

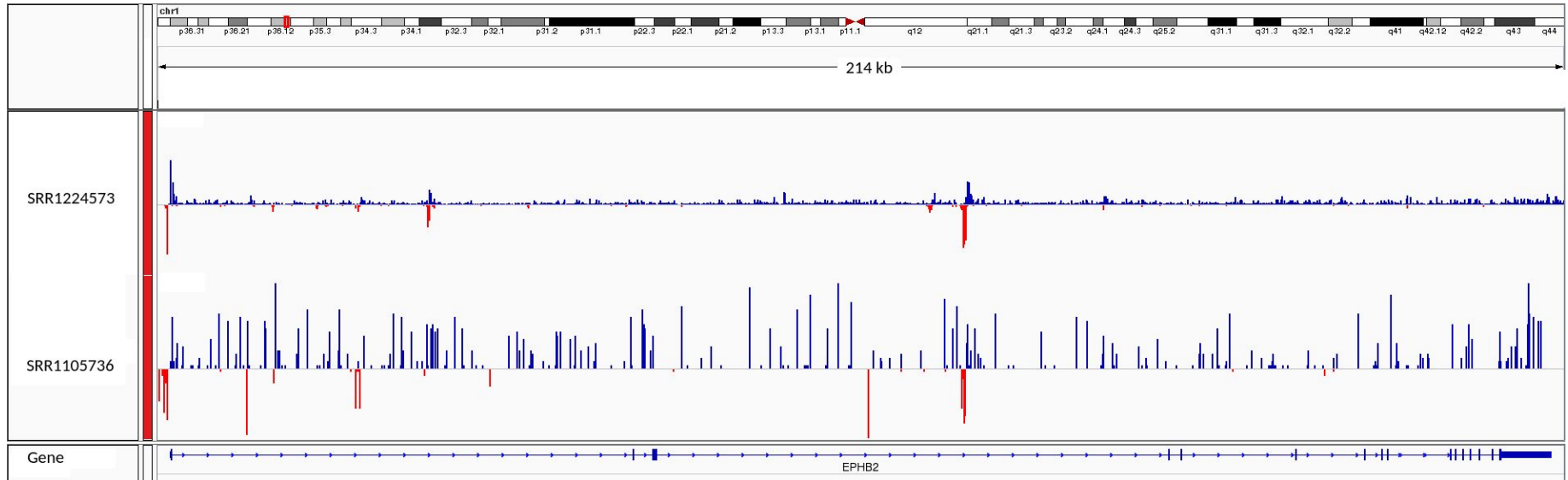
# Day 9: DNA enrichment analysis

## Part 1: Quality Control

Rutendo Sigauke

# Part 1: Quality Control

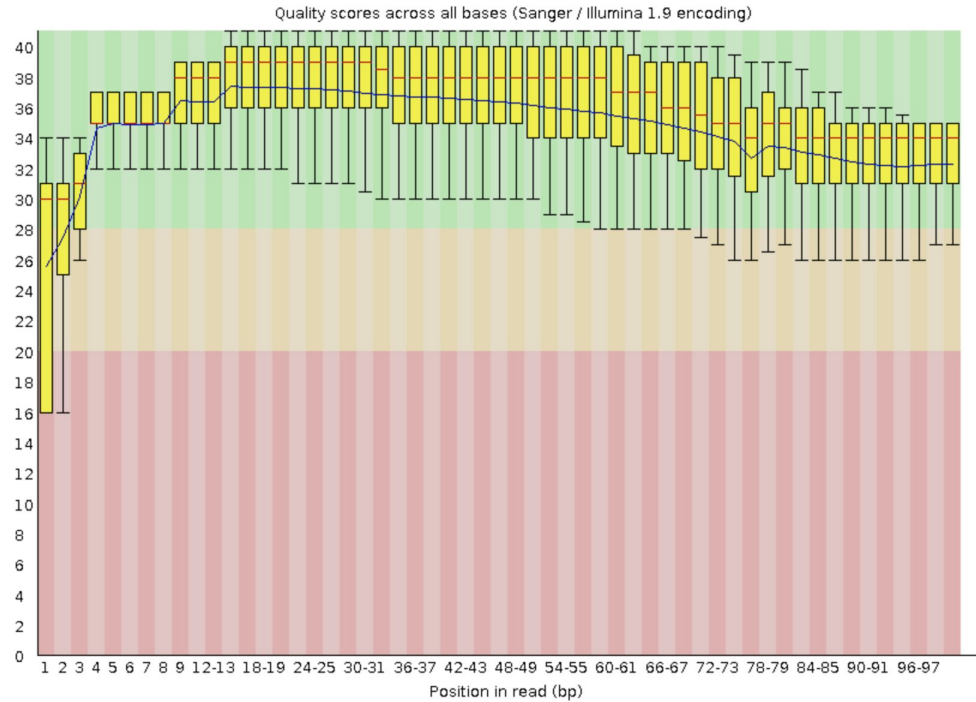
- Data quality affects the type and quality of information that can be obtained from a sequencing experiment
- What do we notice about the two samples below?



