Grokking the genome

Aaron Quinlan
quinlanlab.org

Department of Public Health Sciences
Center for Public Health Genomics
Biochemistry and Molecular Genetics

What do we work on?

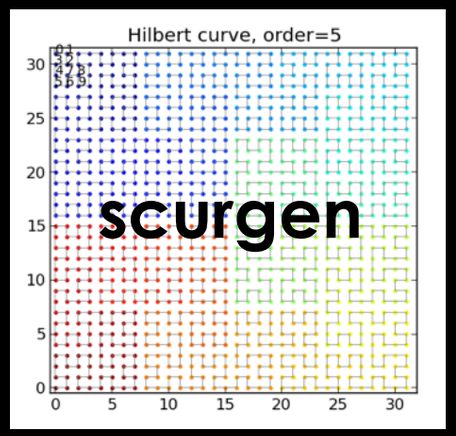
Software tools for genome research

gemini

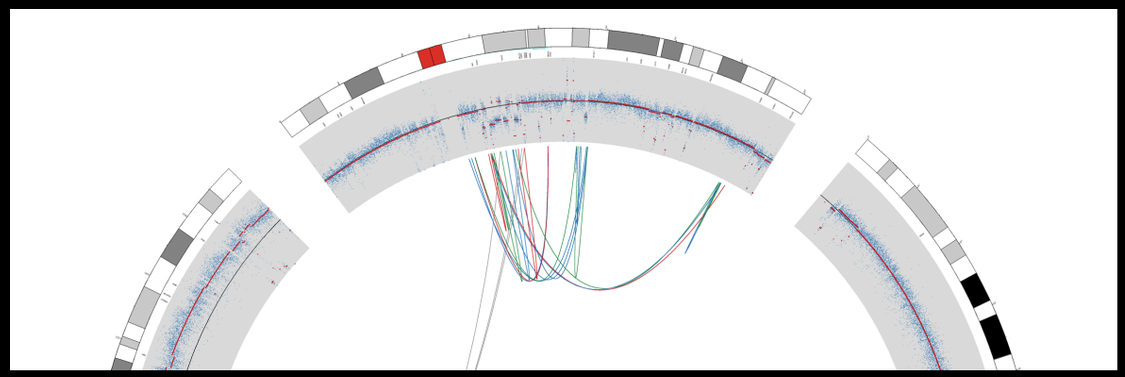
hydra

lumpy

GQL



Cancer genomics



Cellular and molecular origins of GBM

Initiating mutations underlying OV?

Tumor evolution and mutational mechanisms.

Disease genetics

Type 1 diabetes

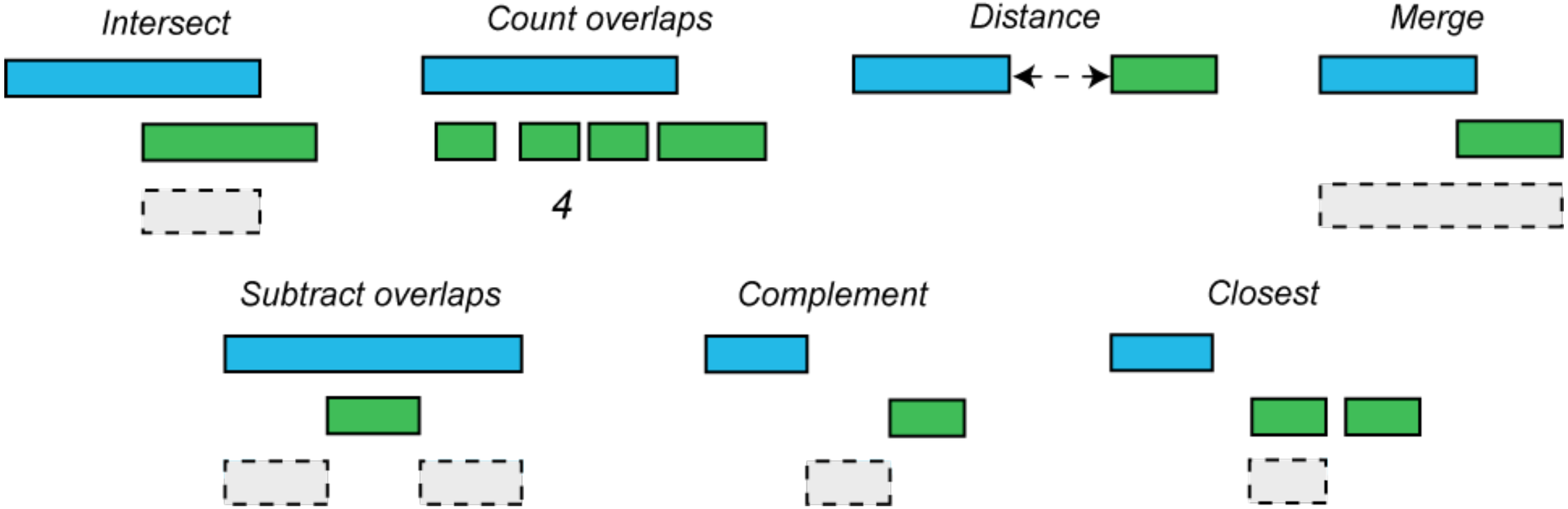
Lupus

Unexplained development disorders

Radiation hypersensitivity

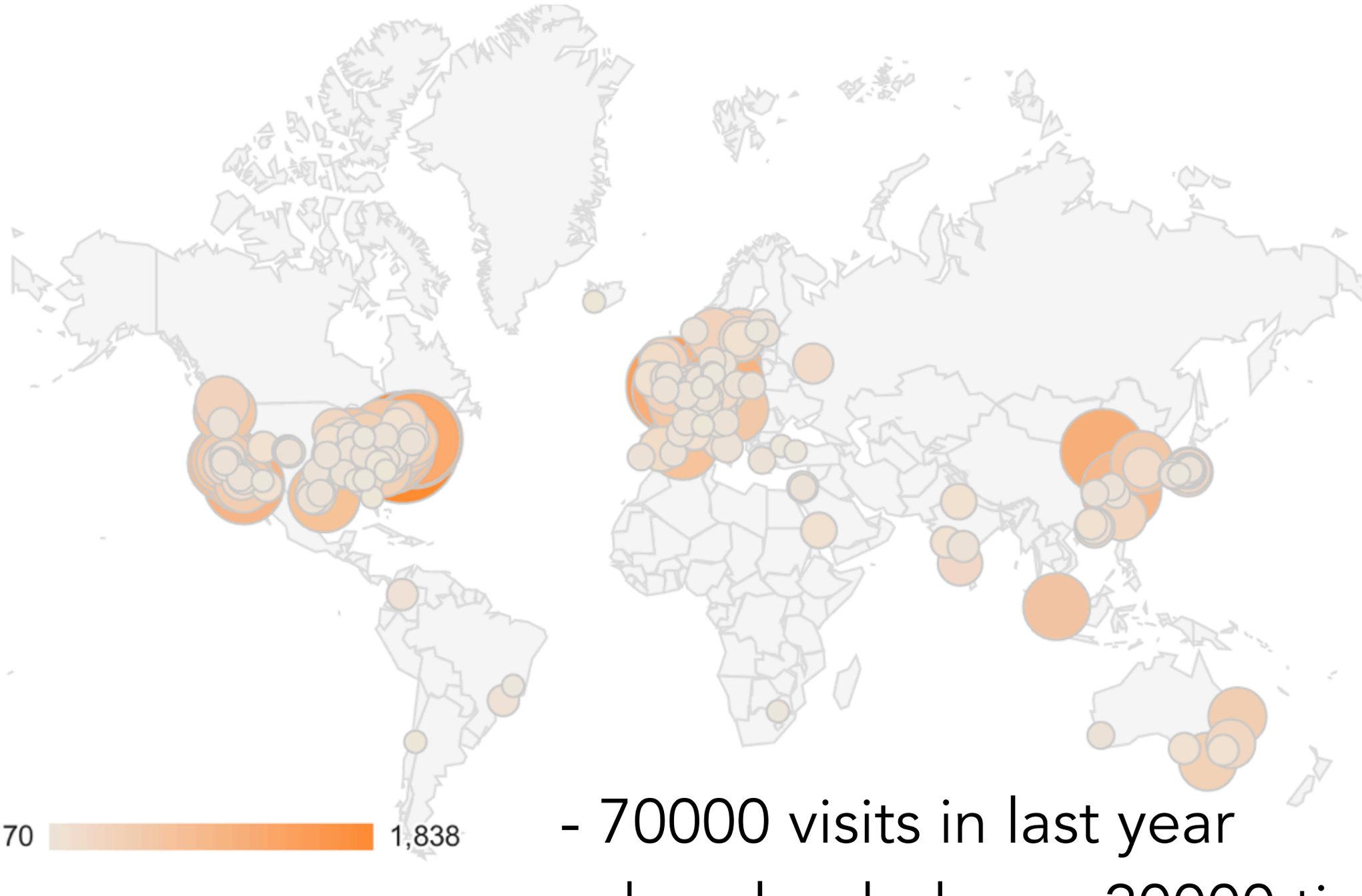
1. Making sense of large-scale, complex genomic datasets

a swiss army knife for genome arithmetic



Started in Ira Hall's lab in 2009

bedtools is very useful...

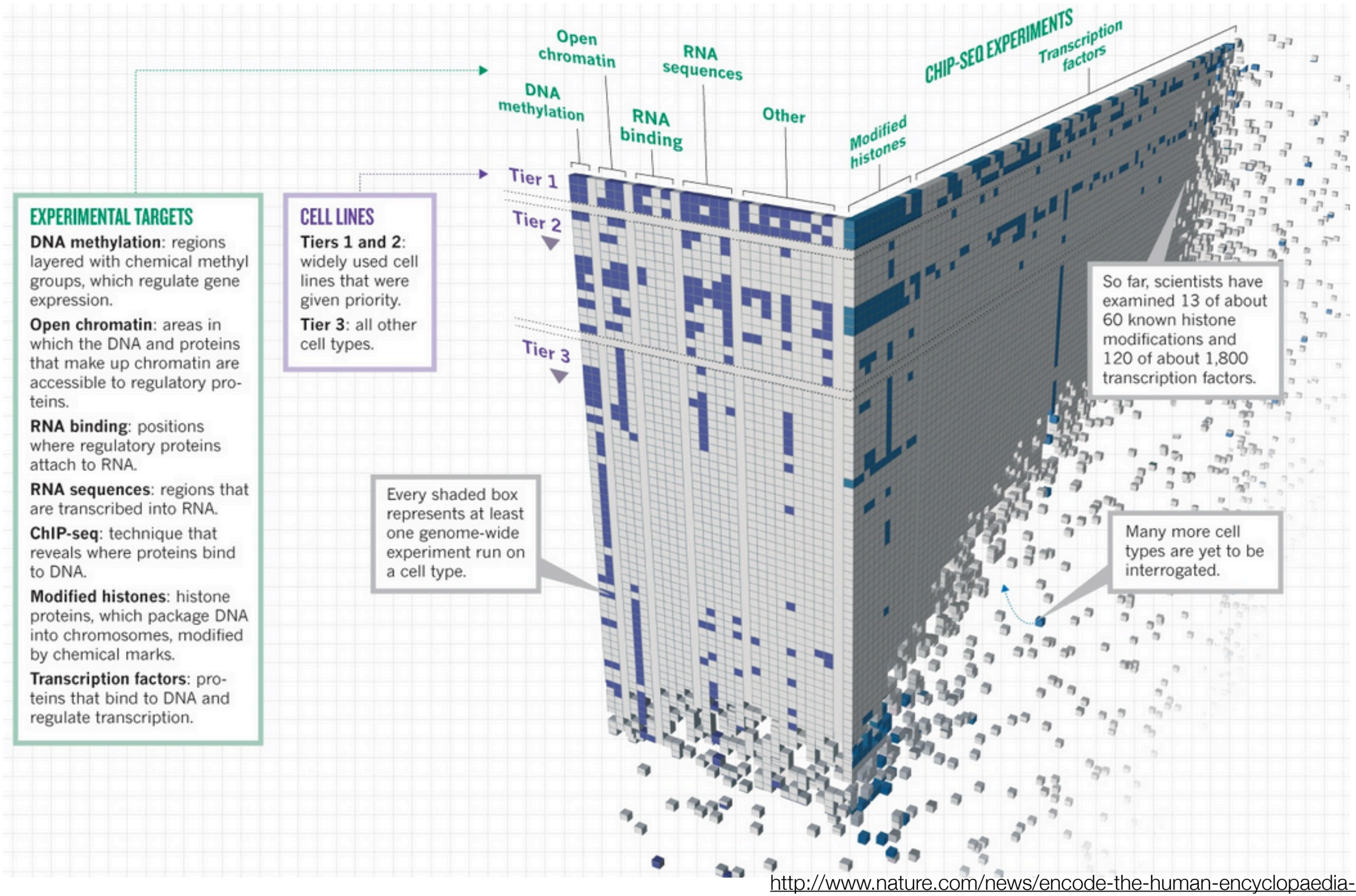


- 70000 visits in last year
- downloaded over 20000 times
- cited over 200 times in 2 years

...but bedtools and its ilk have problems

- Existing methods such as bedtools work primarily on pairs of datasets
- These methods exhaustively list the **individual details** of relationships between genome annotations and experimental datasets.
- How do we reduce the huge dimensionality to something we can understand and compare?

How do we make sense of **complex, multi-dimensional** datasets to gain insights into genome biology?



http://www.nature.com/news/encode-the-human-encyclopaedia-1.11312

“Low dimensional representations and human understanding are synonymous.”

Ewan Birney

“Things that are complex are not useful; things that are useful are simple.”

Michail Kalashnikov
(AK47s are simple)

“For every problem there is a solution which is simple, clean and wrong.”

Henry Mencken

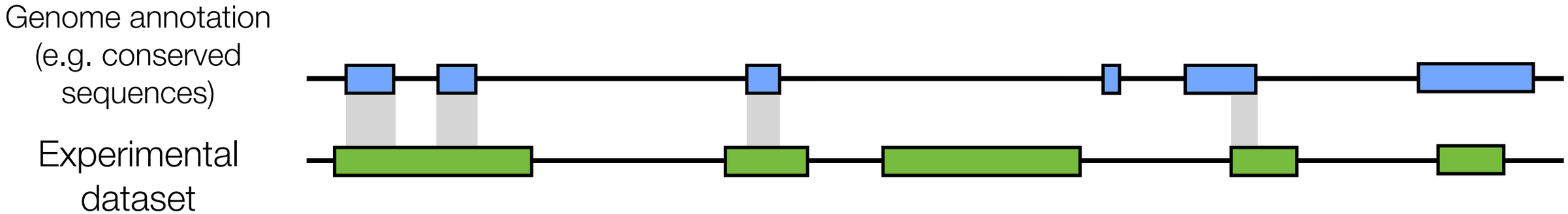
Scale
chr21:
200 kb
hg19
32,700,000
32,800,000
32,900,000
33,000,000
33,100,000
33,200,000
33,300,000
33,400,000
33,500,000
Uniqueness of 35bp Windows from ENCODE/OpenChrom(Duke)
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
TIAM1
TIAM1
TIAM1
TIAM1
BC014150
BC041449
SOD1
HUNK
LINC00159
SCAF4
SCAF4
SCAF4
SCAF4
SCAF4
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
GM12878 Syn Pk
H1-hESC Syn Pk
K562 Syn Pk
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
GM12878 Pk
GM12878 DS
GM12878 OS
H1hESC Pk
H1hESC DS
H1hESC OS
K562 Pk
K562 DS
K562 OS
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
GM12878 FAIRE Pk
GM12878 FAIRE DS
GM12878 FAIRE OS
H1-hESC FAIRE Pk
H1-hESC FAIRE DS
H1-hESC FAIRE OS
K562 FAIRE Pk
K562 FAIRE DS
K562 FAIRE OS
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
GM12878 Ht 1
GM12878 Pk 1
GM12878 Sg 1
GM12878 Ht 2
GM12878 Pk 2
GM12878 Sg 2
H1hESC Ht 1
H1hESC Pk 1
H1hESC Sg 1
K562 Ht 1
K562 Pk 1
K562 Sg 1
K562 Ht 2
K562 Pk 2
K562 Sg 2
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
K562 Sig
HepG2 Sig
HUVEC Sig

This is a hard yet important problem.

- Understand the function (or lack) of every base pair in different cell types and contexts.
- Challenges (among many):
  - Basic exploratory data analysis: slicing and dicing very large, heterogenous datasets.
  - Visualization: unbiased exploration; let the data tell its story.
  - **Testing for significant spatial relationships**

How do we detect genomic co-association?

That is, do two genomic features co-occur (overlap or have spatial consistency) more than expected?

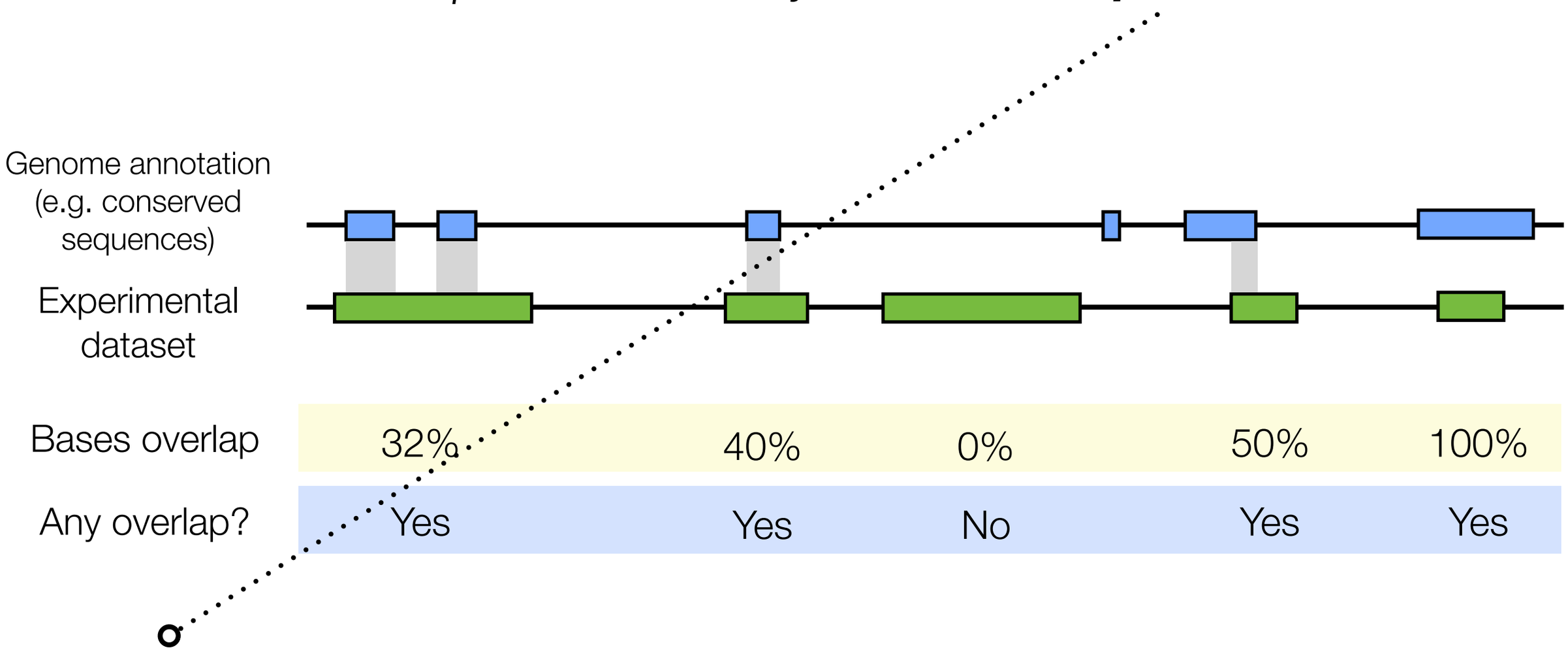


Bases overlap	32%	40%	0%	50%	100%
Any overlap?	Yes	Yes	No	Yes	Yes

Adapted from slide by Ben Brown

How do we detect genomic co-association?

That is, do two genomic features co-occur (overlap or have spatial consistency) more than ***expected****?*



How do we develop a proper null expectation in order to reduce the dimensionality of the data to an informative statistic?

Adapted from slide by Ben Brown

Monte-Carlo simulation

Are the observed feature overlaps more common than what you would expect by chance?

1. **Observed.** Count the number of overlaps between the two sets of features (e.g. ChIP peaks for TF 1 v. TF 2)
2. **Expected.**
    2.1. Randomly reassign the features in each set to new genomic locations.
    2.2. Count the number of overlaps.
    2.3. Repeat 1000 or more times. **SLOW!**
3. **Return either:**
    3.1. **P-value:** how many times were the shuffled overlaps > observed?
        3.1.1. e.g., if answer is 3 and there were 1000 simulations, P = 3e-3
    3.2. **Enrichment score:** log2(observed / median expected)

How do we speed this up?

We (Ryan) invent a new algorithm, of course!

