



MSU Next Generation Sequencing – Summer 2012

ChIP-Seq Studies

István Albert

Bioinformatics Consulting Center
Penn State

What's New! Hot off the press!



The screenshot shows the homepage of BMC Bioinformatics. At the top left is the BMC logo and the journal title "BMC Bioinformatics". To the right is a yellow box displaying the "IMPACT FACTOR 3.03". On the far right is a search bar with the text "BMC Bioinformatics". Below the header is a navigation menu with links: Home, Articles, Authors, Reviewers, About this journal, and My BMC Bioinformatics. The "Articles" link is highlighted. In the main content area, there is a section titled "Software" with a blue "Open access" button. A blue banner highlights an article titled "BioWord: a sequence manipulation suite for Microsoft Word" by Laura J Anzaldi, Daniel Muñoz-Fernández and Ivan Erill. Below the banner, a note says "For all author emails, please [log on](#)". At the bottom, publication details are provided: "BMC Bioinformatics 2012, 13:124" and "doi:10.1186/1471-2105-13-124", with the date "Published: 7 June 2012". A blue link at the bottom left reads "Abstract (provisional)".

BMC
Bioinformatics

IMPACT
FACTOR
3.03

Search BMC Bioinformatics

Home Articles Authors Reviewers About this journal My BMC Bioinformatics

Software Open access

BioWord: a sequence manipulation suite for Microsoft Word

Laura J Anzaldi, Daniel Muñoz-Fernández and Ivan Erill

For all author emails, please [log on](#).

BMC Bioinformatics 2012, 13:124 doi:10.1186/1471-2105-13-124
Published: 7 June 2012

Abstract (provisional)

What's New! Hot off the press!

The screenshot shows a web browser window with the URL www.biomedcentral.com/1471-2105/13/124/abstract. The page is for the article "BioWord: a sequence manipulation suite for Microsoft Word" by Laura J Anzaldi, Daniel Muñoz-Fernández and Ivan Erill, published in BMC Bioinformatics 2012, 13:124. The page includes sections for Software, Abstract (provisional), Background, and Related literature. A large red watermark diagonally across the page reads "Only works with Windows version of Word!".

IMPACT
FACTOR
3.03

Log on

BioMed Central Journals

BMC Bioinformatics

Search BMC Bioinformatics for

Home Articles Authors Reviewers About this journal My BMC Bioinformatics

Software

BioWord: a sequence manipulation suite for Microsoft Word

Laura J Anzaldi, Daniel Muñoz-Fernández and Ivan Erill

For all author emails, please [log on](#).

BMC Bioinformatics 2012, **13**:124 doi:10.1186/1471-2105-13-124 Published: 7 June 2012

Abstract (provisional)

Background

Manipulation, editing and basic processing of DNA and protein sequences has rapidly become a necessary skill for practicing biologists across a wide swath of disciplines. In spite of this, most everyday sequence manipulation tools are distributed across several programs and web servers, sometimes requiring installation and typically involving frequent switching between applications. To address this problem, here we have developed BioWord, a macro-enabled self-installing template for Microsoft Word.

Open access

BMC Bioinformatics Volume 13

Viewing options Abstract Provisional PDF Associated material About this article Readers' comments Related literature Cited by on Google Scholar Other articles ▶ on Google Scholar ▶ on PubMed

Only works with Windows version of Word!

PSU: Bioinformatics Consulting Center

<http://bcc.bx.psu.edu>

The screenshot shows a web browser window with multiple tabs open, including one for 'NGS 2012 Information' and another for 'BioStar'. The main content area displays the homepage of the Bioinformatics Consulting Center at Penn State. The page features a large banner with the text 'Bioinformatics Consulting Center at Penn State' overlaid on a background of various bioinformatics data visualizations like pie charts and bar graphs. Below the banner, there's a navigation menu with links to 'Home page', 'Analysis Reports', 'Request Analysis', and 'Table of contents'. The main content section starts with a 'Welcome!' heading and a brief description of the center's services. It then has two main sections: 'Analysis Reports' and 'Education'. The 'Analysis Reports' section includes a list of links for pipelines, requests, and examples. The 'Education' section mentions BioStar and bioinformatics courses. On the right side, there's a 'Table Of Contents' sidebar with links to 'Welcome!', 'Analysis Reports', 'Education', and 'Contact us', along with a 'Quick search' bar.

Bioinformatics Consulting Center at Penn State

Home page | Analysis Reports | Request Analysis | Table of contents > index

Welcome!

The Bioinformatics Consulting Center (BCC) at Penn State offers bioinformatics data analysis services that cover multiple application domains of high throughput sequencing.

Our services range from routine data manipulation steps: quality filtering, genomic alignments, variant calling to in-depth data analyses and developing novel, domain specific methodologies.

Analysis Reports

The results that we generate are available via the [Analysis Reports](#) web page.

- Learn more about the [Analysis Pipelines](#)
- Information on how to [Request Analysis](#)
- Consult the [Analysis Examples](#) for examples of what each report includes.

Education

Our center spearheaded the creation of [BioStar](#) a community driven question-and-answer website for bioinformatic research. Members of the center also teach credit bearing bioinformatics related courses :

- [Bioinformatics Courses](#) offered in Fall and Spring semesters at the UP and Hersey campuses

Table Of Contents

- [Welcome!](#)
- [Analysis Reports](#)
- [Education](#)
- [Contact us](#)

Quick search

Enter search terms or a module, class or function name.

BioStar: <http://www.biostars.org>

The screenshot shows a web browser window with several tabs open at the top, including "ngs-2012] TODO this mon", "Day 2: BLAST, Python, and", "NGS 2012 Information page", "Show questions - BioStar", and "Ialbert/ngs-install". The main content area is the BioStar website, which has a dark header bar with links for "BioStar", "Tags", "Users", "Badges", "About", "FAQ", "Search", and a user profile for "Istvan Albert". Below the header, there are navigation links for "Recent", "My Tags", and "Questions", with "Questions" being the active tab. A "Unanswered (1)" badge is visible. The main content area displays a list of unanswered questions:

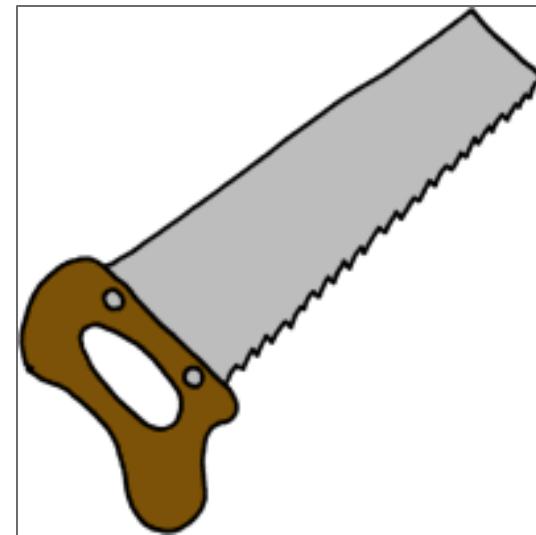
- Command line Blastplus strange results**
0 votes, 0 answers, 1 view. Tags: blastn, blast. Posted 45 min ago by [nupurgupta0806](#) 0•1.
- Meaning of HG and NA prefix for 1000 Genomes Project sample notation**
0 votes, 1 answer, 7 views. Tags: 1kg, hg, na, notation, prefix. Posted 2 hrs ago by [mubozi](#) 0•1.
- non-uniquely mapped reads**
2 votes, 1 answer, 38 views. Tags: bwa, non-unique, alignment. Posted 8 hrs ago by [Wen.Huang](#) 58•1•6.
- Standalone BLAST issue**
1 vote, 0 answers, 16 views. Tags: blast. Posted 8 hrs ago by [chazaris](#) 0.
- How does a Bioinformatics Scientist document his/her work?**
56 votes, 9 answers, 128 views. Tags: workflow, documentation, bioinformatics. Posted 1 days ago by [anjan.purkayastha](#) 17•4.
- Why are almost all protein or mRNA abundance measures transformed by log?**
3 votes, 1 answer, 31 views. Tags: protein, mrna, statistics, log, normalization. Posted 14 hrs ago by [zhilongjia](#) 1•3.

At the bottom of the list, there is a "Sort by: rank" dropdown menu.

Bioinformatics tools

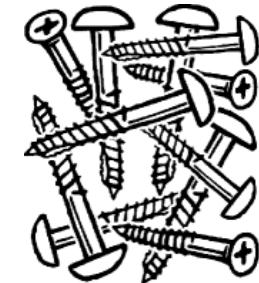
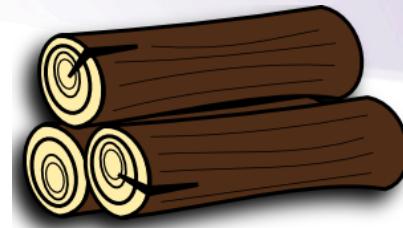


vs



Bioinformatician's Toolbelt

- Lumber → data
- hammer, saw, pliers → (mapper, samtools, bedtools)
- Glue, nails, screws, ductape → scripts → shell, python, perl, awk



You can do just about anything with these!
You don't need specialized tools for most tasks!
Common sense and a good understanding of what each does

Great ChIP-Seq video series see the Videos tab

Screenshot of a web browser showing a BioStar post about ChIP-seq data analysis.

Address bar: www.biostars.org/post/show/45206/chip-seq-data-analysis/

User navigation: Gmail, Google, Loco, GH, BioStar, BCC, Calendar, 597D, NGS

User info: Istvan Albert (1.5k) | Logout

Post navigation: Recent, My Tags, Questions, Unanswered, Tools, Videos, Planet, Forum, Tutorials, Search, New Post!

Section title: Video: Chip-seq data analysis

Video thumbnail: A diagram illustrating the major steps of ChIP-seq data analysis. It shows the workflow from crosslinking (using formaldehyde) to sequencing. Key steps include shearing (sonication or MNase digestion), immunoprecipitation (using antibodies), and sequencing. The diagram also shows DNA purification and library preparation (using PCR).

