# Integrative Genomics

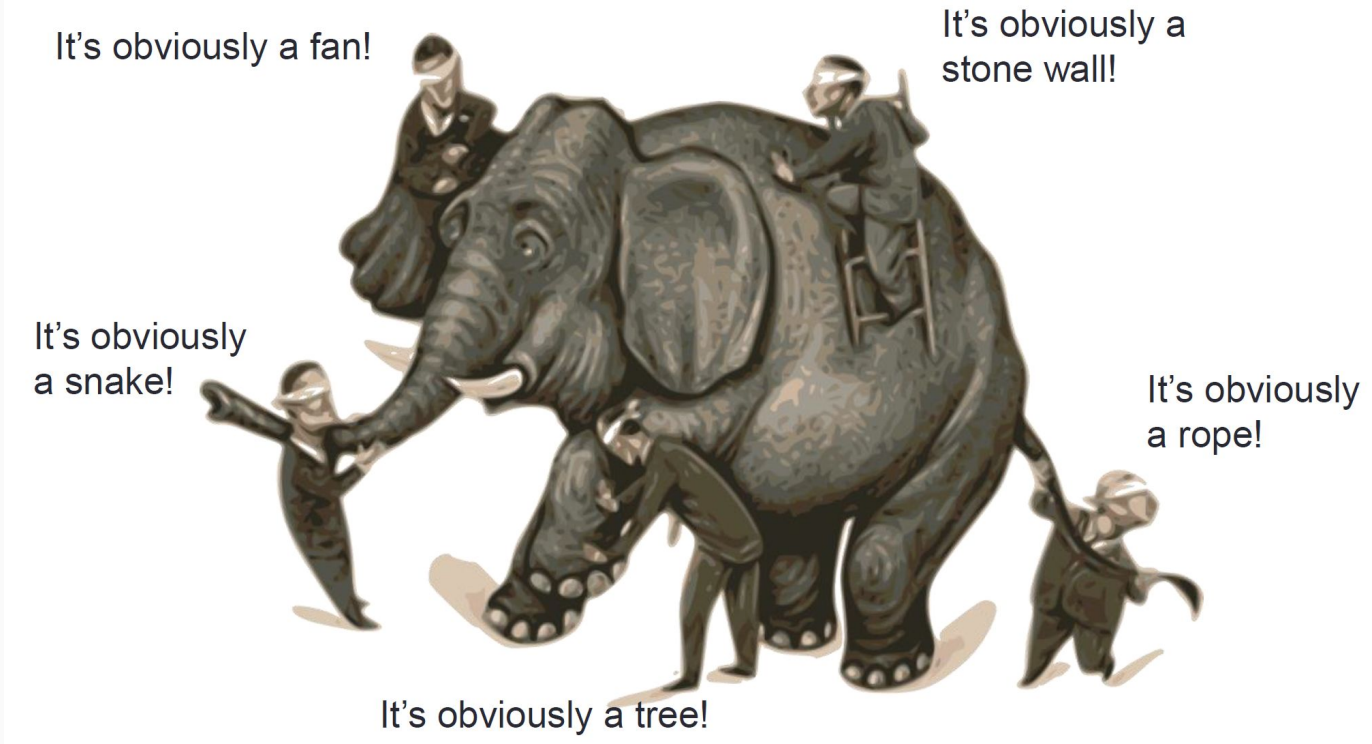
# Introduction

- A **single –omics experiment** may only provide a **limited view** without further context
  - Example: **ChIP-seq**. What's so special about where a protein binds?
  - Example: **A single RNA-seq experiment**. What can you learn using relative transcript abundance alone?

# Introduction

- Only looking at one component of a cell (transcripts, protein binding, etc.) can also lead to **false conclusions**
- Ideally, results from different types of experiments should be **integrated** to **support your conclusions**