# Two Important Points

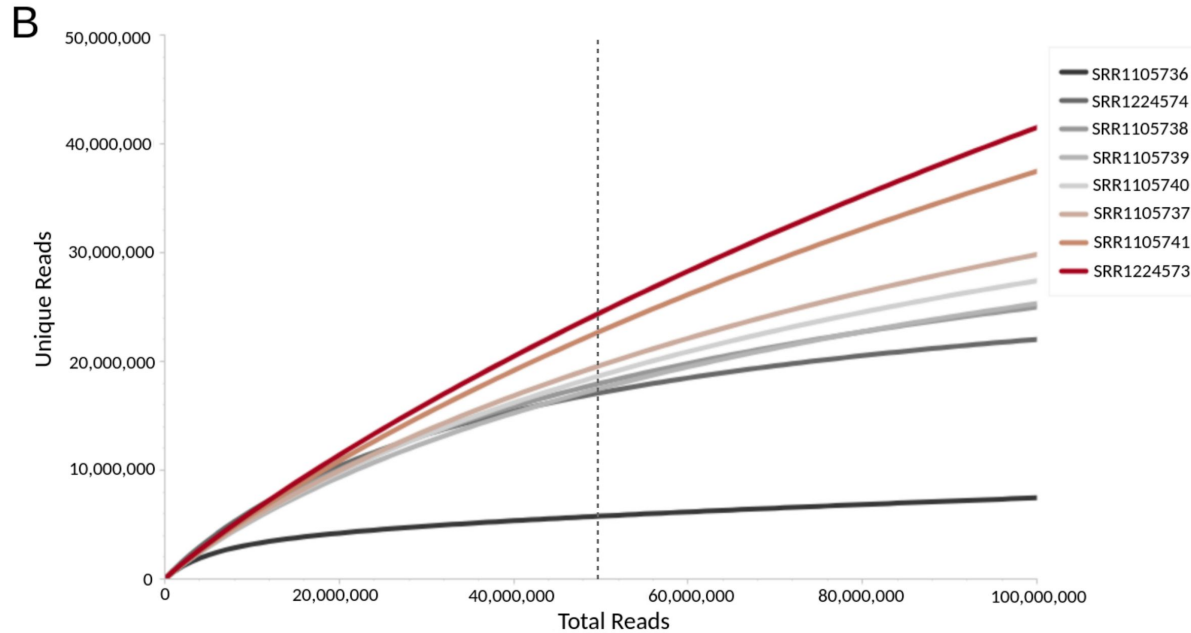
- 1. Coverage and complexity are *correlative* but are NOT the same**
- 2. Increased coverage/complexity does not guarantee better data**

# FastQC : Check overall quality of reads in a library (recap)



```
$ fastqc input.fastq output_directory
```

# Preseq : Measures library complexity



**Complexity** = # of unique reads / total reads sequenced

```
$ preseq c_curve -o complexity_output.txt input.sorted.bam
```

