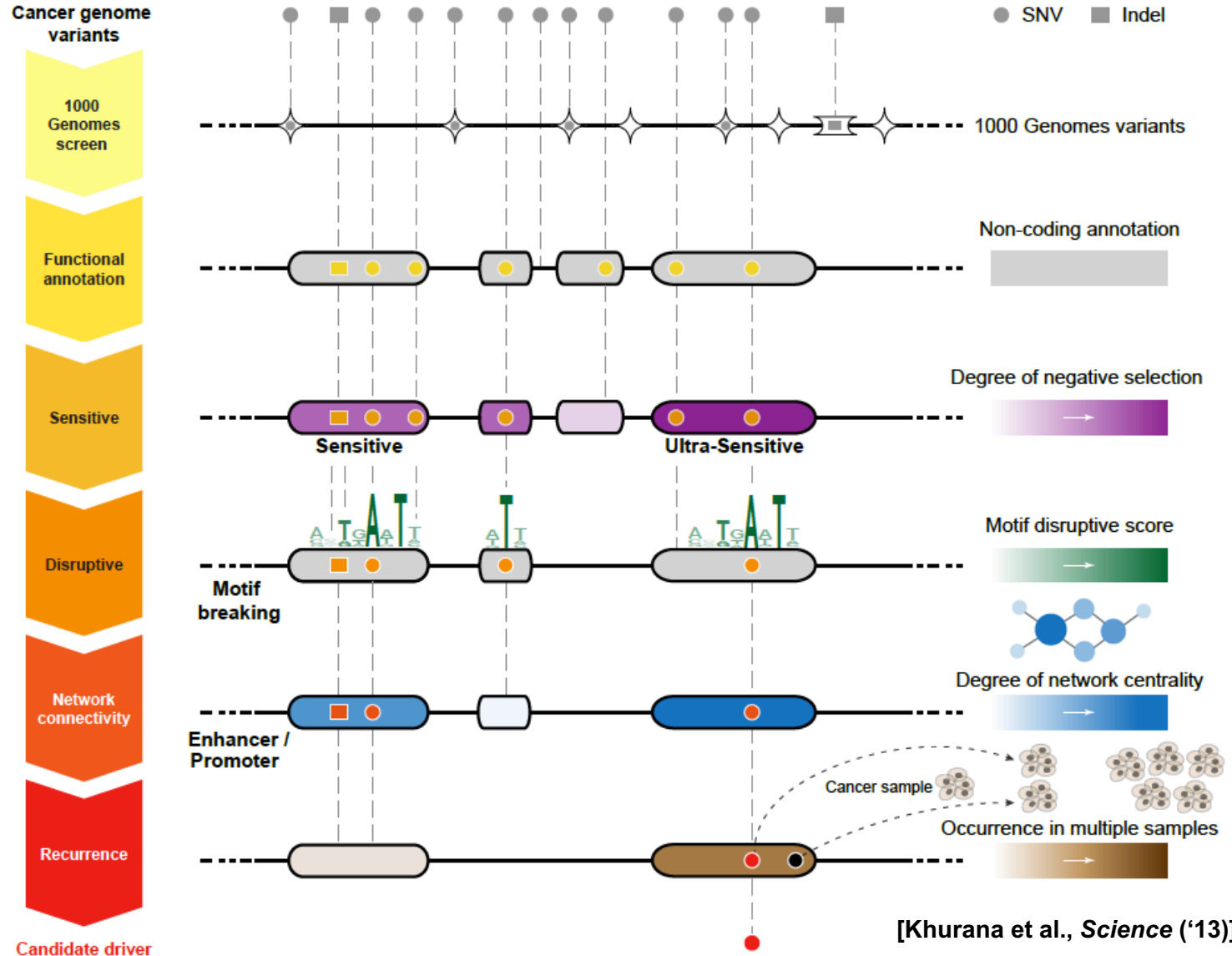
3. Read depth



[Snyder et al. Genes & Dev. ('10)]