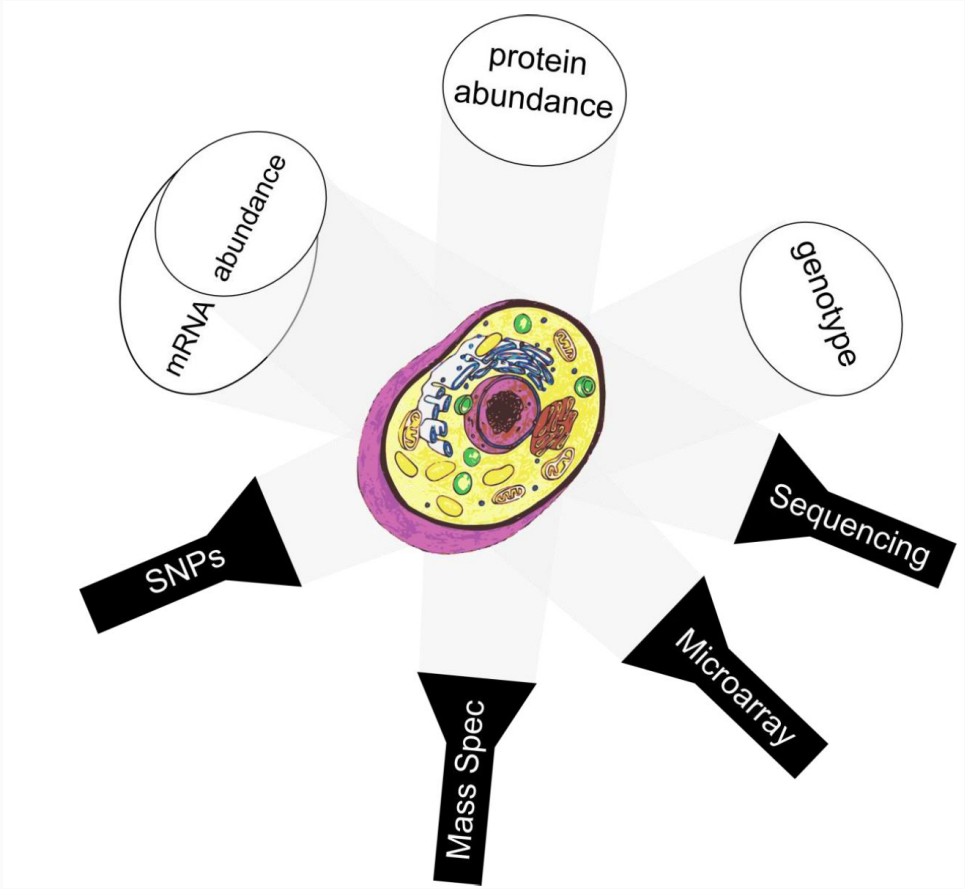
# Metaphor 1: Blind men and an elephant



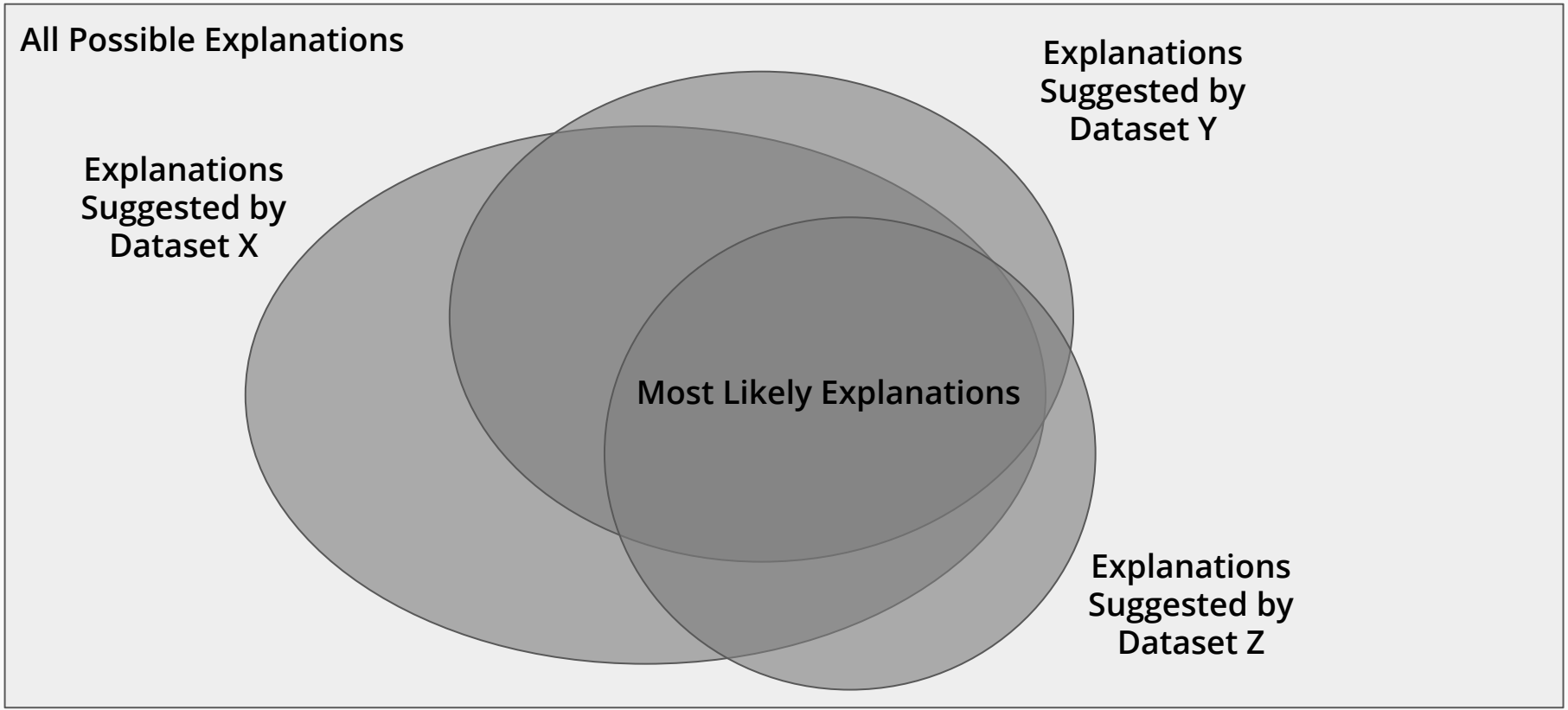
# Metaphor 2: Shadow Art

http://graphics.stanford.edu/~niloy/research/shadowArt/shadowArt_sigA_09.html

# Each experiment provides a different view



# Why Does Some Biological Event Occur?

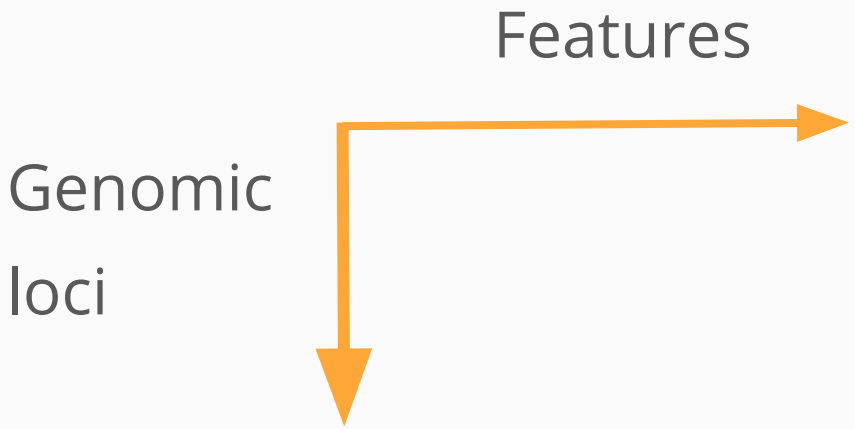


# Integrative Genomics

- One of the **most valuable bioinformatics skills** is to be able to **combine** several **genome-scale experiments** together
- This practice is generally called **integrative genomics**
- **How do we combine** genome-scale experiments together?

# Reference Genomes

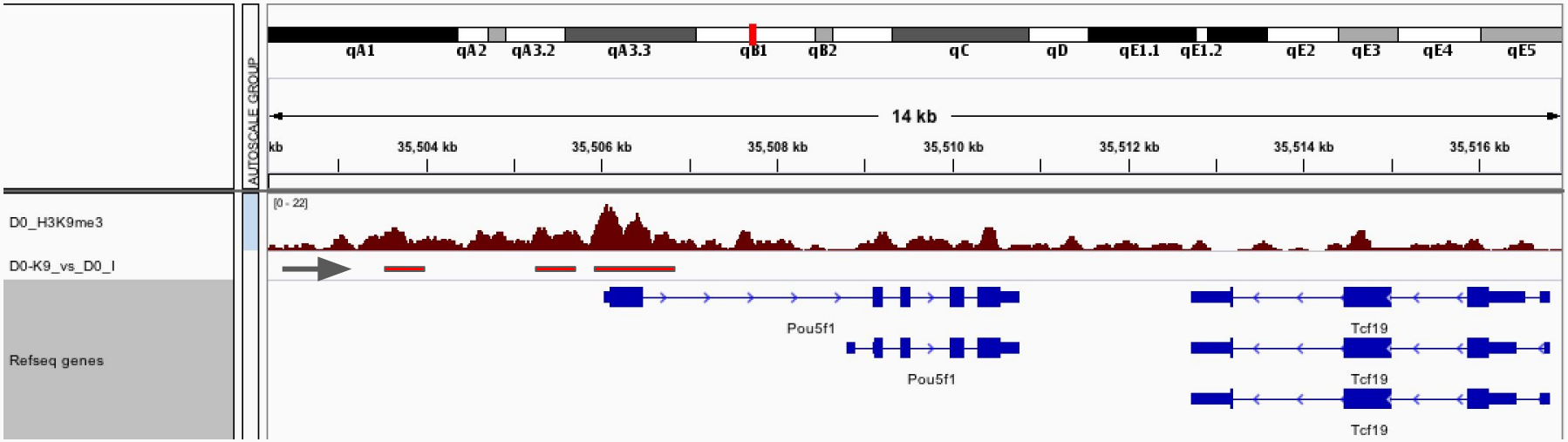
- **Consistent coordinate system** across experiments allows **different data** to be **integrated together**



