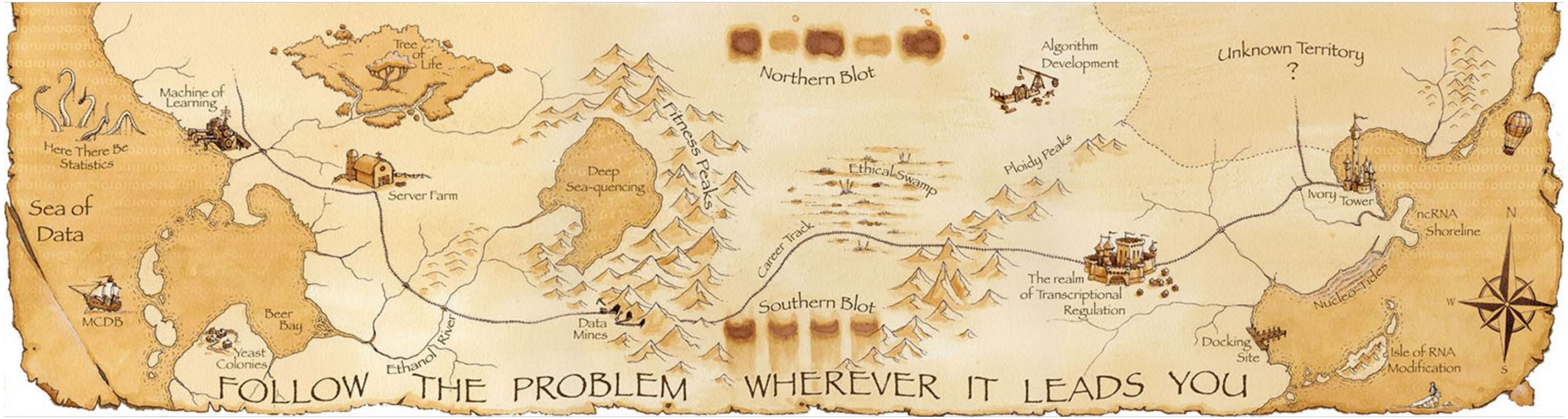
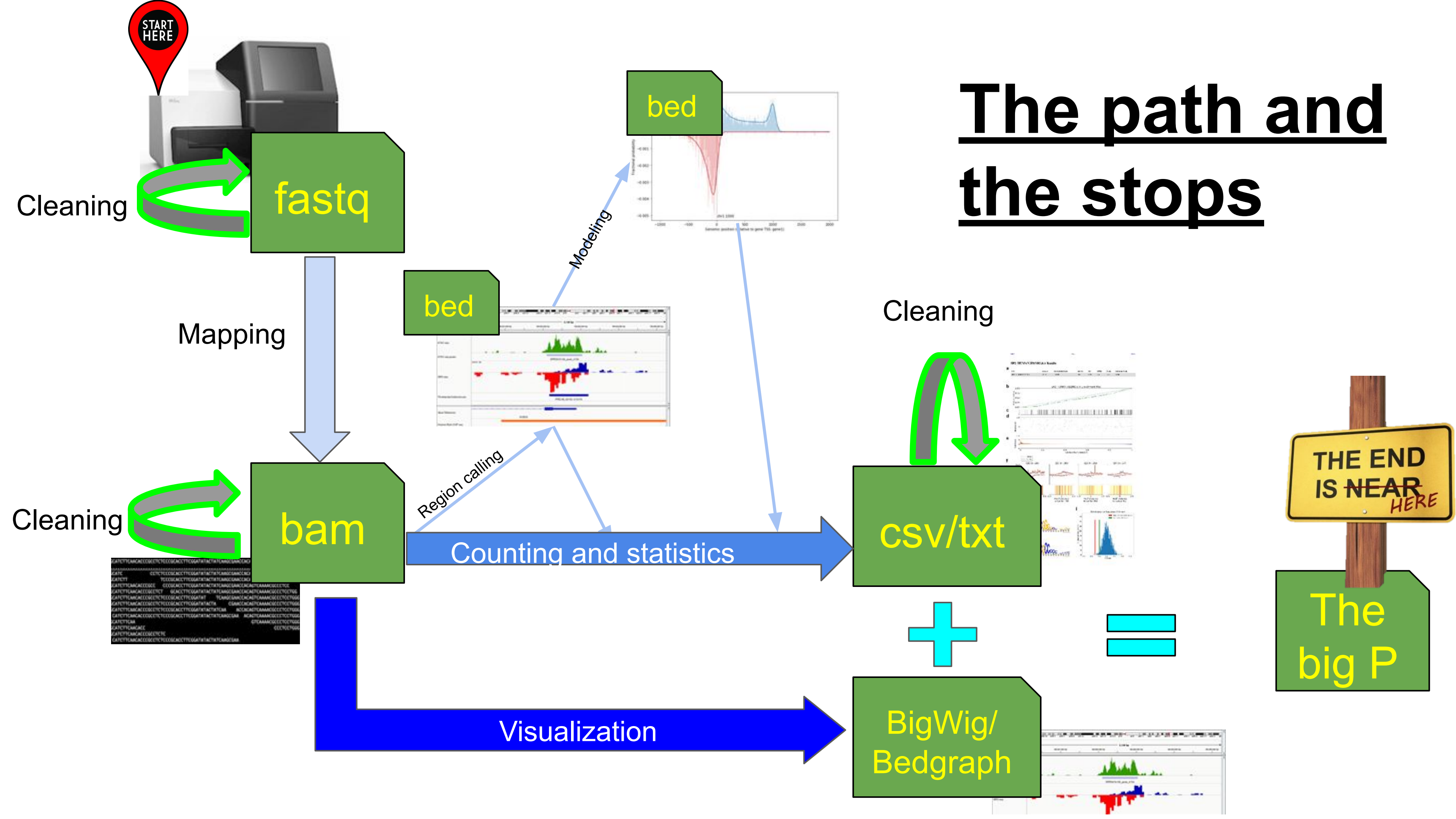