Binary Interval Search (BITS):
A Scalable Algorithm for Counting Interval Intersections

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3*]

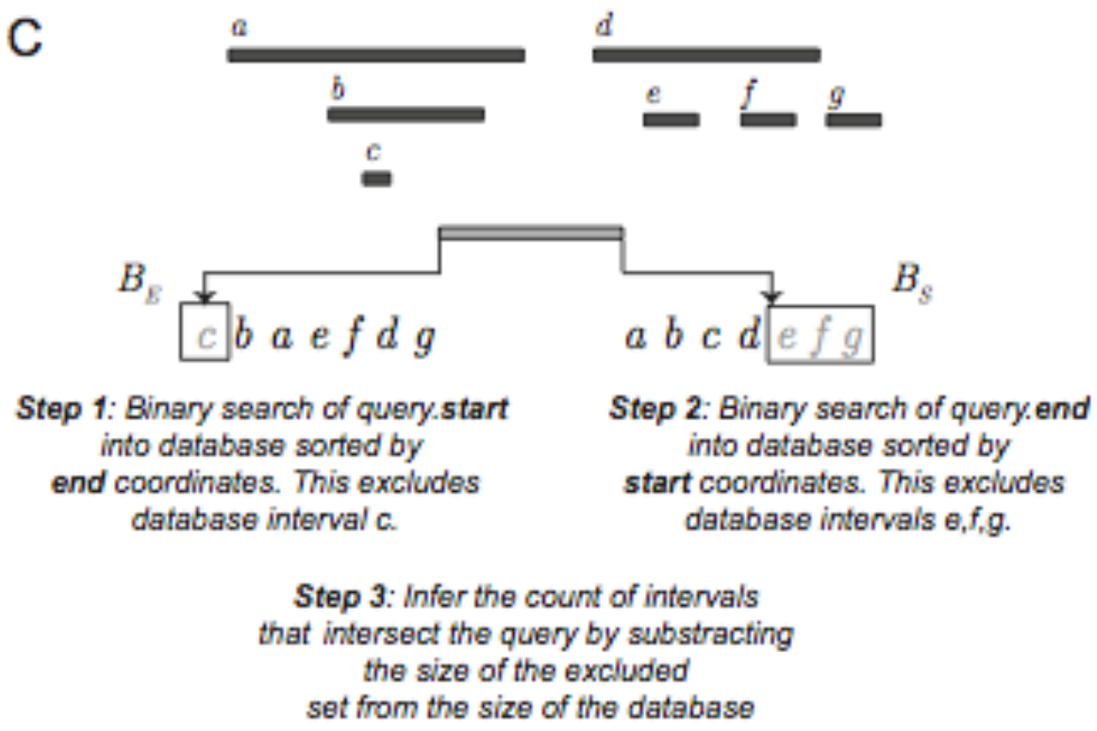
[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA

Ryan Layer

https://github.com/arq5x/bits

Advance access at Bioinformatics

Binary InTerval Search BITS



***Step 1**: Binary search of query.**start** into database sorted by **end** coordinates. This excludes database interval c.*

***Step 2**: Binary search of query.**end** into database sorted by **start** coordinates. This excludes database intervals e,f,g.*

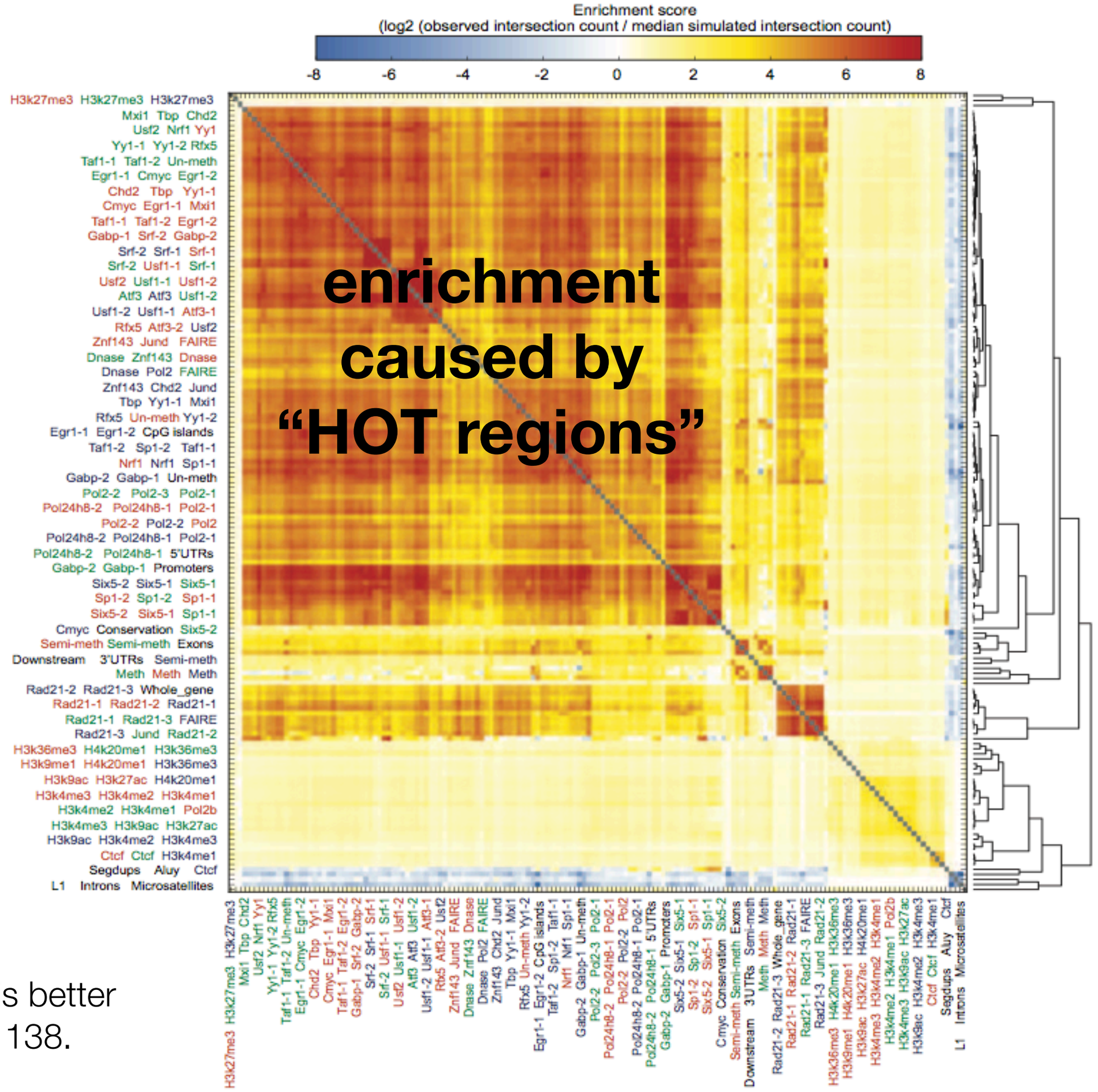
***Step 3**: Infer the count of intervals that intersect the query by substracting the size of the excluded set from the size of the database*

Algorithm 1: Single interval intersection counter

Input: Sorted interval starts and ends B_S and B_E, query interval a
Output: Number of intervals c intersecting a

Function ICOUNT(B_S, B_E, a) **begin**
    $first \leftarrow$ BINARYSEARCH($B_S, a.end$)
    $last \leftarrow$ BINARYSEARCH($B_E, a.start$)
    $c \leftarrow first - last$ /* $= |B| - (last + (|B| - first))$ */
    return c

- Entirely novel algorithm for detecting genomic interval intersections.
- Clever aspect: unlike any other algorithm, it can deduce the **count** of overlaps without having to **enumerate** each individual intersection.
- Uses two binary searches. Very fast. ***Spaceballs* fast on a GPU**.
- If you haven't heard, faster is better.



6 days is better
than 138.

How do we develop a proper null expectation?

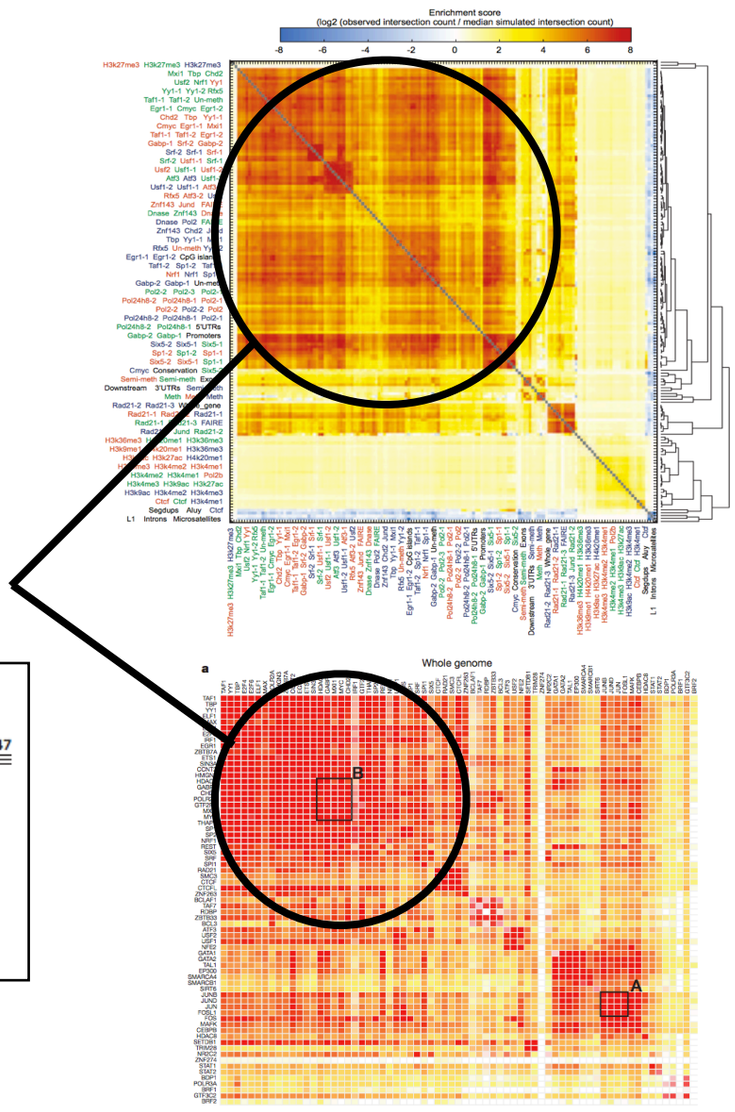
1. Monte-Carlo simulation

Binary Interval Search (BITS):
A Scalable Algorithm for Counting Interval Intersections

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3]*

[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA

"HOT regions"
Gerstein et al; modENCODE; 2010



2. Block bootstrap sampling (Peter Bickel)

ARTICLE
doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

NON PARAMETRIC METHODS FOR GENOMIC INFERENCE

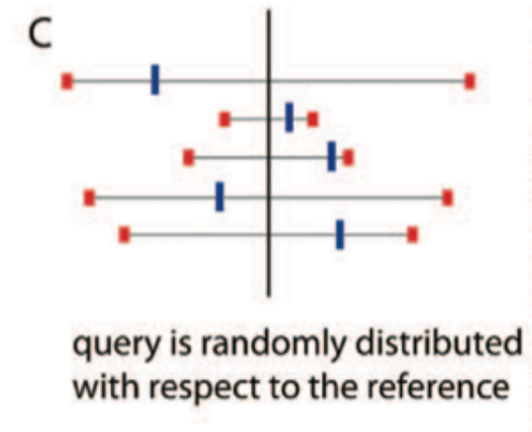
BY PETER J. BICKEL*,†, NATHAN BOLEY*,†, JAMES B. BROWN*,†, HAIYAN HUANG*,†, NANCY R. ZHANG*,‡

3. Spatial methods (Sarah Wheelan)

Exploring Massive, Genome Scale Datasets with the GenometriCorr Package

Alexander Favorov[1,2,3]*, Loris Mularoni[1]¤a, Leslie M. Cope[1], Yulia Medvedeva[2,3]¤b, Andrey A. Mironov[4,5], Vsevolod J. Makeev[2,3], Sarah J. Wheelan[1]*

1 Department of Oncology, Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **2** Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia, **3** Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia, **4** Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, **5** Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia



Integrating these and other, novel association statistics into ***bedtools2***

What is the role of **conserved, yet non-coding** (CNC) elements?

ARTICLE

doi:10.1038/nature10530

A high-resolution map of human evolutionary constraint using 29 mammals

4% of the genome is under constraint

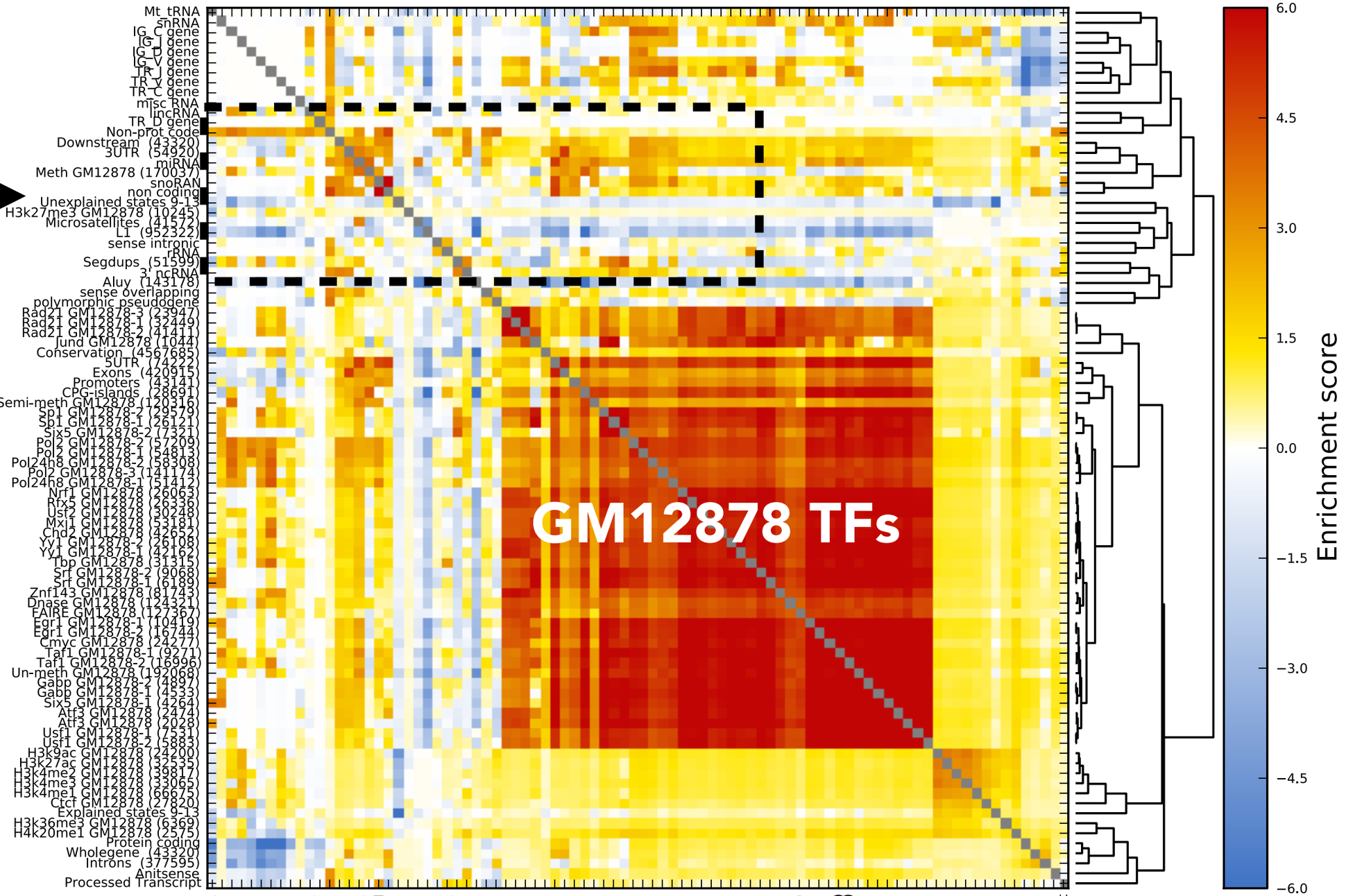
60% coding or regulatory (e.g., enhancers)

40% is ***unexplained.***

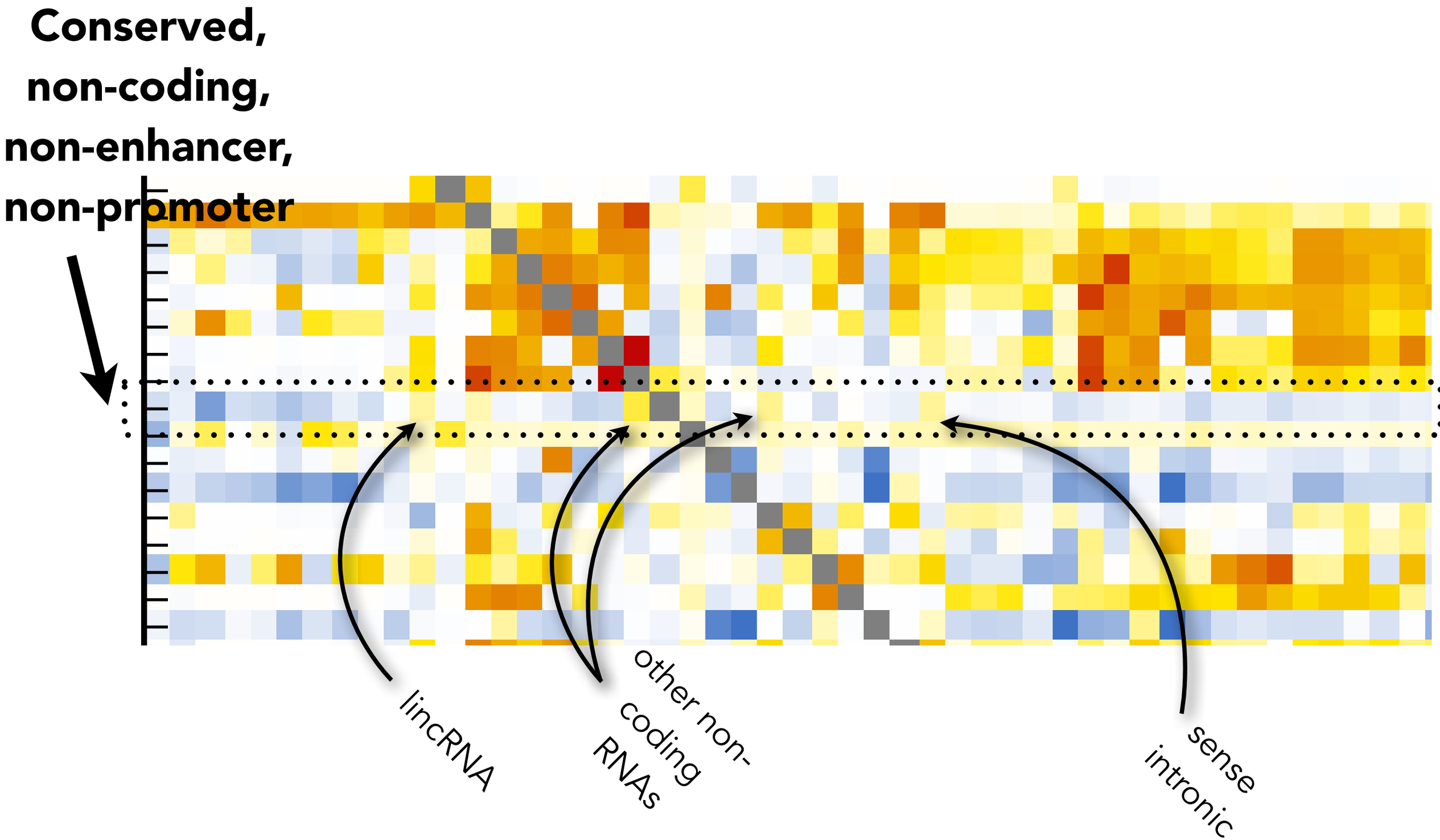
John Kubinksi
Undergraduate
(Biology)

What is the role of conserved, non-coding (CNC) elements?

CNC elements



Making sense of conserved, yet non-coding, and non-"functional" elements



2. Genome Query Language (GQL)

Genome Query Language (GQL)

(coming to your lab in 2013)

- All new, simple, & expressive language for exploring complex datasets. Inspired by simplicity of SQL.
- Set theory, correlations, statistics, visualization
- **Abstract the language from the engine and file management.**
  - Parallel. Automated. No file tracking.
  - Meant for both simple & massive/complex analyses.
  - Laptop, cloud, clusters, grid, etc.
- **Goals**: standardized, powerful, fast, facile. Empower biologists.
- **Explore 1000s of datasets with a single command.**

Ryan Layer

https://github.com/ryanlayer/gql

Neil Kindlon

Genome Query Language (GQL)

Datasets

BED3
BED4
BED6
BED12
BAM
VCF
BCF
GTF
GFF
BEDGRAPH
TABIX

Attributes

SCORE
NAME
CHROM
START
END
STRAND
GENOME

Operations

INTERSECT
SUBTRACT
MERGEMAX
MERGEFLAT
MERGEMIN
FILTER
LOAD (∞ files or directories)
CAST
PRINT
SAVE
COUNT
PEAK
COMPLEMENT
PLOT
FETCH
MEASURE

Quickly retrieve datasets
from UCSC, Ensemble, etc.
Also our own hosted annotations.

genes = FETCH ucsc.hg19.gencode

Stat. Tests

JACCARD
MONTECARLO
GSC
RELATIVE_DIST
ABSOLUTE_DIST

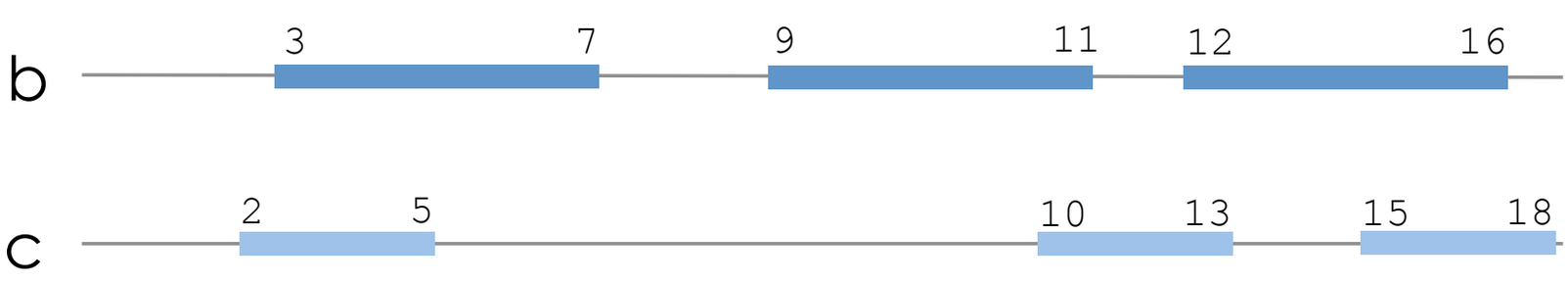
Meas'mnts

DISTANCE
MIN
SUM
MAX
MEAN
MEDIAN
MODE
ANTIMODE
COLLAPSE
STDEV

A GQL example

(find all intervals where at least 2 datasets overlap)

(e.g., find "hot regions" for TF ChIP-seq peaks)