# But... what if I have a lot of samples?

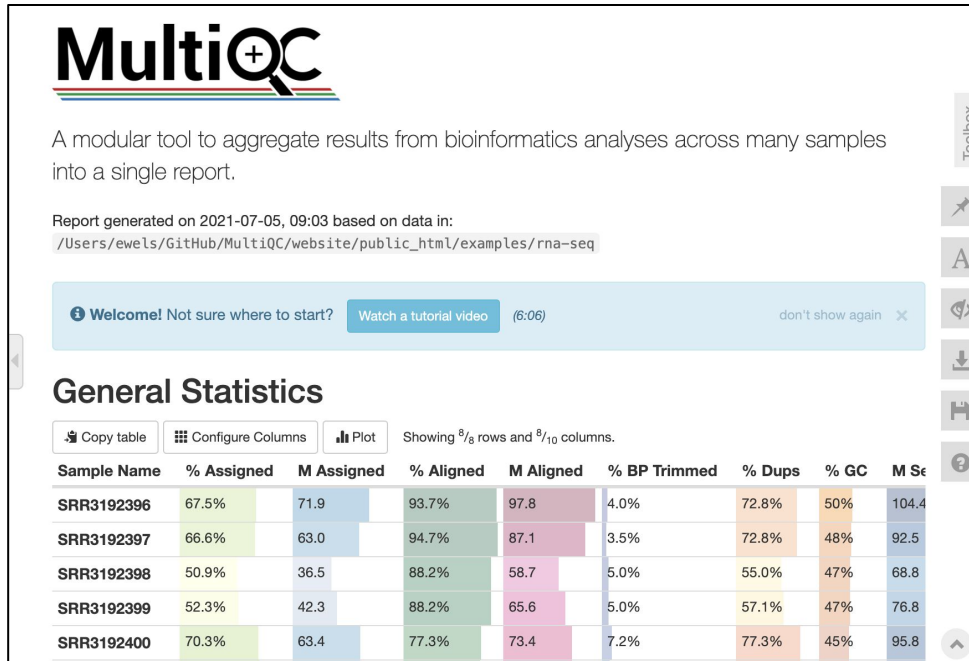
It would be nice to be able to view the stats on all of your samples at once....

Fortunately, there's a tool for that: **MultiQC**

<https://multiqc.info/>



# MultiQC : Summarizes all the QC metrics



**MultiQC**

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2021-07-05, 09:03 based on data in:  
/Users/ewels/GitHub/MultiQC/website/public\_html/examples/rna-seq

Welcome! Not sure where to start? [Watch a tutorial video](#) (6:06) [don't show again](#) ✕

### General Statistics

[Copy table](#) [Configure Columns](#) [Plot](#) Showing 5/8 rows and 8/10 columns.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% BP Trimmed	% Dups	% GC	M S <sub>e</sub>
SRR3192396	67.5%	71.9	93.7%	97.8	4.0%	72.8%	50%	104.4
SRR3192397	66.6%	63.0	94.7%	87.1	3.5%	72.8%	48%	92.5
SRR3192398	50.9%	36.5	88.2%	58.7	5.0%	55.0%	47%	68.8
SRR3192399	52.3%	42.3	88.2%	65.6	5.0%	57.1%	47%	76.8
SRR3192400	70.3%	63.4	77.3%	73.4	7.2%	77.3%	45%	95.8

```
$ multiqc qc_directory
```