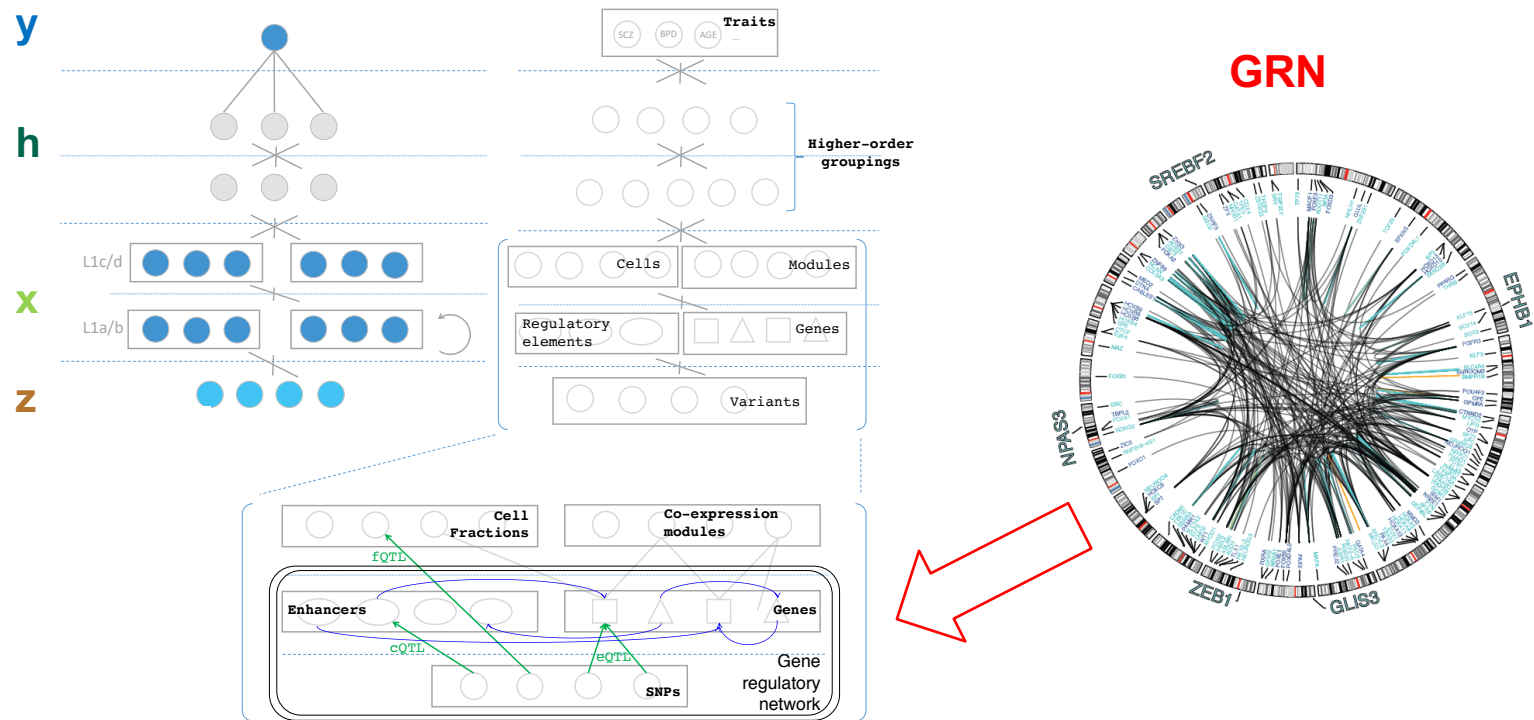
Characterizing Variants in the Genome

Identification of non-coding candidate drivers amongst somatic variants, using genome annotation & patterns of natural variation



Putting it all together in Neuro-genomics

- Embed **Gene Regulatory Network** in deep neural network
- Allows transcriptome (+other) imputation & trait prediction



y: phenotypes

h: hidden units (e.g., circuits)

x: intermediate phenotypes (e.g. expression, enhancers)