Video controls: Share, More info

Video rating: 2 (highlighted)

Video author: Raony Guimarães (created 21 days ago)

Video views: 15 • 6

Text below video: Click to go to YouTube

Text below video: Standard ChIP-seq data processing I

Todays' goal

- Understand the principles of ChIP-Seq analysis
- Run through an entire mapping procedure – “view from the trenches”
- We provide scripts to download the data
- Characterize the results by their density around other genomic features
- Compare results by different peak callers

Next big Topic: Protein - DNA interactions

ChIP-Chip and ChIP-Seq studies

- **ChIP** → Chromatin Immuno-Precipitation
(used during sample preparation)
- **Chip** → microarray technology to detect bound genomic locations
- **Seq** → high throughput sequencing to detect bound genomic locations

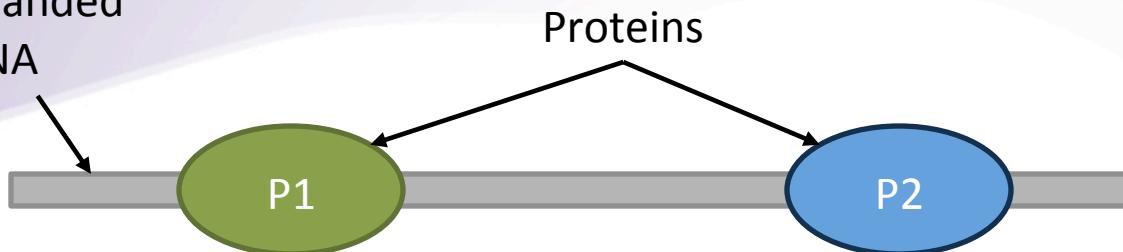
Chromatin Immuno-Precipitation

It is a well known methodology to detect:

- transcription factor binding
- polymerase binding
- chromatin structure and modifications
- etc...

ChIP: Quick Overview

Double
stranded
DNA



Crosslink bound
proteins



Fragment/digest DNA
around bound locations



Isolate with protein
specific antibody



Reverse cross link

This is the DNA fragment
that gets sequenced

Sample origins

Understanding the sample preparation is essential for analysis

- **WGS** whole genome sequencing (shotgun) → random DNA fragments covering the entire genome
- **Chip-Seq** → DNA fragments covering bound locations in the genome

(we will keep adding to this list as we cover new techniques)

The ChIP output

- a DNA sample **enriched*** for fragments associated with the events under study
- We are measuring an ensemble of cells that may be in different states
- Coverage depends on the number of sites, efficiency of the IP (precipitation) step.
- Accuracy depends on **fragmentation** strategy: sonication, MNASE digestion, lambda-nuclease digestion

Note: Lots of **other DNA fragments** can make it through!

Chip-Seq → High throughput sequencing

- Fragments are sequenced
- Aligned against genome (Bowtie is a good choice for Chip-Seq)
(Note: today at 4pm we have a presentation in the Berg auditorium by the PI directing the Bowtie development!)
- The output is in the form of the intervals (start/end) where each read matches the genome

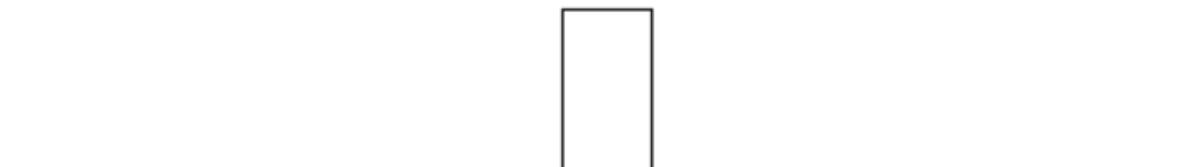
Type of events of interest



Long-range
(e.g., histone
modifications)



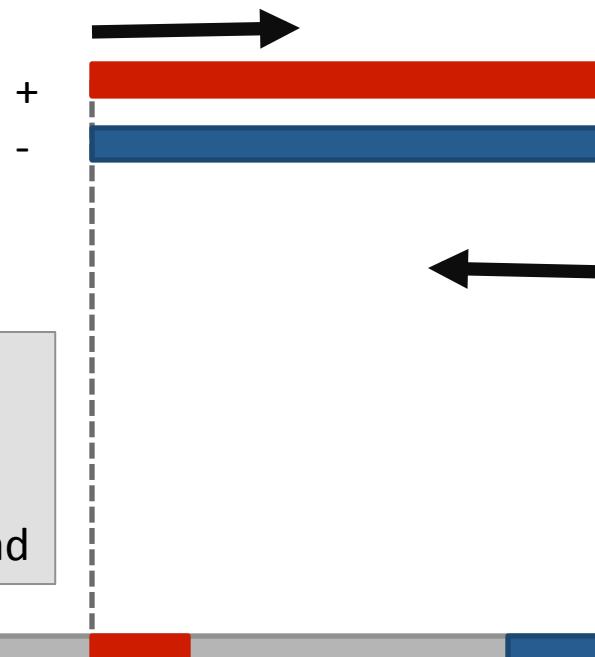
Mid-range
(e.g. polymerase
binding)



Punctate
(e.g. TF
binding)

Single End sequencing

Original two stranded DNA fragment

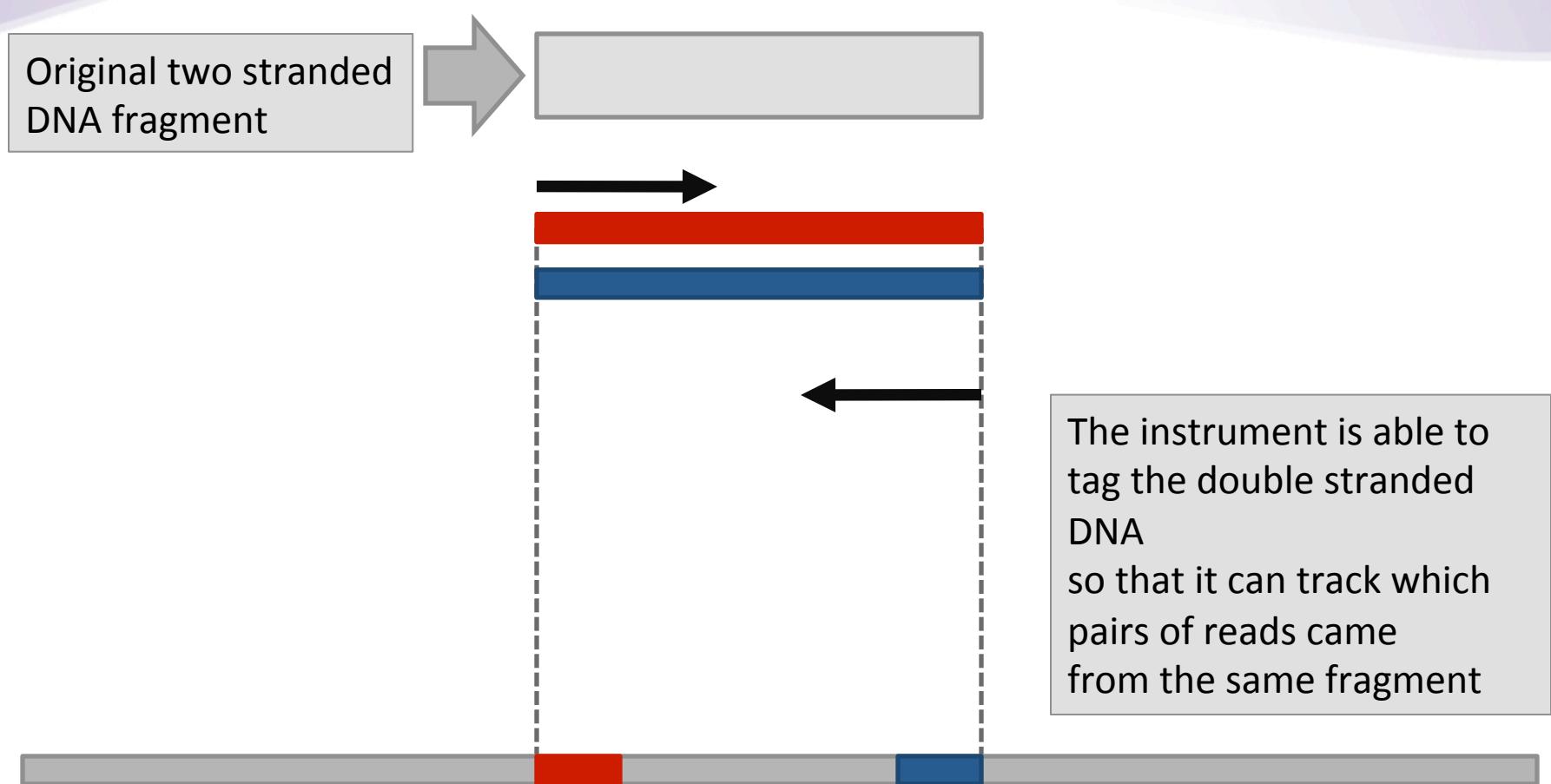


the original DNA fragment could be longer/shorter than the length of the read

Sequencing proceeds from the 5' to 3'
This is where we get reads from.

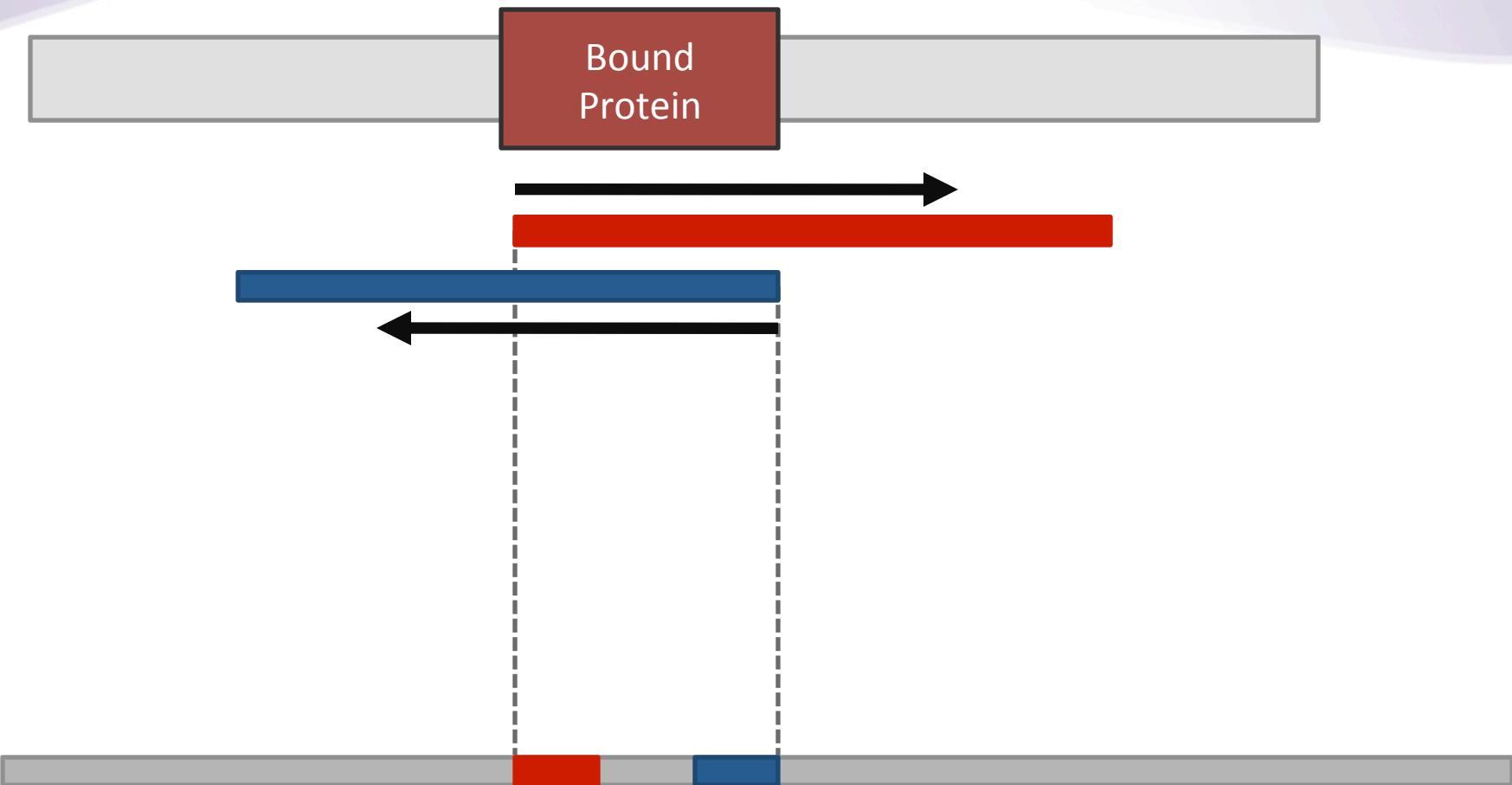
Single end read: we don't know which left border is paired up with which right border.

