The wet part of RNAseq

MBL

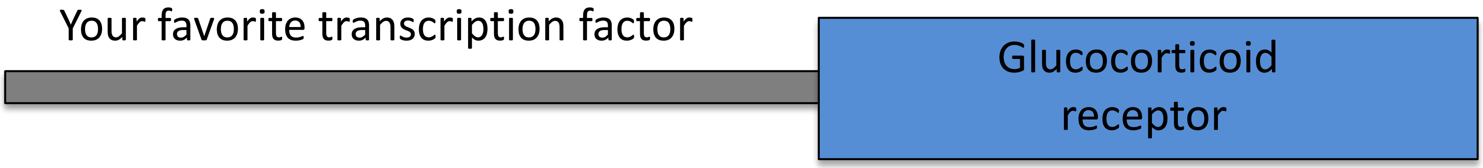
Ian Quigley

iquigley@gmail.com/iquigley@salk.edu

Experimental design

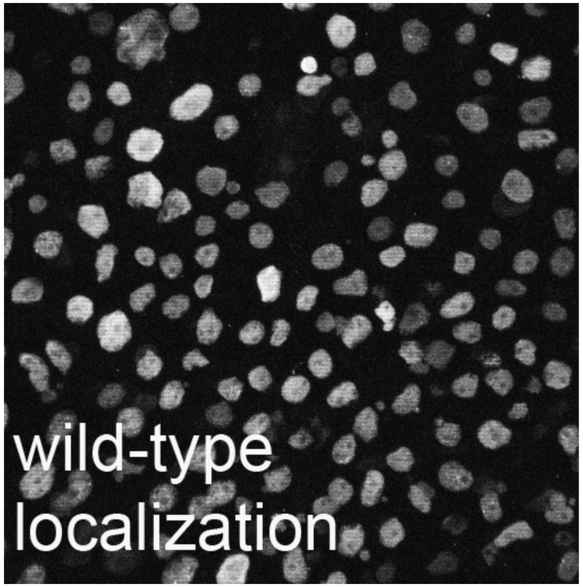
- You generally want to compare two states
- Overexpression, knockdown
  - Also can use inducible constructs! (Next slide)
- Whole embryo can be messy, sometimes dissection better
- You don’t need much! One cap will do. I like to do 20, so if I screw up a few injections most of them are (hopefully) good

Inducible constructs allow temporal control

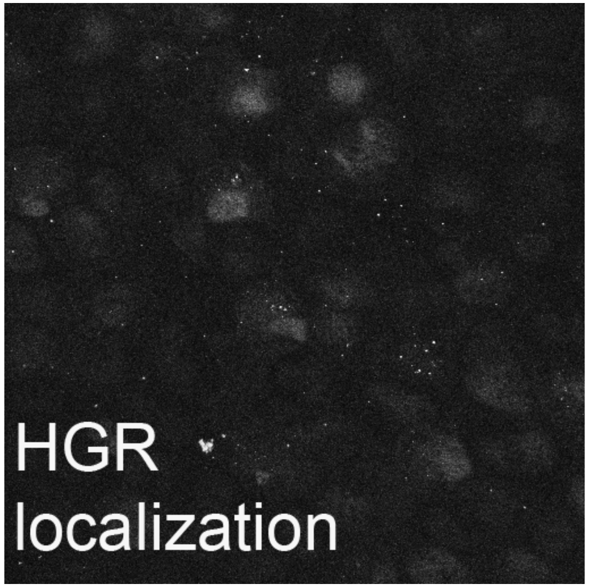


Your favorite transcription factor

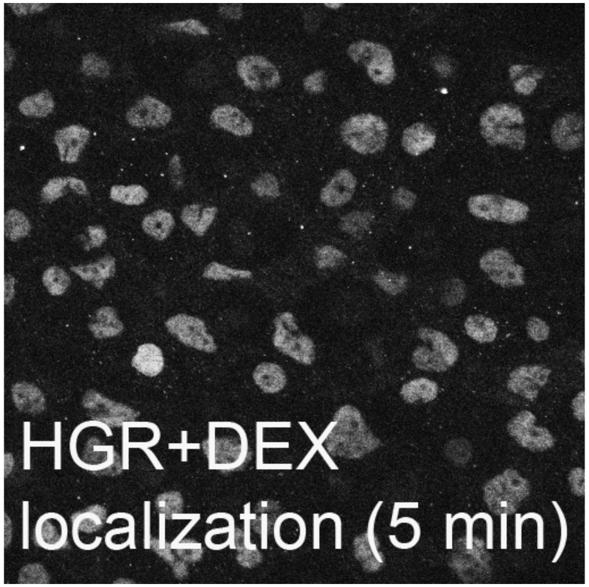
Glucocorticoid receptor



wild-type localization

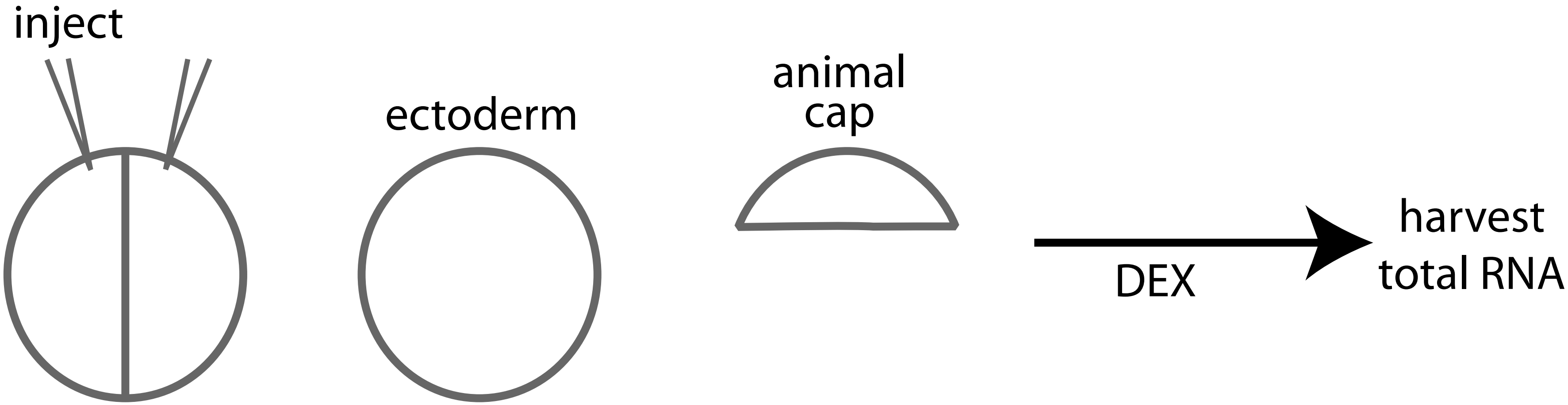


HGR localization



HGR+DEX localization (5 min)