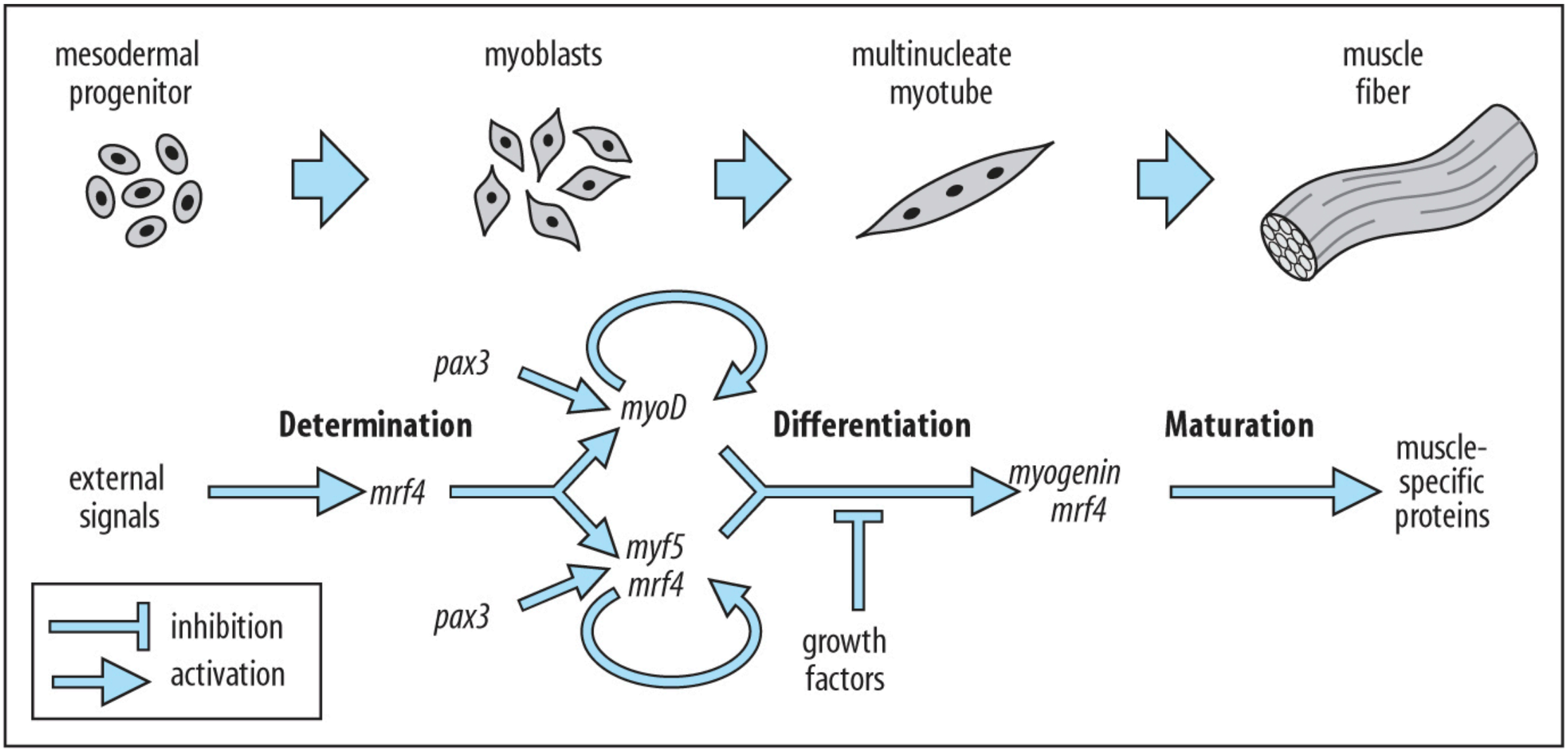

```
chr1  1    4    a1   1    +
chr1  6    8    a2   2    +
chr1  14   17   a3   3    +
```

```
chr1  3    7    b1   1    -
chr1  9    11   b2   2    +
chr1  12   16   b3   3    +
```

```
chr1  2    5    c1   1    +
chr1  10   13   c2   2    +
chr1  15   18   c3   3    +
```

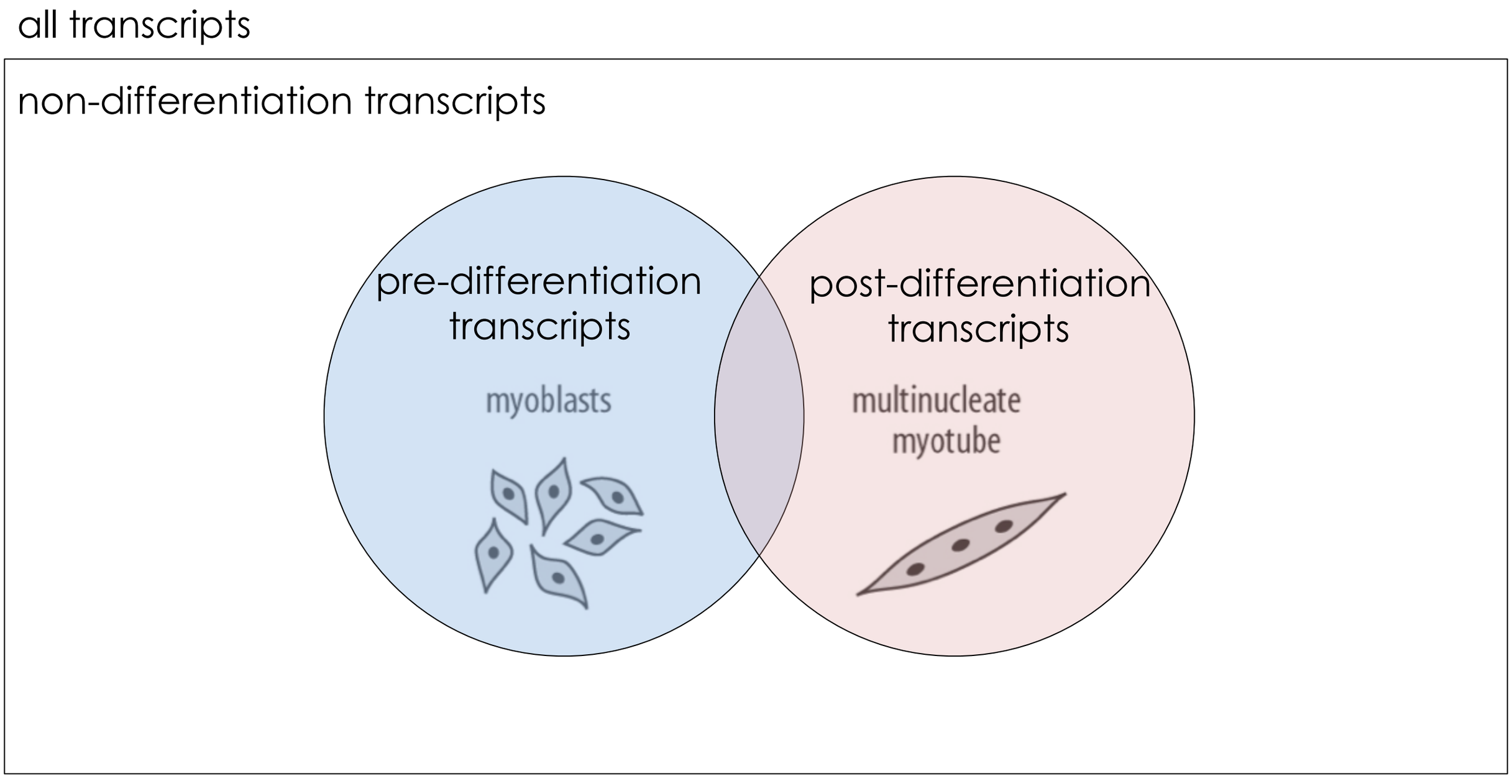
```
> result = SELECT a,b,c WHERE COUNT(>1);
> PRINT result
> mc = MEASURE result WITH(MONTE_CARLO) AND a,b,c
> PLOT mc
```

GQL Example 2: Differentiation in Muscle Cells

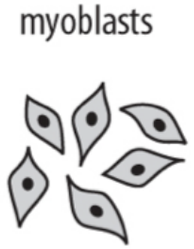


http://www.mun.ca/biology/desmid/brian/BIOL3530/DEVO_10/ch10f19.jpg

Do Novel lncRNAs Have a Role in Differentiation?



Pre-/Post-Transcription Data Sets



Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms

Cole Trapnell[1,2,5], Brian A. Williams[3], Geo Pertea[2], Ali Mortazavi[3], Gordon Kwan[3], Marijke J. van Baren[4], Steven L. Salzberg[1,2], Barbara J. Wold[3], and Lior Pachter[5,6,7]

MB_rnaseq

MT_rnaseq

Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq

Hao Sun[1], Jiejun Wu[2], Priyankara Wickramasinghe[1], Sharmistha Pal[1], Ravi Gupta[1], Anirban Bhattacharyya[1], Francisco J. Agosto-Perez[1], Louise C. Showe[1], Tim H.-M. Huang[2] and Ramana V. Davuluri[1,*]

MB_polII

MT_polII

Known Transcript Data Sets

UCSC Genome Bioinformatics

Known Exons	mm9_exons
Known Transcription Start Sites (+/- 5KB Genes)	mm9_5kb_tss
Known miRNA	mm9_mirna

Transcript Data Sets

mm9_exons
mm9_5kb_tss
mm9_mirna

all transcripts

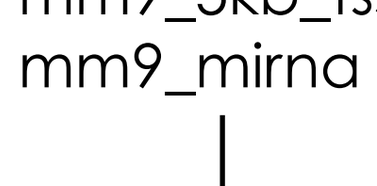
non-differentiation transcripts

pre-differentiation transcripts

myoblasts

post-differentiation transcripts

multinucleate myotube



MT_rnaseq
MT_polII

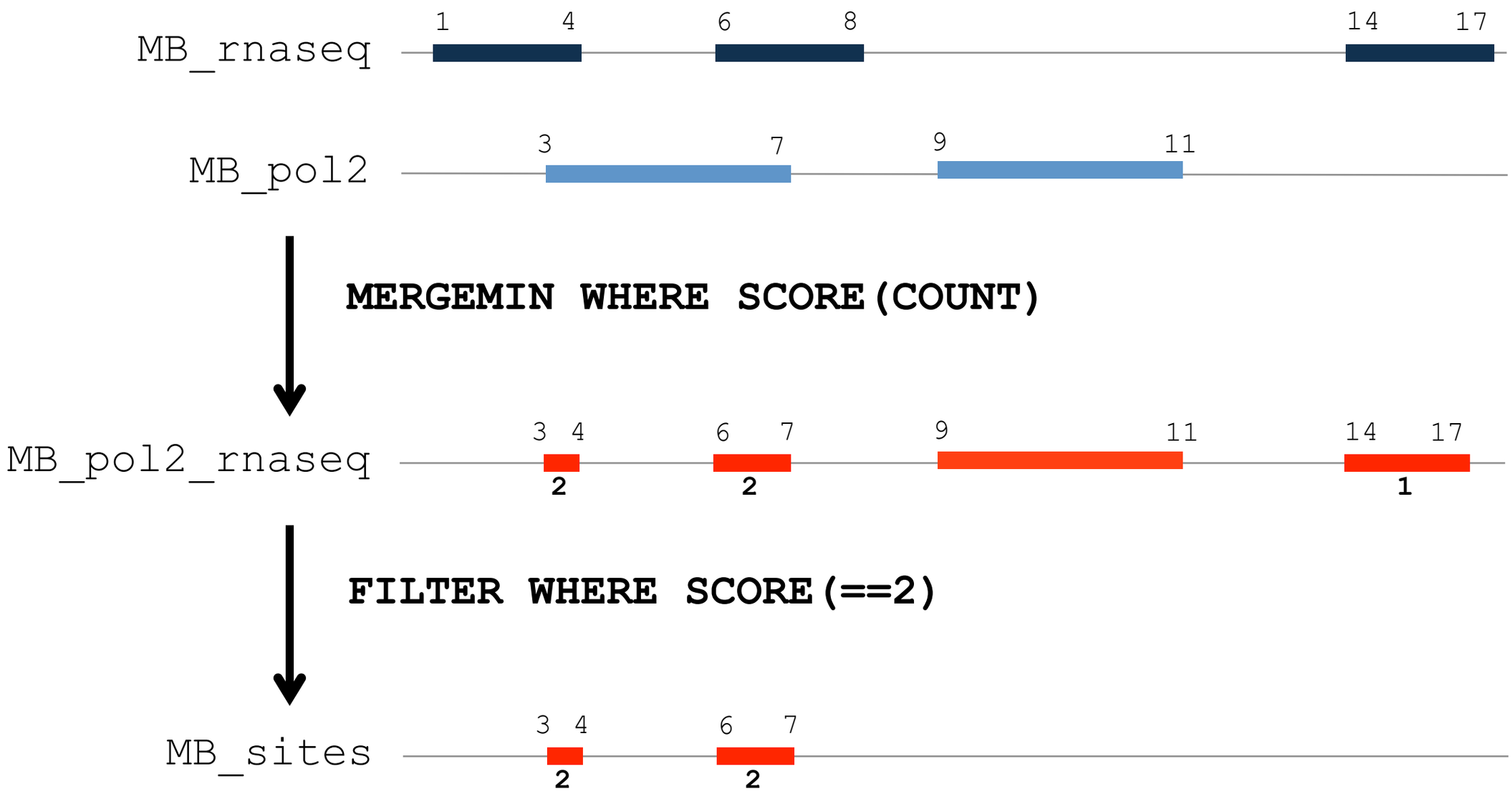
MB_rnaseq
MB_polII

LOAD Data Sets

```
# Known transcriptions
mm9_exons       = LOAD "data/mm9_exons.bed";
mm9_miRNA       = LOAD "data/mm9_miRNA.bed";
mm9_5KB_tss     = LOAD "data/mm9_tss_5KB.bed";

# Pre-Differentiation Data
MB_rnaseq       = LOAD "data/MB_rnaseq.bed";
MB_pol2         = LOAD "data/PolII_MB-Sonicated_input_MB.bed";

# Post-Differentiation Data
MT_rnaseq       = LOAD "data/MT_rnaseq.bed";
MT_pol2         = LOAD "data/PolII_MT-Sonicated_input_MT.bed";
```

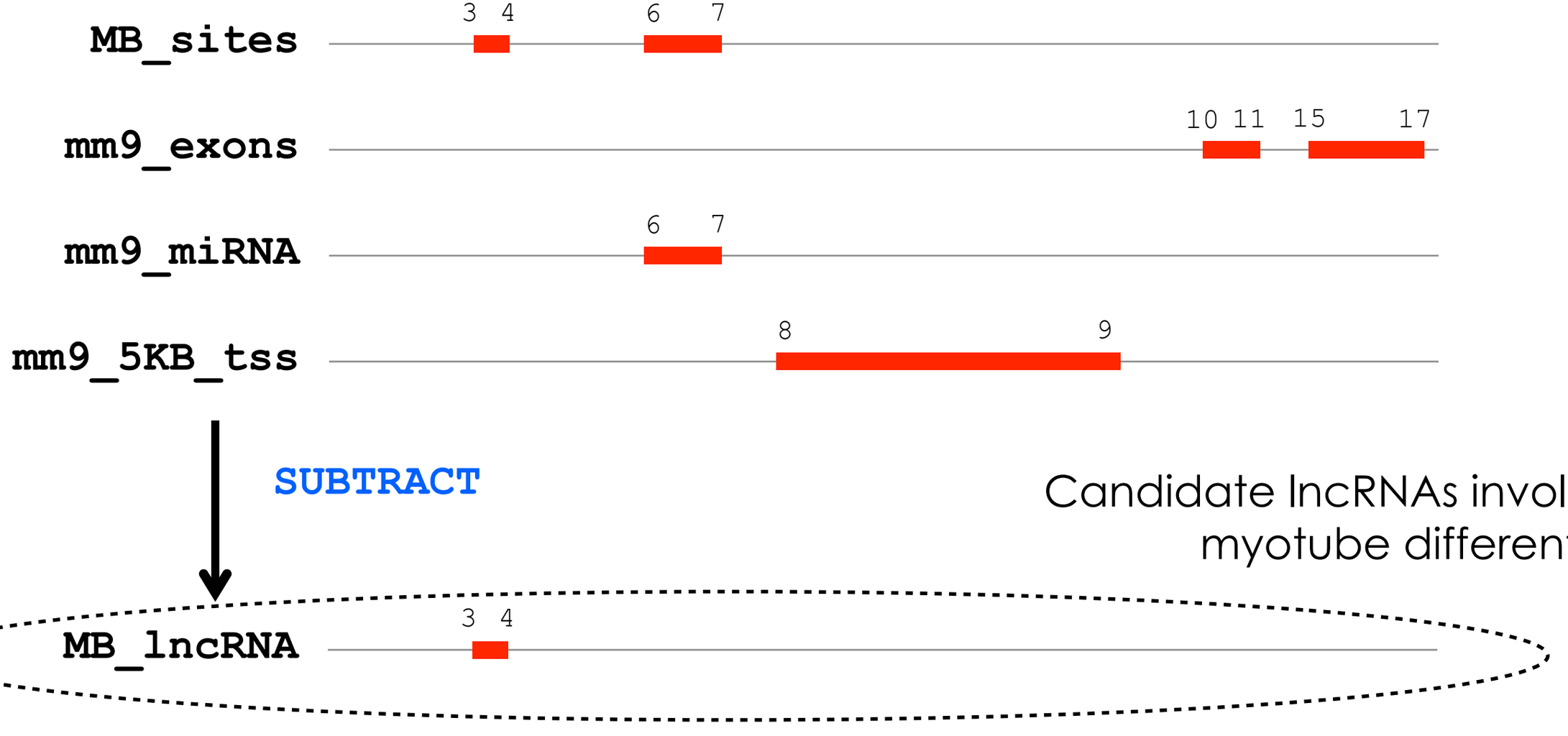
Find Pre/Post-Transcription Sites with Pol II signal

```
# Pre-Differentiation Data
MB_pol2_rnaseq = MERGEMIN MB_rnaseq,MB_pol2 WHERE SCORE(COUNT);
MB_sites       = FILTER MB_pol2_rnaseq WHERE SCORE(==2);

# Post-Differentiation Data
MT_pol2_rnaseq = MERGEMIN MT_rnaseq,MT_pol2 WHERE SCORE(COUNT);
MT_sites       = FILTER MT_pol2_rnaseq WHERE SCORE(==2);
```



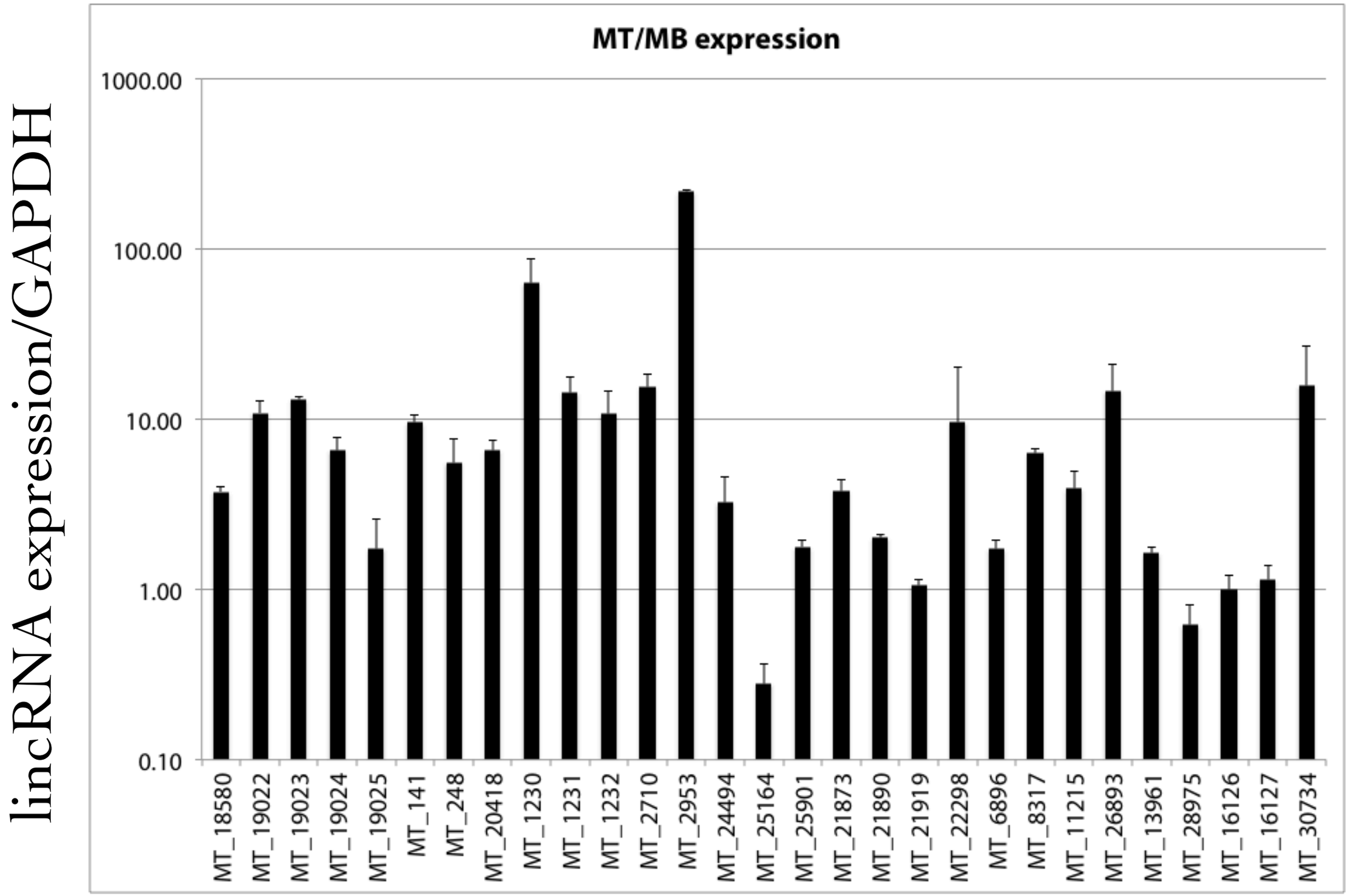
Remove Known Transcripts to leave putative lncRNAs

```
# Pre-Differentiation Predictions
SAVE MB_lncRNA AS "MB_lncRNA.bed";

# Post-Differentiation Predictions
SAVE MT_lncRNA AS "MT_lncRNA.bed";
MT_lncRNA = MT_sites SUBTRACT
                     MB_sites,mm9_exons,mm9_miRNA,mm9_5KB_tss;
```



Candidate lncRNAs involved in myotube differentiation

Validated MT lncRNA candidates identified by Dutta lab



courtesy of Adam Mueller and Anindya Dutta

WANTED: collaborators to test GQL and to use it in active research projects.