# Running QC for multiple samples

1. Run fastqc (recap)
2. Map reads to genome (recap)
3. Summarize QC using MultiQC

Day9 QC Worksheet/Homework

# Part 2: ChIP-seq and MACS

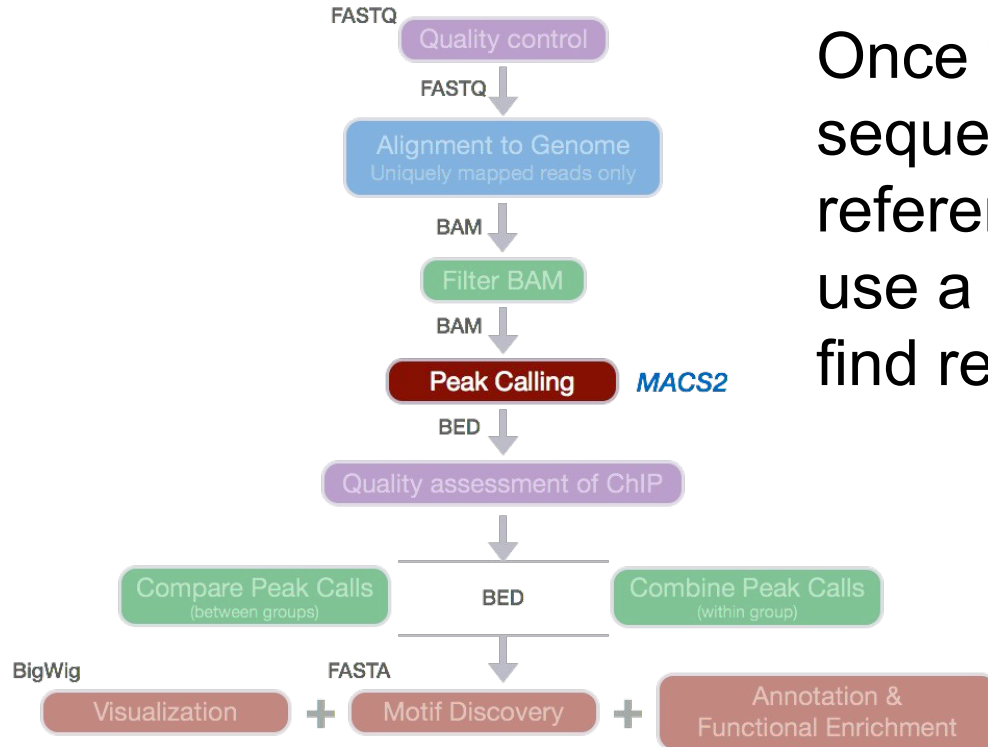
Jessica Westfall and Gilson Sanchez

# Learning Objectives

## Downstream analysis of ChIP-seq and ATAC-seq data

- Demonstrate the use of a peak calling program MACS2 to identify genomic regions with robust signal in each of these data types
  - control/input
  - ENCODE Blacklist
- Visualize the raw data and corresponding called peaks
- Additional tools for downstream differential analysis: DAStk

# Peak calling pipeline



Once you have your sequencing data aligned to reference genome, we will use a peak caller (MACS) to find region with enrichment

# ChIP-seq peak calling for enrichment

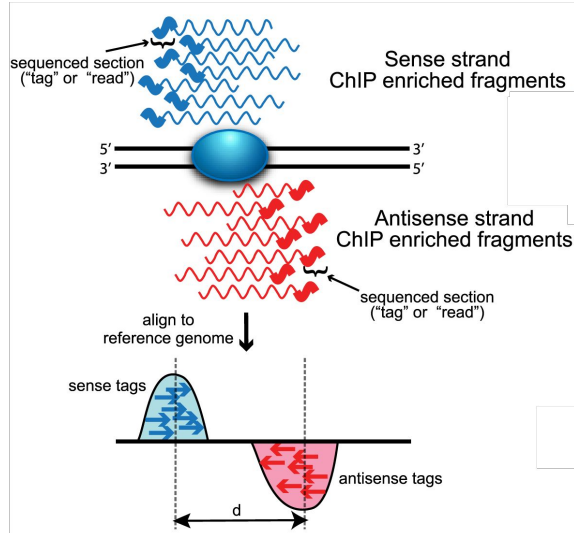


Image source: [Wilbanks and Facciotti, PLoS One 2010](#)

ChIP-seq identifies two type of enrichment

- **Broad peaks:** eg., histone modification. Here we are looking for broad peaks that cover entire gene bodies
- **Narrow peak:** eg., transcription factor binding. Here we are looking for regions of higher amplitude compared to background

# Model-based Analysis of ChIP-Seq (MACS)

## Introduction

---

With the improvement of sequencing techniques, chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq) is getting popular to study genome-wide protein-DNA interactions. To address the lack of powerful ChIP-Seq analysis method, we present a novel algorithm, named Model-based Analysis of ChIP-Seq (MACS), for identifying transcript factor binding sites. MACS captures the influence of genome complexity to evaluate the significance of enriched ChIP regions, and MACS improves the spatial resolution of binding sites through combining the information of both sequencing tag position and orientation. MACS can be easily used for ChIP-Seq data alone, or with control sample with the increase of specificity.