One experimental design



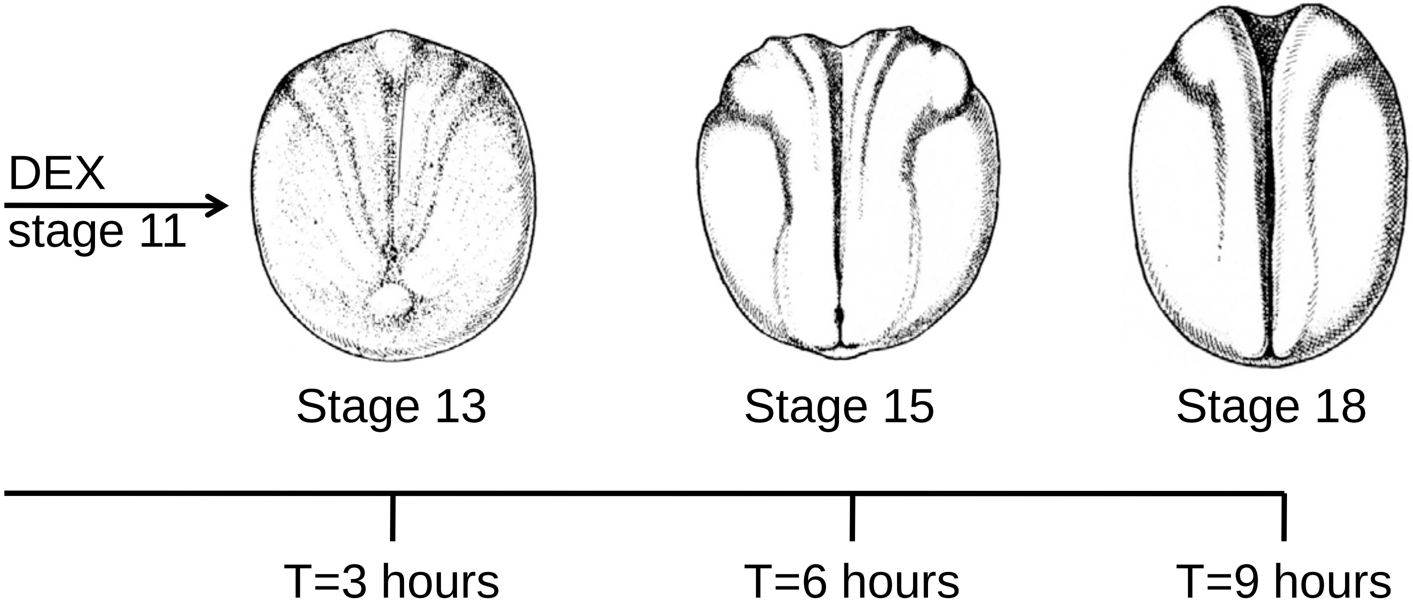
inject

ectoderm

animal cap

DEX

harvest total RNA



DEX

stage 11

Stage 13

Stage 15

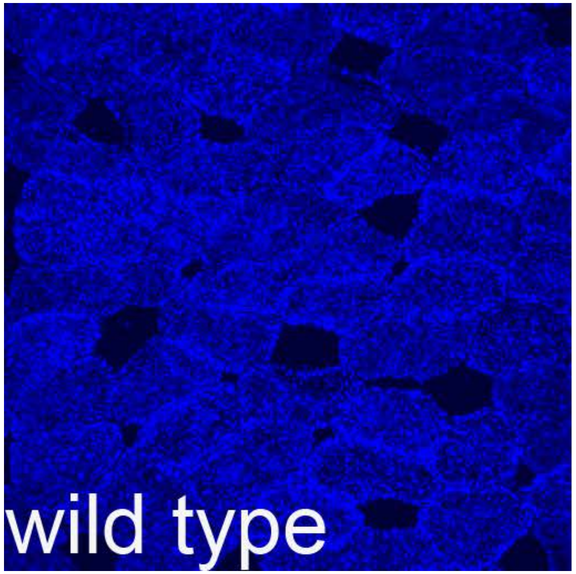
Stage 18

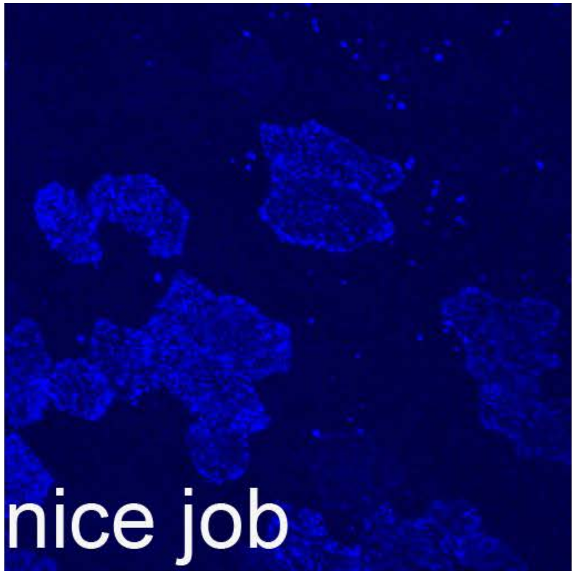
T=3 hours

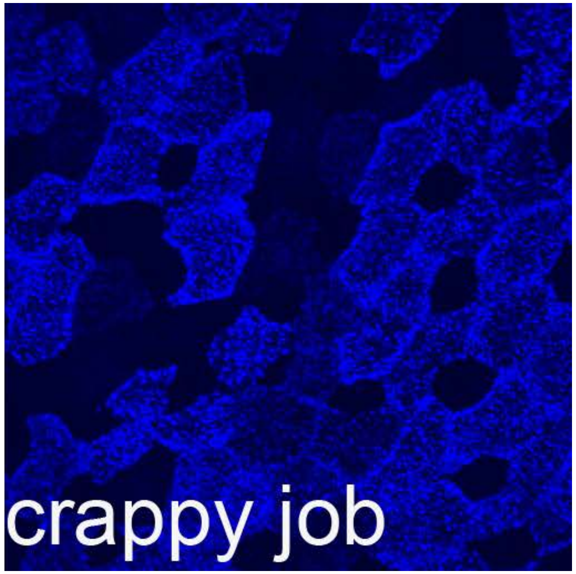