Paired End sequencing



Data will come in pairs of right-left border
We know the width of each fragment!

Other techniques: Chip-Exo



The reads are longer than the bound location – allows one to measure short-punctate locations. Other sample preparation techniques are always been developed. It is essential to know the sample preparation.

Alignment to the genome

After alignment we get genomic intervals for each read, minimally:

chrom, start, end, strand

The fragment 5' end locations will then

the **start** coordinate for the + strand
the **end** coordinate for the – strand

Other considerations

- For single end sequencing **each fragment** may correspond to **0, 1 or 2 reads**. If it has **two reads** we don't know which two formed the fragment
- For paired end sequencing each fragment corresponds to **0, 1 or 2 reads**. (1 no mate). We know which **two reads** correspond to one another.

Peak Calling

- Process of finding the locations enriched due to events of interest

We will need to define

- **Peak Region** - contiguous set of basepairs that belong to a peak
- **Enrichment Level** - read-based measure of supporting evidence

Peak calling: base pair level measurements

- Number of fragments overlapping that position
(need to go from reads to fragments)
- Number of reads (fragment ends, midpoints)
reported at that position (possibly, taking
strandedness into account)

Variation: **kernel-smoothed read density**. This is closer to overlap approach.

(we will cover the peak calling in next lecture)

Data/script location

Collected all into a single installer, get the repository:

```
git clone git://github.com/ialbert/kbs-install.git
```

Then switch to the kbs directory and run the install-chipseq.sh script.

```
cd kbs-install  
sh install-chipseq.sh
```

Running the install script

Look inside the script to see what it does

```
more install-chipseq.sh
```

Then run it

```
sh install-chipseq.sh
```

Downloads a series of files and programs that will be compiled on your system. It will create the following directory structure:

```
~/work/data  
~/work/src  
~/work/refs
```

Practical exercise

- We have two datasets → the same binding factor was simulated as if it had a short or a long footprint (**long.fq, short.fq**)
- We will visualize and investigate these: detect bound locations, binding size, peaks etc

Script automation

The screenshot shows a software interface with two windows. The left window is a code editor titled "test.sh" with the file path "C:\cygwin\home\ialbert\docs\web\bioinfo-courses\source\597D-2...". It contains the following script:

```
1 # create variables
2 name=WORLD
3
4 echo Hello $name!
5 echo Goodbye $name!
```

The right window is a terminal window titled "~/2011/down/17/misc". It shows the output of running the script:

```
$ sh test.sh
Hello WORLD!
Goodbye WORLD!
$
```

The screenshot shows a software interface with two windows. The left window is a code editor titled "test.sh" with the file path "C:\cygwin\home\ialbert\docs\web\bioinfo-courses\source\597D-2...". It contains the following script:

```
1 # first argument to script
2 name=$1
3
4 echo Hello $name!
5 echo Goodbye $name!
```

The right window is a terminal window titled "~/2011/down/17/misc". It shows the output of running the script with arguments "Jack" and "Jill":

```
$ sh test.sh Jack
Hello Jack!
Goodbye Jack!
$
$ sh test.sh Jill
Hello Jill!
Goodbye Jill!
$
```



Start Page install-bwa.sh install-chipseq.sh generate-chipseq.sh

```
1 # this ensures that the script
2 # exists immediately on error or undefined variables
3 set -ue
4
5 # a variable that stores the reference
6 REF=~/work.refs/yeast.fasta
7
8 # this is the name of the program
9 BWA=~/work/src/bwa-0.5.9/bwa
10
11 # this is the first argument to the program
12 INPUT=$1
13
14 # the alignment file
15 ALN=$INPUT.aln
16
17 # the resulting sam file
18 SAM=$INPUT.sam
19
20 # indexing only needs to happen once
21 #$BWA index $REF
```



Start Page install-bwa.sh install-chipseq.sh generate-chipseq.sh *

```
15 # the alignment file
16 ALN=$INPUT.aln
17
18 # the resulting sam file
19 SAM=$INPUT.sam
20
21 # the bam format
22 BAM=$INPUT.bam
23
24 # indexing only needs to happen once
25 #$BWA index $REF
26
27 # this performs the alignment
28 $BWA aln $REF $INPUT > $ALN
29 $BWA samse $REF $ALN $INPUT > $SAM
30
31 # now transform into sorted BAM format
32 $SAMTOOLS view -Sb $SAM > tempfile
33 $SAMTOOLS sort tempfile $INPUT
34 $SAMTOOLS index $BAM
35
```

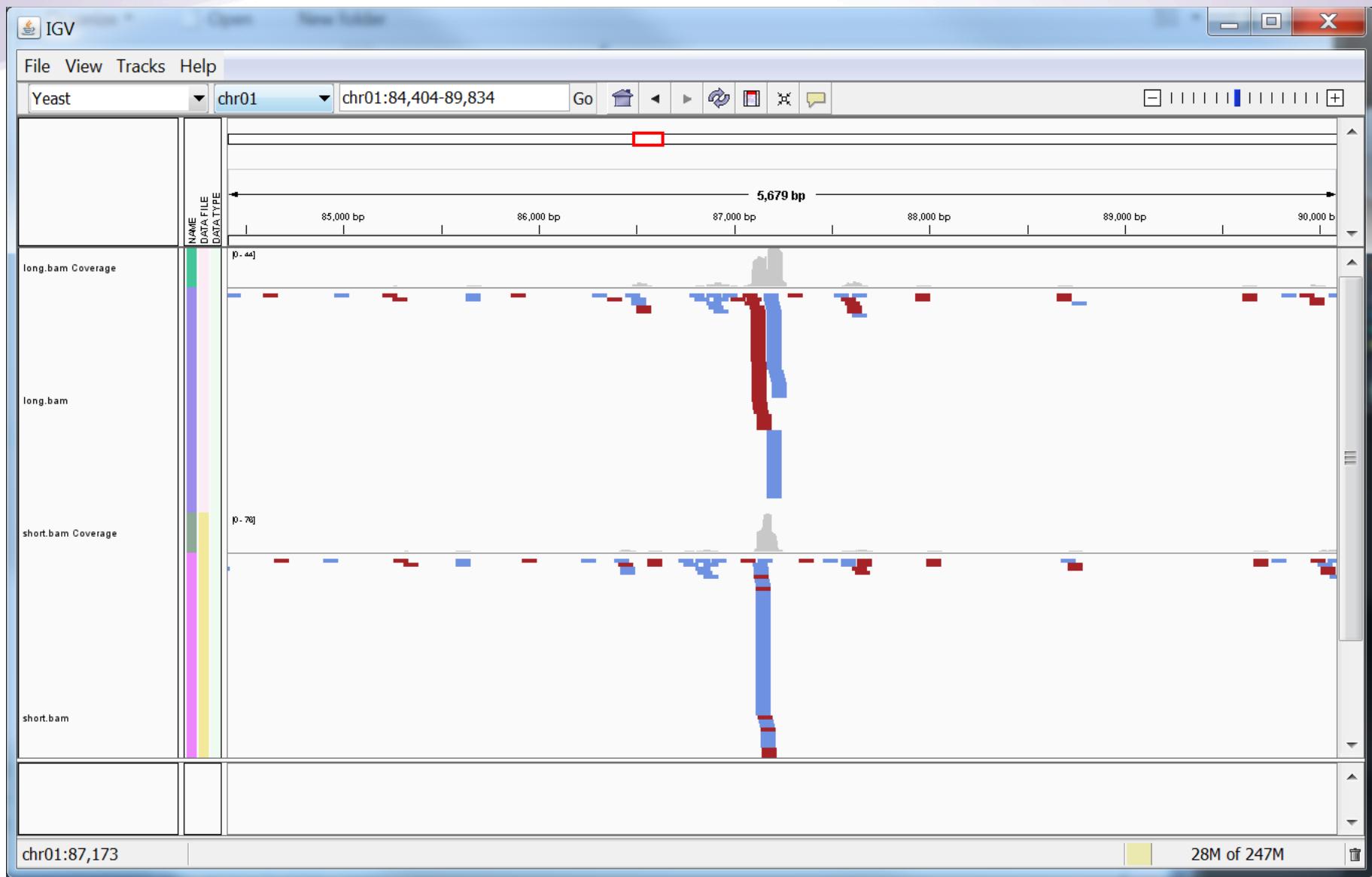
Run the analysis for both files

```
sh ~/kbs-install/generate-chipseq.sh data/short.fq
```

```
sh ~/kbs-install/generate-chipseq.sh data/long.fq
```