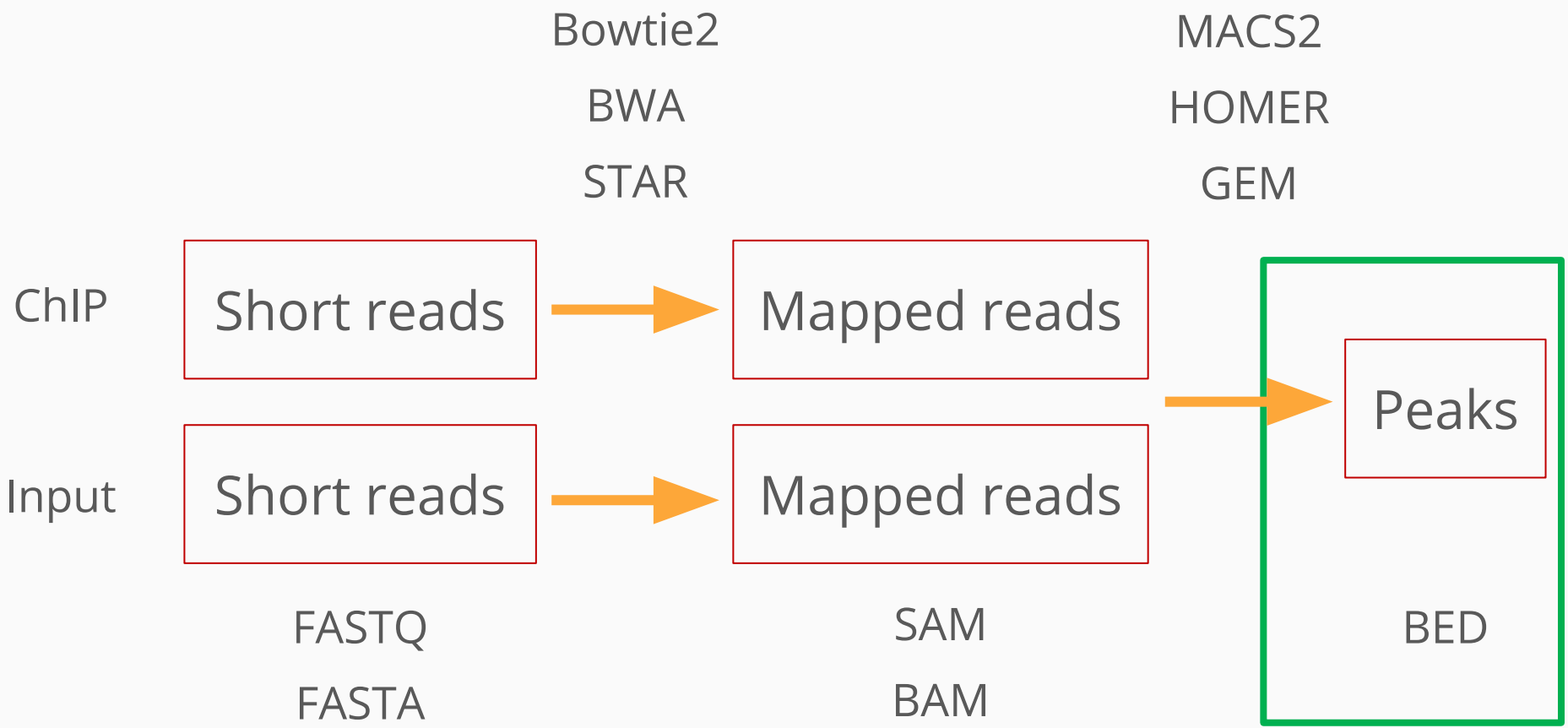
# Today – Integrative Genomics

- BED format redux
- Set Theory & Genome arithmetic
  - Useful tools (BEDtools, bedops, GAT)
  - Basic operations
  - Enumeration vs. association
- Practical examples

# Review: ChIP-Seq Peaks



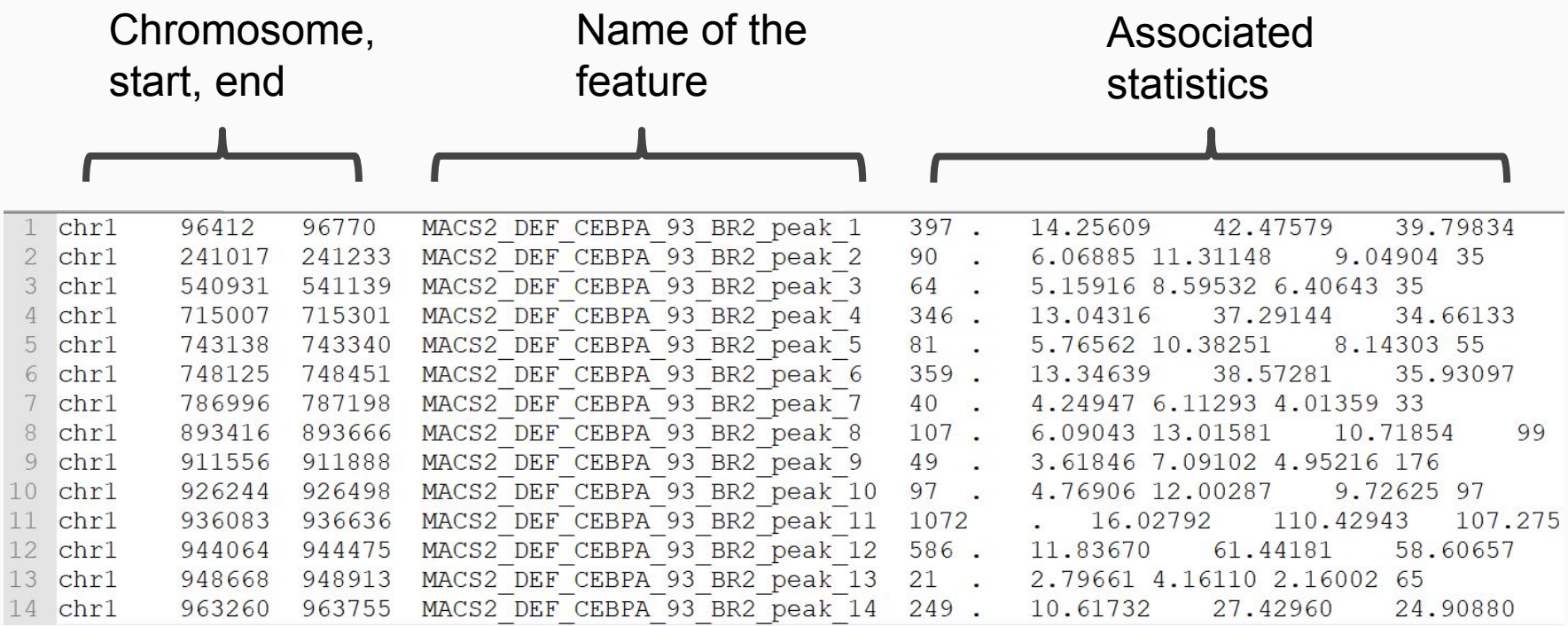
# Review: ChIP-seq analysis workflow



# The BED format

- A loose **tab-delimited text format** that defines **locations and information** related to a **genomic feature** of interest
- Columns **can be variable** but always start with these 3 (or 4):
  - chromosome, start, end, (and name)