T=6 hours

T=9 hours

Always always always check phenotype!



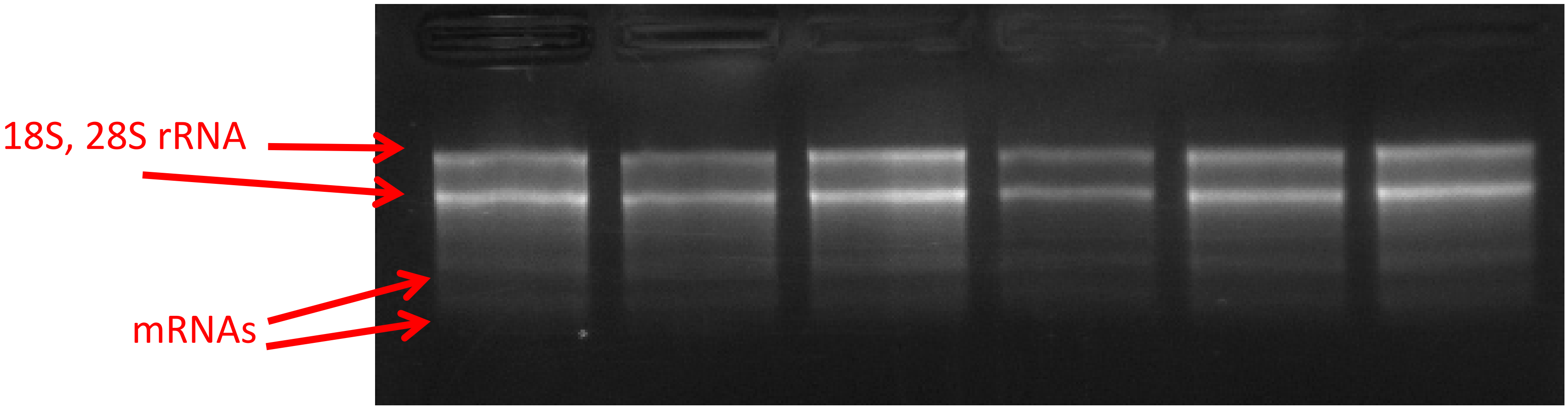




Inject a few extra embryos, grow them up and look at them

I have regretted not doing this.

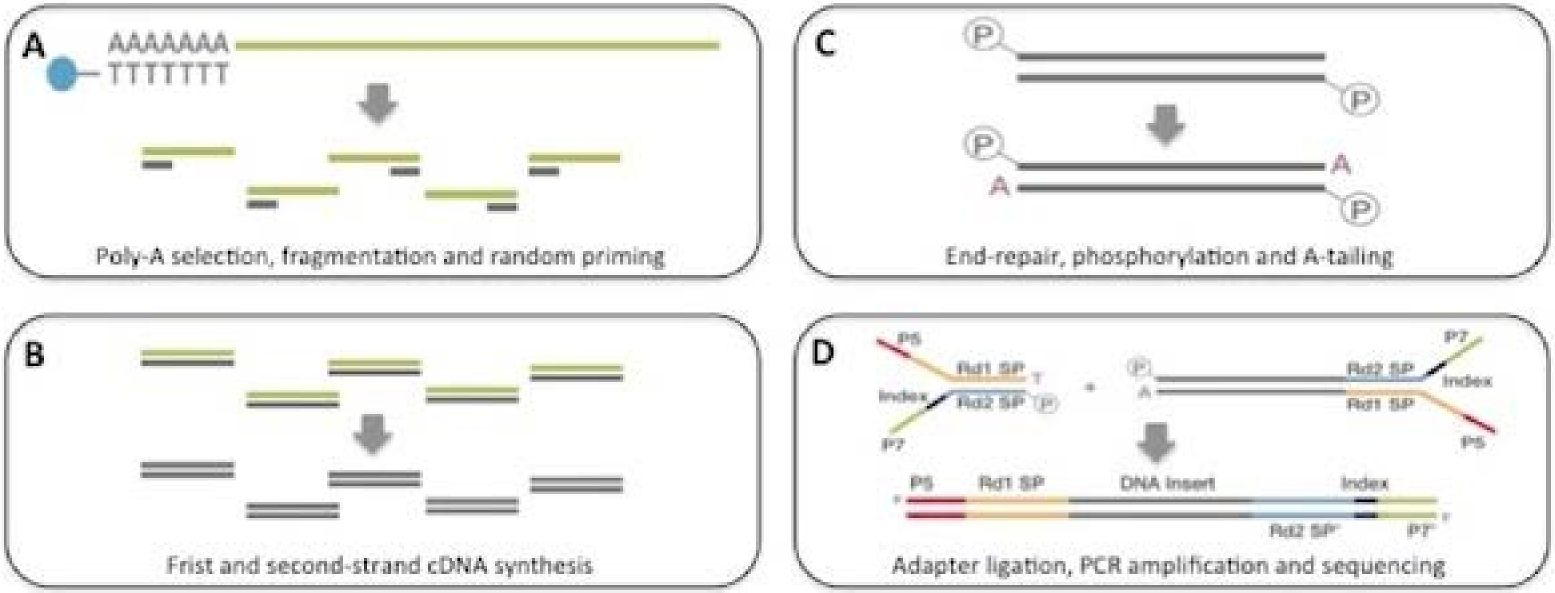
How good is your RNA?