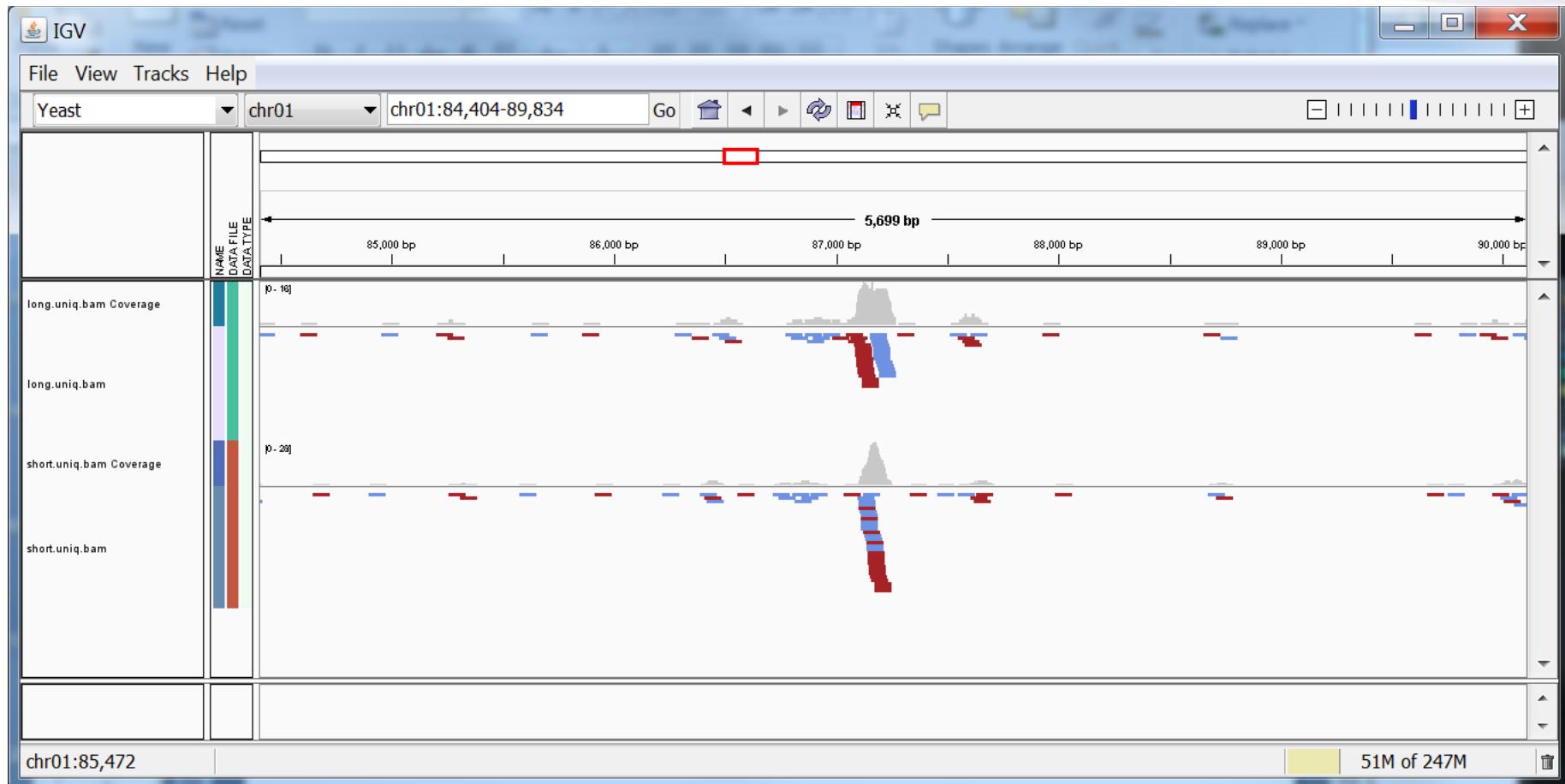
You have to pass the right path to the script and the right path of the file

Visualize in IGV



After de-duplication

(add the necessary steps to your pipeline and rerun)

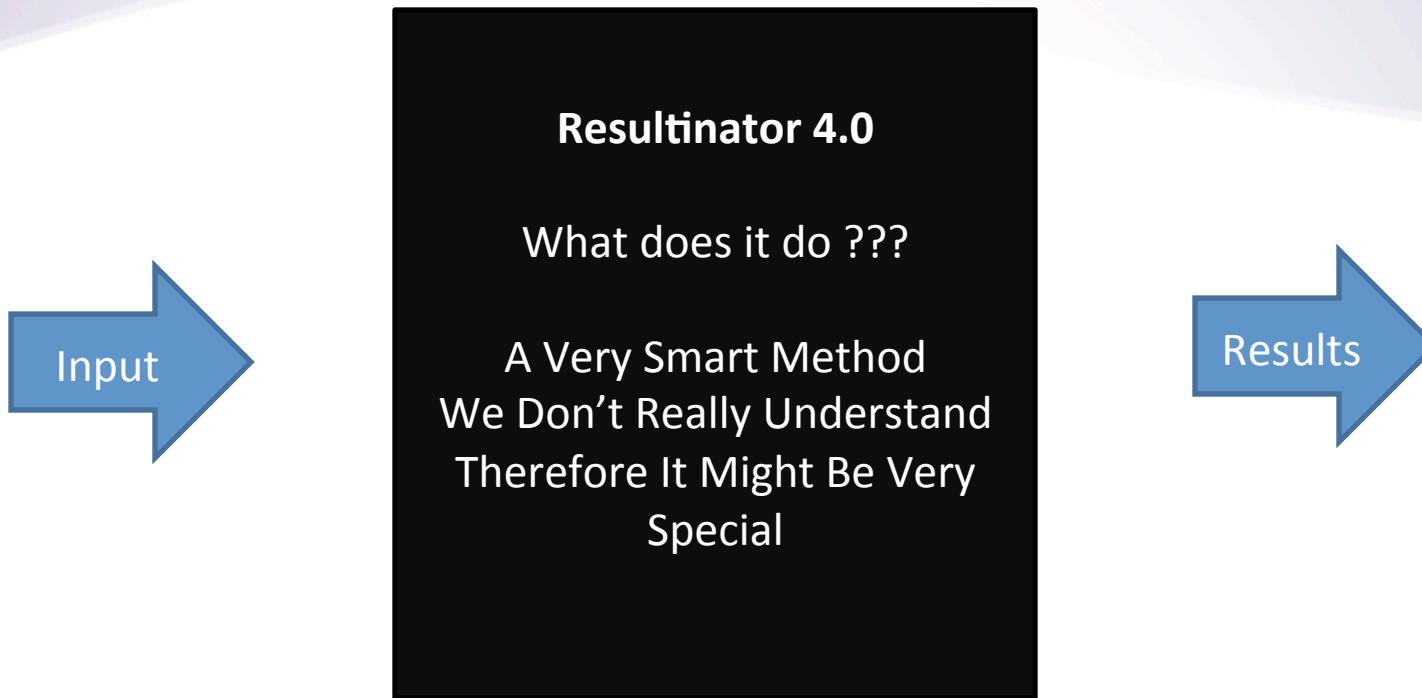


Back to Chip-Seq

Identify bound locations then

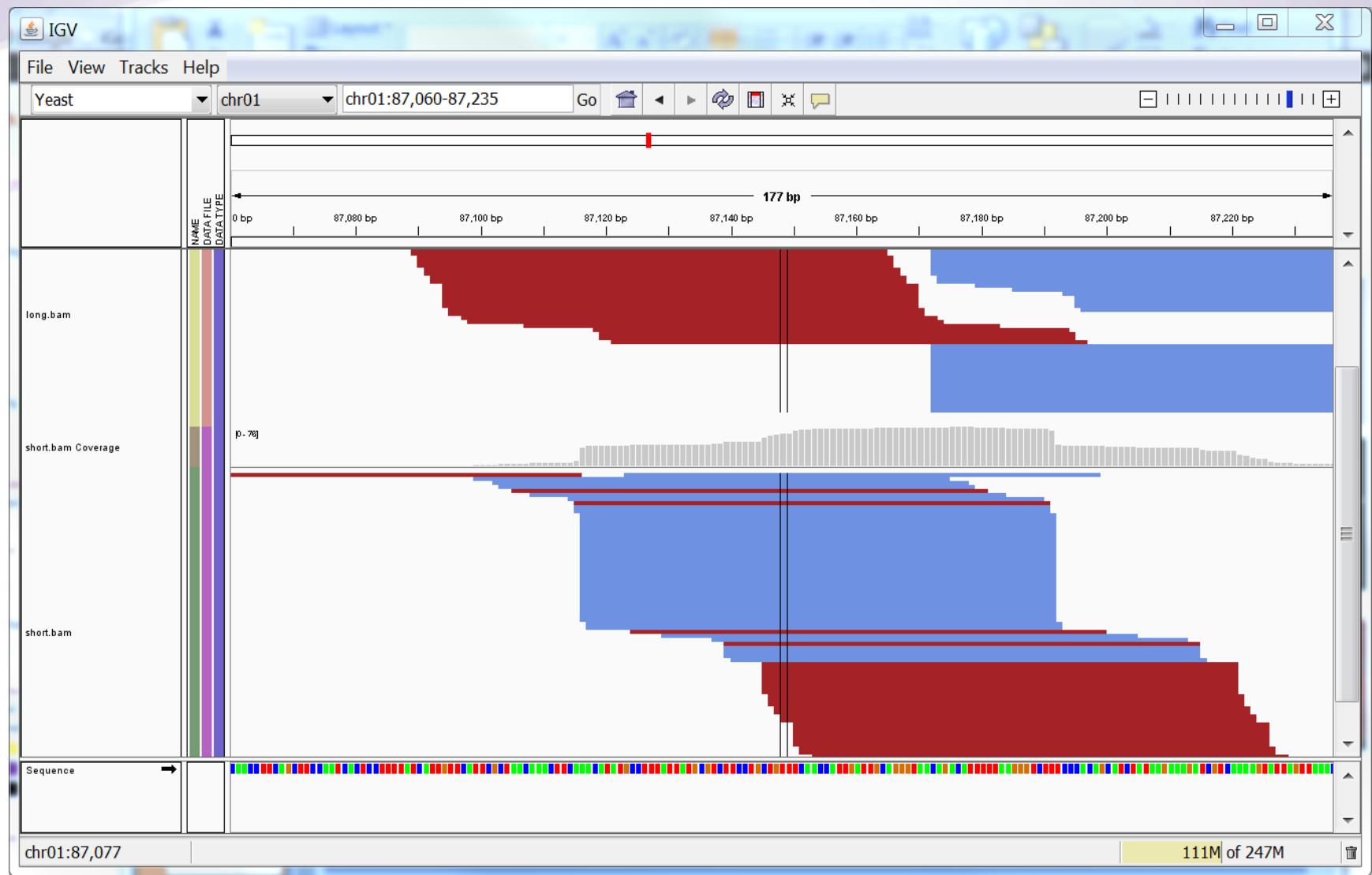
1. tabulate the observations → protein X binds at locations Y
2. compare between samples to detect different level of occupancy
3. extract sequences for binding locations and find motifs
4. study common attributes of genomic features nearby bound locations