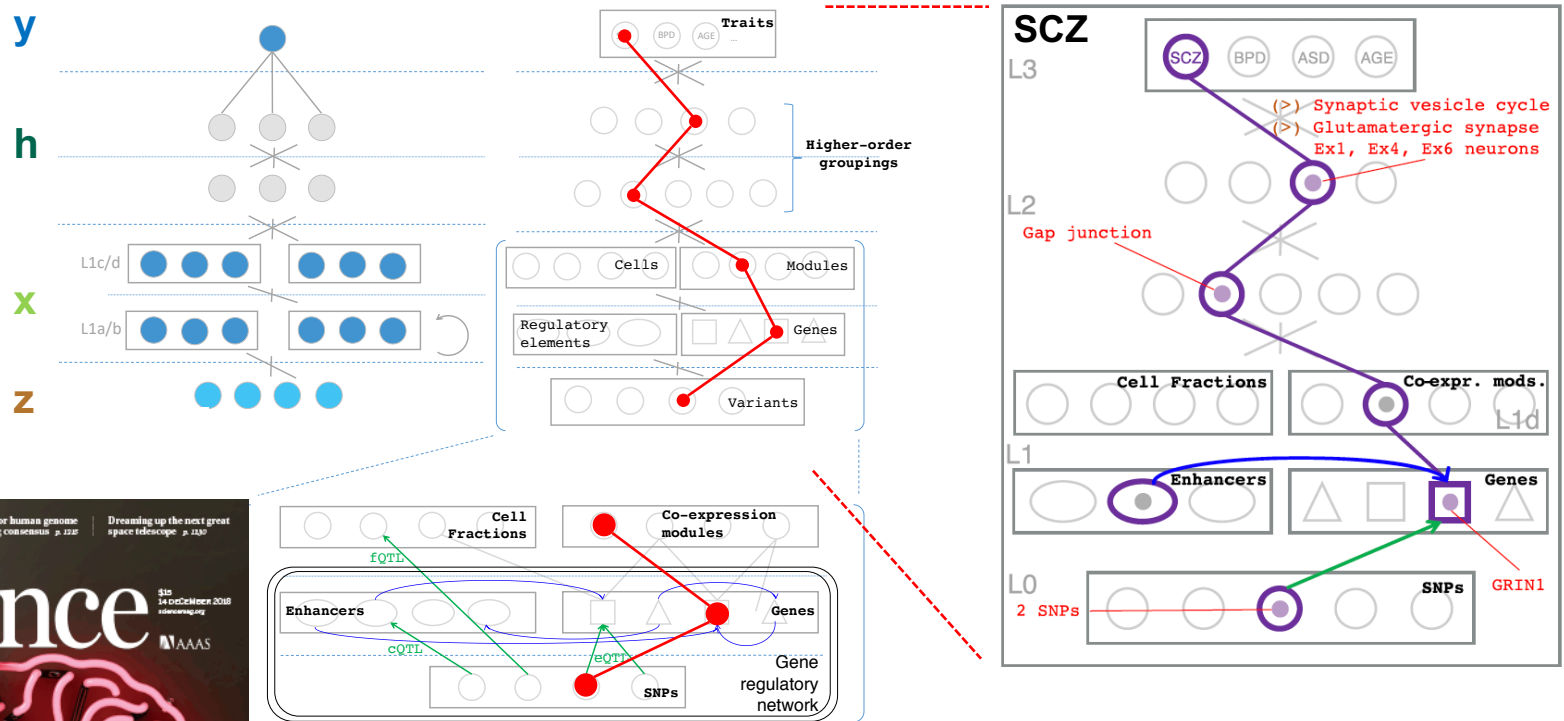
z: genotypes (e.g., SNPs)

Deep Boltzmann Machine Energy model:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}) \propto \exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}))$$

Putting it all together in Neurogenomics

- Allows prioritization of genes / modules through network interpretation (using path tracing)



types
n units (e.g., circuits)

X: intermediate phenotypes (e.g. expression, enhancers)

Z: genotypes (e.g., SNPs)