- Formaldehyde in sample (only the part you load!)
- Look for separate bands below rRNAs
- Bad RNAs will look more smeary
- Also check 260/280, 260/230 ratios
- Some people use Bioanalyzer. I never have.

Here’s how you make the library

Illumina Tru-Seq RNA-seq protocol



Poly-A selection, fragmentation and random priming

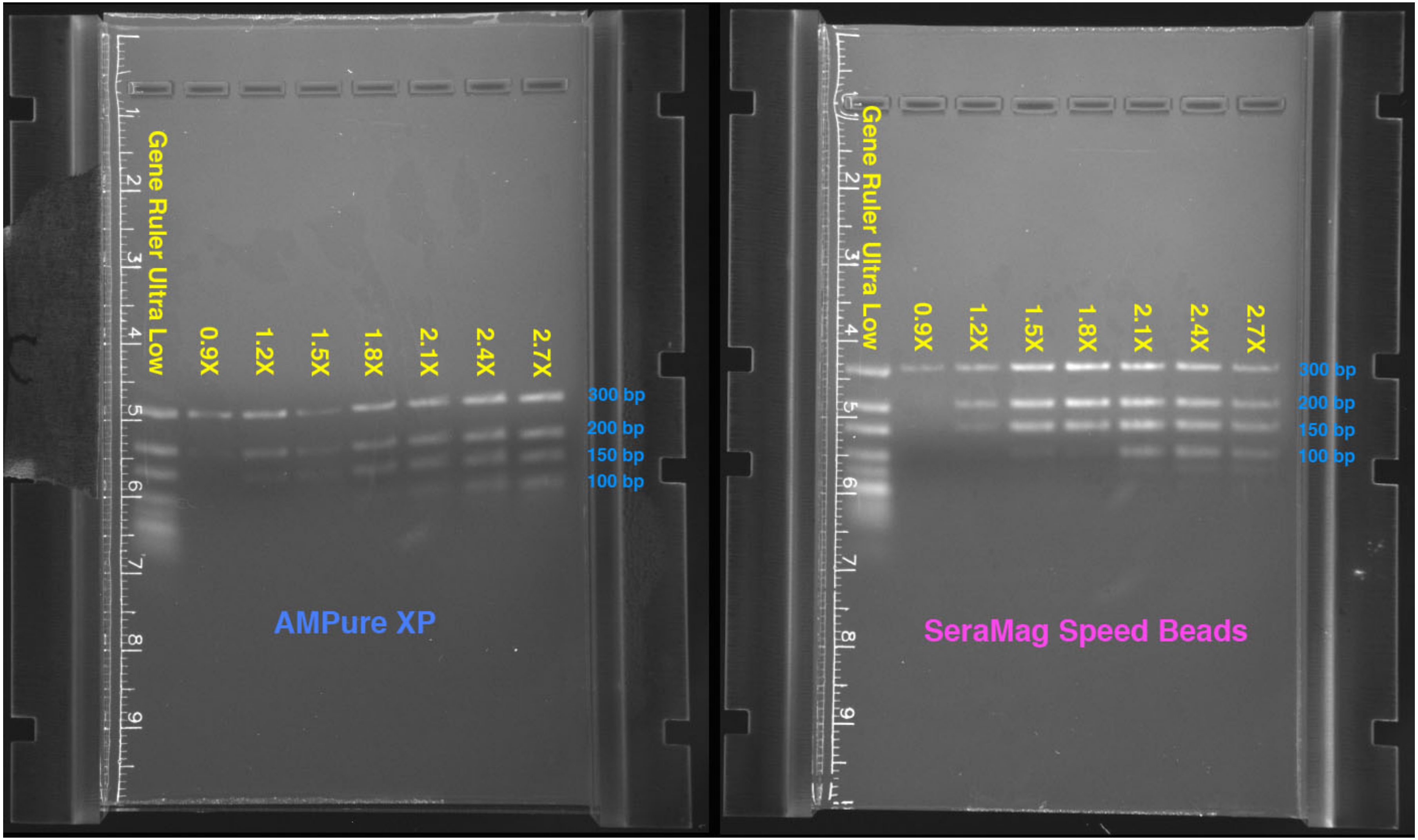
End-repair, phosphorylation and A-tailing