Beware of Black Boxes



- The **more sophisticated** a method the **less likely** is that we understand its limitations and applicability !
- In general always **do a simple analysis first** – do common sense, straightforward comparisons – you'll be repeatedly surprised just how far you can get that way
- Then move to more complicated approaches

Visualization can be counterintuitive



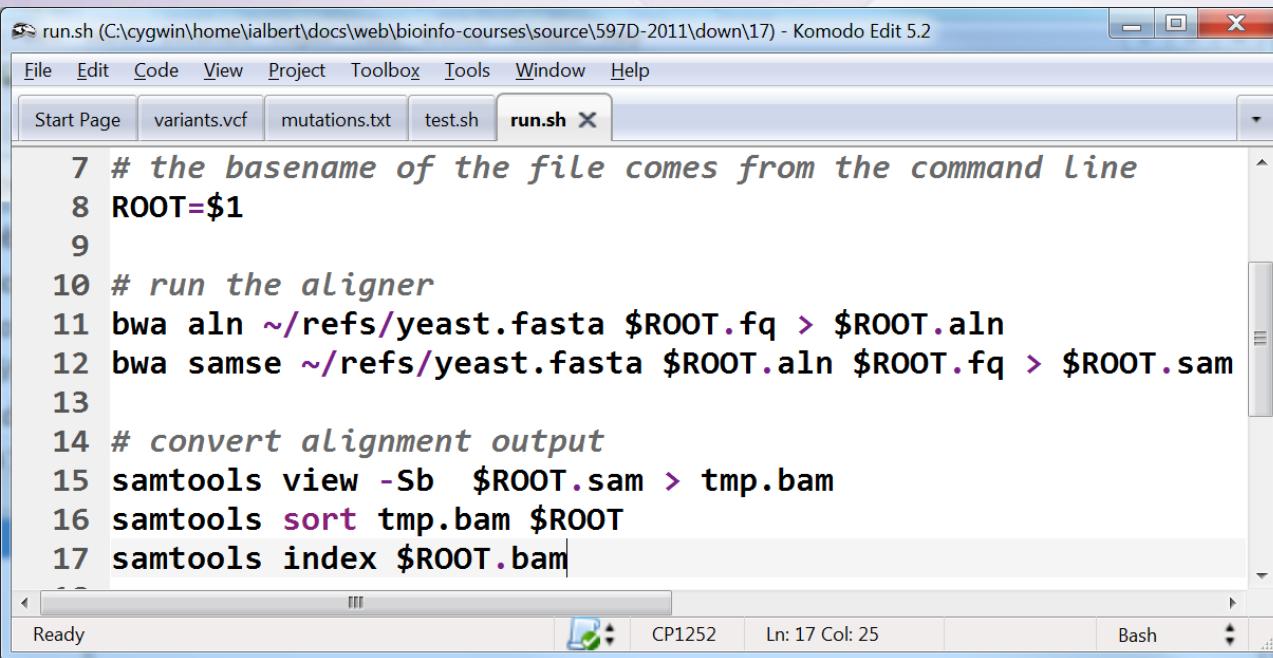
Read extension/trimming

- Extend to full fragment:
 - operate on the actual fragments that were under study
 - we need to know how far to extend
- Trim to 5' end:
 - use the two ends as markers of the binding location
 - twice as much data to keep track of (+/-) borders

In both cases since strands are sequenced independently we can treat the data as technical replicates → allows us to evaluate errors/noise

Get to know your data → compute the coverage

- Convert to wiggle format (WIG) → useful and required for some tools
- Convert to coverage → we can query it with tabix
- We will continue editing our pipeline from lecture 17

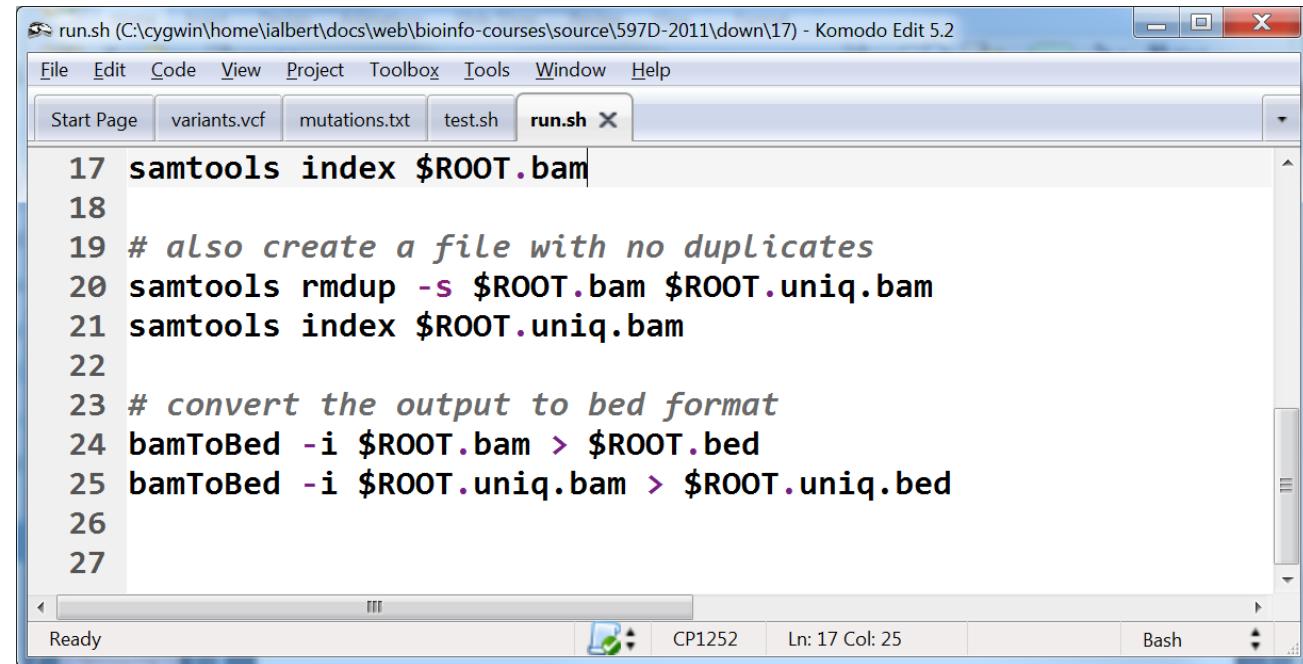


```
7 # the basename of the file comes from the command line
8 ROOT=$1
9
10 # run the aligner
11 bwa aln ~/refs/yeast.fasta $ROOT.fq > $ROOT.aln
12 bwa samse ~/refs/yeast.fasta $ROOT.aln $ROOT.fq > $ROOT.sam
13
14 # convert alignment output
15 samtools view -Sb $ROOT.sam > tmp.bam
16 samtools sort tmp.bam $ROOT
17 samtools index $ROOT.bam
```

All steps in one file
allows us to repeat
the analysis

Don't copy it!

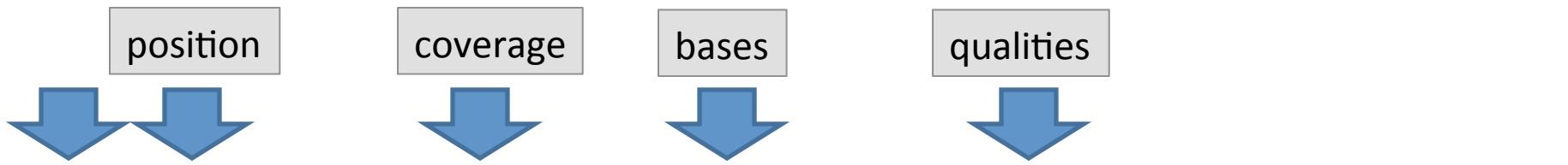
Build it one step at
a time! Then comment
out the previous step



```
17 samtools index $ROOT.bam
18
19 # also create a file with no duplicates
20 samtools rmdup -s $ROOT.bam $ROOT.uniq.bam
21 samtools index $ROOT.uniq.bam
22
23 # convert the output to bed format
24 bamToBed -i $ROOT.bam > $ROOT.bed
25 bamToBed -i $ROOT.uniq.bam > $ROOT.uniq.bed
26
27
```

Samtools mpileup output

samtools mpileup long.bam



position	coverage	bases	qualities
chr01 56	N	9	AAaAAAaAAA
chr01 57	N	9	CCcCCCCCCC
chr01 58	N	9	AAaAAAaAAA
chr01 59	N	9	CCcCCCCCCC
chr01 60	N	9	AAaAAAaAAA
chr01 61	N	10	CCcCCCCCCC^FC
chr01 62	N	10	AAaAAAaAAAA
chr01 63	N	10	TTtTTtTTTT
chr01 64	N	10	CCcCCCCCCCC
chr01 65	N	10	CCcCCCCCCCC
chr01 66	N	13	TTtTTtTTTT^FT^FT^FT
chr01 67	N	13	AAaAAAaAAAAAAA
chr01 68	N	16	AAaAAAaAAAAAAA^FA^FA^FA
chr01 69	N	16	CCcCCCCCCCCCCCC
chr01 70	N	17	AAaAAAaAAAAAAAAA^FA
chr01 71	N	17	CCcCCCCCCCCCCCC
chr01 72	N	19	TTtTTtTTTTTTTTT^FT^FT
chr01 73	N	20	AAaAAAaAAAAAAAAA^Fa

--More--(byte 1445)

Lets transform the BAM file to wiggle

The image shows a Windows desktop environment. At the top, there is a taskbar with several icons. Below the taskbar, there are two windows: a Komodo Edit 5.2 code editor and a terminal window.