3. Gemini:

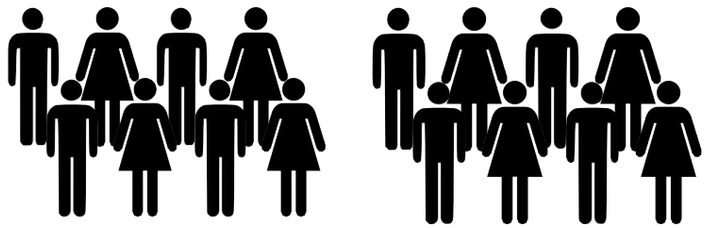
a flexible framework for mining genome variation

Uma Paila

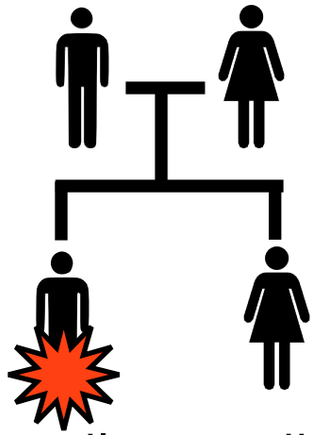
Typical sequencing studies of disease

phenotype

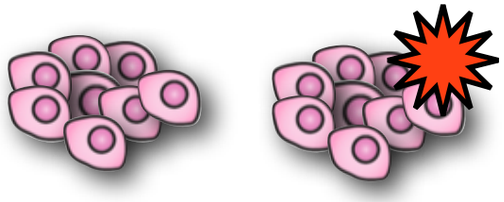
no phenotype



case / control



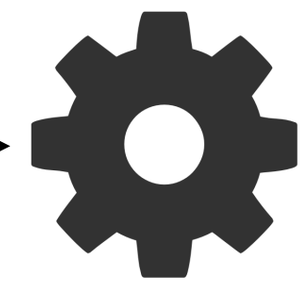
family studies



normal tumor

cancer genomics

Genome sequencing



Align sequences
Scrub data
Call variants

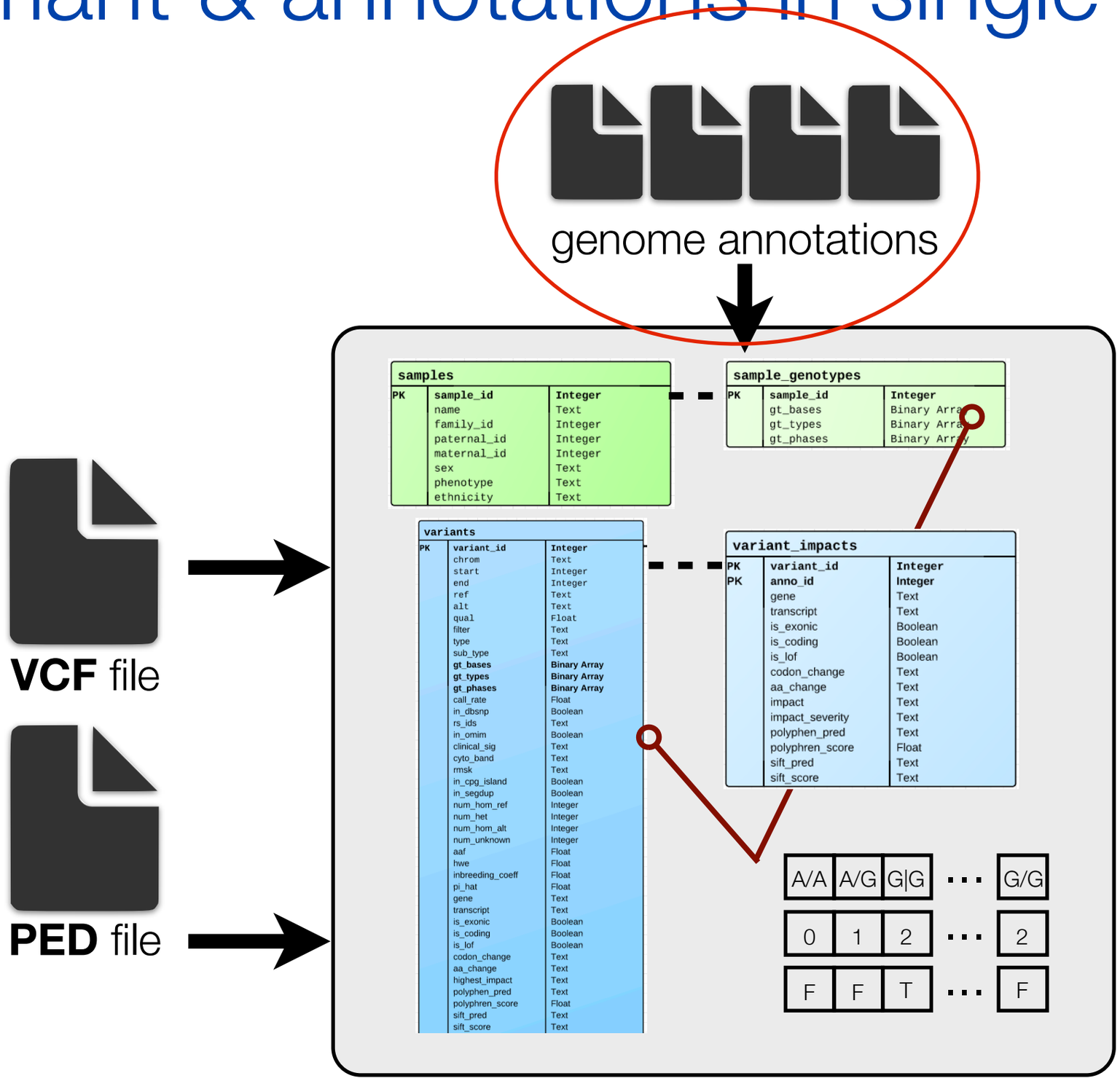
Analysis

Gemini overview

we have genetic variants. now what?

- **Goal**: unified framework for exploring genetic variation for disease and population genetics.
- Samples ➤ Genomes ➤ Variants ➤ VCF ➤ Gemini ➤ Test **H**
- **Annotate w/ extensive genome annotations**
- Structured querying interface
- Extensible for new tool development
- Powerful, yet easy to use

Variant & annotations in single framework



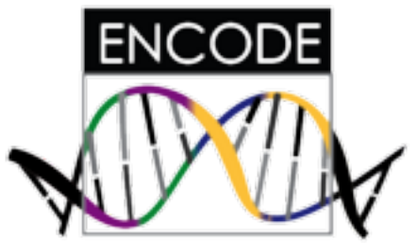
Compressed arrays of genotype information accommodates 1000s of genotypes per variant.

```
$ gemini load -v my.vcf -t VEP my.db
```

Genome annotations place variants in context

OMIM

CpG
Repeats
SegDups
Conservation
(more)



DNase
FAIRE
TF Binding
chromatin states

Recombination

COSMIC

dbSNP

ESP

pathways

protein interactions

Variants, genotypes, & sample info all in a single database

samples		
PK	**sample_id**	**Integer**
	name	Text
	family_id	Integer
	paternal_id	Integer
	maternal_id	Integer
	sex	Text
	phenotype	Text
	ethnicity	Text

sample_genotypes		
PK	**sample_id**	**Integer**
	gt_bases	Binary Array
	gt_types	Binary Array
	gt_phases	Binary Array

variants		
PK	**variant_id**	**Integer**
	chrom	Text
	start	Integer
	end	Integer
	ref	Text
	alt	Text
	qual	Float
	filter	Text
	type	Text
	sub_type	Text
	gt_bases	**Binary Array**
	gt_types	**Binary Array**
	gt_phases	**Binary Array**
	call_rate	Float
	in_dbsnp	Boolean
	rs_ids	Text
	in_omim	Boolean
	clinical_sig	Text
	cyto_band	Text
	rmsk	Text
	in_cpg_island	Boolean
	in_segdup	Boolean

variant_impacts		
PK	**variant_id**	**Integer**
PK	**anno_id**	**Integer**
	gene	Text
	transcript	Text
	is_exonic	Boolean
	is_coding	Boolean
	is_lof	Boolean
	codon_change	Text
	aa_change	Text
	impact	Text
	impact_severity	Text
	polyphen_pred	Text
	polyphren_score	Float
	sift_pred	Text
	sift_score	Text

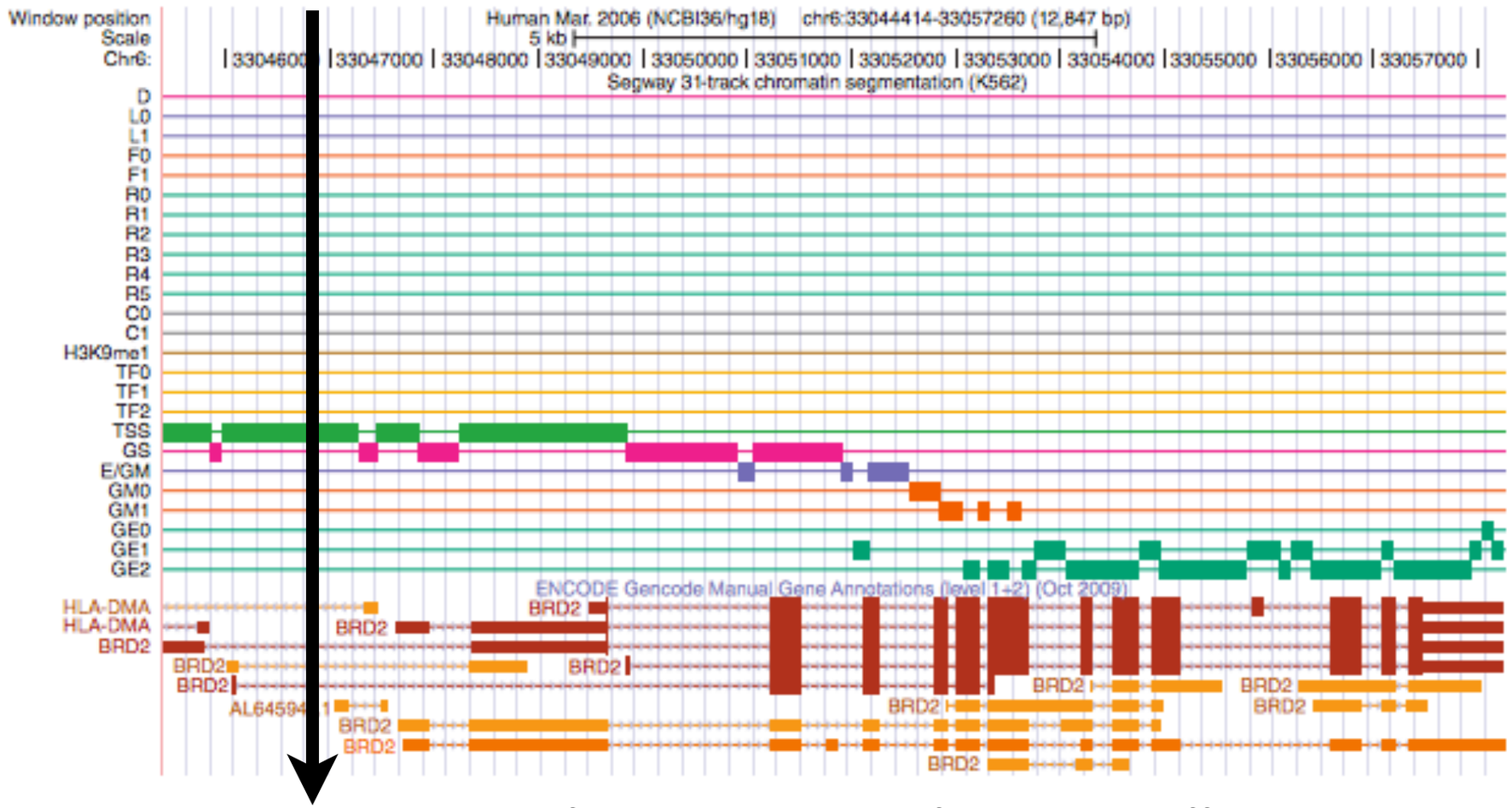
How are the annotations useful?

More nuanced interpretation of **coding** variation.

- Is it benign of loss-of-function (LoF)?
- *Not all LoF variants are created equal*
- Does it have clinical significance?
- Does it form a compound heterozygote?
- What's the freq. in 1000G? ESP?
- With what proteins does the altered gene interact?
- In what pathways is the gene involved?

How are the annotations useful?

Assessing the relevance of **non-coding** variation.



What is the role of this segment of DNA in different tissues?

TSS?

Enhancer?

Promoter?

Silencer?

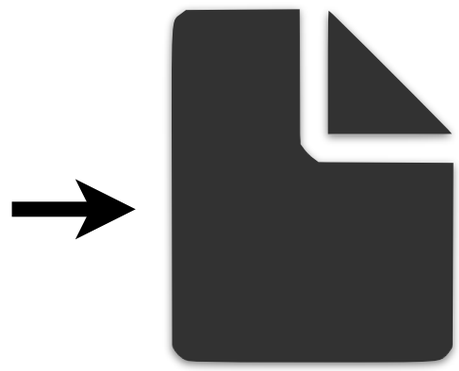
gemini incorporates ENCODE chromatin state maps

So what can we do with gemini?

Ad hoc data exploration

Enhanced SQL engine allows **selection** and **filtering** based on individual genotypes. **Scales to 1000s of samples.**

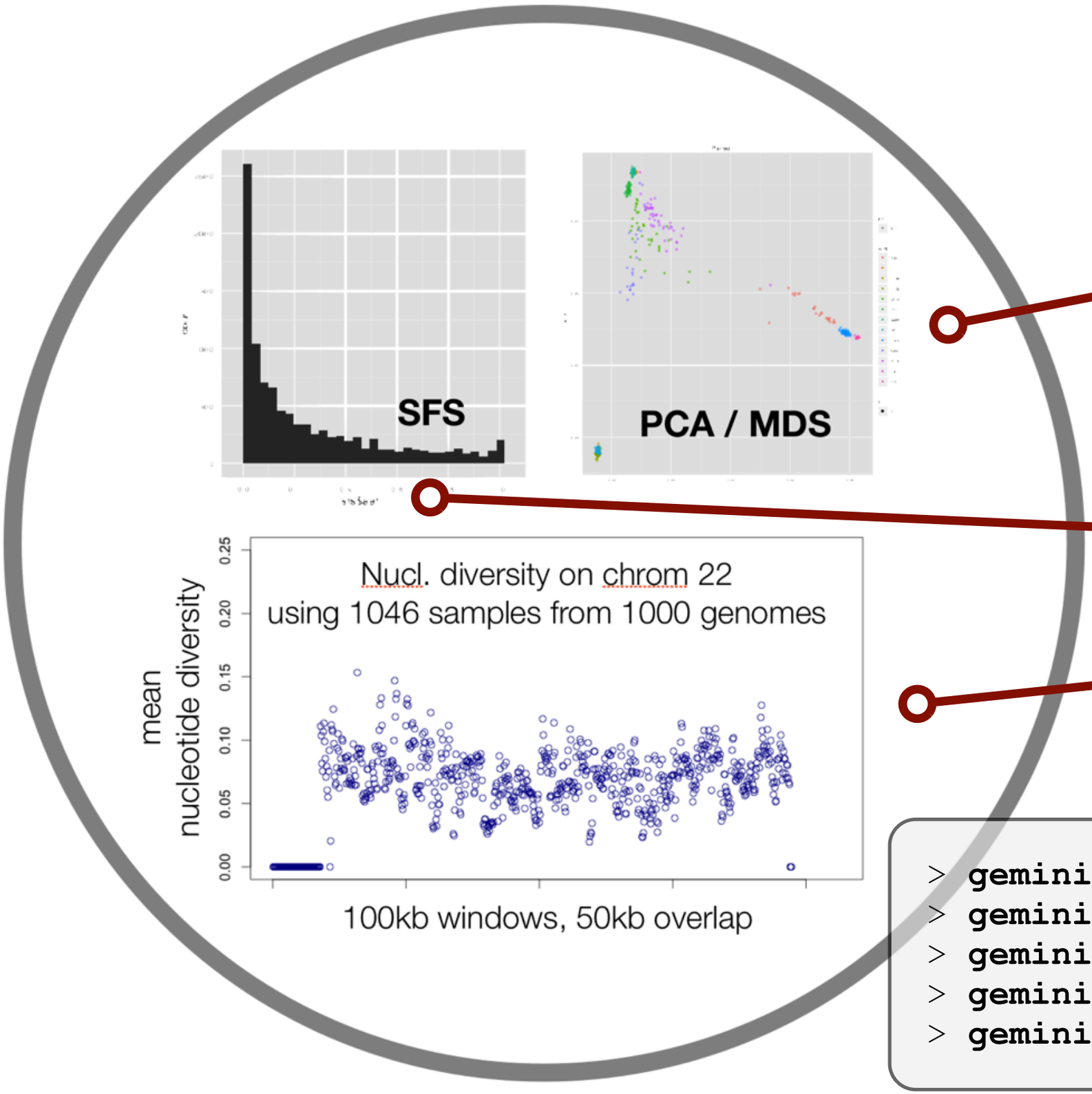
```
SELECT chrom,
start, gene,
impact, hwe,
in_dbsnp, in_omim,
gts.NA12878
FROM variants
WHERE is_lof = 1
AND aaf <= 0.01
AND call_rate \
    >= 0.95
```



Output files in standard formats (BED, etc.)

```
> gemini query -q [QUERY] my.db
```

Population genetics



Multidimensional scaling,
Kinship coefficients
PCA

Variant QC and profiling

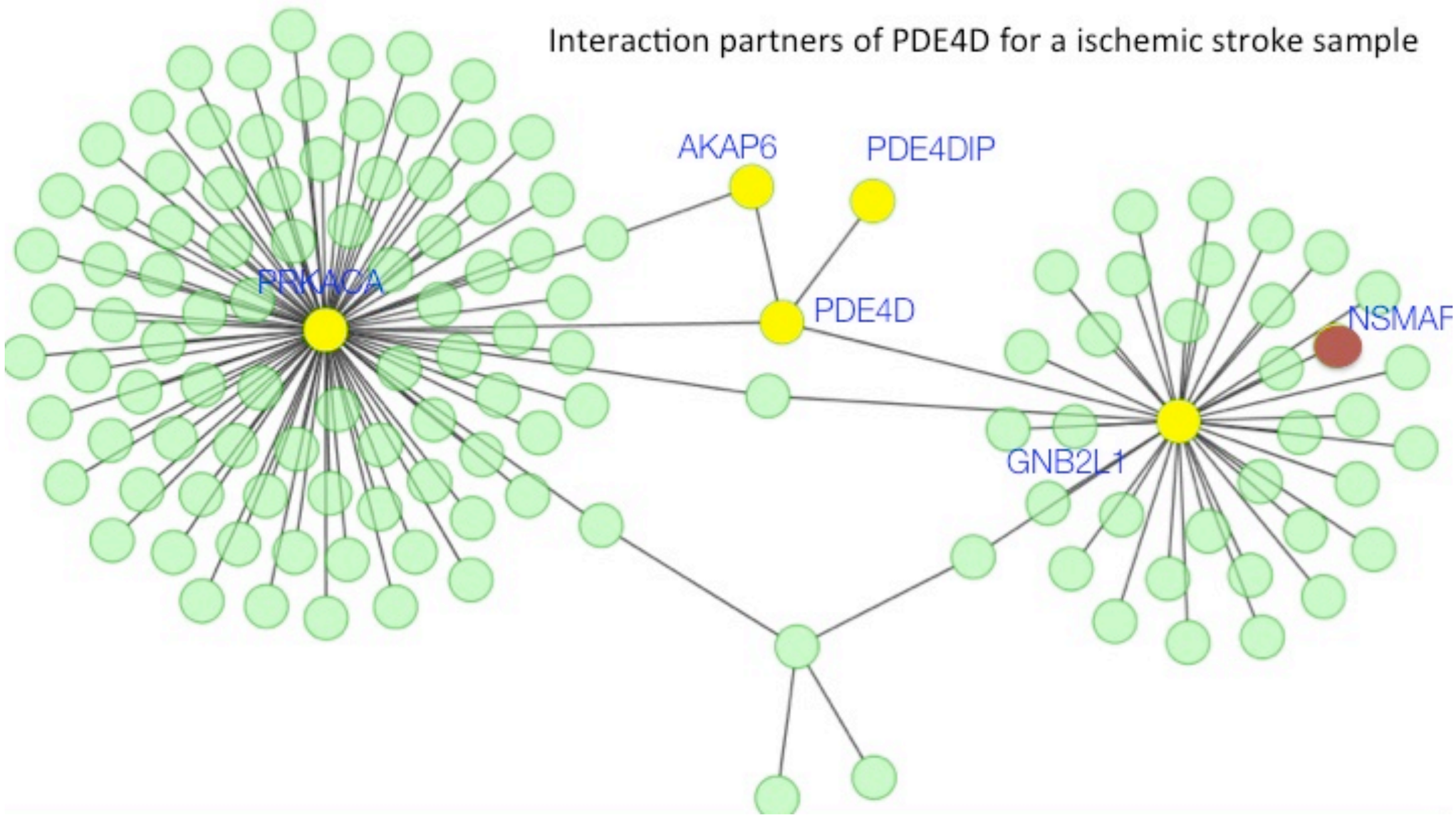
“windowing” tools

```
> gemini stats --sfs my.db
> gemini stats --mds my.db
> gemini stats --tstv my.db
> gemini stats --vars-by-sample my.db
> gemini stats --gts-by-sample my.db
```