Frist and second-strand cDNA synthesis

Adapter ligation, PCR amplification and sequencing

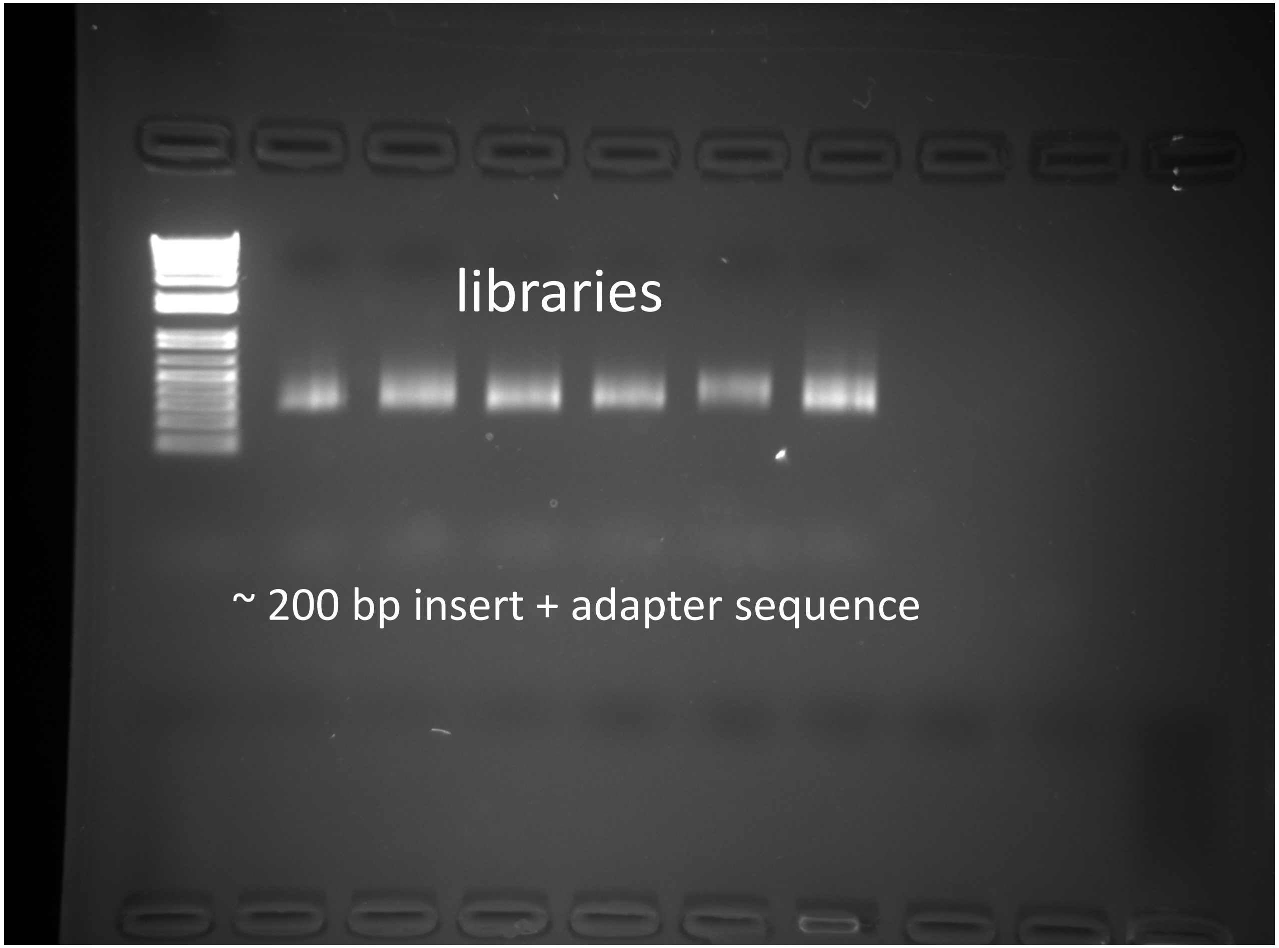
Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

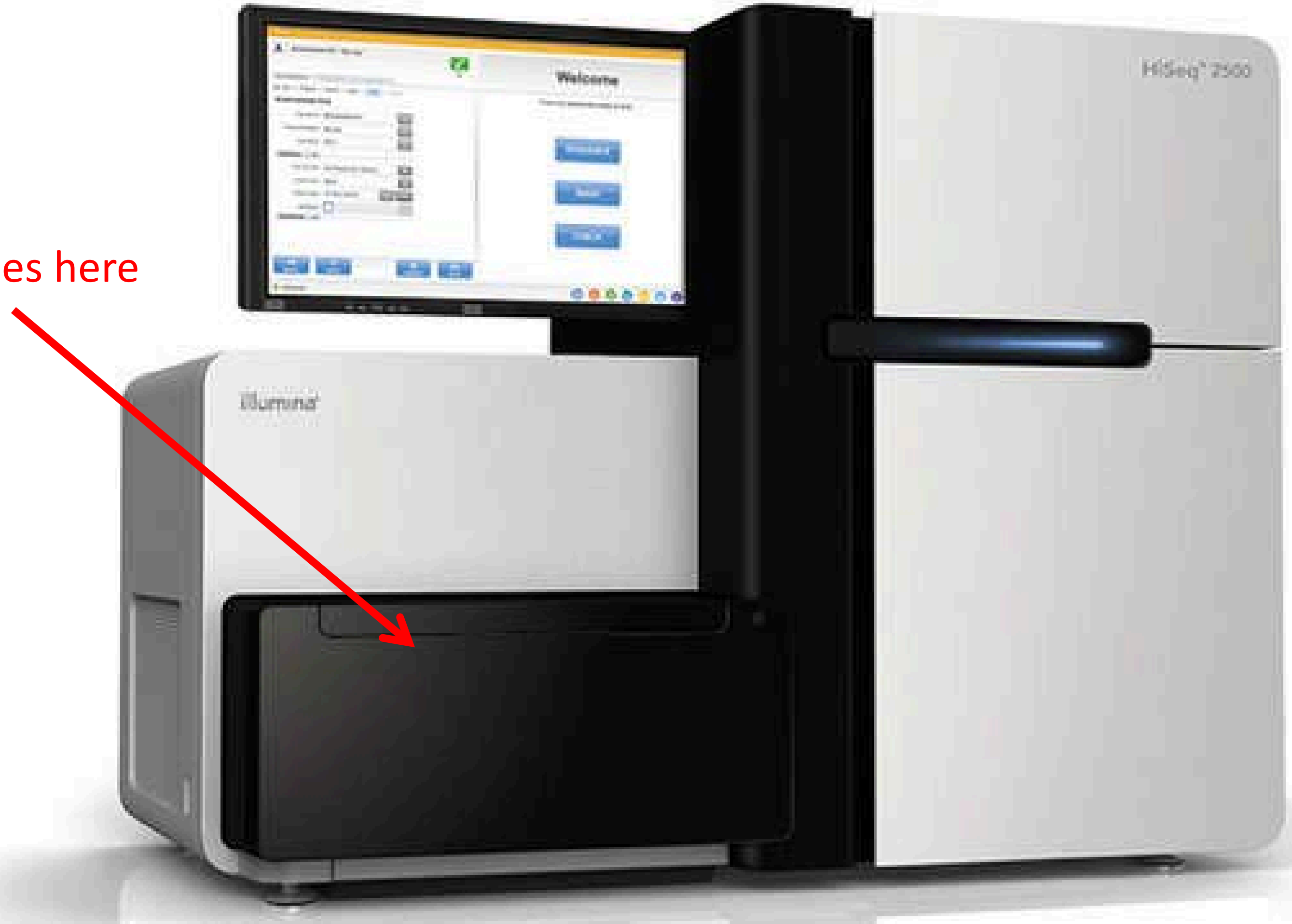
Size selection with ampure (or homemade!)