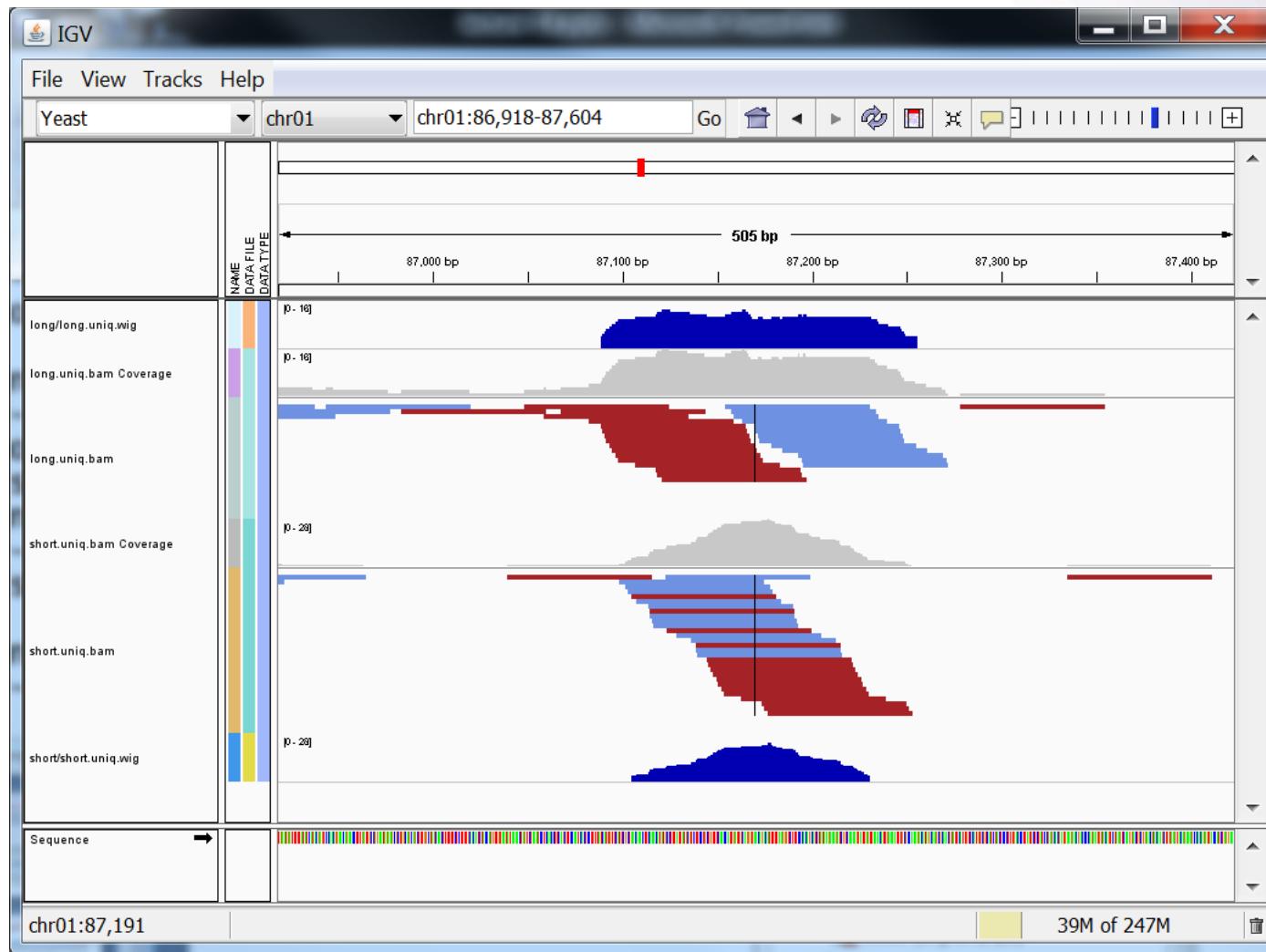
Komodo Edit 5.2 Window:

- Title Bar:** run.sh (C:\cygwin\home\ialbert\docs\web\bioinfo-courses\source\597D-2011\down\17) - Komodo Edit 5.2
- Menu Bar:** File, Edit, Code, View, Project, Toolbox, Tools, Window, Help
- Tab Bar:** Start Page, variants.vcf, mutations.txt, test.sh, run.sh (active tab)
- Code Area:** The code is a shell script named run.sh. It includes lines 5 through 13, which involve setting variables and running scripts to convert BAM files to wiggle format.
- Status Bar:** Ready, CP1252, Ln: 15 Col: 1, Bash

Terminal Window:

- Title Bar:** ~/2011/down/17
- Content:** The terminal shows the execution of the run.sh script. It outputs the command run.sh long/long, followed by the conversion process for long/long.bam and long/long.uniq.bam to their respective wiggle files. It also shows the conversion of long/long.uniq.bam to long/long.uniq.wig. The terminal ends with a \$ prompt.

Visualizing the filtered wiggle file



Peak calling tools

- Very large selection of tools and techniques:
ERANGE, FindPeaks, MACS, QuEST, CisGenome, SISSRS, USeq, PeakSeq, SPP, ChIPSeqR, GLITR, ChIPDiff, T-PIC, BayesPeak, MOSAiCS, CCAT, CSAR,

Featured today (historically precedes many of the tools above)

GeneTrack—a genomic data processing
and visualization framework

*Istvan Albert, Shinichiro Wachi,
Cizhong Jiang and B. Franklin Pugh*

Bioinformatics, 2008

The nomenclature is a bit hazy

- The term “peaks”, “signal”, and “enriched regions” are used interchangeably → leads to a lot of confusion

Proper definition would be

- Peaks → regions with a well defined shape, a midpoint and a spread that describe one or more measurements
- Enrichment → selecting those peaks that show statistically significant properties when compared to background/control

GeneTrack

- It performs error corrections and peak predictions
- Visualizes data as tracks in a ‘browser’ interface.
- Also integrated into our own LIMS: Laboratory Data Management System LionDB3: with projects, users, data sharing etc.

<http://genetrack.googlecode.com> the LIMS is at <http://liondb3.atlas.bx.psu.edu/>

GeneTrack Features

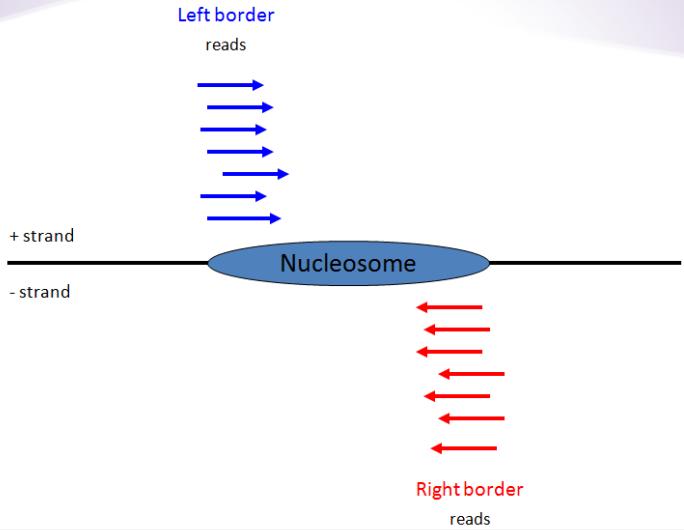
Unique features:

- Peak detection and visualization at the same time.
- User sees the effect of the peak prediction parameters right away

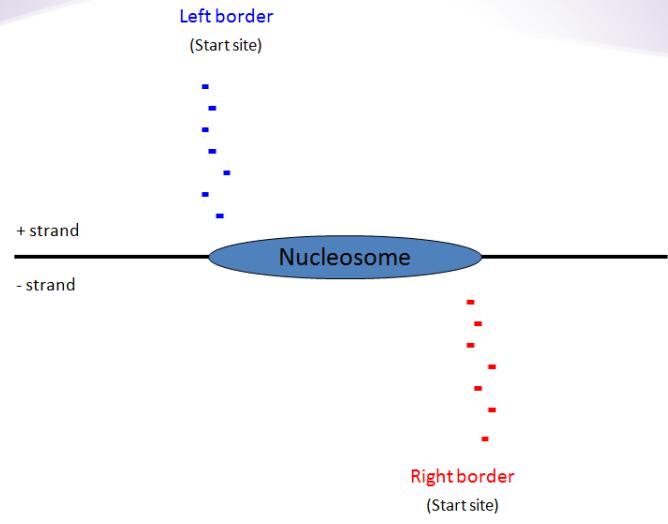
Much criticized missing component

- Lacks the statistical modeling of the background (expected frequencies)