<https://pypi.org/project/MACS2/>

# MACS genomic input/control

## Controls are important!

- ChIP-seq and ATAC-seq are protocols that produce background noise as well as meaningful data, need controls to not call background noise as peaks
- p/q value cutoffs matter and should vary based on your experiment
- Know your data type: your experiment should inform the parameters of the peak caller
- Blacklist regions: some genomic regions almost always show up in these protocols so remove these regions using a Blacklist

# MACS2 peak calling recommendations

Data type	q value	--broad and --control flags	Reasoning
ChIP-seq for TF	<0.01	--control <INPUT>	TF ChIP-seq often has very abrupt, small peaks that are well defined, so narrow peaks is necessary, and a less stringent adjusted p value is likely needed than for other data types
ChIP-seq for histone marks	<0.0001	--broad --control <INPUT>	Histone marks are often broadly dispersed without very well defined edges so a broad peak tag is useful but a very low p value helps differentiate between background and data
ATAC-seq	<0.0001	--control <INPUT>	ATAC-seq should show peaks at open chromatin across the genome similarly to histone ChIP-seq data, but with more abrupt peaks, so no broad peak tag is needed



# MACS

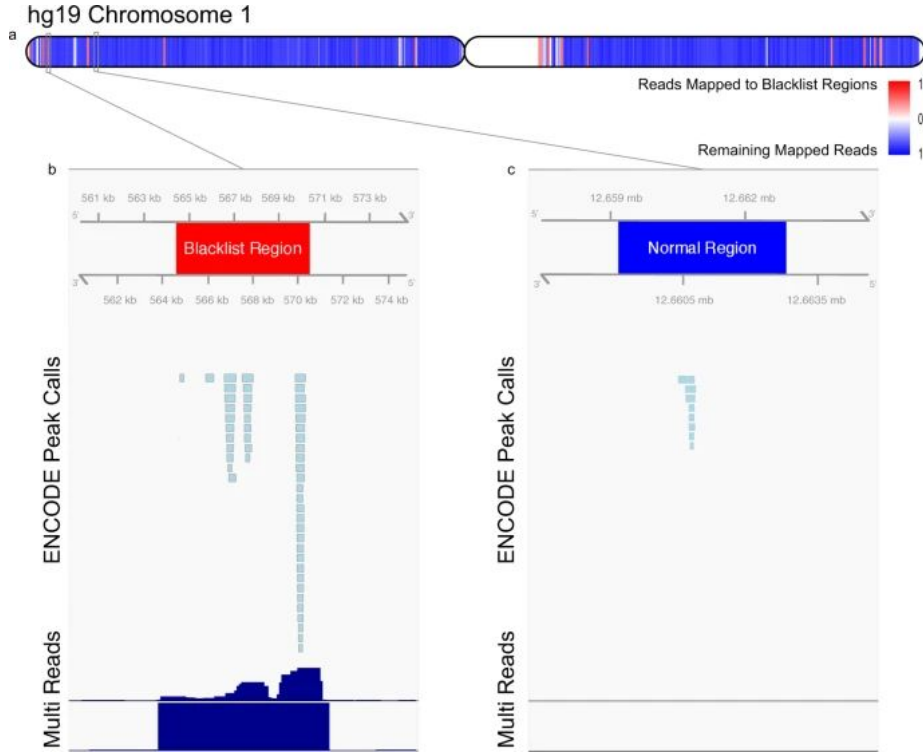
## Usage

```
macs2 [-h] [--version]
      {callpeak,bdgpeakcall,bdgbroadcall,bdgcmp,bdgopt,cmbreps,bdgdiff,filterdup,predictd,pileup}
```

Example for regular peak calling: `macs2 callpeak -t CHIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01`

Example for broad peak calling: `macs2 callpeak -t CHIP.bam -c Control.bam --broad -g hs --broad-cutoff 0.1`

# Blacklist regions should be remove



These regions contain repetitive regions across the genome and almost always are enriched in ChIP-seq data.

# Bedtools suite

Genomic analysis tool that has multiple tools to explore, process and manipulate genomic interval files.

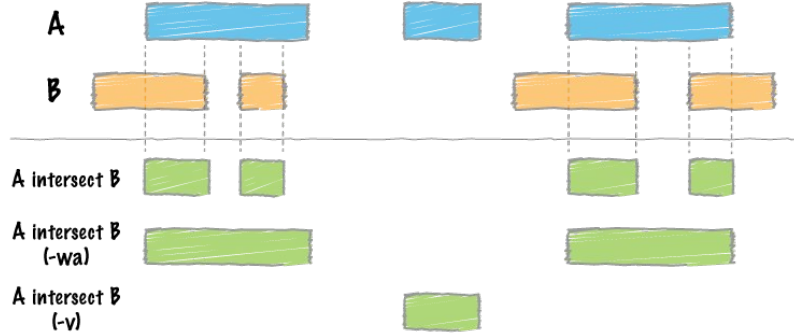
<https://bedtools.readthedocs.io/en/latest/>