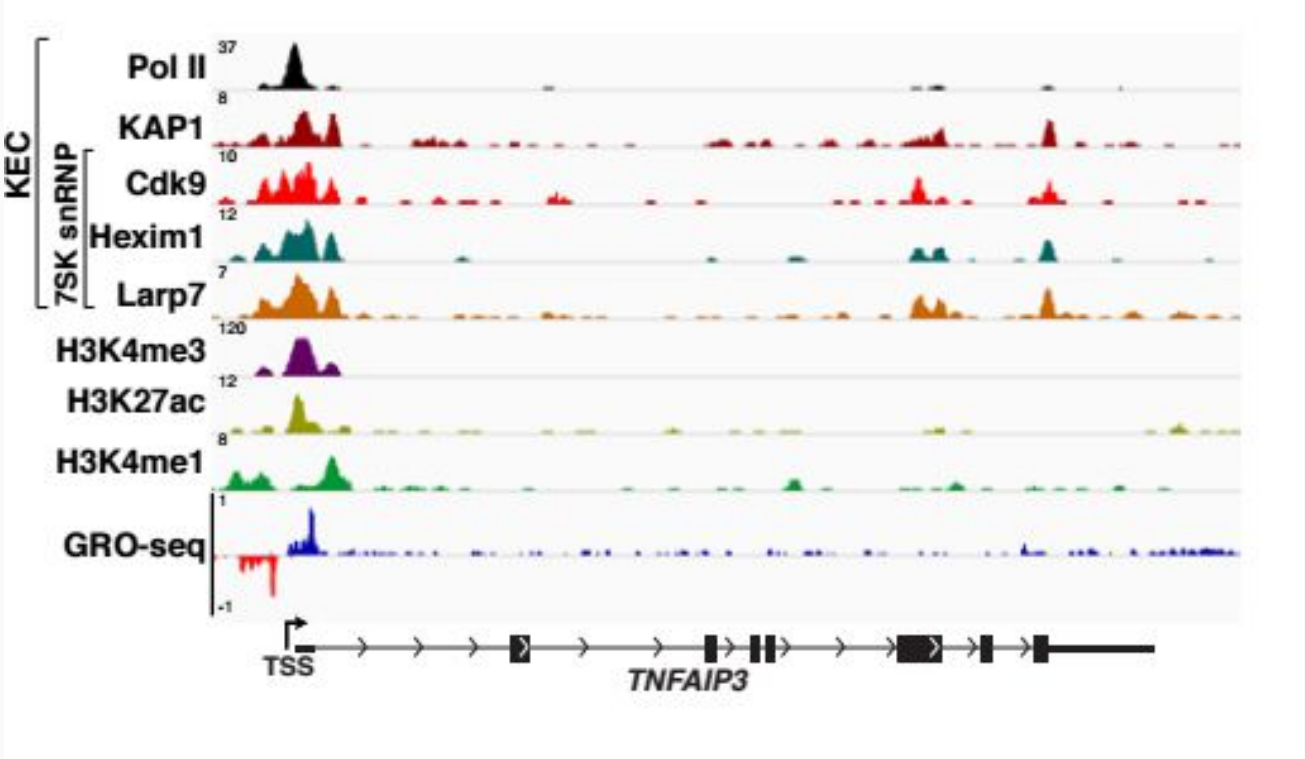
# BED format example




```
1  chr1    96412   96770   MACS2_DEF_CEBPA_93_BR2_peak_1   397 .   14.25609    42.47579    39.79834
2  chr1    241017  241233  MACS2_DEF_CEBPA_93_BR2_peak_2   90  .   6.06885 11.31148    9.04904 35
3  chr1    540931  541139  MACS2_DEF_CEBPA_93_BR2_peak_3   64  .   5.15916 8.59532 6.40643 35
4  chr1    715007  715301  MACS2_DEF_CEBPA_93_BR2_peak_4   346 .   13.04316    37.29144    34.66133
5  chr1    743138  743340  MACS2_DEF_CEBPA_93_BR2_peak_5   81  .   5.76562 10.38251    8.14303 55
6  chr1    748125  748451  MACS2_DEF_CEBPA_93_BR2_peak_6   359 .   13.34639    38.57281    35.93097
7  chr1    786996  787198  MACS2_DEF_CEBPA_93_BR2_peak_7   40  .   4.24947 6.11293 4.01359 33
8  chr1    893416  893666  MACS2_DEF_CEBPA_93_BR2_peak_8   107 .   6.09043 13.01581    10.71854    99
9  chr1    911556  911888  MACS2_DEF_CEBPA_93_BR2_peak_9   49  .   3.61846 7.09102 4.95216 176
10 chr1    926244  926498  MACS2_DEF_CEBPA_93_BR2_peak_10  97  .   4.76906 12.00287    9.72625 97
11 chr1    936083  936636  MACS2_DEF_CEBPA_93_BR2_peak_11  1072    .   16.02792    110.42943   107.275
12 chr1    944064  944475  MACS2_DEF_CEBPA_93_BR2_peak_12  586 .   11.83670    61.44181    58.60657
13 chr1    948668  948913  MACS2_DEF_CEBPA_93_BR2_peak_13  21  .   2.79661 4.16110 2.16002 65
14 chr1    963260  963755  MACS2_DEF_CEBPA_93_BR2_peak_14  249 .   10.61732    27.42960    24.90880
```

- ENCODE formats: tagAlign, narrowPeak, broadPeak

# Integrating Multiple Genomic Datasets



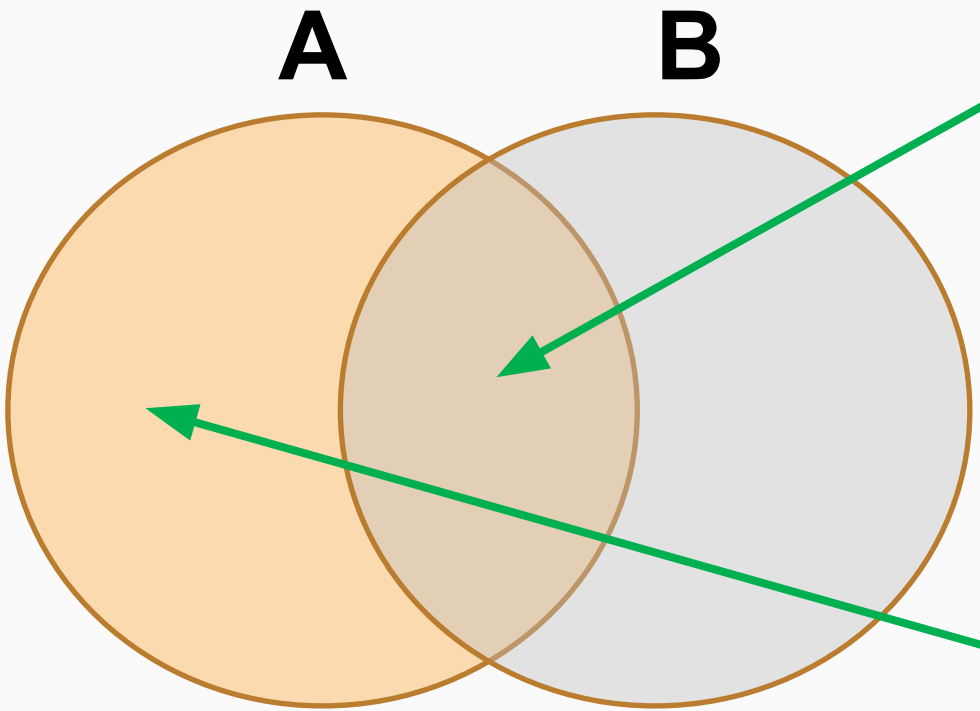
# Integrative Genomics on BED files

- How do we learn information about 2 or more BED files?
- There exists a core set of operations that can be applied to multiple BED files – **genome arithmetic**
- These generally correspond to **set theory operations** but at the **genome-scale**

# Set Theory

- Branch of math/logic that studies **sets** which can be **collections of any objects**
- **Traditional arithmetic** (+, -, x, /) is binary operations on **numbers**
- Set theory encompasses binary operations performed on sets
- Core operations can be visualized with **Venn diagrams**

# Set Operations




**Intersection:**

$A \cap B$

**Union:**

$A \cup B$

**Complement:**

$A \setminus B$ **One way!**

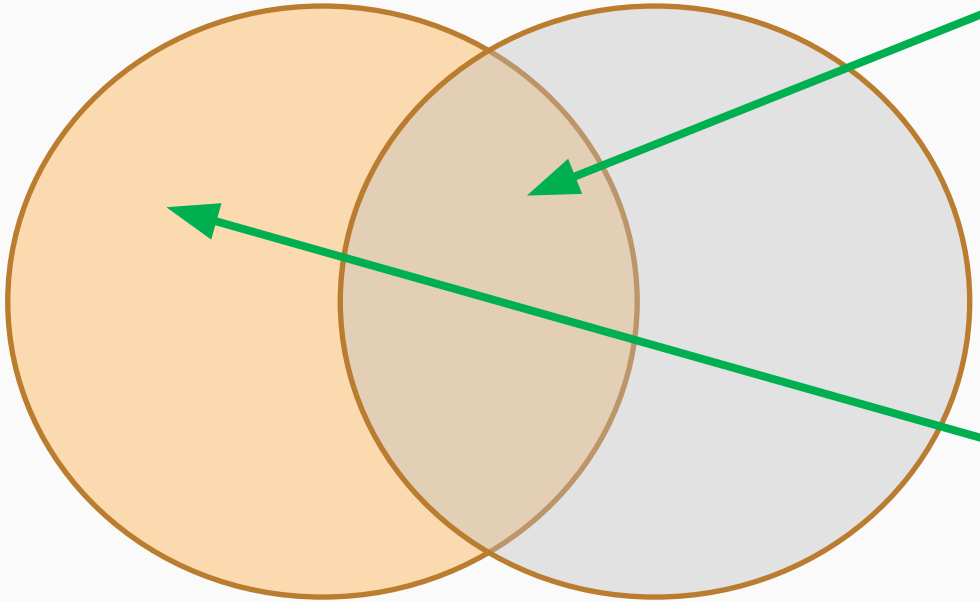
# Genome Arithmetic

- Most basic operations in **genome arithmetic** are **set theory operations**
- In this case, the sets we're interested in are **genomic features/loci** from **different BED files**
- A surprising amount of **integrative genomics** can be accomplished using these basic **set theory operations**