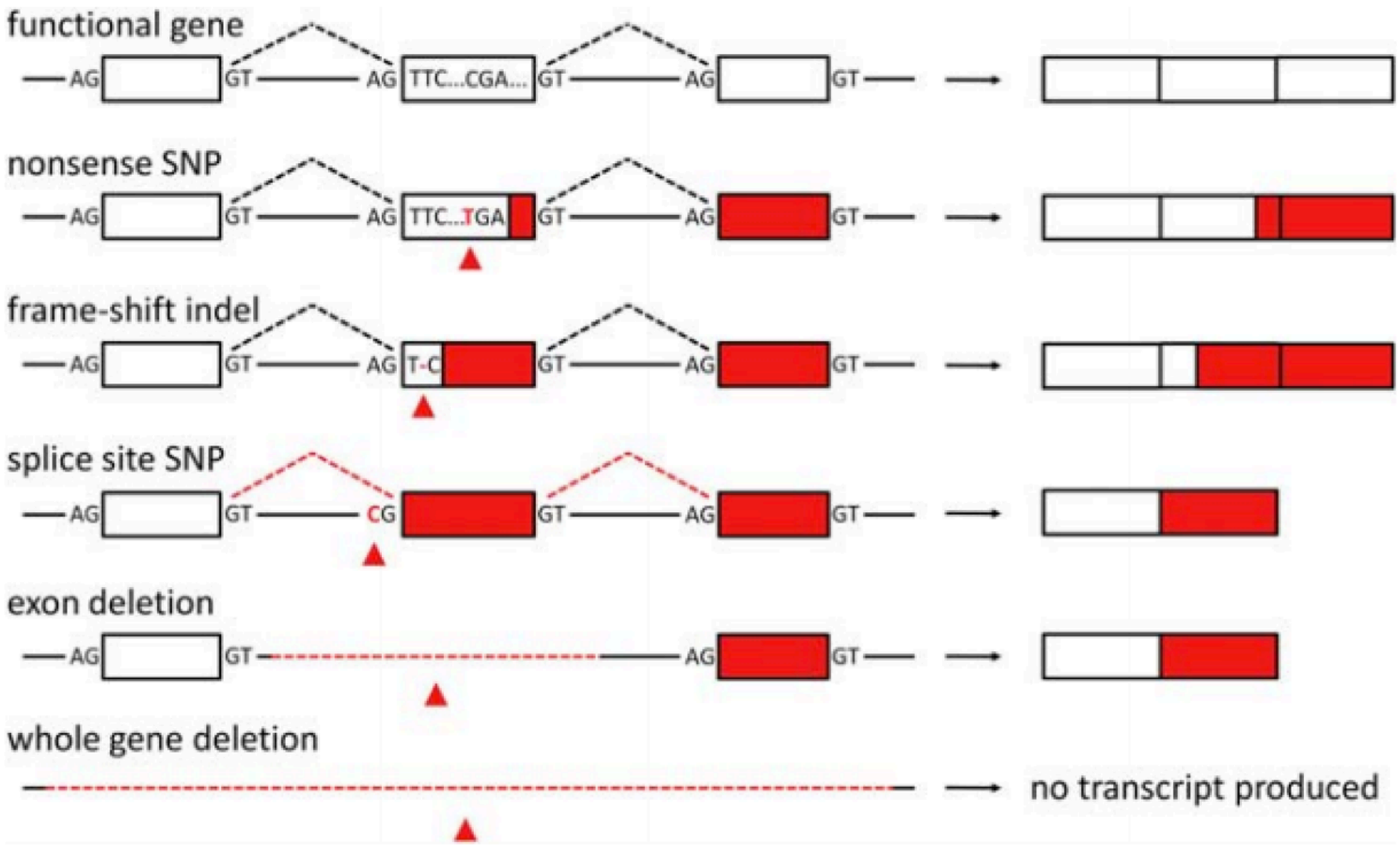
Pathway and interaction analysis

pathways

protein interactions



Loss-of-function (LoF) variants



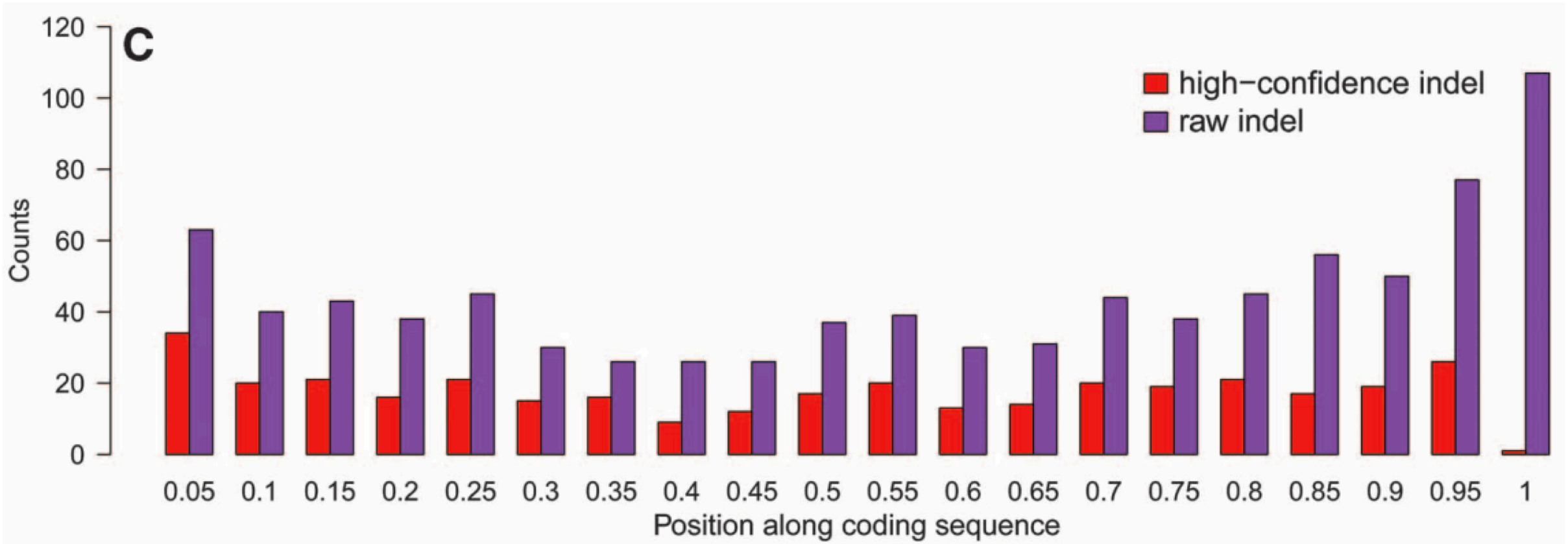
Loss-of-function variants in the genomes of healthy humans

Daniel G. MacArthur* and Chris Tyler-Smith

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 **R125–R130**
doi:10.1093/hmg/ddq365
Advance Access published on August 30, 2010

Loss-of-function (LoF) variants

Not all LoF variants are created equal.



1. Consider the position of the LoF variant in the polypeptide.
2. Test whether a second frameshift restores reading frame.

A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes
Daniel G. MacArthur *et al.*
Science **335**, 823 (2012);
DOI: 10.1126/science.1215040

Burden tests for rare variants

- GWAS based on the notion that common variants underly common disease.
- There are many, many more rare variants than common
- Standard, GWAS-like single-locus tests will be underpowered unless there is a huge effect size.
- Burden tests and Collapsing methods.
  - C-alpha, SKAT, KBAT, Morris-Zeggini, Hotelling's T, ...
  - Active area of research. Gemini is a common framework.

~Simple and extensible.

Framework for new tool development

```python
# gemini imports
import gemini_utils as util

def my_tool(c, args):
    """
    Execute a query against the gemini database and
    conduct a custom analysis on the results
    """
    # build and execute the relevant query against the da
    query = "SELECT * FROM variants \
             WHERE is_coding = 1"
    c.execute(query)

    # loop through the results.
    for row in c:
        gt_types  = np.array(unpack(row['gt_types'])))
        gt_phases = np.array(unpack(row['gt_phases'])))
        gt_bases  = np.array(unpack(row['gts'])))

        # MAGIC HAPPENS HERE

def run(parser, args):
    if os.path.exists(args.db):
        conn = sqlite3.connect(args.db)
        conn.isolation_level = None
        conn.row_factory = sqlite3.Row
        c = conn.cursor()
```

Development in Python with C and C++ code for heavy lifting.

1. Issue a query against DB
2. Iterate through results
3. Work with genotypes
4. Your genius.

```
> gemini my_tool -arg1 -arg2 my.db
```

Case study: Miller syndrome

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng[1,10], Kati J Buckingham[2,10], Choli Lee[1], Abigail W Bigham[2], Holly K Tabor[2,3], Karin M Dent[4], Chad D Huff[5], Paul T Shannon[6], Ethylin Wang Jabs[7,8], Deborah A Nickerson[1], Jay Shendure[1] & Michael J Bamshad[1,2,9]

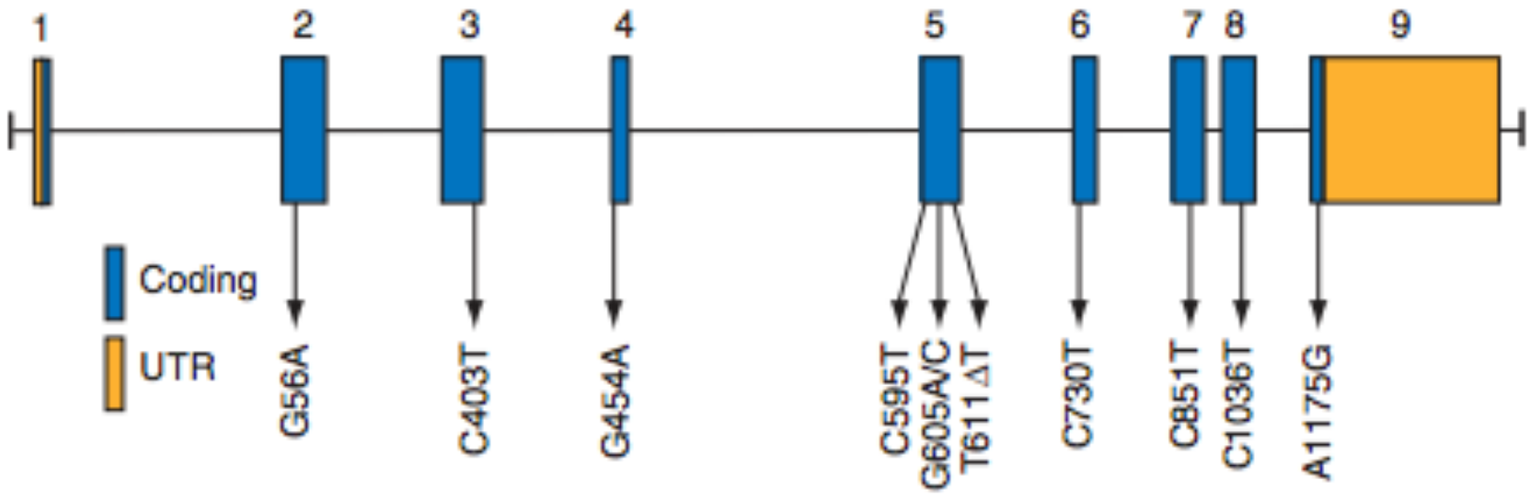
- Exome sequencing 2 siblings with Miller syndrome and 2 unrelated individuals with Miller syndrome (three kindreds total).

Ng et al, 2010 Nature Genetics

How to find the causal gene?

Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

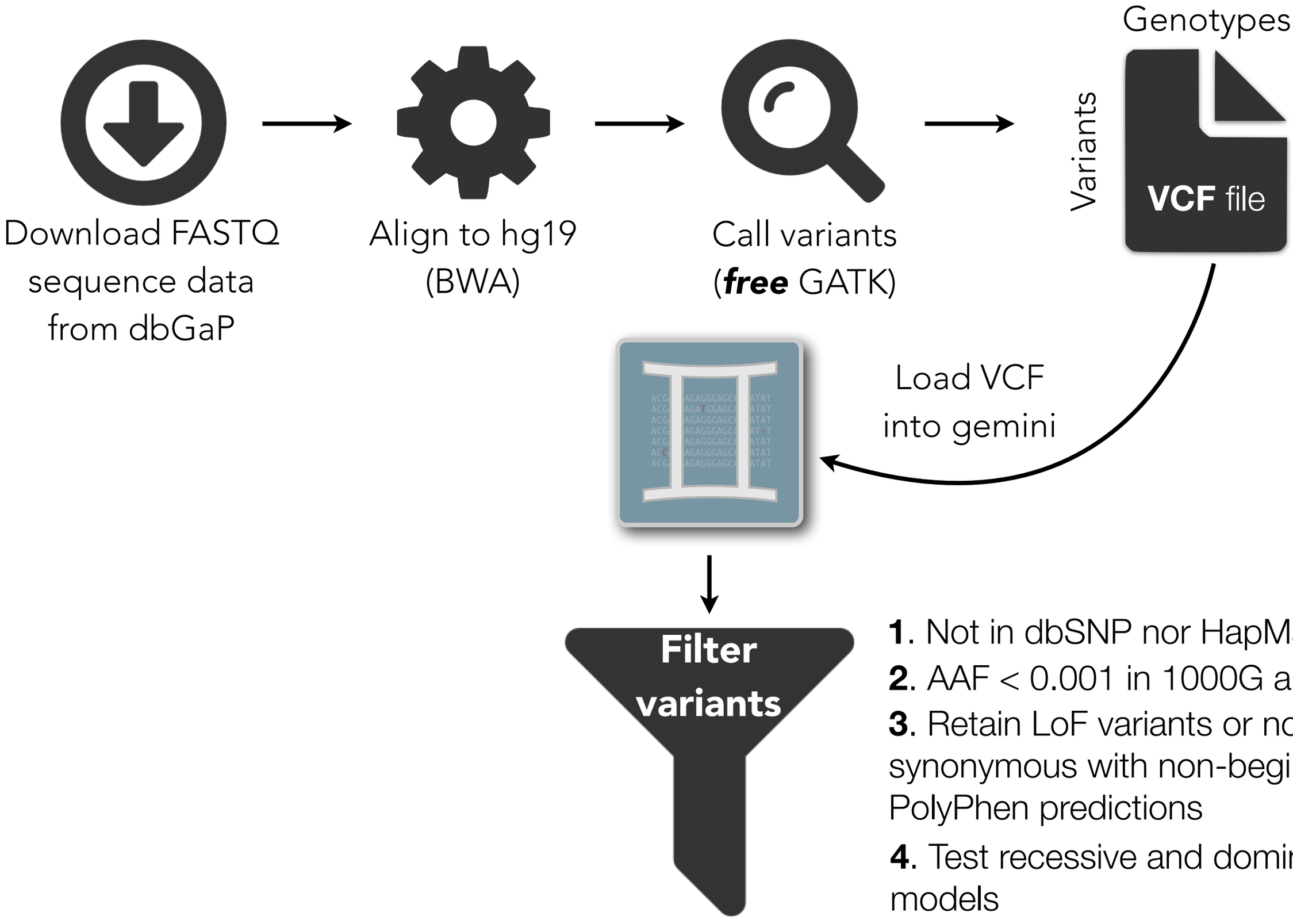
	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)	
Filter	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362
Not in dbSNP129	641	102	647	114	369	53
Not in HapMap 8	898	123	923	128	506	46
Not in either	456	31	464	33	228	9
Predicted damaging	204	6	204	12	83	1



DHODH

Can we identify DHODH with gemini?

Gemini workflow for Miller syndrome



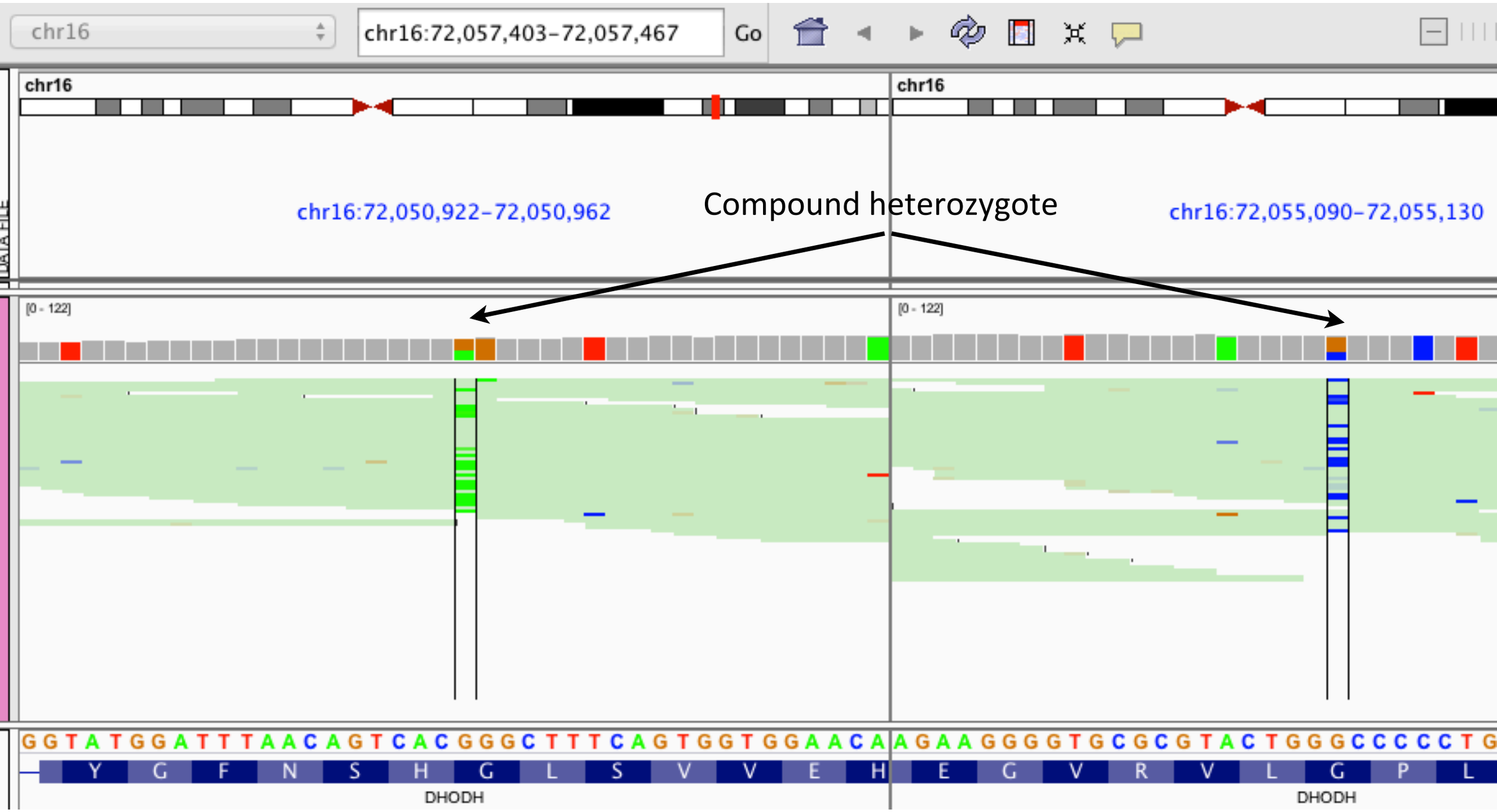
Predictions

- Ng et al were looking for compound heterozygotes owing to previous failures for simple recessive mutations.
- Just like Ng et al, there were no genes under the recessive model that >=2 LoF mutations in each of the 3 kindreds.
- If we allow PolyPhen “benign” predictions, DHODH has compound hets in all 3 kindreds.

Kindred 1: M10478

DHODH: Exon4: Gly->Arg (***benign***)

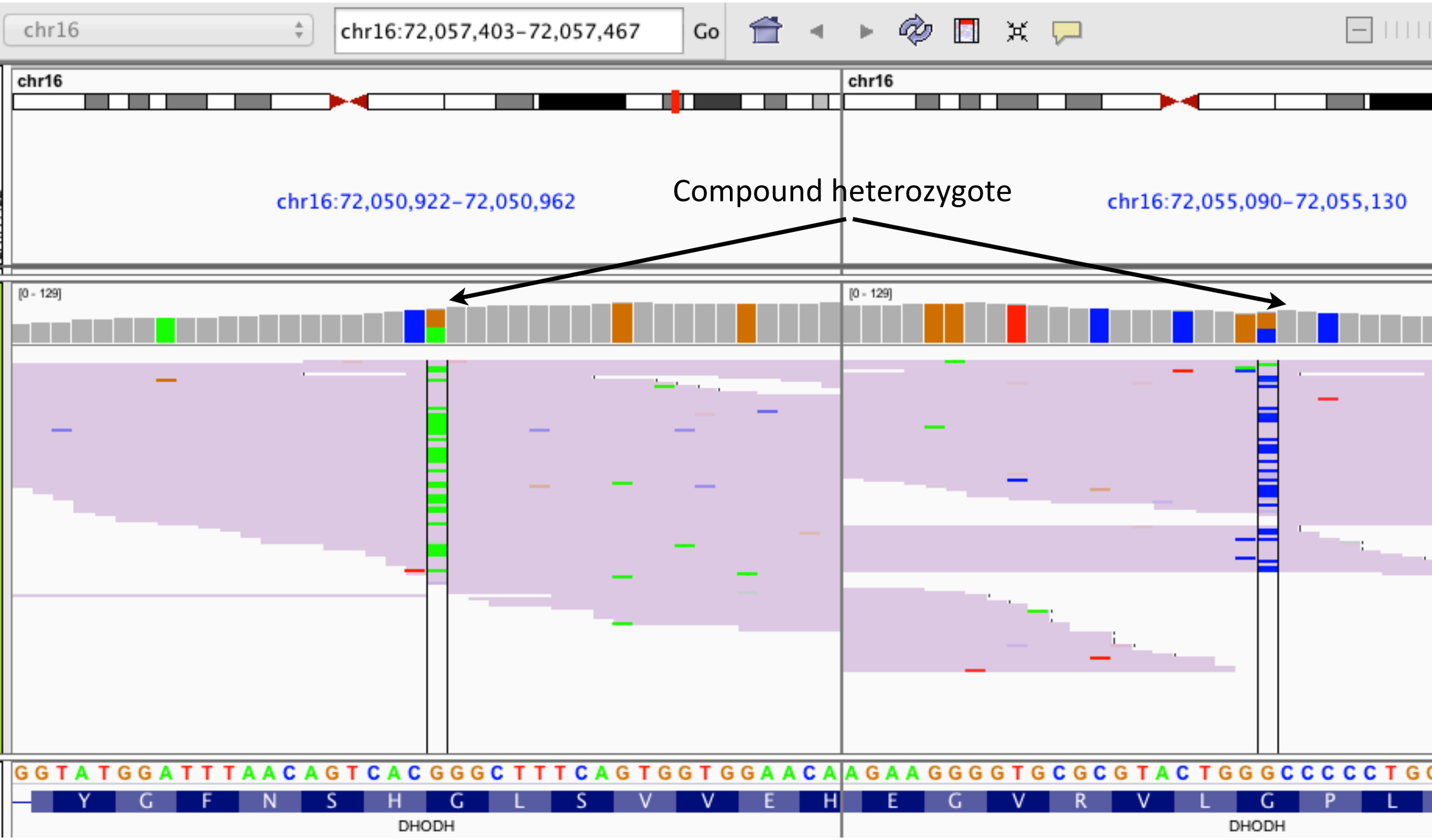
DHODH: Exon5: G->C (**damaging**)



Kindred 1:M10500

DHODH: Exon4: Gly->Arg (***benign***)