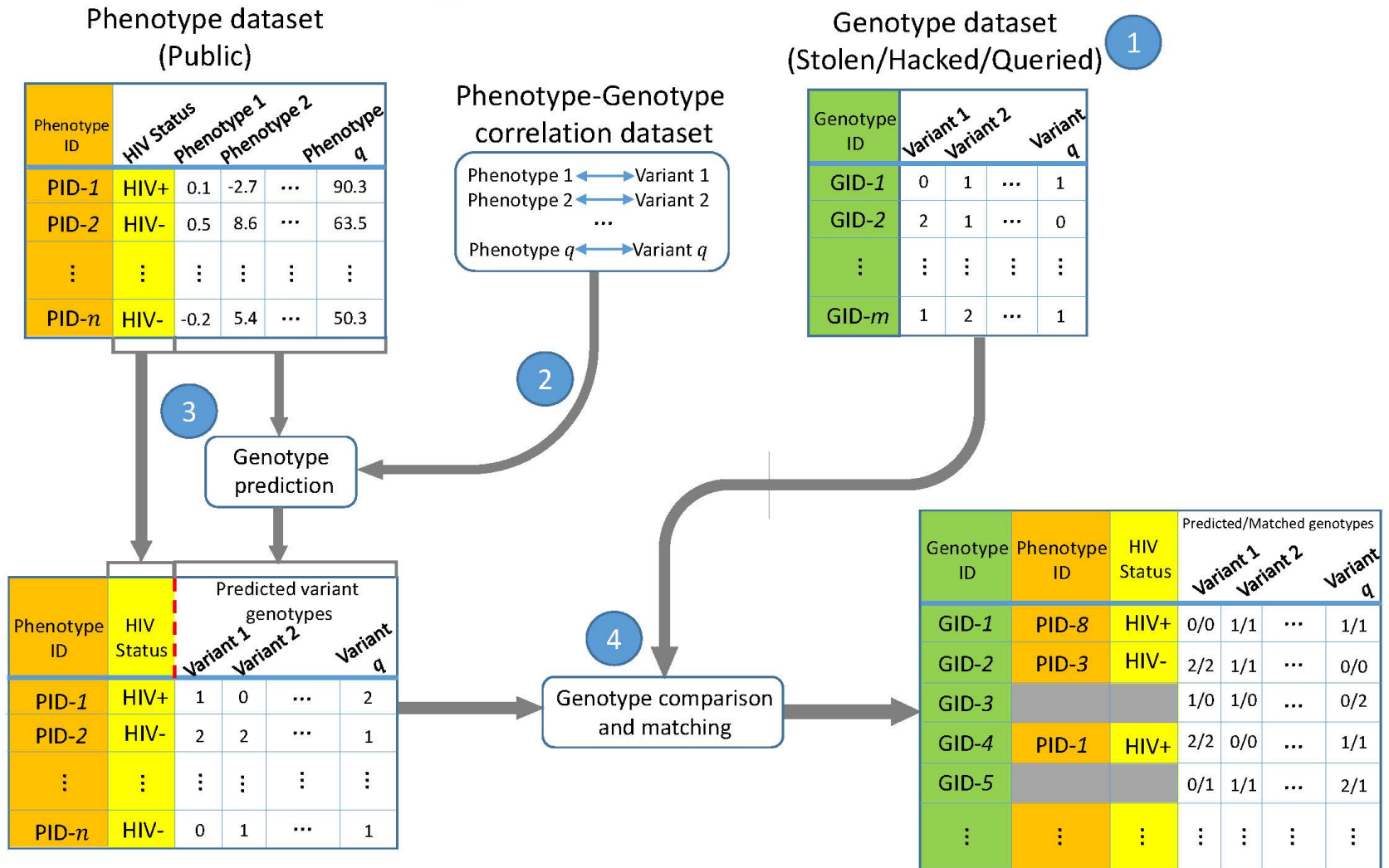
Deep Boltzmann Machine Energy model:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}) \propto \exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}))$$

[Wang et al. ('18) Science]