# Genome Arithmetic

Transcription Factor "A" Binding sites

Transcription Factor "B" Binding sites



**Intersection:**

$A \cap B$

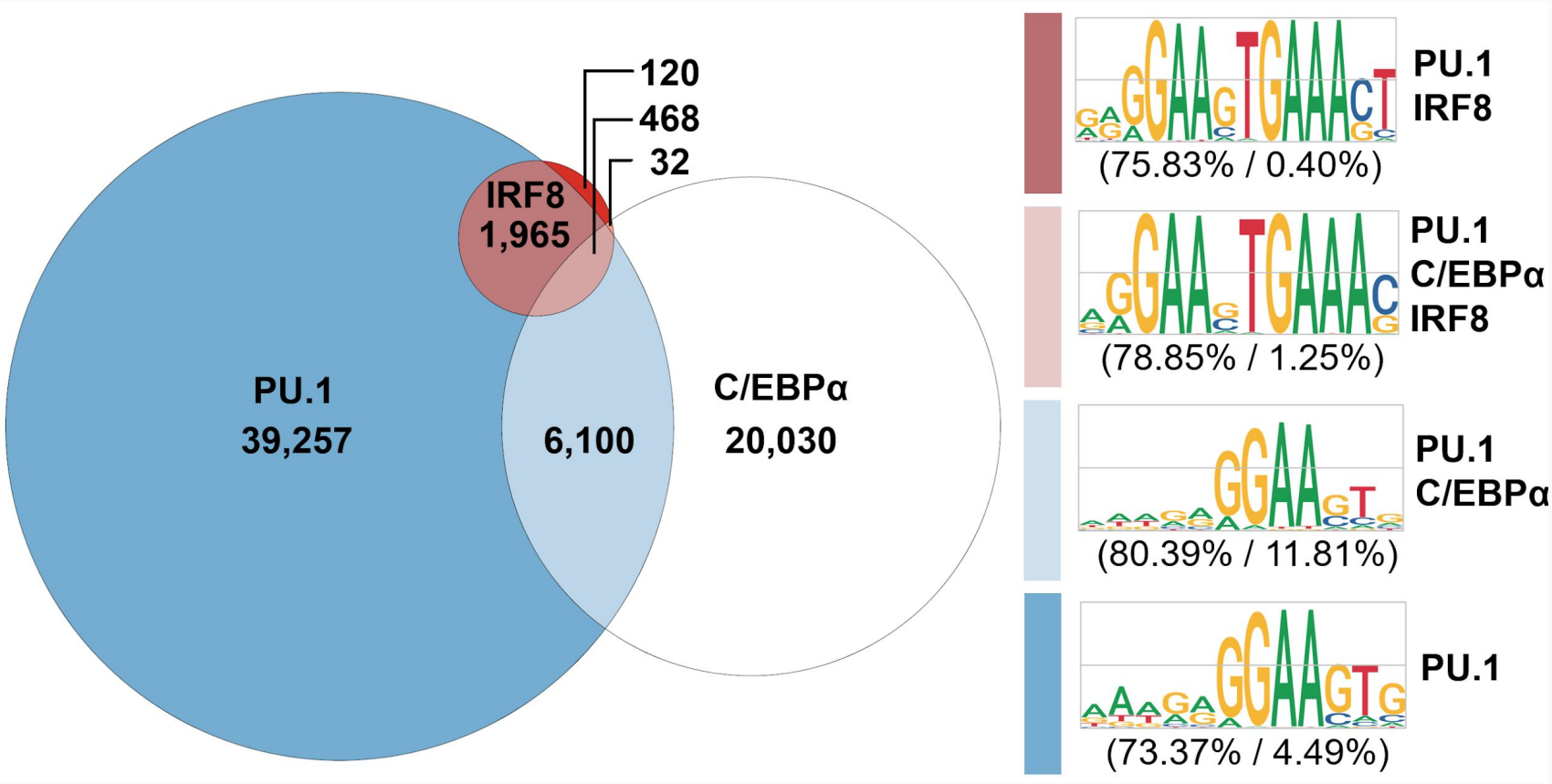
**Union:**

$A \cup B$

**Complement:**

$A \setminus B$ **One way!**

# Integrative Genomics Example

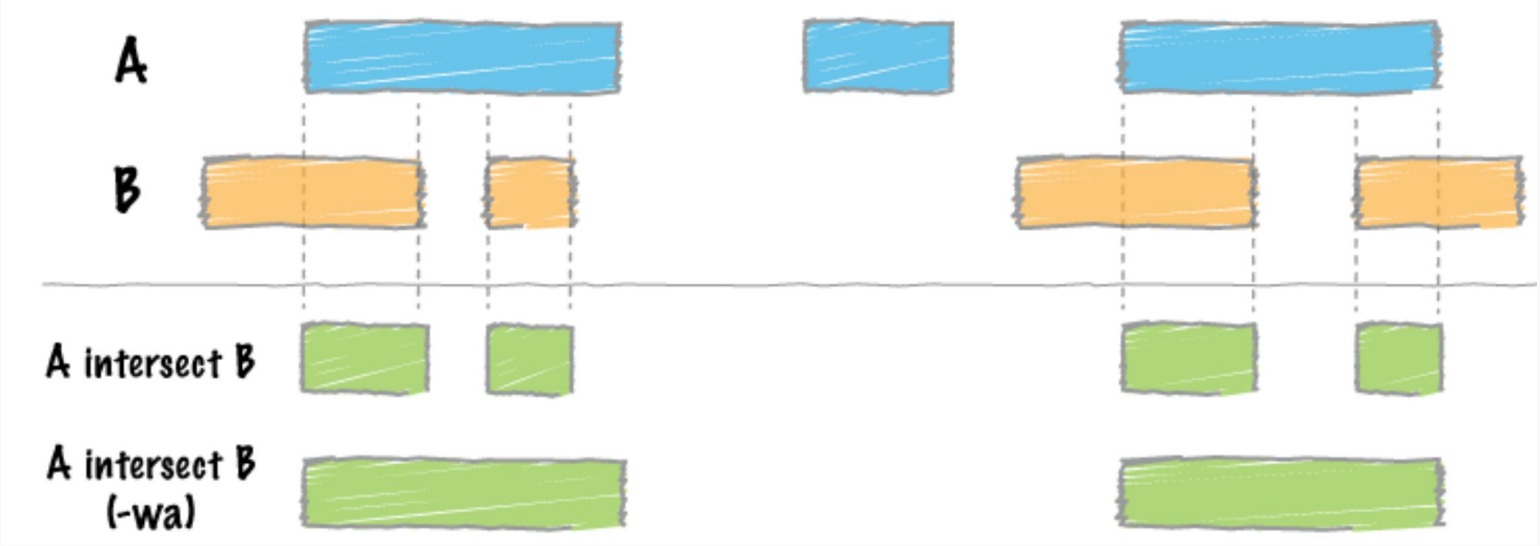


# BEDtools

- The self-proclaimed **“swiss army knife”** of genomic arithmetic
- A set of **command-line operations** that can **be performed on BED files** – sets of genomic features (chrom, start, end)
- Includes the **basic operations (intersections, unions, set differences)** as well as more **advanced functionalities**