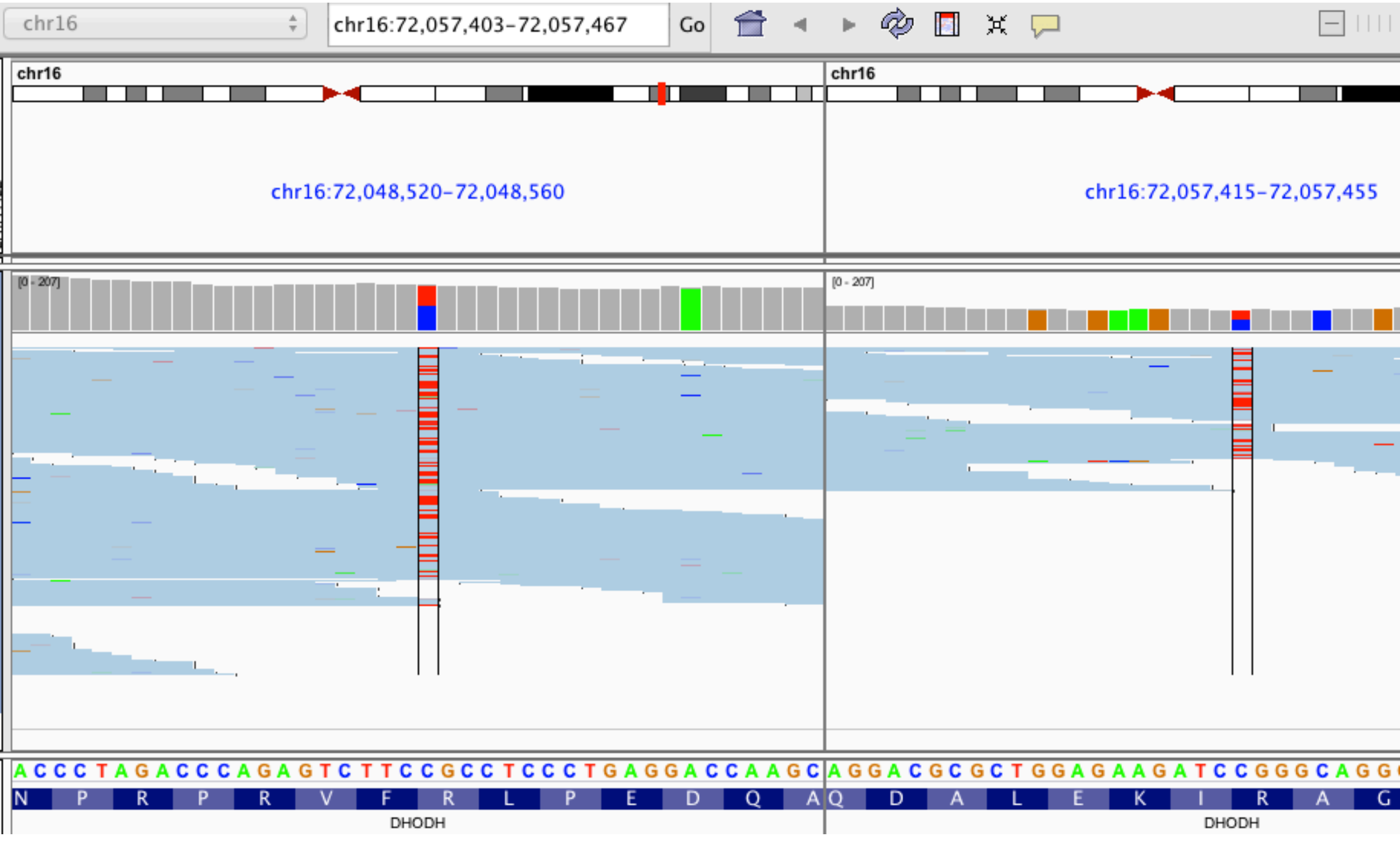
DHODH: Exon5: G->C (**damaging**)



Kindred 2: M10475

DHODH: Exon3: C->T

DHODH: Exon8: C->T



Kindred 3: M128215

DHODH: Exon5: C->T

Missed by GATK:
DHODH: Exon5: ΔT

Summary

- Integrates extensive genome annotations to facilitate analysis and interpretation
- Gemini is currently ideal for family-based studies
- Future
  - Cancer-focused version
  - Scale to 1000s of genomes
  - Visualization and locus-inspection.

Quinlan lab

Uma Paila, Ph.D.

Postdoctoral Research Associate

udp3f @ virginia.edu

Research Projects and Interests: Investigation of the genetic basis of extreme sensitivity to ionizing radiation; development of new analytical tools for exploring genetic variation identified through next-generation sequencing projects.

John Kubinksi
Undergraduate
(Biology)

Neil Kindlon, M.S.

Staff Scientist and Software Engineer

nek3d @ virginia.edu

Research Projects and Interests: Software development for genomic analysis. Structural variation discovery and interpretation using DNA sequencing technologies.

Gift Sinthong
Undergraduate
(Comp. Science)

Co-mentored with Ira Hall and Gabe Robins

Ryan Layer

Graduate student

rl6sf @ virginia.edu

Research Interests: Scalable algorithm development for high-throughput genomic analysis; genome data mining and analysis; structural variation discovery and interpretation.

Acknowledgements

Ira Hall *Univ. of Virginia*

Ankit Malhotra *Univ. of Virginia*

Michael Lindberg *Univ. of Virginia*

Royden Clark *Univ. of Virginia*

Svetlana Sokolova *Univ. of Virginia*

Mitchell Leibowitz *Univ. of Virginia*

Pat Concannon Univ. of Virginia

Steve Rich Univ. of Virginia

Suna Onengut-Gumuscu Univ. of Virginia

Shu-Man Fu Univ. of Virginia

Gabor Marth Boston College

Jim Robinson Broad Institute

Nik Krumm *Univ. of Washington*

Evan Eichler *Univ. of Washington*

Debbie Nickerson *Univ. of Washington*

Chris Carlson *Fred Hutchinson CRC*

Mark Rieder *Univ. of Washington*

Josh Smith *Univ. of Washington*

Peter Sudmant *Univ. of Washington*

Funding

NHGRI: R01 HG006693-01

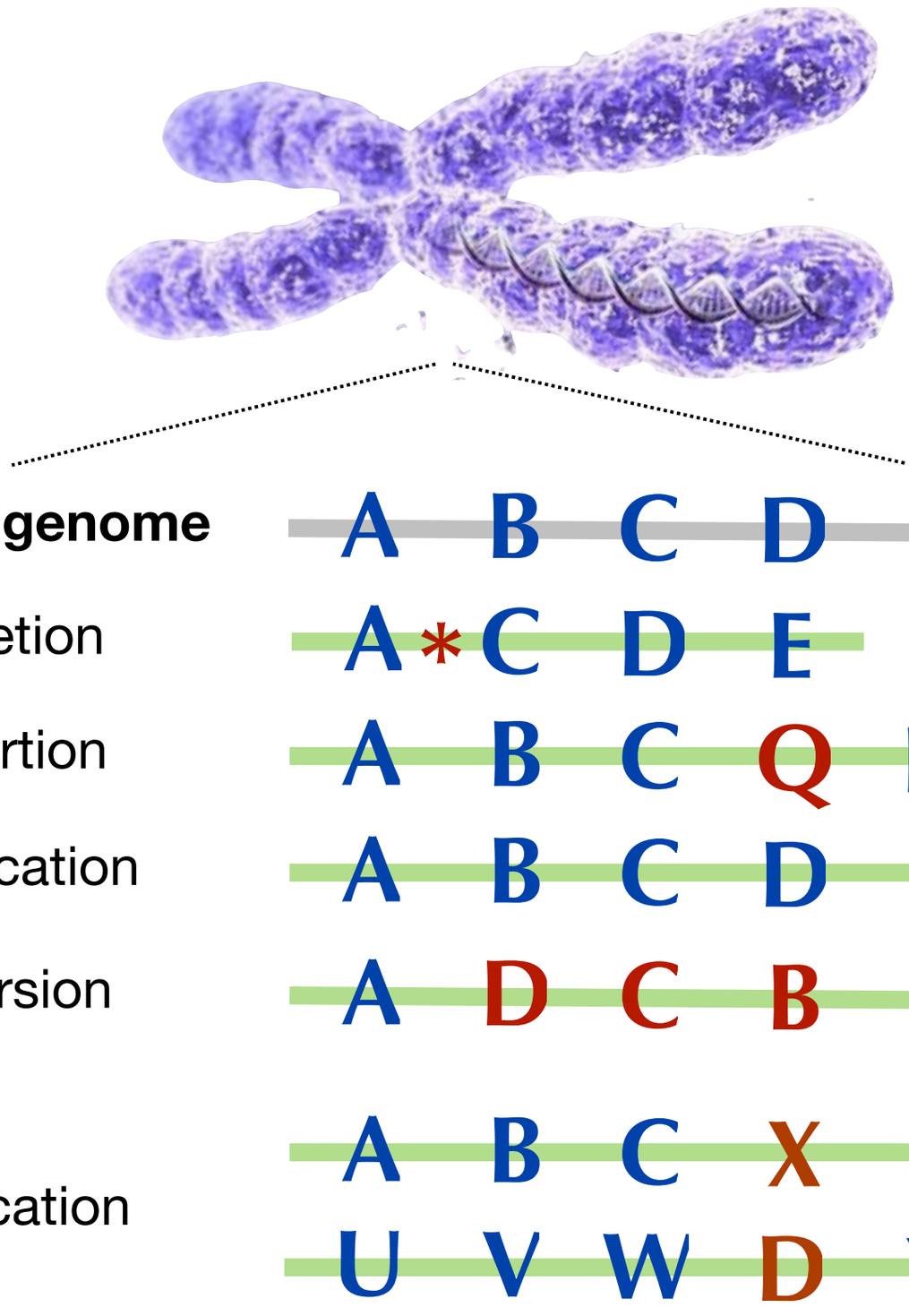
NIEHS: R21 ES020521-01

UVA FEST award

UVA Cancer Center Pilot Program

1. Advances in structural variation discovery

A brief overview of structural variation



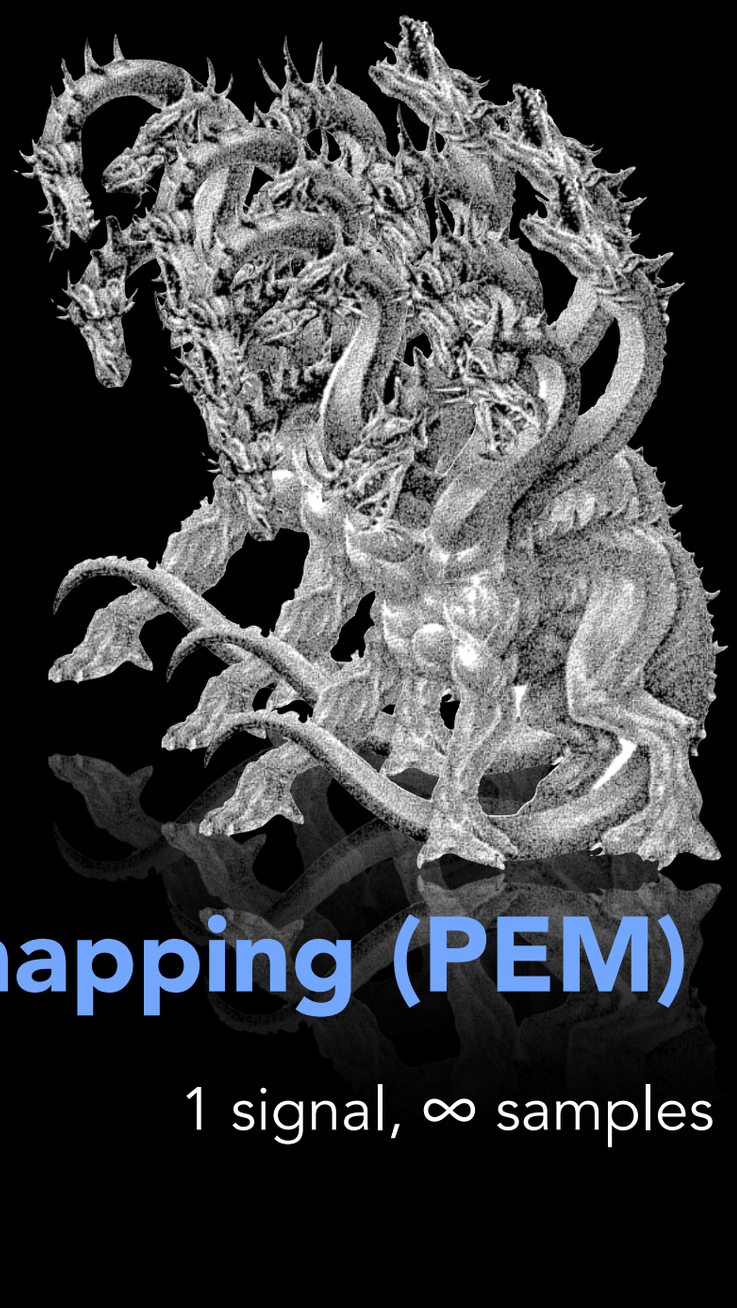
Evolution of our SV discovery tools

Hydra
(2010)

Paired-end mapping (PEM)

1 signal, 1 sample

Hydra_Multi
(2011)



1 signal, ∞ samples

LUMPY
(2012)

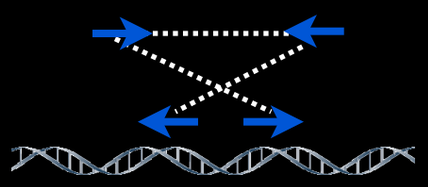
lumpy

A Probabilistic Framework for Structural Variant Discovery

many signals,
many samples

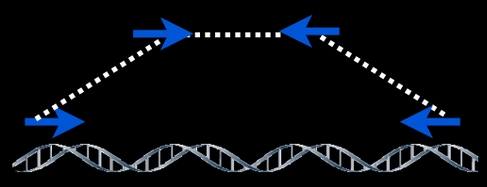
Hydra clusters ***discordant*** PEM mappings