Recommended tutorial to learn more about the available tools

<http://quinlanlab.org/tutorials/bedtools/bedtools.html>

# Bedtools intersect

Intersect w/  
1 database



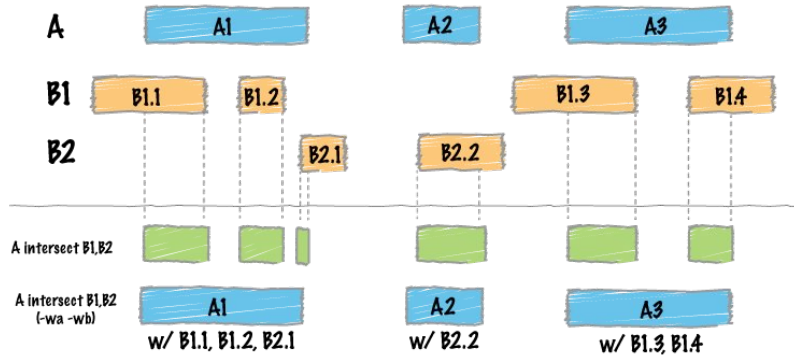
Usage:

```
bedtools intersect [OPTIONS] -a <FILE> \  
                        -b <FILE1, FILE2, ..., FILEN>
```

(or):

```
intersectBed [OPTIONS] -a <FILE> \  
                        -b <FILE1, FILE2, ..., FILEN>
```

Intersect w/  
2 or more databases



# Let's practice

1. Let's run MACS to call peaks on our bam files with a genomic input.  
\*Note there is an input and anti-BACH1/GABPA CHIP-seq bam files.
2. Remove problematic regions using bedtool intersect and a Blacklist region list

# MACS output

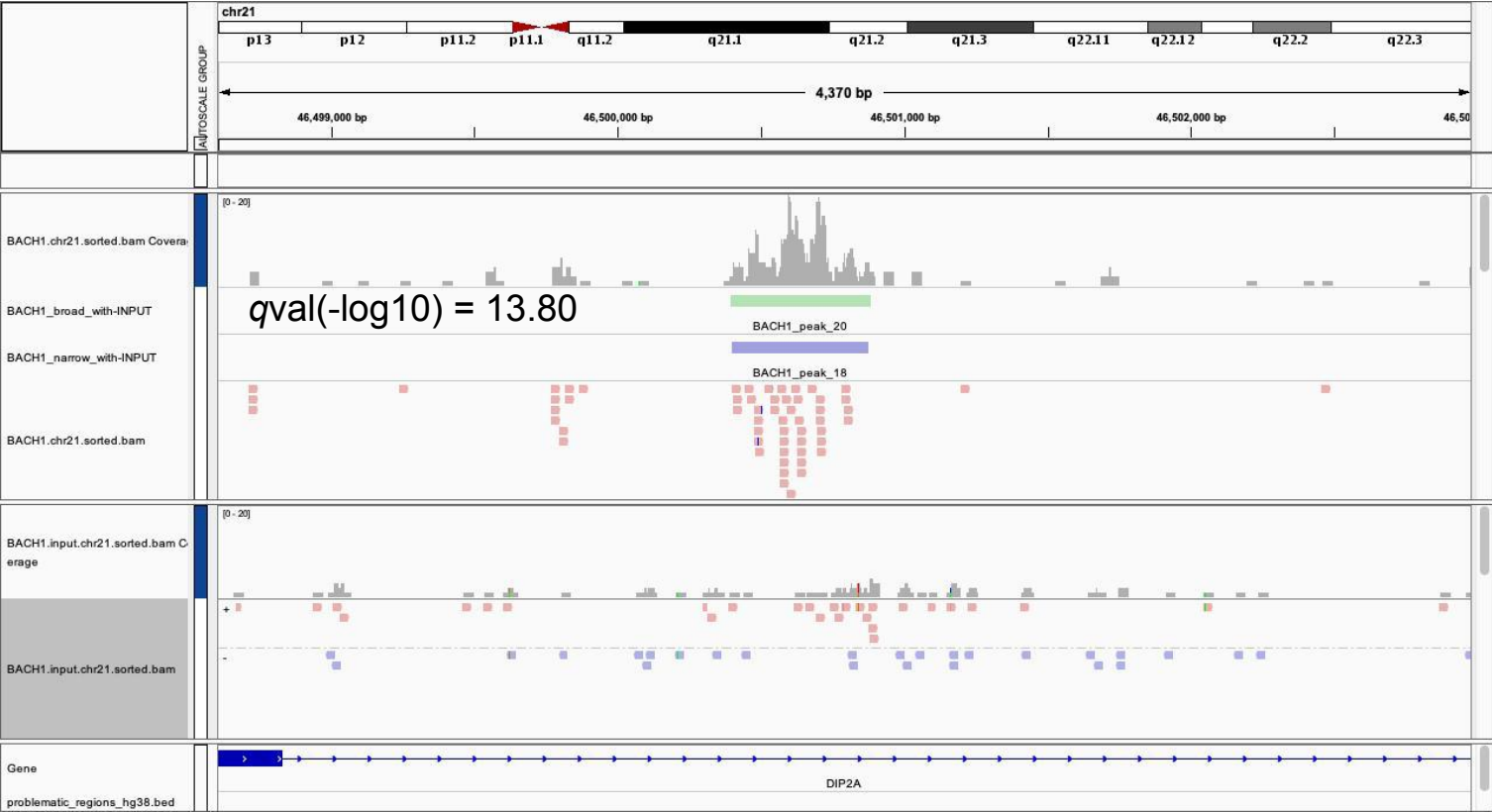
1. chromosome
  2. start coordinate
  3. end coordinate
  4. name
  5. score
  6. strand
  7. **signalValue** - Measurement of overall enrichment for the region
  8. **pValue** - Statistical significance (-log10)
  9. **qValue** - Statistical significance using false discovery rate (-log10)
  10. **peak** - Point-source called for this peak; 0-based offset from chromStart
- 
- The diagram illustrates the MACS output fields. A black bracket on the left groups fields 1 through 6, labeled 'Standard BED file fields'. A red bracket on the right groups fields 7 through 10, labeled 'narrowPeak specific fields'.