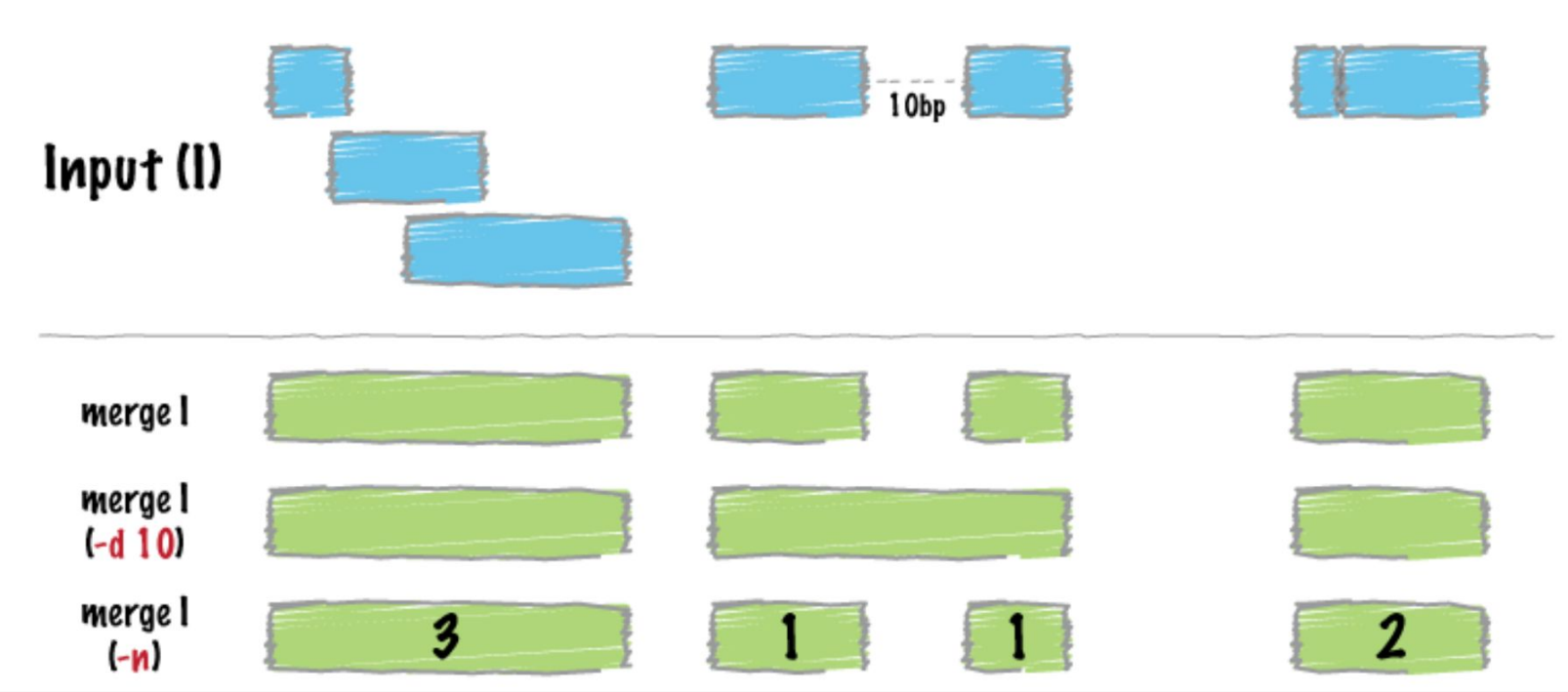
# Genome arithmetic in BEDtools

- The intersection of 2 sets of genomic features (ChIP-seq peaks for example):

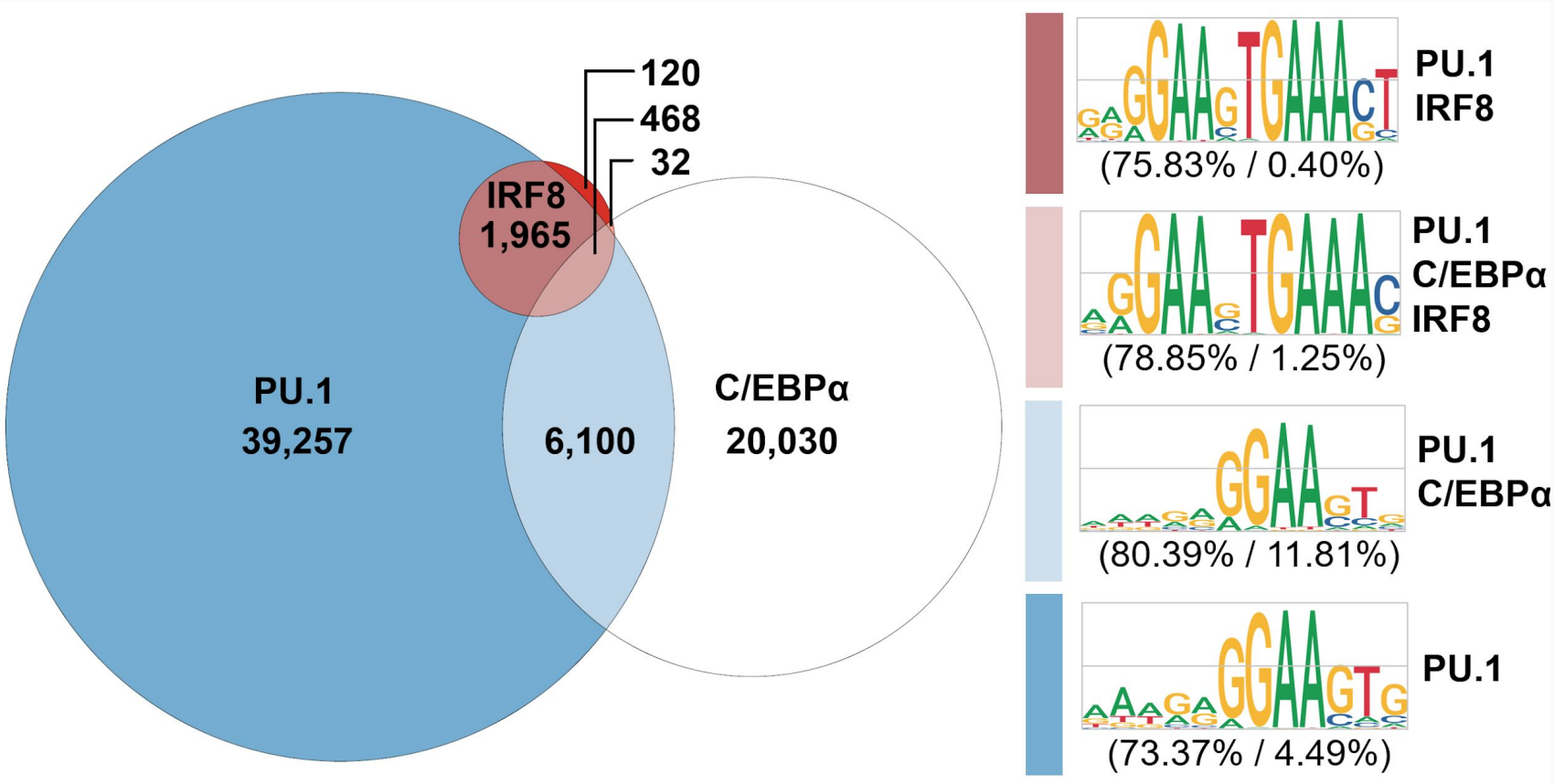


# Genome arithmetic in BEDtools

- The union set of genomic features (merge)