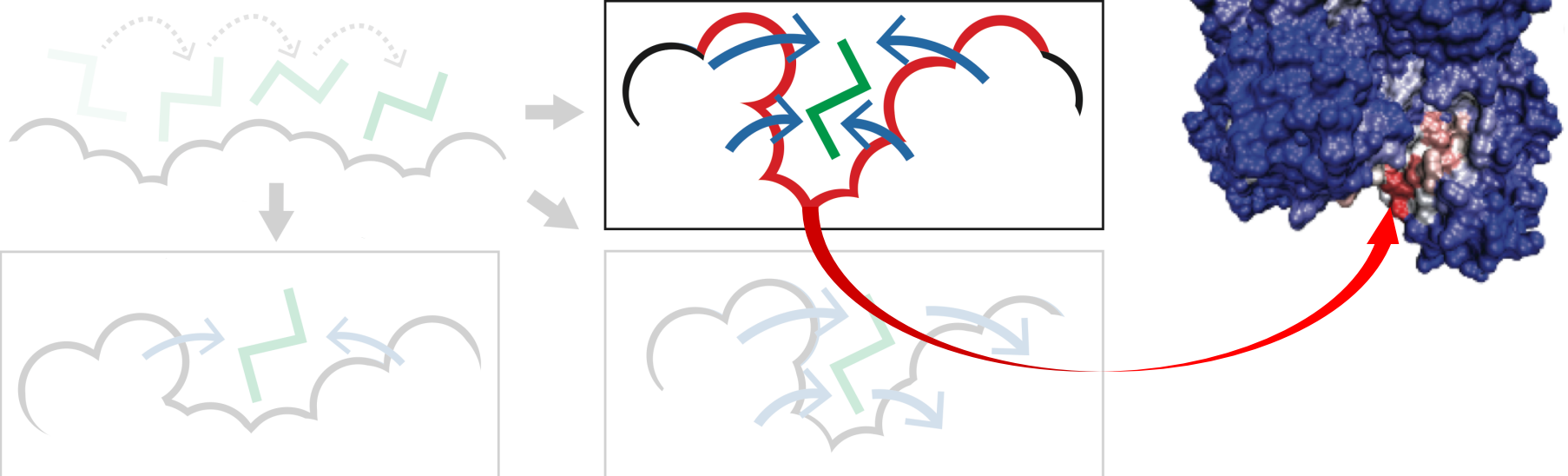


Biological Data Science: Protecting Genomic Privacy from linking attacks

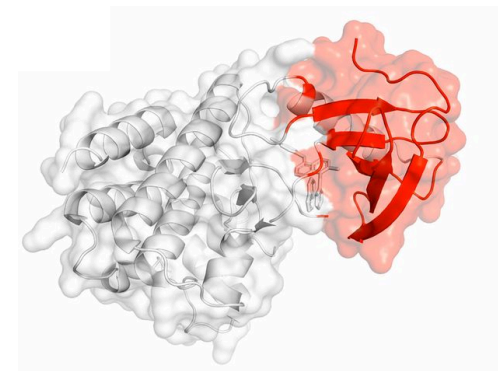


Predicting Important Residues at the Surface via **simulation**, to characterize deleterious variants