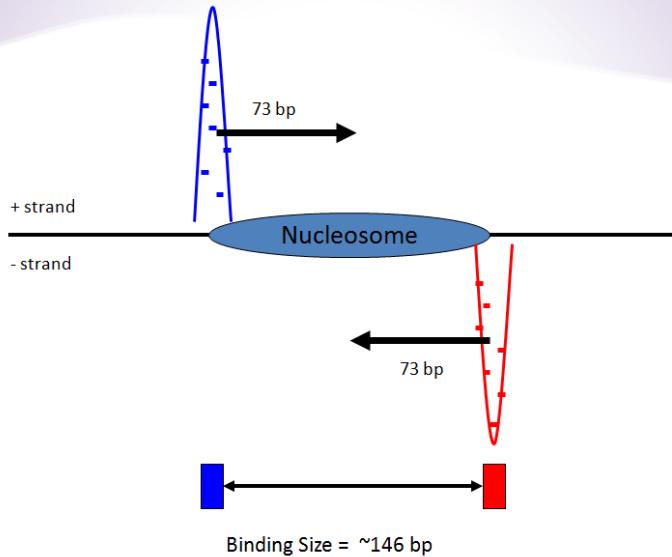
Aligning to reference genome



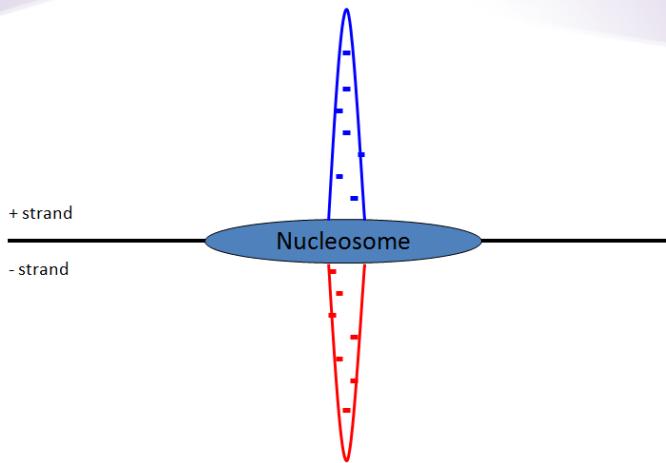
Find the borders



Finding the binding sizes



Shift to the center

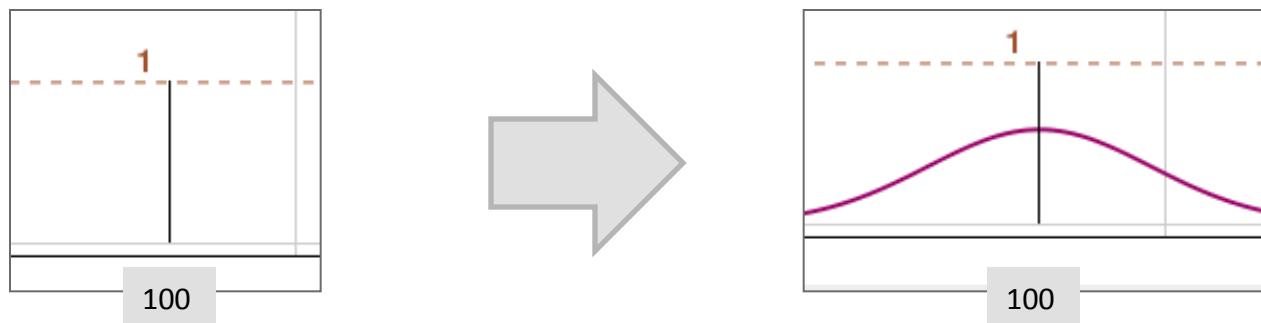


GeneTrack: reads on both strands



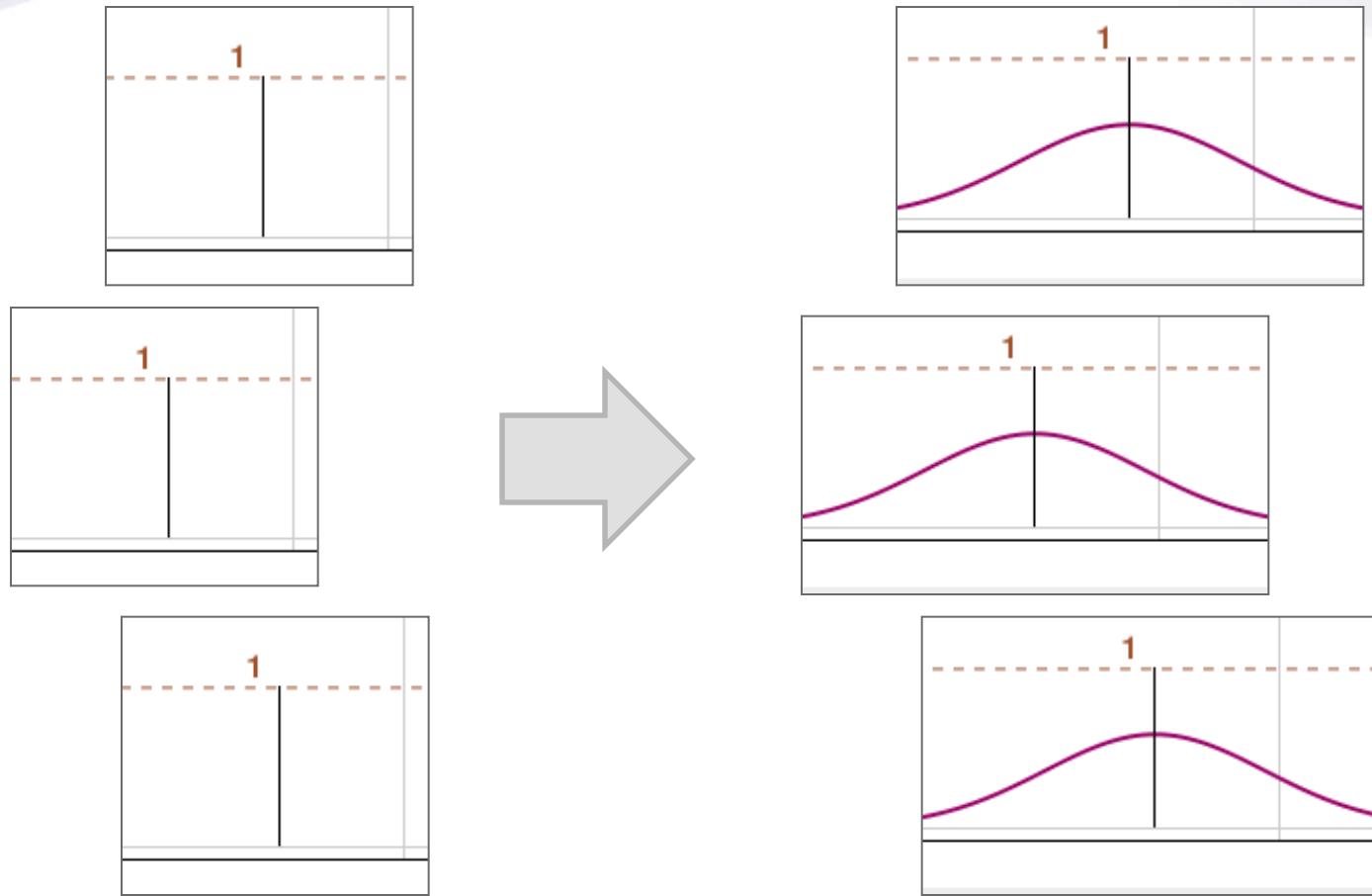
Error Model

- Replace every read with the probability of the read falling onto the observed location.



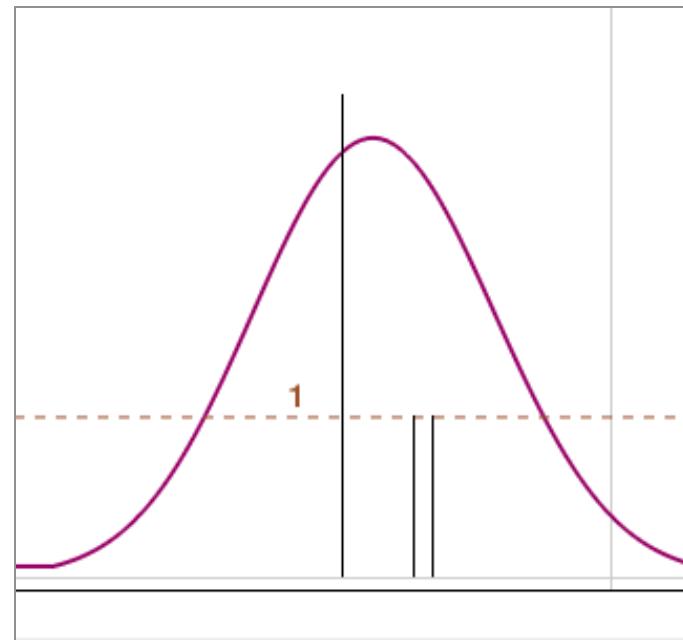
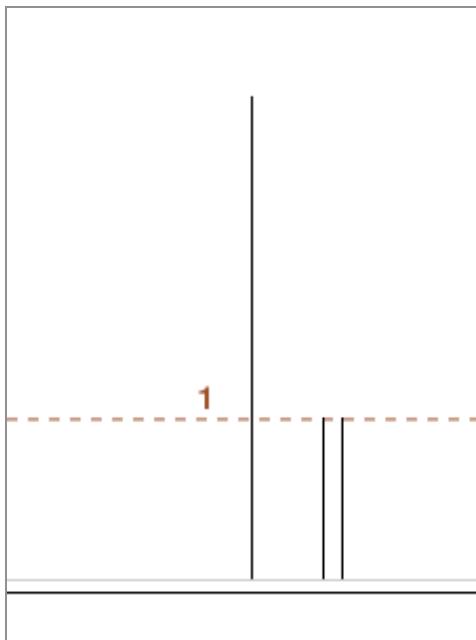
Each single coordinate will be replaced with a “smudge” of values (hundreds of them).

Sum up your “smudges”