# Example: Differential Binding Motifs



# Genomic loci co-enriched with TFs

1. First need a unioned set of loci as a comparator
   - `cat PU1.bed CEBPA.bed IRF8.bed > loci.bed`
   - `sort -k1,1 -k2,2n loci.bed > loci_sorted.bed`
   - `bedtools merge -i loci_sorted.bed > loci_merged.bed`
2. Check which loci are co-enriched for binding of multiple TFs
   - `bedtools closest -a loci_merged.bed -b PU1.bed …`
   - `bedtools closest -a loci_merged.bed -b CEBPA.bed …`
   - `bedtools closest -a loci_merged.bed -b IRF8.bed …`
3. Examine proximity pattern for overlap of PU.1, C/EBPa, and IRF8 sites

# Genomic loci co-enriched with TFs

- Overlapping features are given a closest distance of “0”
- Non-overlapping features have a signed distance

loci_merged.bed

| | | |
|---|---|---|
| chr1 | 594823 | 595234 |
| chr1 | 724309 | 724623 |
| chr1 | 928347 | 928702 |
| chr1 | 992918 | 993253 |

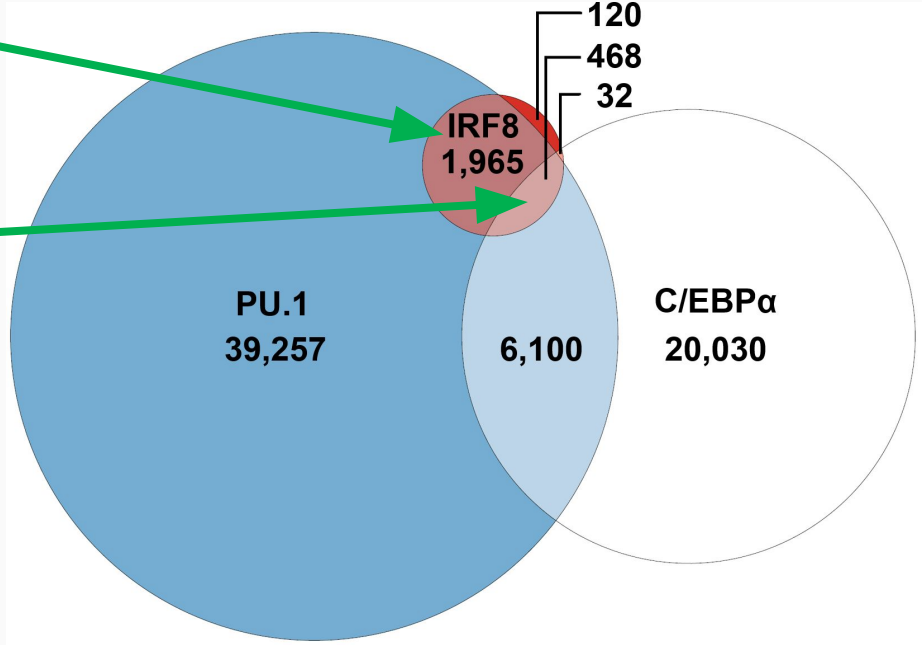
⋮

| PU.1 | C/EBPa | IRF8 |
|---|---|---|
| 0 | 0 | -58437 |
| 0 | 9283 | 0 |
| 10293 | 0 | 20394 |
| 0 | 0 | 0 |

Signed distance from given feature

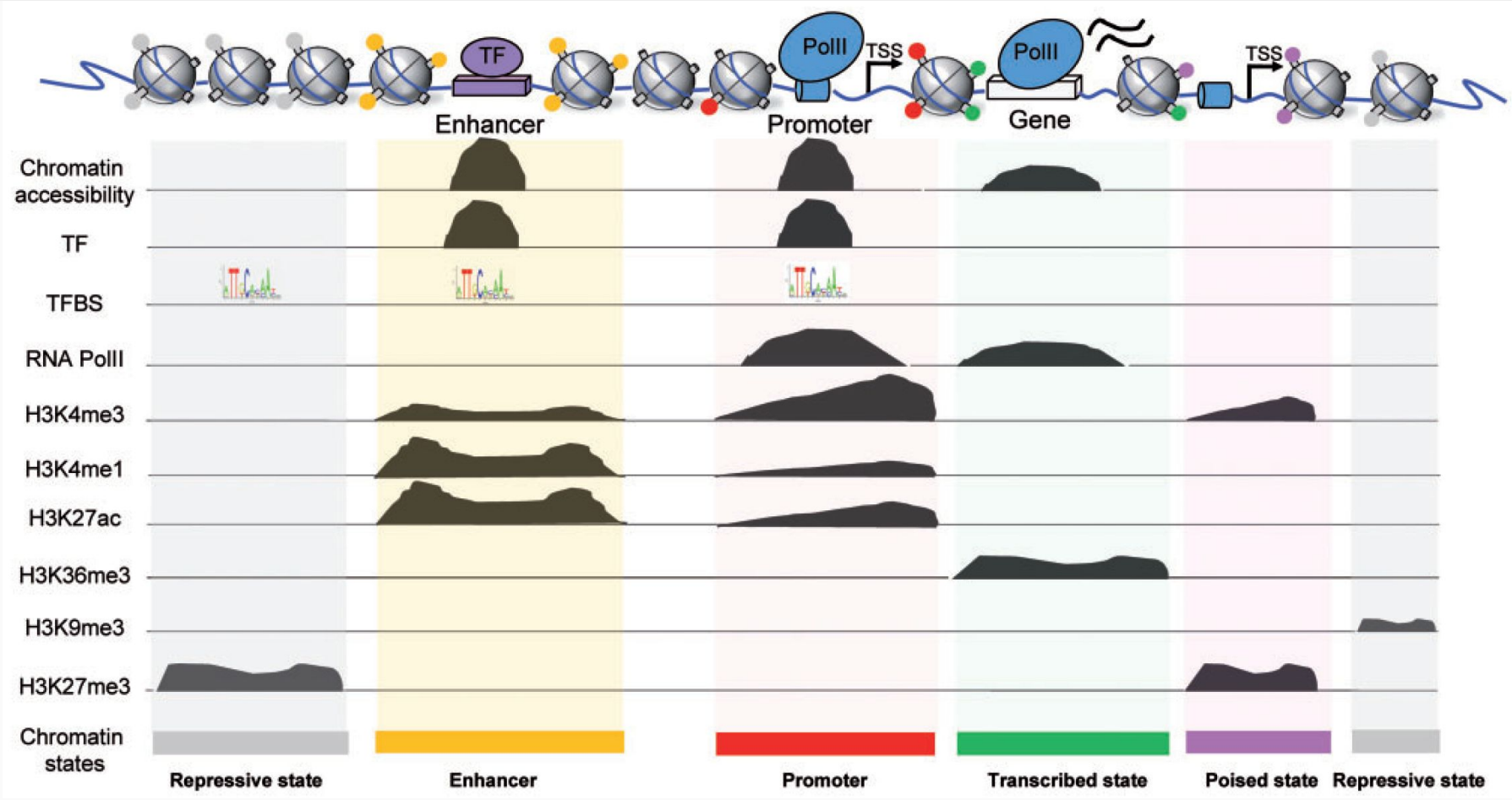
# Deconstructing the previous example

| PU.1 | C/EBPa | IRF8 |
|---|---|---|
| 0 | 0 | -58437 |
| 0 | 9283 | 0 |
| 10293 | 0 | 20394 |
| 0 | 0 | 0 |




- Count number of occurrences of each

# A larger example – Chromatin States



# Genome Associations

- Sometimes it's **not enough** to find and count **which features overlap others**
- Often times you would like to know **if the amount of overlap is surprising**
- For example, do 2 types of features overlap **more/less than should be expected**?

# Practical Example – SRF binding

- Given a complete set of genomic SRF binding sites, where does SRF bind?
- 2 ways to approach this problem:
  - Annotate the binding site locations (intergenic, intronic, UTR3, UTR5, CDS, etc.) and enumerate them
  - Does SRF preferentially bind certain locations more or less than we would expect?