Image: [https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05\\_peak\\_calling\\_mac.html](https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_mac.html)

# Broad vs Narrow Peaks



# Broad vs Narrow Peaks

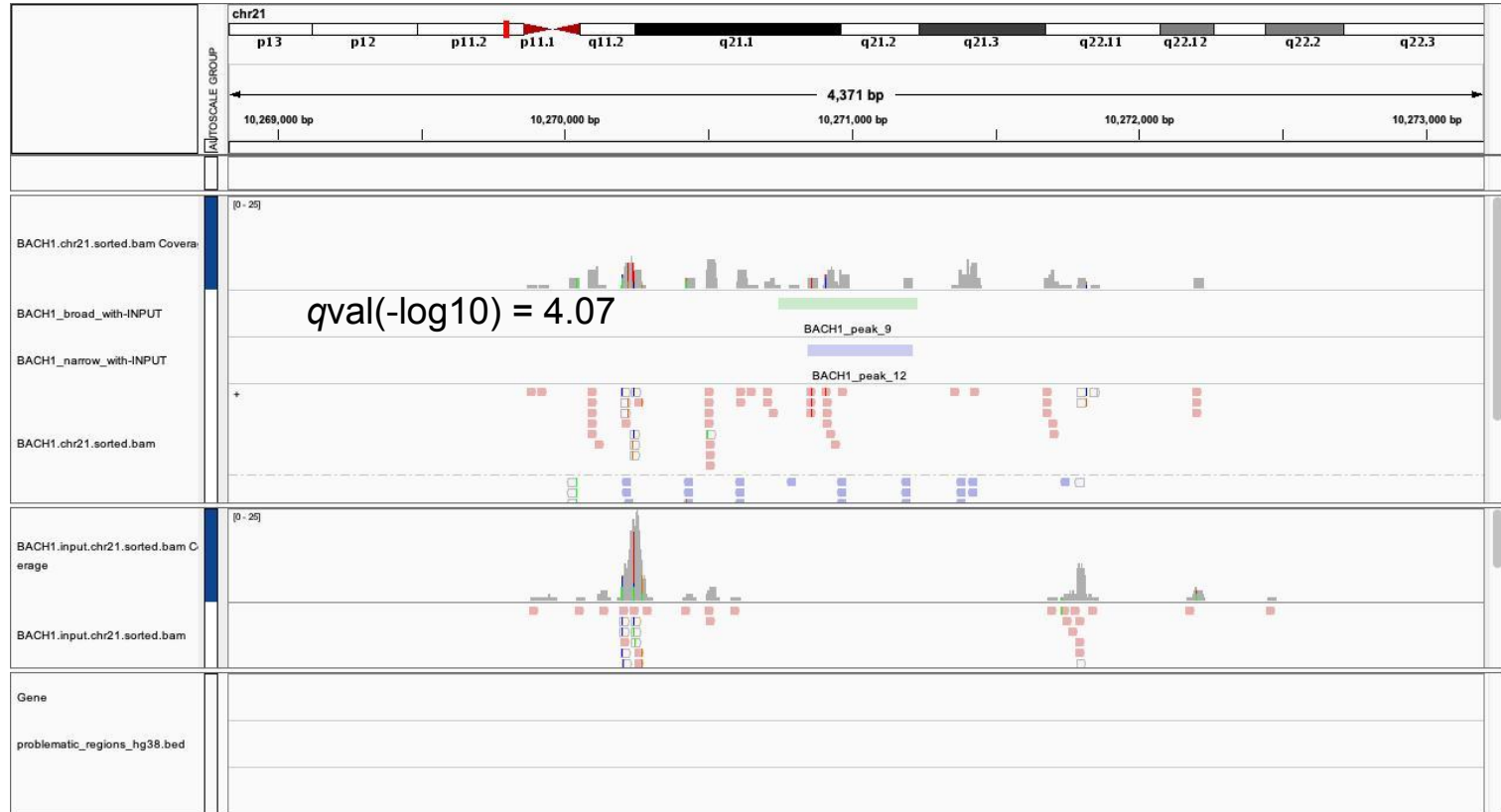




# Broad vs Narrow Peaks

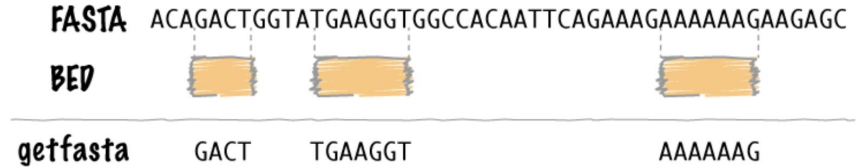


# Broad vs Narrow Peaks



# MEME and TOMTOM

1. Convert peaks to fasta file

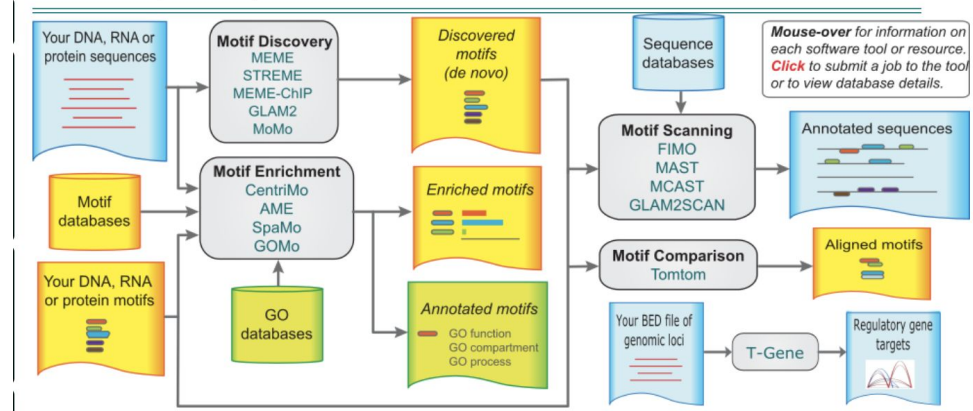


2. MEME suite

- Upload the fasta file to [MEME](#)
- Push MEME to [TOMTOM](#)

## The MEME Suite

Motif-based sequence analysis tools



# Additional Peak Caller

Depending on your experiment, there may be other peak callers that are more suitable options

Fstitch <https://github.com/Dowell-Lab/FStitch>

SICER <https://zanglab.github.io/SICER2/>

PeakSeq <https://www.nature.com/articles/nbt.1518>

Hpeak <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-369>

PeakRanger <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-139>