Here they are on a gel

(2.5 μl of 30 μl of library)





DNA goes here

Here’s what the output looks like:

```
@HWI-D00220:61:C2RBCACXX:3:1101:1473:1951 1:N:0:TAGCTT

NTTCAACTTGAACTGTTACCTGTAATGTCAGTTTGTATCAATTTTTGTTCC

+

#0<FFFFFFFFFFIIIIIIIIIIFIIIFIFFIIIIFIFIFIIIIIIIIIIF
```

Here’s what the output looks like:

Exact position on flowcell of read

Barcode of sample

```
@HWI-D00220:61:C2RBCACXX:3:1101:1473:1951 1:N:0:TAGCTT
```

Sequence of read (50 bases here)

```
NTTCAACTTGAACTGTTACCTGTAATGTCAGTTTGTATCAATTTTTGTTCC
```

```
+
```

they put an empty line in the format just in case

```
#0<FFFFFFFFFFIIIIIIIIIIFIIIFIFFIIIIFIFIFIIIIIIIIIIF
```

Sanger quality scores:

worst

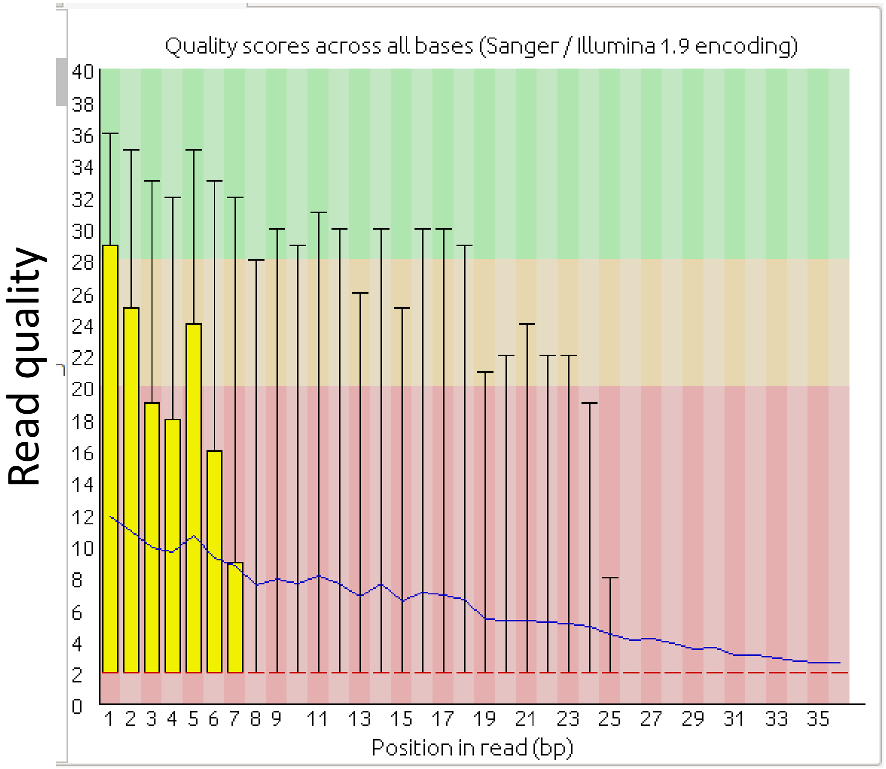
best

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

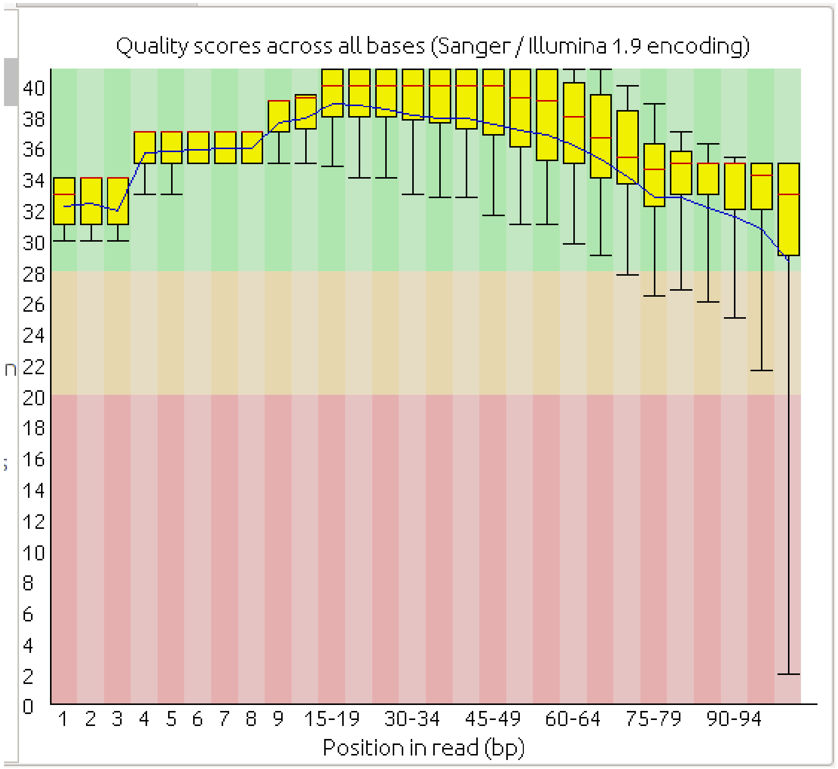
Each read has four lines

- We can use “head” to get a small number of reads to try out some tools
- Just be sure to do it in multiples of 4!
- $head –n 40000 my.fastq > baby.fastq

Is the output any good?