# GAT – the Genome Association Tester

- GAT is a flexible, easy-to-use command line program to test genome associations
- Takes 3 files as input: **segments of interest**, **annotation segments**, and **genome workspace**

```
chr5    60627981        60628031        SRF.1
chr5    137801055       137801105       SRF.2
chr5    137800766       137800816       SRF.3
chr7    5570273 5570323 SRF.4
chr5    137827838       137827888       SRF.5
...
```

```
chr13   0       115169878       ws
chr12   0       133851895       ws
chr11   0       135006516       ws
chr10   0       135534747       ws
chr17   0       81195210        ws
...
```

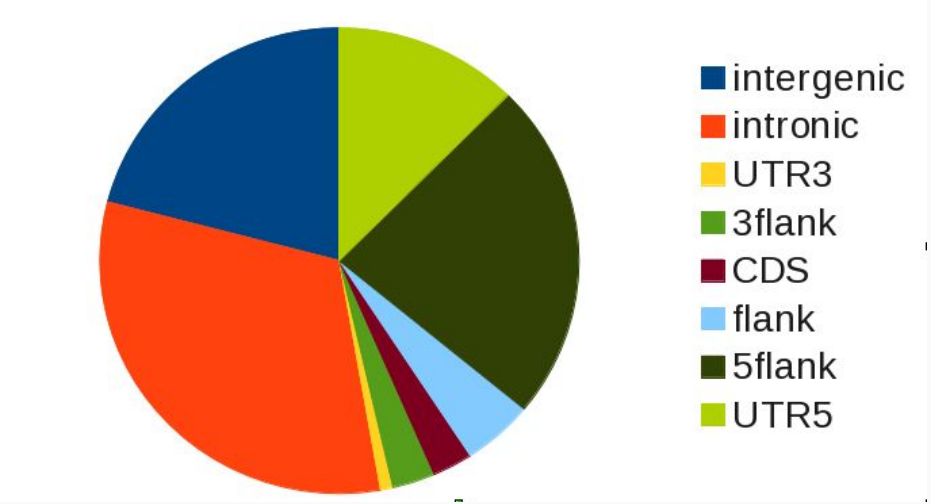
```
...
chr1    362640  367639  5flank
chr1    367640  367658  UTR5
chr1    367659  368594  CDS
chr1    368595  368634  UTR3
chr1    368635  373634  3flank
chr1    373635  616058  intergenic
chr1    616059  621058  3flank
chr1    621059  621098  UTR3
chr1    621099  622034  CDS
chr1    622035  622053  UTR5
chr1    622054  627053  5flank
...
```

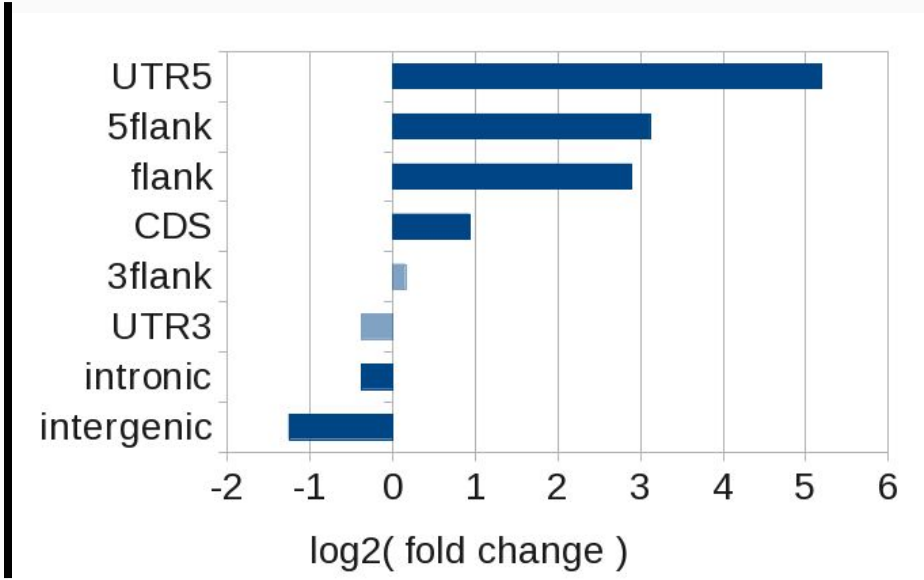
# GAT – association testing steps

- GAT performs the following:
  - Measures **observed nucleotide overlap** between your segments of interest and the annotation track
  - Randomizes the location of your segments of interest (user-specified number of times) while again measuring overlap to empirically obtained the **expected overlap**
  - Reports observed overlap, expected overlap, confidence intervals, p-values, **adj. p-values** for each annotation

# Enumeration vs. Association

**Where does SRF bind in the genome?**




**Enumeration** (via BEDtools/bedops)

**Association** (via GAT)

# Integrative Genomics Summary

- **BED files** are the **standard format** for most “downstream” genomics analyses
- **Set theory (genome arithmetic)** is the basis for operations used in integrative genomics
- **Chaining BEDtools operations** can take you a long way
- **Enumeration and association** can provide different answers (learn when to use each!)

# Recommended Tutorial

- http://quinlanlab.org/tutorials/bedtools/bedtools.html
  - Official bedtools tutorial
  - Walks you through many of the basic operations (intersect, merge, complement, genomecov)
  - Shows you how to do higher order operations (chaining multiple commands, performing PCA using bedtools and R)

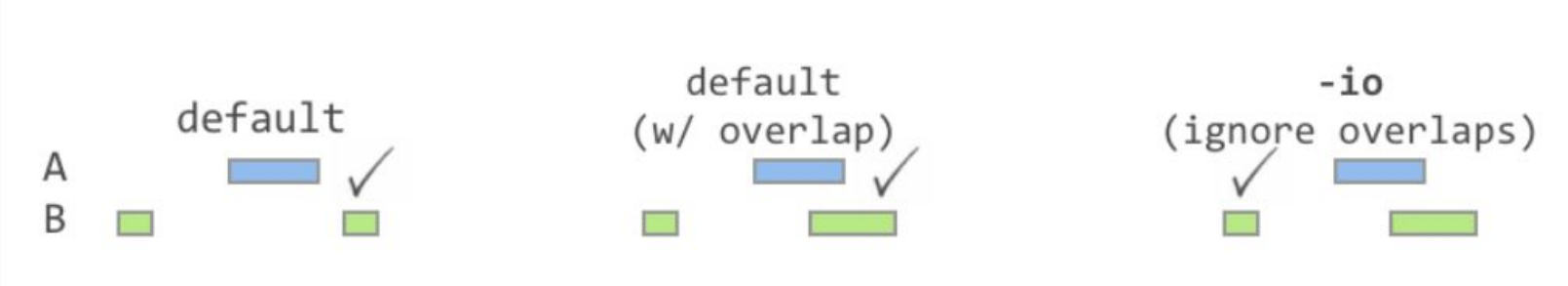
# Acknowledgements


- Much of the content in this lecture is from:
  - Quinlan Lab & bedtools (http://bedtools.readthedocs.io/en/latest/)
  - Heger et al. (2013) – GAT: A simulation framework for testing the association of genomic intervals
  - Jiang & Mortazavi (2018) – Integrating ChIP-seq with other functional genomics data

# Be sure to practice!

# Genome arithmetic in BEDtools

- Similar to intersect: “closest”
- Returns closest feature regardless of whether the feature intersects or not




- In practice, I use this much more than intersect

# Genome arithmetic in BEDtools

- The complement set of features


- Corresponds to every interval in your reference track that isn't covered

# Genome arithmetic in BEDtools

- Subtracting features from a BED file


- Contrasts with “complement” from previous slide

# Other BEDtools functions

- In addition to the basic set operations (intersection, union, set difference):
  - **Shuffle** – samples a genome file and outputs genomic features of the same size as your input with different locations
  - **Random** – generates pseudo-random intervals of a user-specified size
  - Resizing or moving features - **Slop, shift**