everted orientation



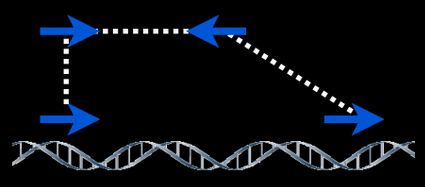
tandem duplication

too big



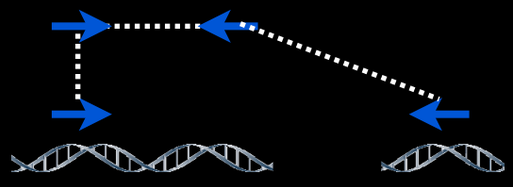
deletion

same-strand



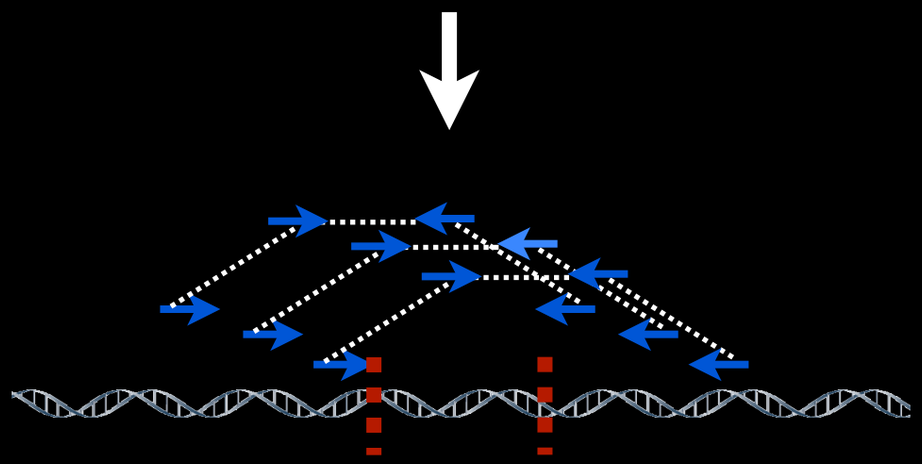
inversion

distant linkage



insertion,
retrotransposition,
translocation

Cluster to localize breakpoints



Hydra

1 signal (PEM), 1 sample

LUMPY integrates ***all*** SV signals

Sources

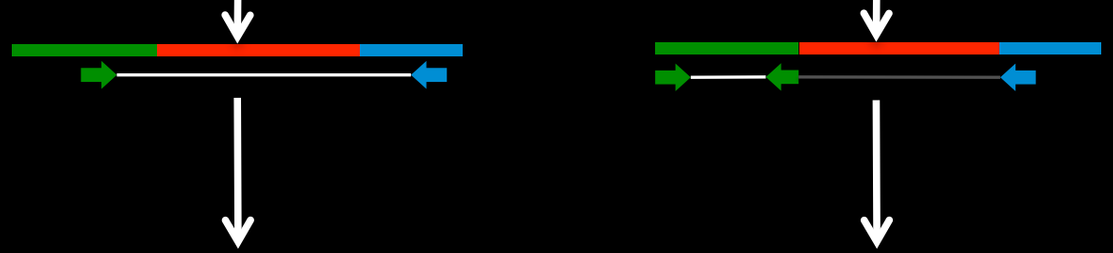
Sample

Prior Result

Pair-End Aligner

Split-Read Aligner

Input



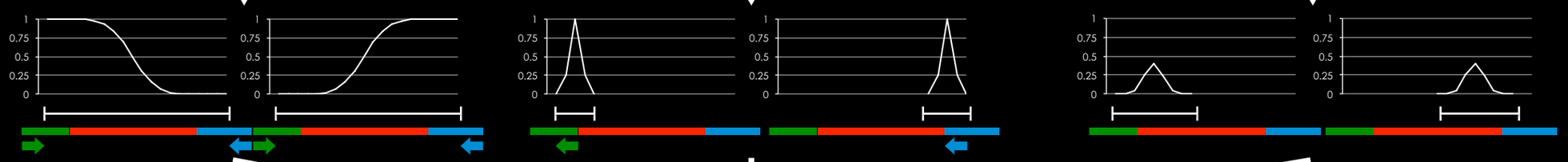
Evidence

Pair-End Module

Split-Read Module

Generic Module

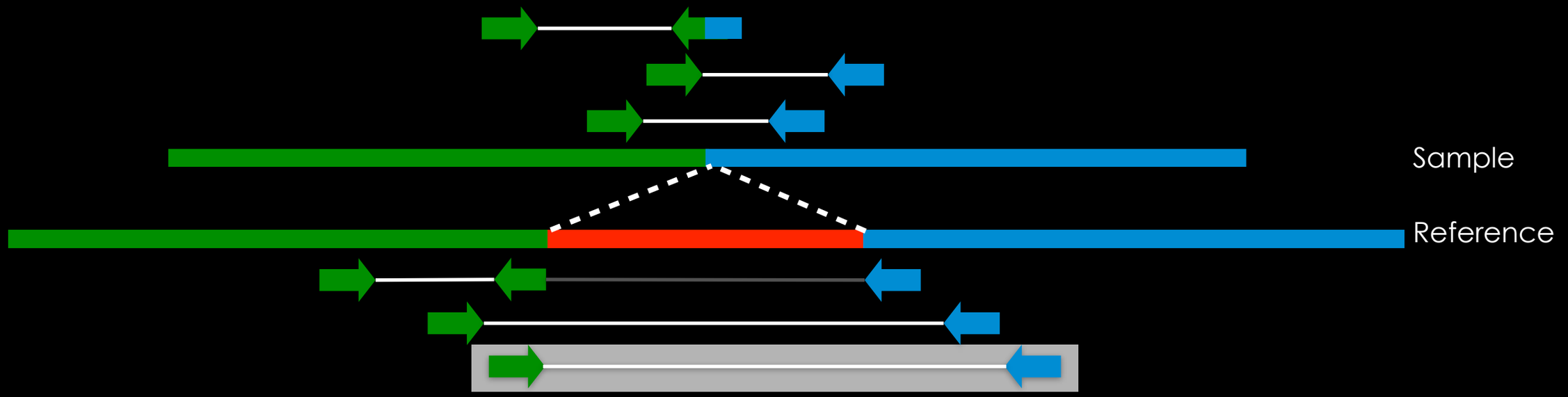
Breakpoints

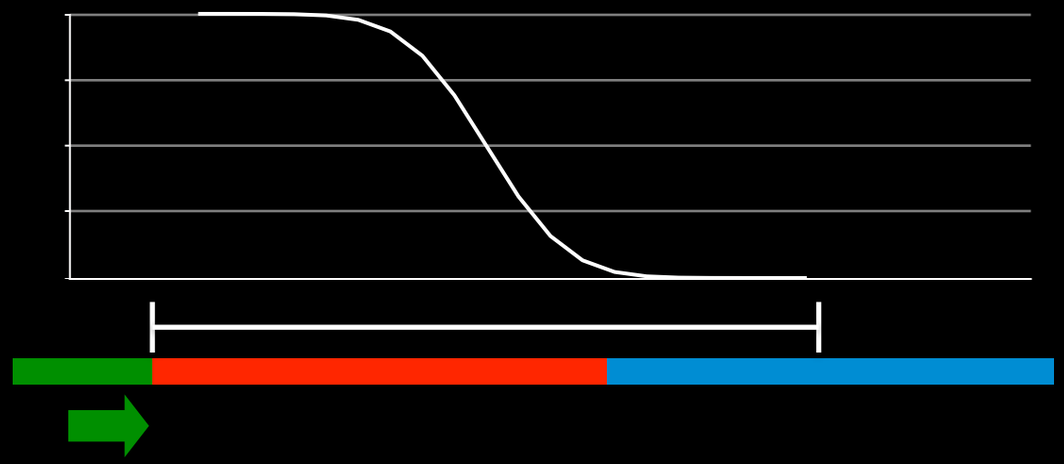


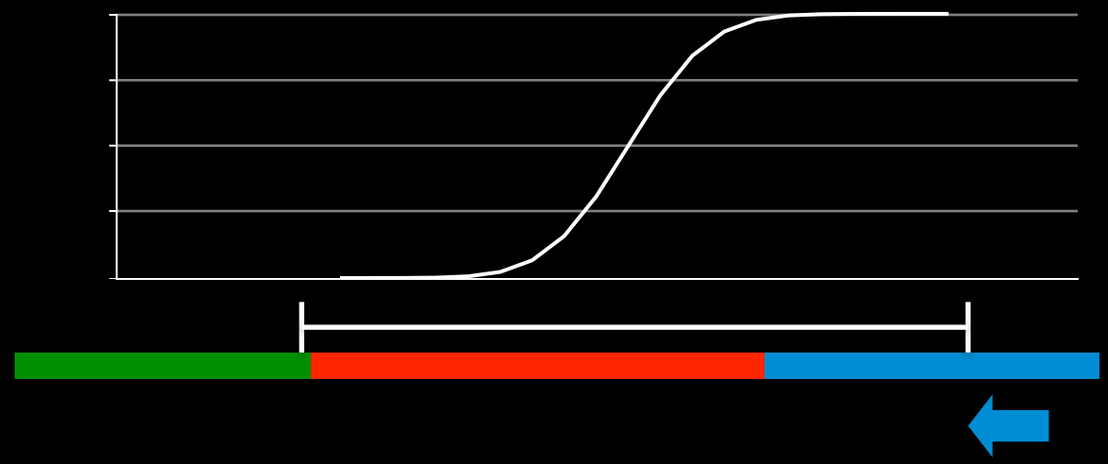
Pool

SV Prediction

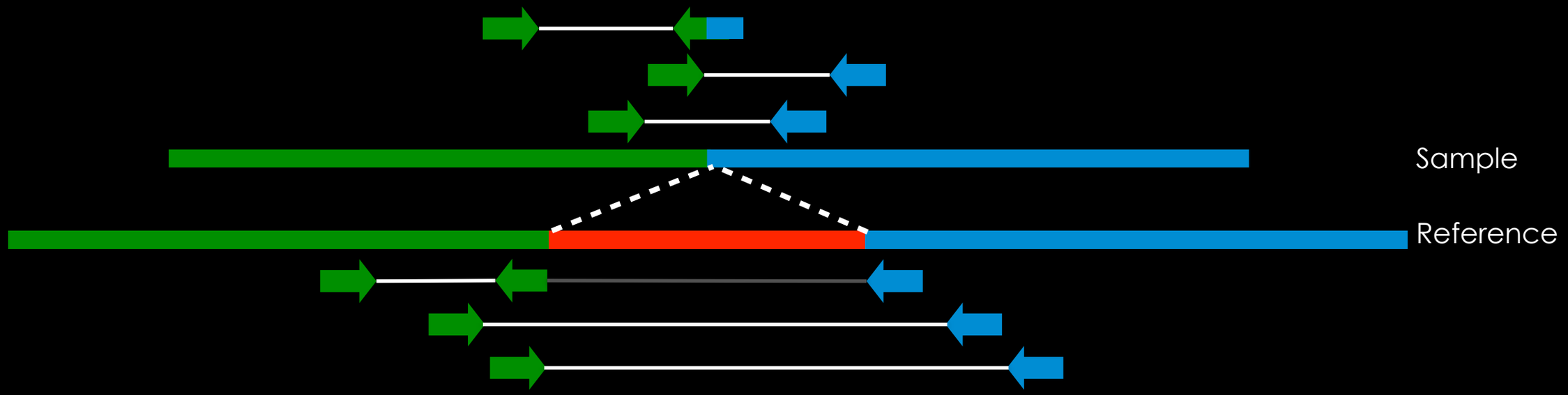
Pooling Evidence

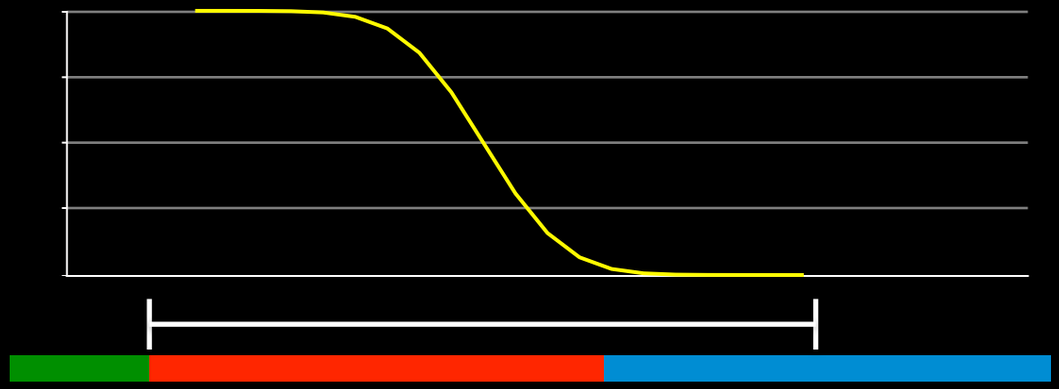


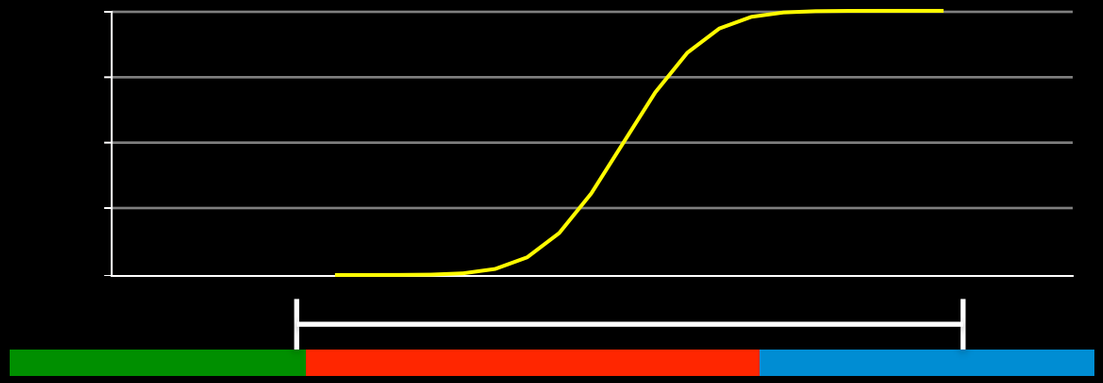




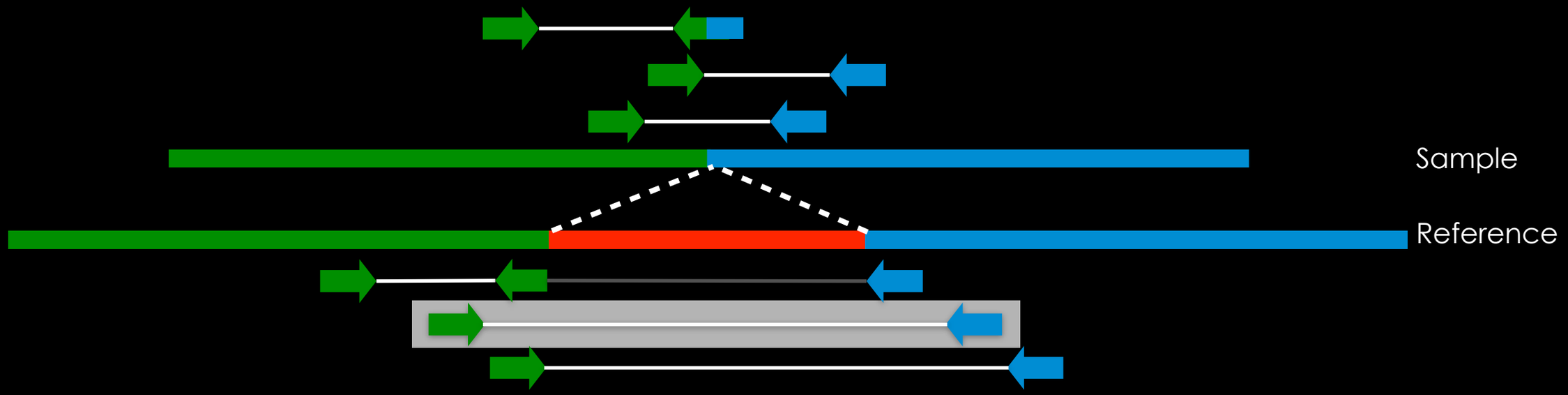
Pooling Evidence

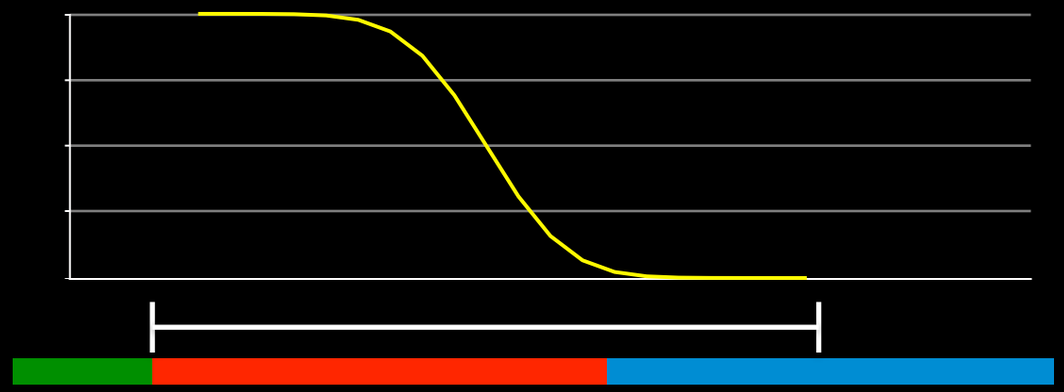


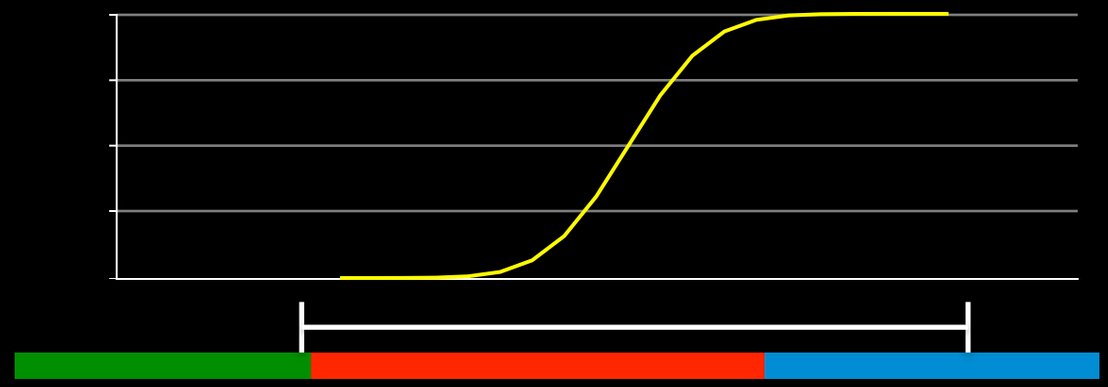




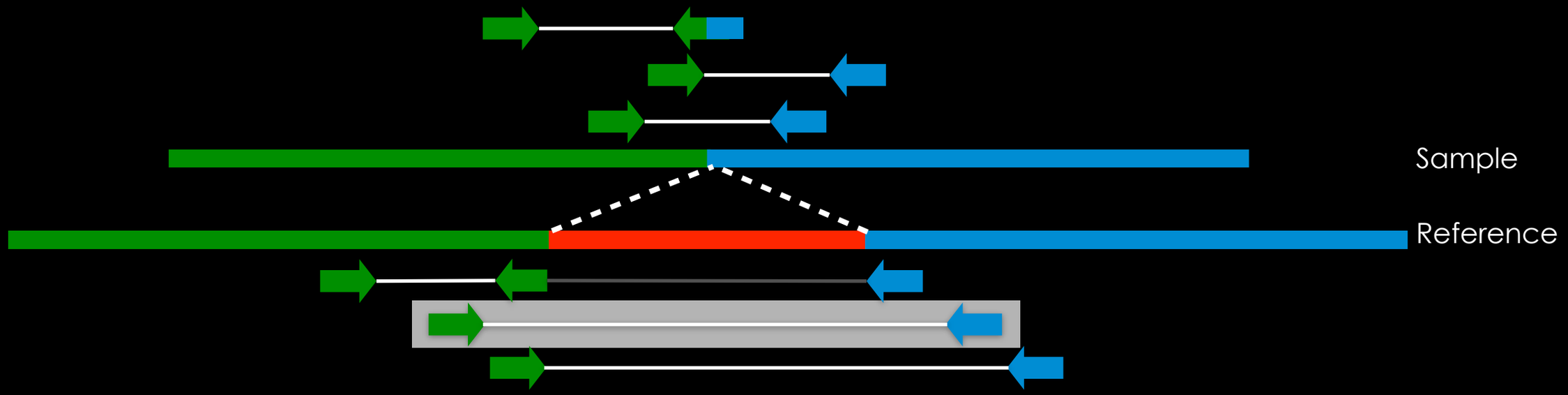
Pooling Evidence

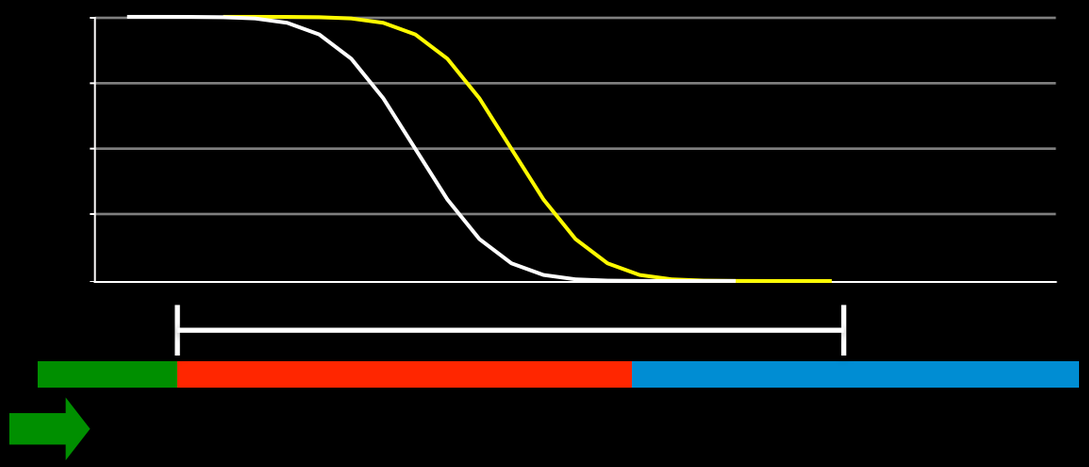


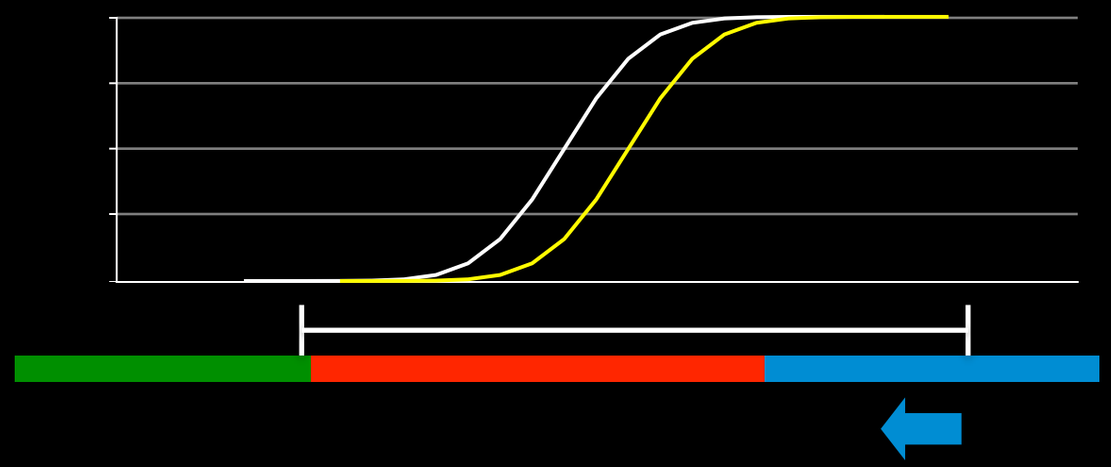