Sucky data

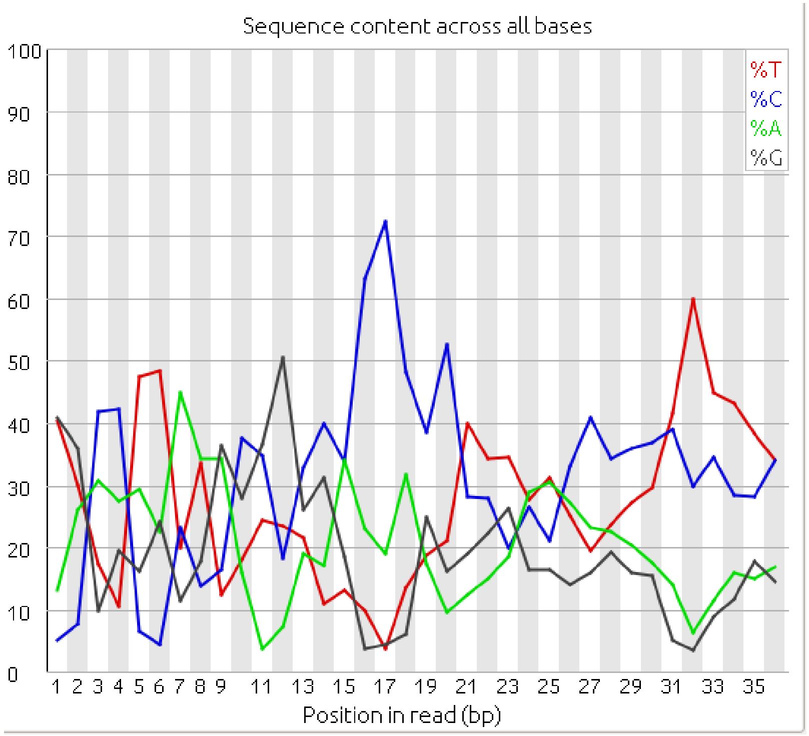


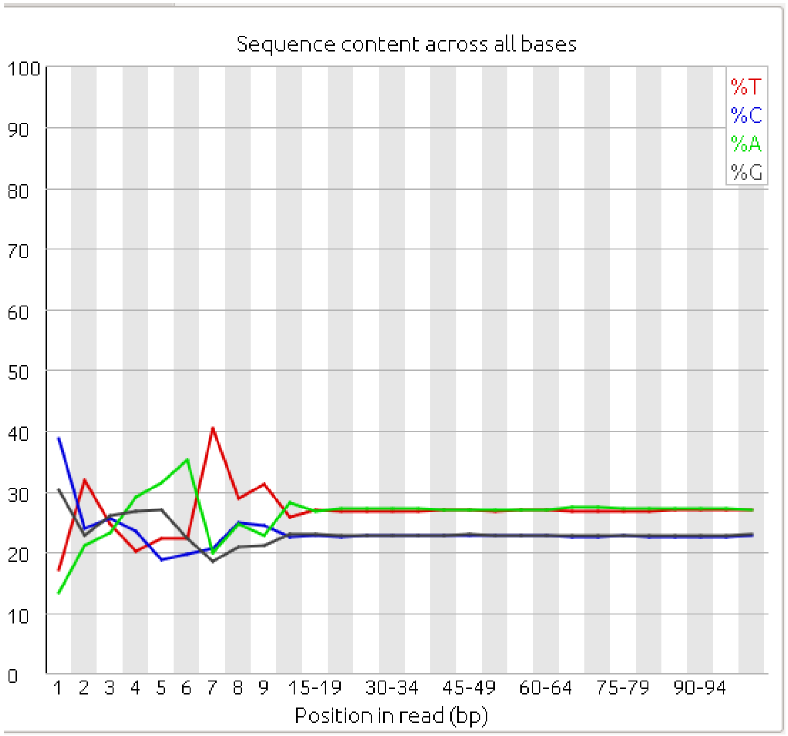
Good data

Quality score of 10 means 90% of bases are correct
20 means 99% of bases are correct
30 means 99.9% of bases are correct, etc.

Genome Res 21: 410-421

Is the output any good?