PDB: 3PFK

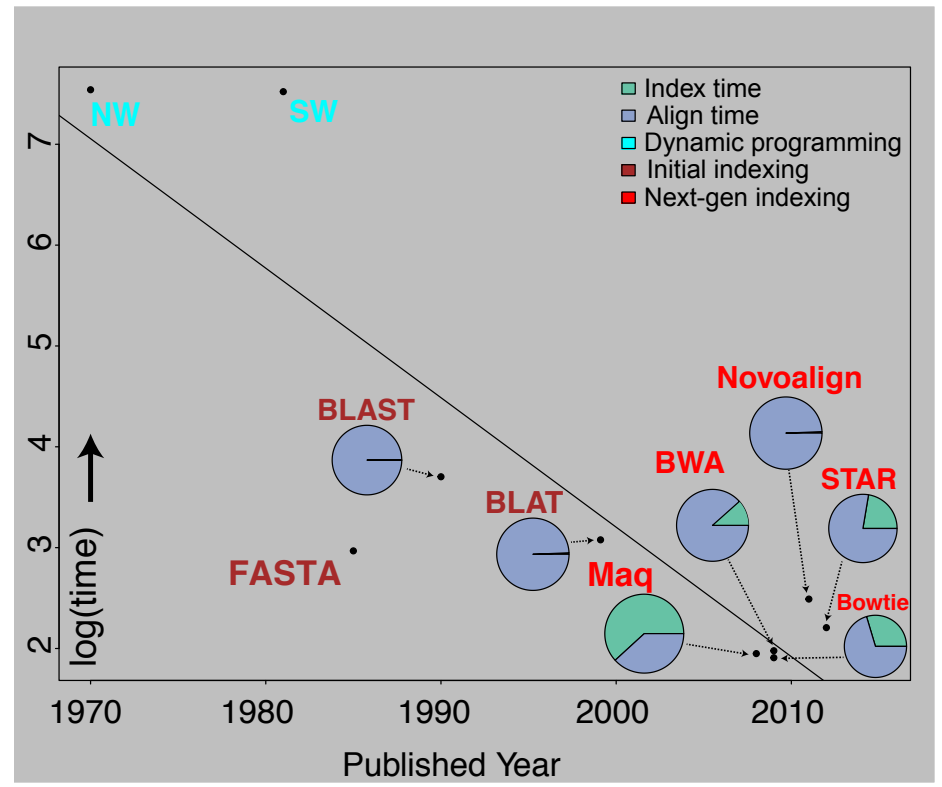
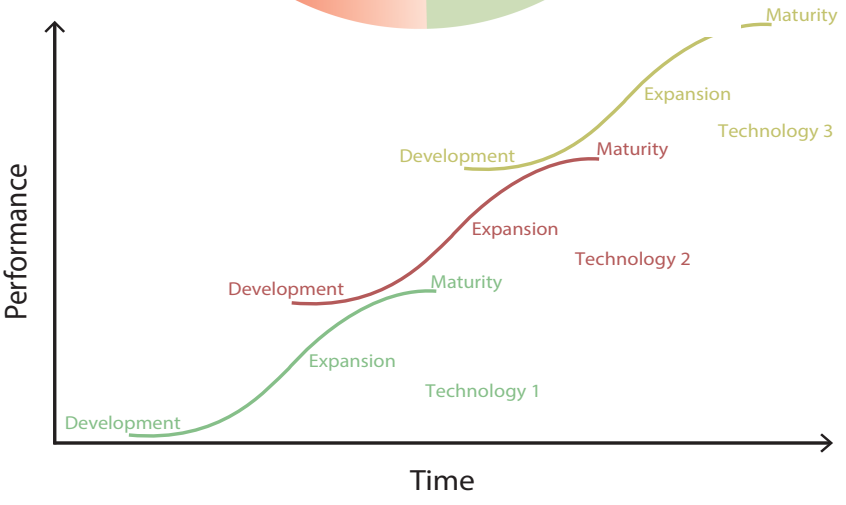
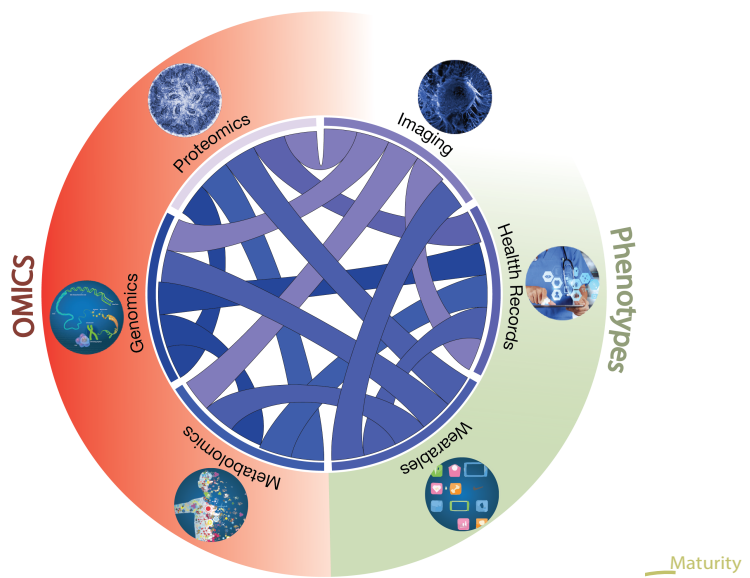


Can be adapted to find
cancer drivers



Clarke*, Sethi*, et al (2016) *Structure*
Kumar, Clarke, Gerstein (2019) *PNAS*

Perspectives: Characterizing the Growth & Scaling in Biomedical Data Science



Alignment algorithms scaling to keep pace with data generation

GersteinLab.org Overview

- Biological Knowledge Representation
 - Literature Mining
- Personal Genomics
 - Calling variants, particularly SVs
 - Genomic Privacy
- Human Genome Annotation
 - Machine learning models, esp. interpretable ones
 - Pseudogenes & Comparative Genomics
- Disease Genomics
 - Variant Interpretation
 - Neurogenomics
 - Cancer Genomics
- Networks of Genes
- Macromolecular Structures & Motions
 - Simulation
- Perspectives on Biomedical Data Science
 - Scaling
- New Directions
 - Wearables & Devices
 - Image Processing