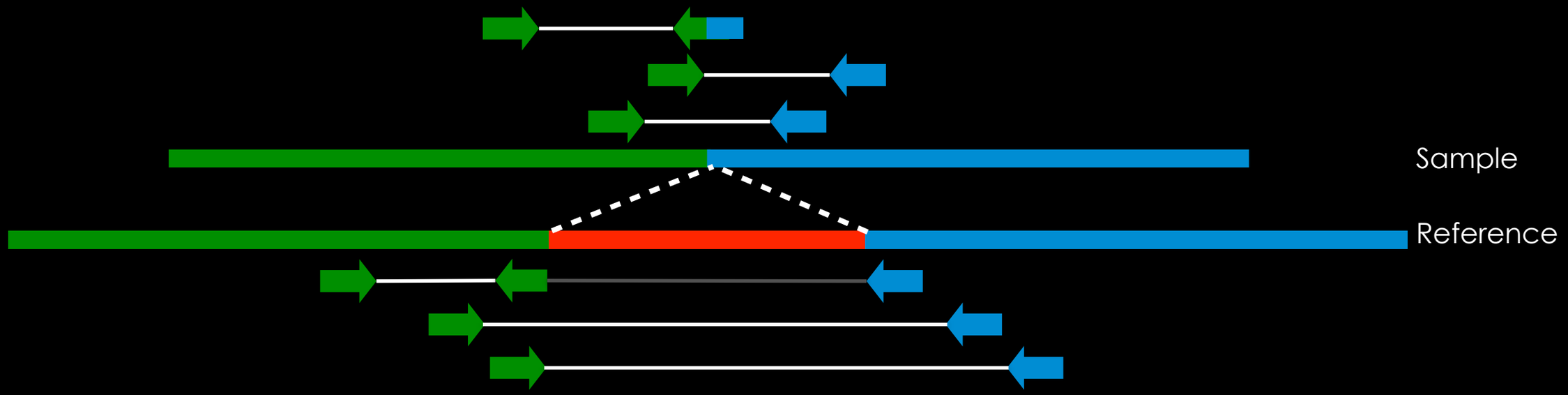
Pooling Evidence

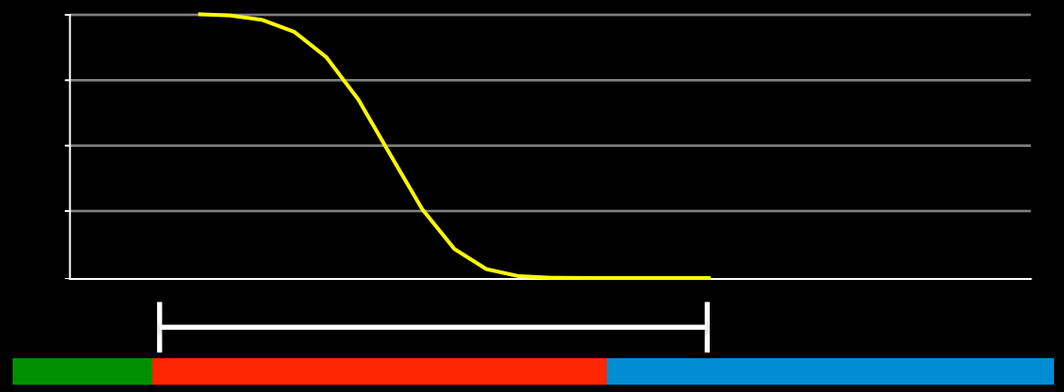


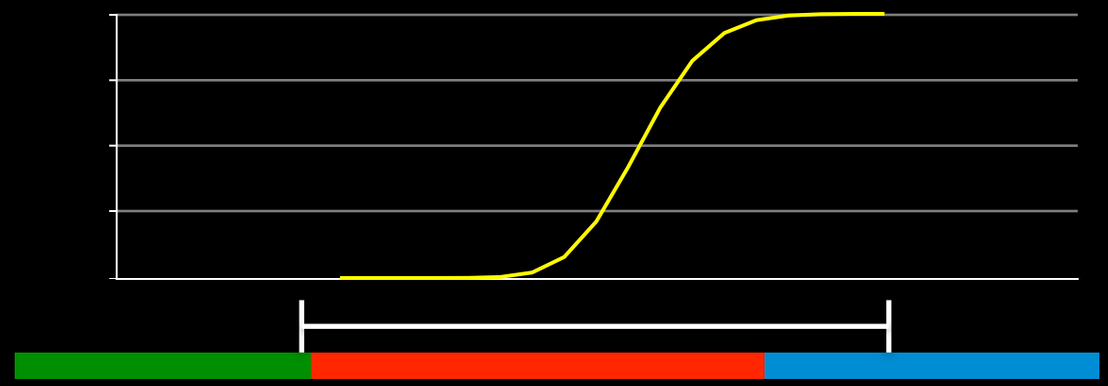




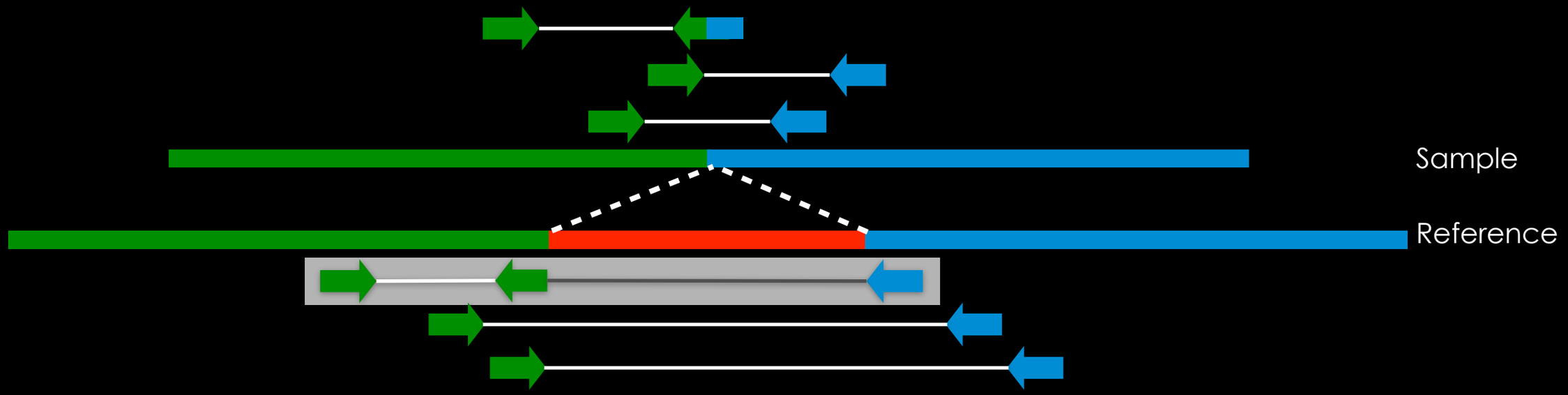
Pooling Evidence

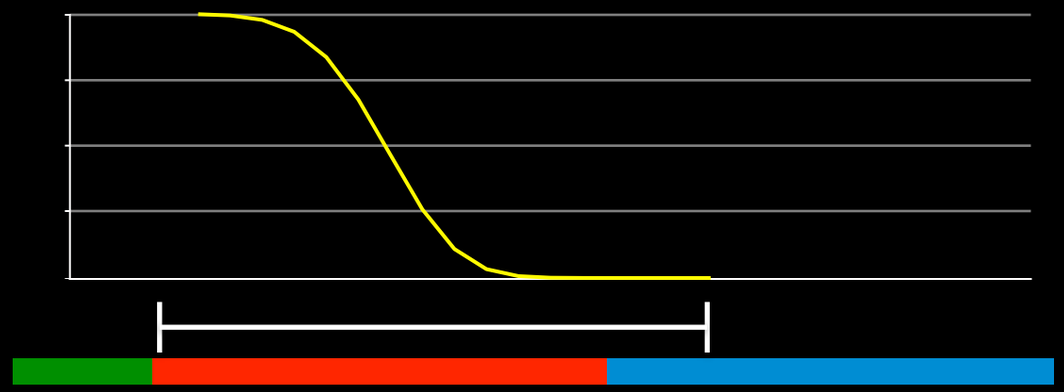


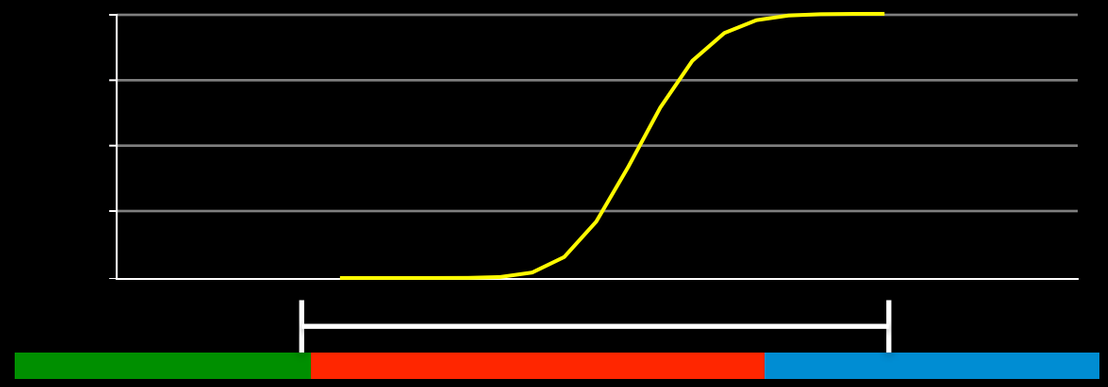




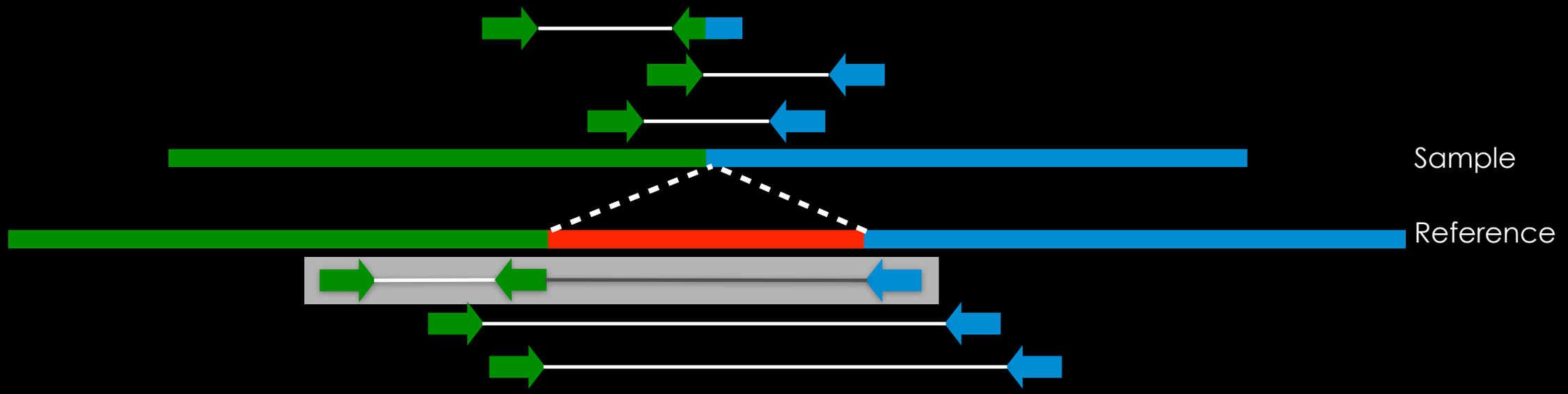
Pooling Evidence

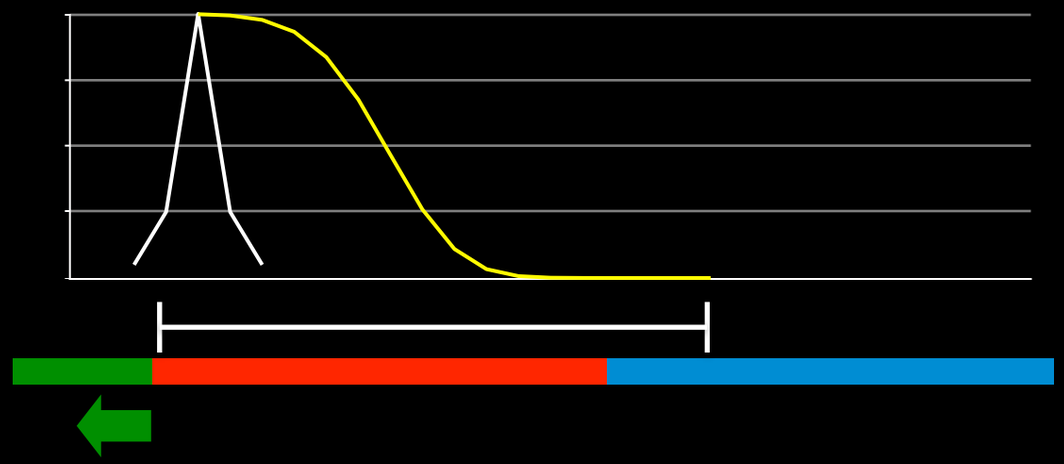


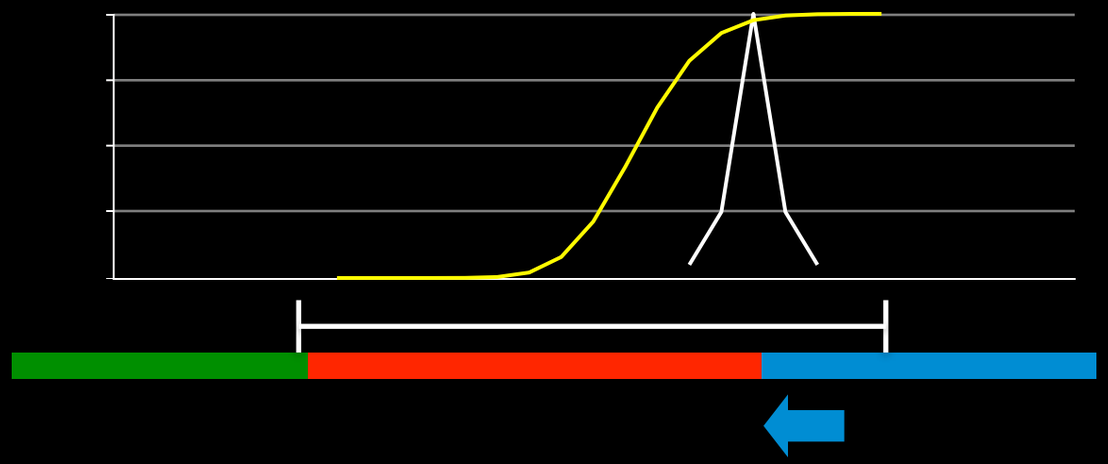




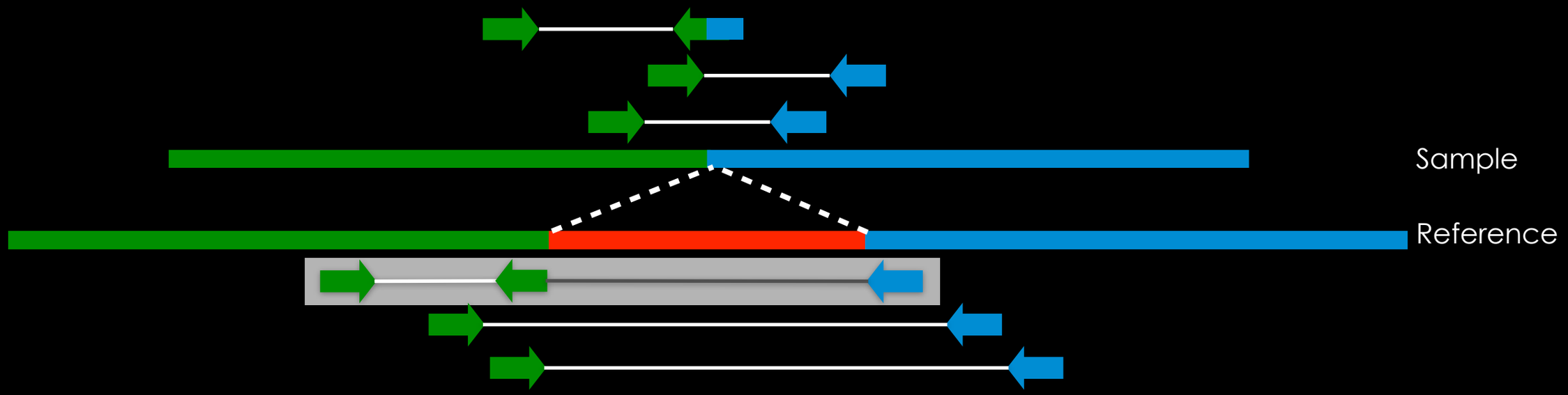
Pooling Evidence

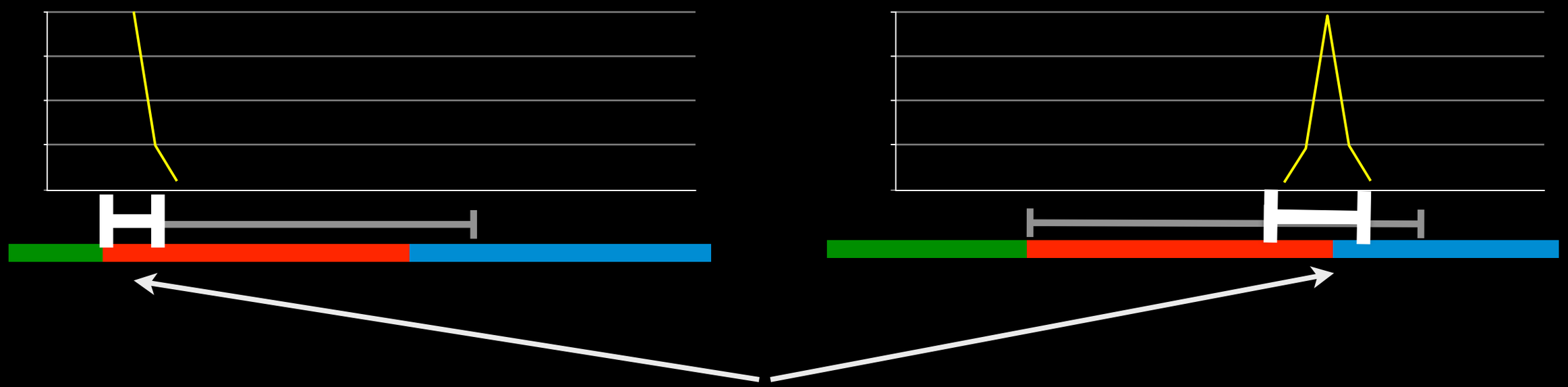






Pooling Evidence



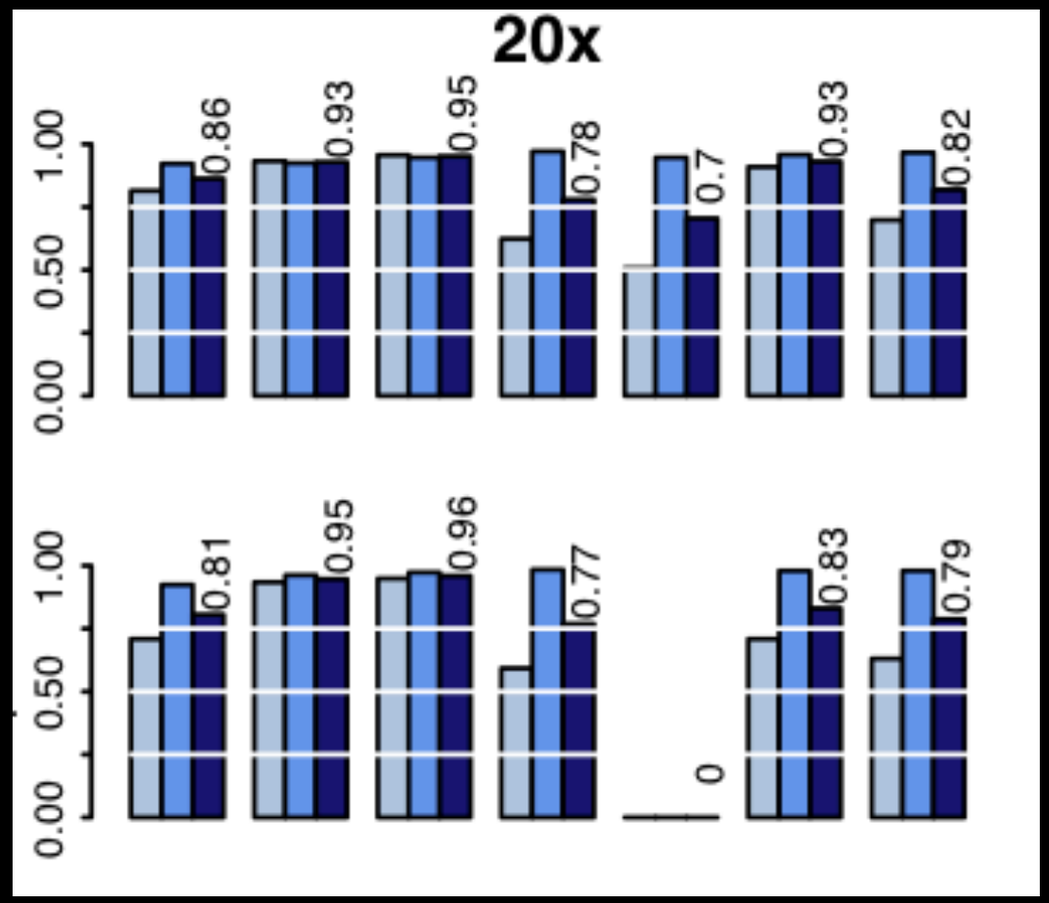


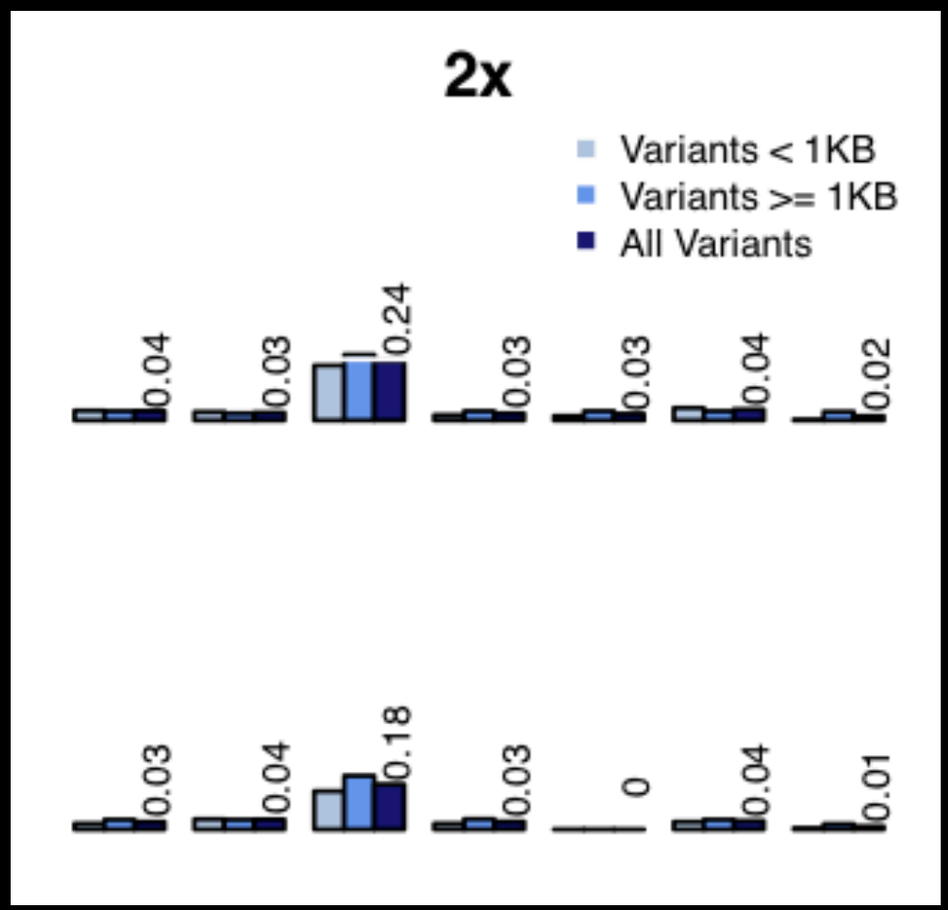
Much greater SV breakpoint resolution and sensitivity

Greater sensitivity v. *lumpless* approaches

Deletions

Duplications

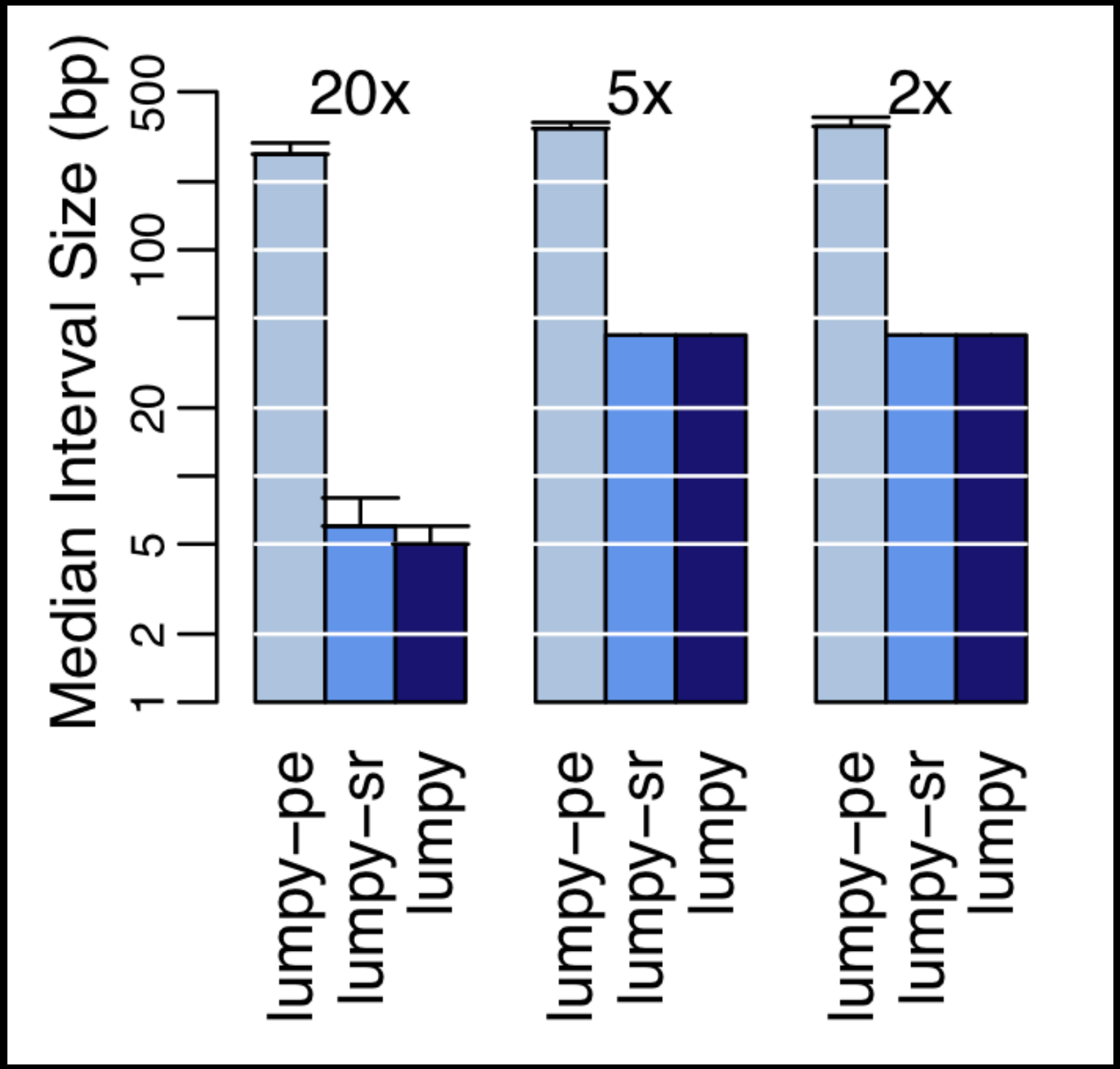


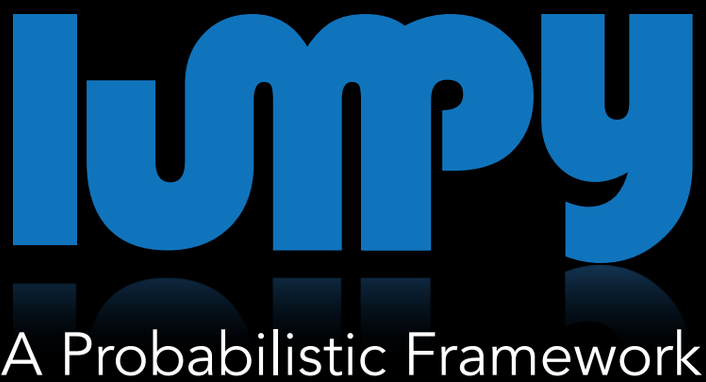


Highest sensitivity. Most profound for smaller variants and at **low coverage**.

Sensitivity is crucial for cancer studies (e.g., tumor heterogeneity)

SV breakpoint resolution





A Probabilistic Framework
for Structural Variant
Discovery

https://github.com/arq5x/lumpy-sv

submitted to RECOMB 2013