DnA Lab
Short read sequencing Workshop

The path and the stops



START HERE

Cleaning

fastq

Mapping

Cleaning

bam

bed

Region calling

Modeling

bed

Counting and statistics

Cleaning

csv/txt

Visualization

BigWig/ Bedgraph

The big P

Check qc for your fastq

What should you do

- How many reads do you have
- GC content
- Quality scores
- Adapter content
- Percent duplicates

Tools

- wc -l <filename> and divide by 4
- Head <filename>
- Fastqc
- Picardtools
- Rseqc

Trim reads in a fastq

What should you do

- Cut off adapters
- Drop low quality sequences

Tools

- Trimmomatic
- bbduk

Map reads

What should you do

- Map to the right genome

Tools

- Hisat2 (this is what DnA lab uses)
- BWA
- Bowtie2
- STAR (this is what Rinn lab uses)

- Salmon (quasi-mapper?)

Sam to bam

What should you do

- You will need a sorted bam/cram file
- You will need a bam.bai (index file) for many things

Tools

- Samtools
- Picard

bam to bed

What should you do

- You will need a sorted bam/cram file
- You will need a bam.bai (index file)

Tools

- Bedtools

bam quality control

What should you do

- How many reads mapped
- How complex is the data
- How many reads are
  - Genic
  - Intergenic
  - In introns

Tools

- Preseq
- Pileup
- Python Rseqc
  - module load python/3.6.3/rseqc
- Mapping programs
  - On mapping slide
- Counting programs
  - On counting slide

Making bams more manageable

What should you do

- Make bigWig
- Make bedGraph
  - Make tdf
- Make a cram

Tools

- Bedtools
- Deeptools
- Igvtools toTDF

Which gene list (gtf/gff/bed) do I use?

What should you do

- Chose a genome
  - NCBI/Ensembl (high false positives)
  - UCSC (high false negative)

Tools

- UCSC genome browser
- Igenomes
- Ensembl and biomart

Counting reads over genes or regions

What should you do

Think about:

Do you want to count reads that map in more than one place?

If you have paired end should it count if only one end maps?

Should spliced reads count?

Tools

- Htseq
- Subread (FeatureCounts)
- Bedtools coveragebed (or multicov)
- sailfish

Differential expression

What should you do/think about

- Do I care about isoforms?
- Did I account for batch effects?

Tools

- DEseq2
- edgeR
- Bayseq
- Ballgown
  - Stringtie
- Limma Voom
- Cuffdiff

What I do with my gene list after that

What should you do

Tools

- Enrichr (first pass)
- GO terms
  - Panther
  - Davidtools
  - topGO (in R)
- GSEA

Find regions to count over

What should you do

- If you are not sequencing genes you have to know where to count
- Do you want to just find read piles or fit a model to the data?
  - What model?

Tools

- Macs2 or Macs3
- Homer
- Fstitch
- Tfit
- Cufflinks (fuzzy if this goes here)

What do I do with my ChiP-seq peaks

What should you do

- Which motifs are in the peaks?
- What genes are my peaks in?

Tools

- Bedtools
  - Interset (with genes or other ChiP-seq peal
  - Jaccard
- Motif finding
  - Meme
  - Dreme
  - TomTom

Plotting your data

What should you do

- Who knows what graph you need

Tools

- R ggplot2
- Python
  - Pandas
  - Matplotlib
  - ploty

Visualize your data

What should you do

- Look at your peaks
- Look at your differential expressed genes
- If you have lots of reads in intergenic regions, look at them!!!!!

Tools

- Igv
- UCSC genome browser
- WashU genome browser