Central Limit Theorem guarantees another normal function

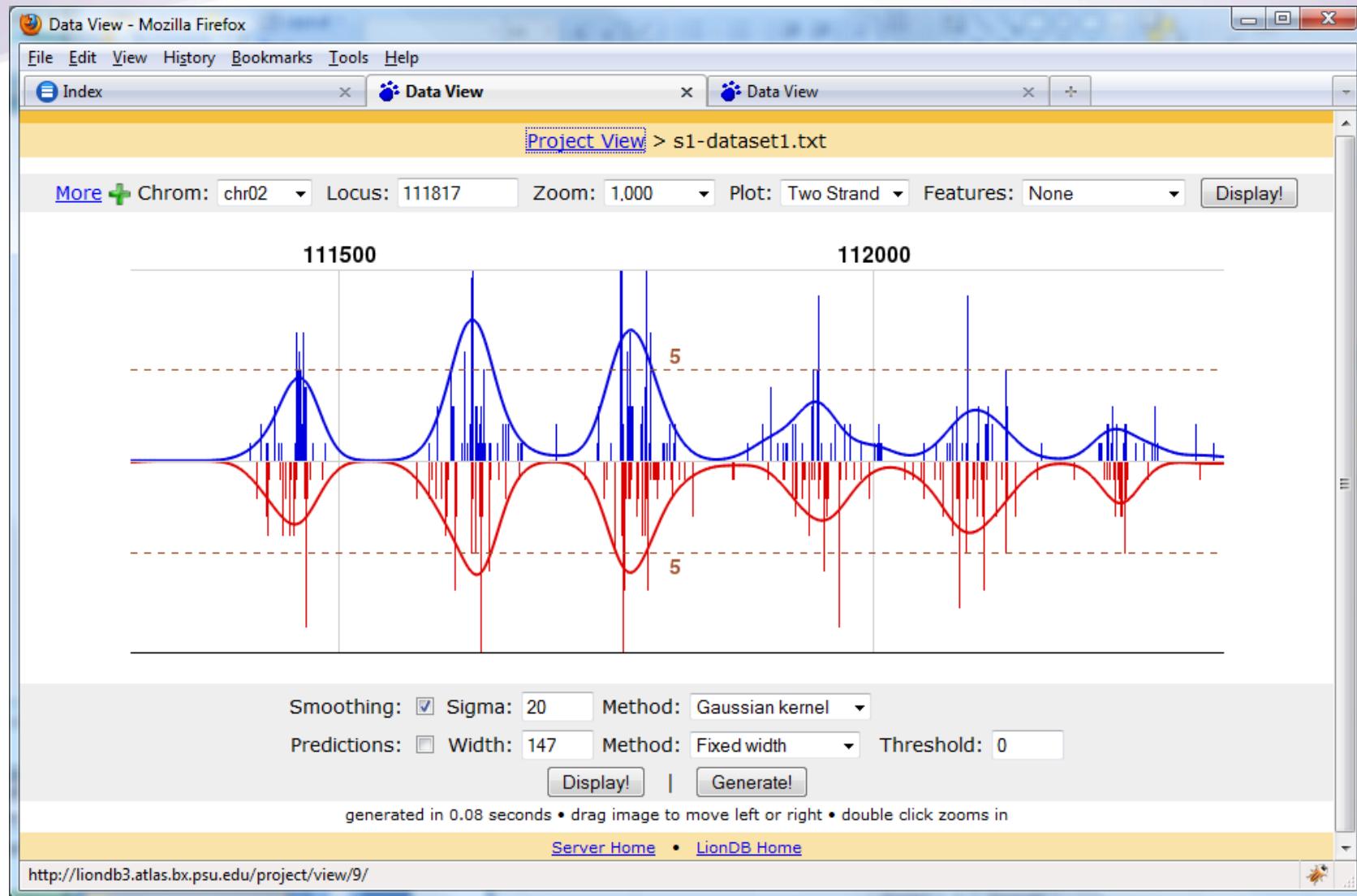
Summing up multiple errors



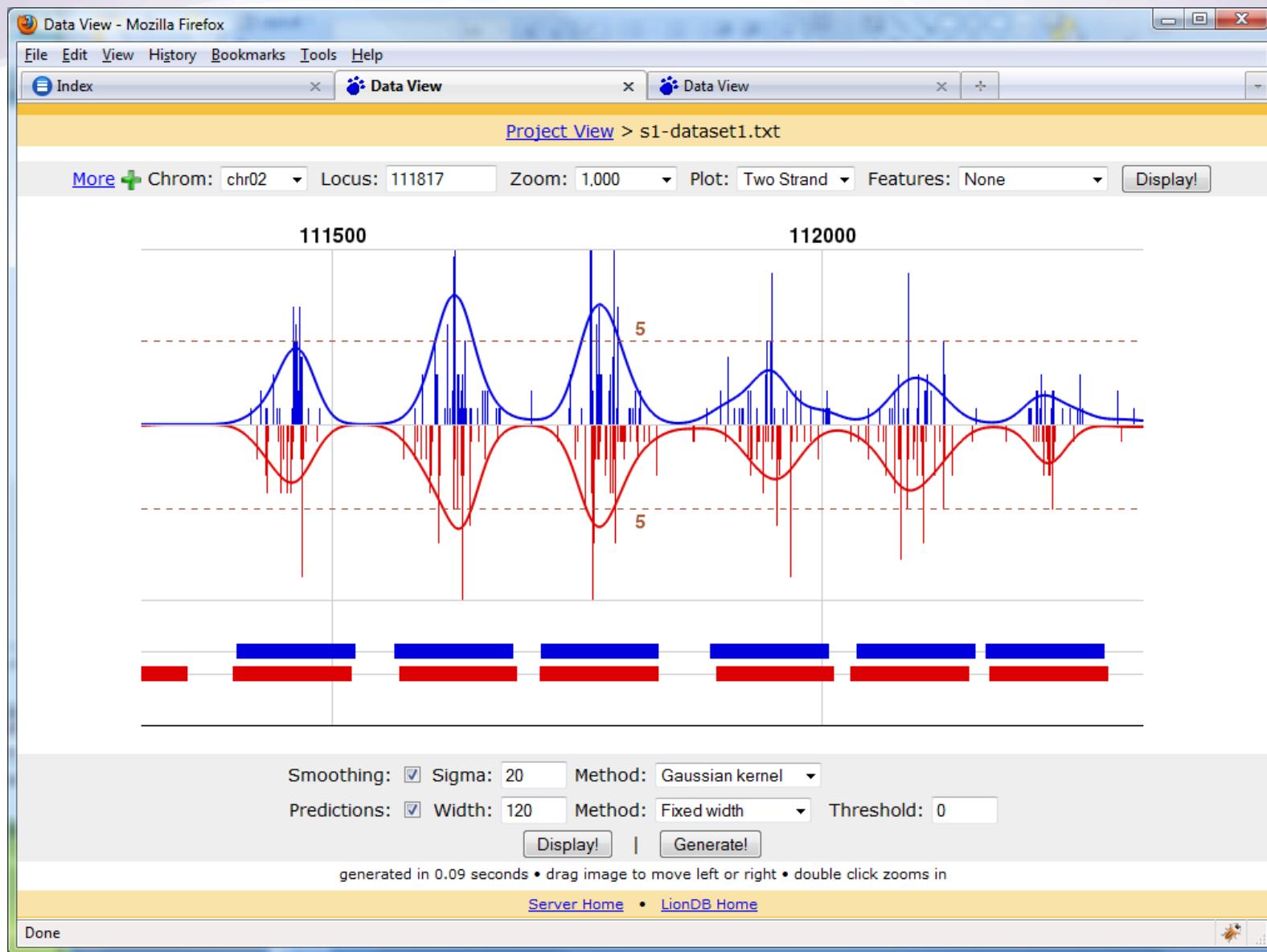
GeneTrack: revisit the reads



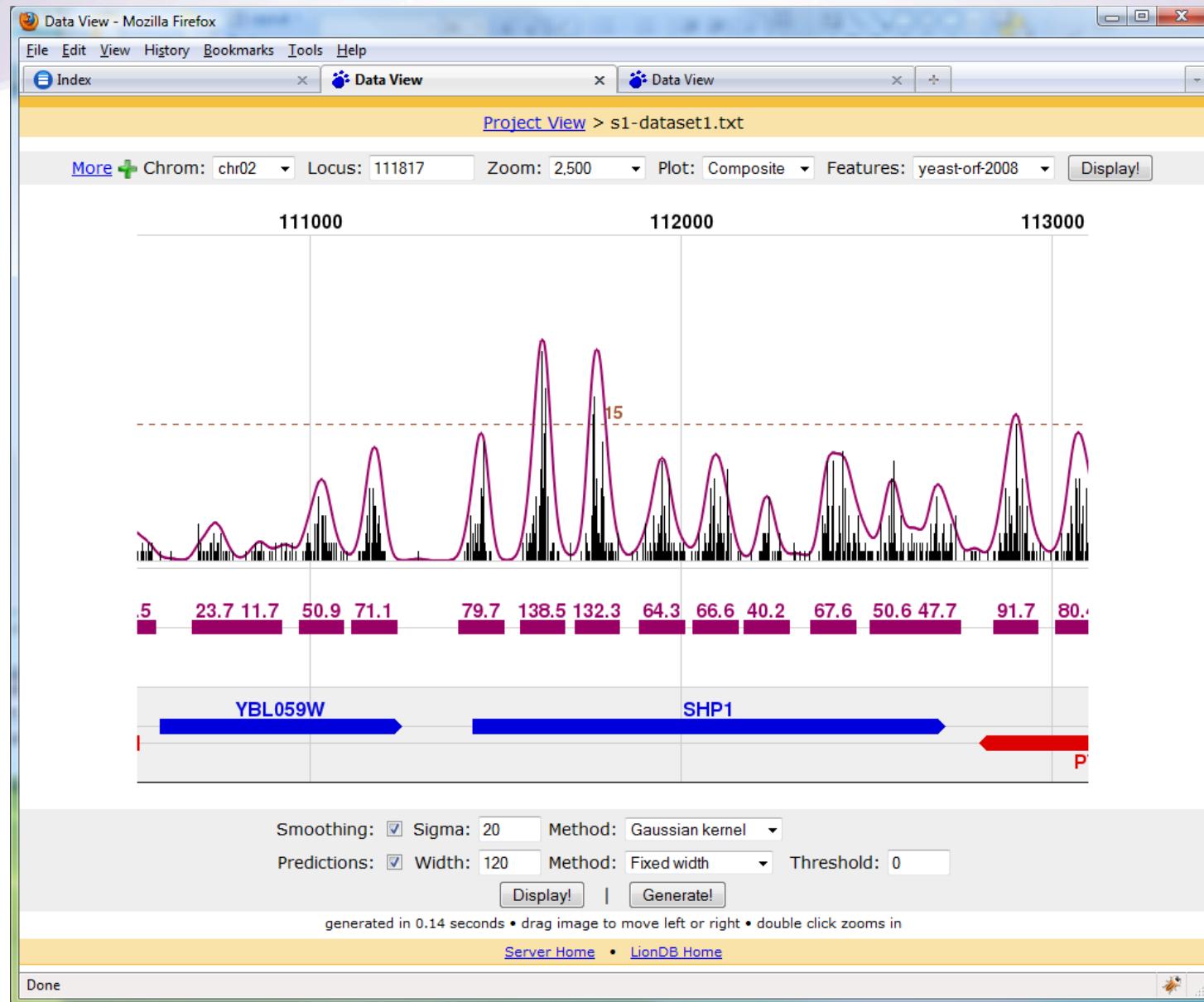
GeneTrack: with smoothing



GeneTrack also does peak prediction

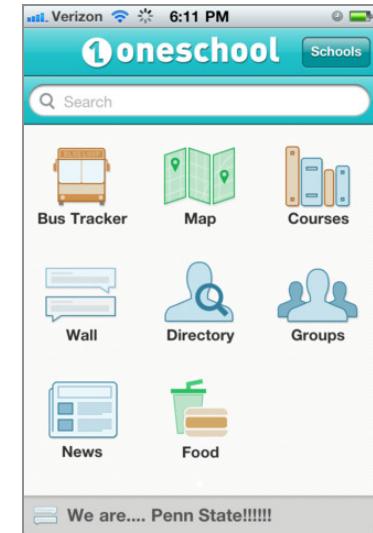


Displayed relative to genomic features



Command line GeneTrack

- Author Pindi Albert (freshman at PSU)
- The functionality of GeneTrack website ported to a command line tool
- Also the author of many cool stuff (not science related alas):



Clone the chipexo and bioawk

- Clone my “**fork**” of it rather than the original since I made some changes for this class

```
$ ~/2011
$ python ~/src/chipexo/genetrack/genetrack.py
Usage: genetrack.py [options] input_paths

input_paths may be:
- a file or list of files to run on
- a directory or list of directories to run on all files in the
- "." to run in the current directory

example usages:
python genetrack.py -s 10 /path/to/a/file.txt path/to/another/f
python genetrack.py -s 5 -e 50 /path/to/a/data/directory/
python genetrack.py .

Options:
-h, --help      show this help message and exit
-s SIGMA       Sigma to use when smoothing reads to call peaks
```

Running the GeneTrack peak caller

```
$ ~/2011
$ bamToBed.exe -i long.uniq.bam > long.uniq.bed
$ awk -f ~/src/bioawk/bed2gff.awk < long.uniq.bed > long.uniq.gff
$
```

```
$ ~/2011
$ # make a shortcut to genetrack
$ # so that the commands are shorter
$ ln -s ~/src/chipexo/genetrack/genetrack.py
$ python genetrack.py -F 5 long.uniq.gff > long.peaks.gff
```