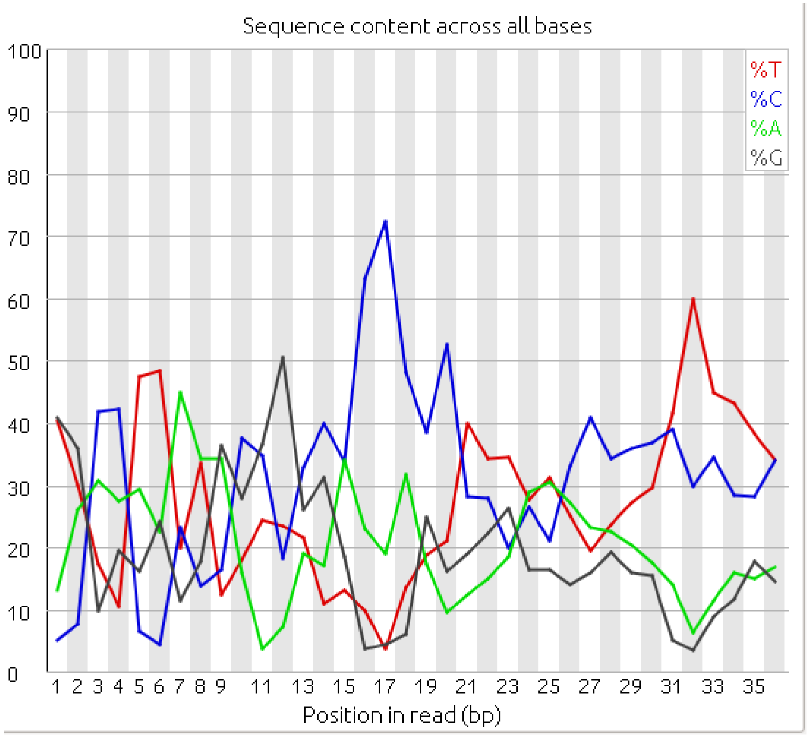


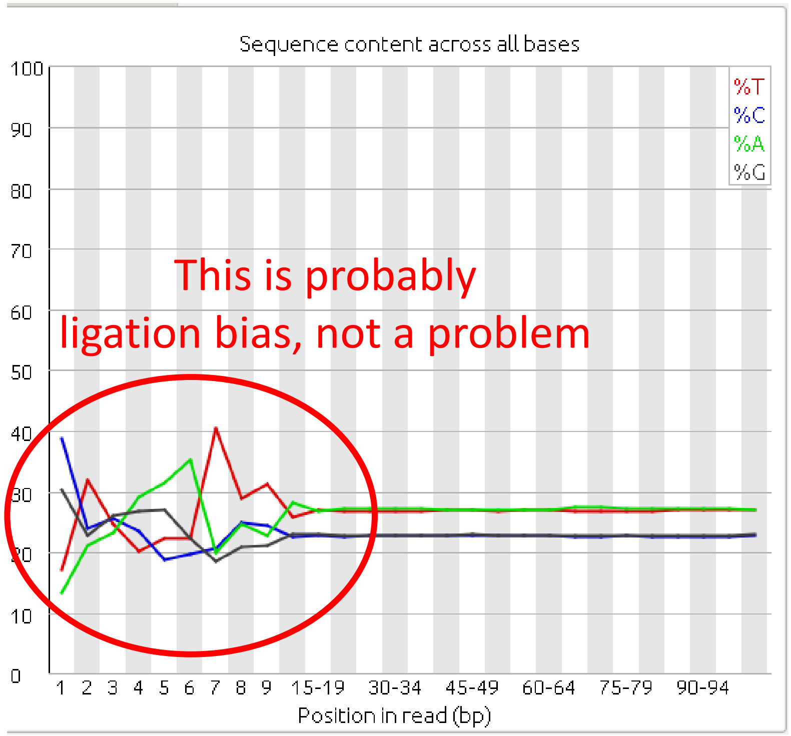
Sucky data

Good data

Genome Res 21: 410-421

Is the output any good?





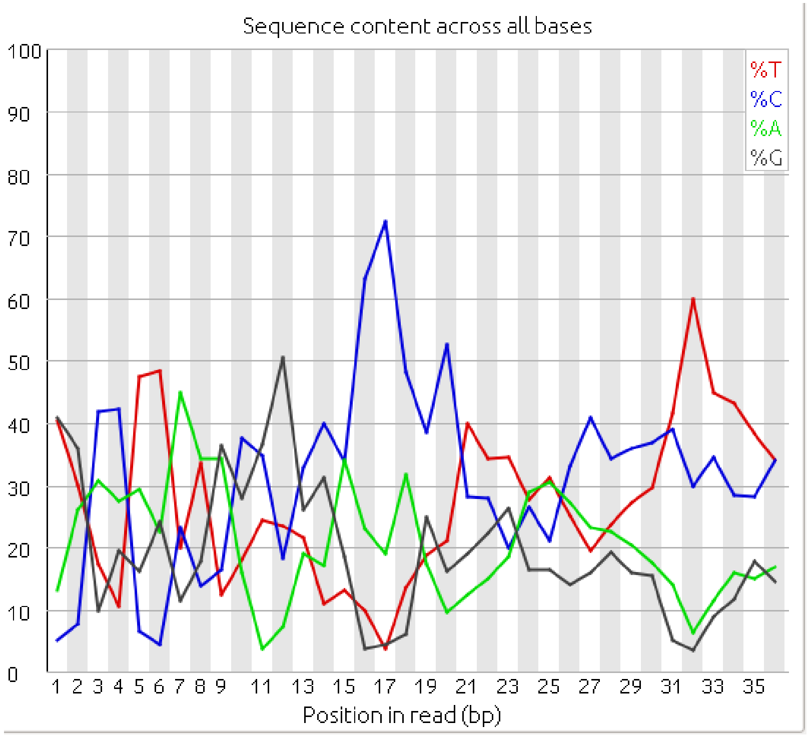
This is probably ligation bias, not a problem

Sucky data

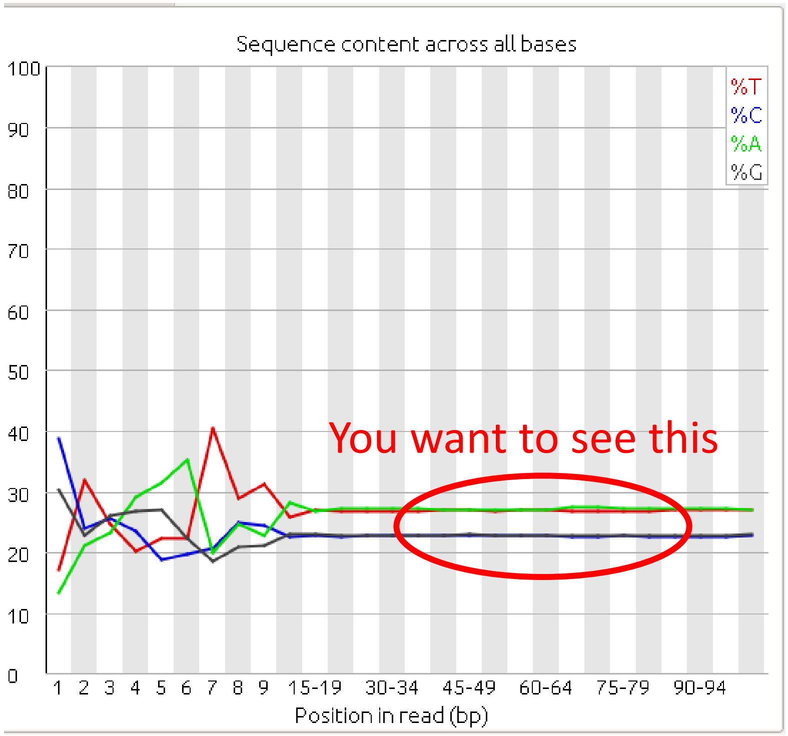
Good data

Genome Res 21: 410-421

Is the output any good?



Sucky data



You want to see this

Good data

Genome Res 21: 410-421

Let’s all install fastqc together

(google fastqc)

What genomic resource to align to?

- Transcriptome: which one?
  - Xenopus laevis EST collection
  - X. laevis JGI project predicted models
  - Univ of Texas Oktoberfest models
  - Univ of Texas Mayball models
- Genome: which one?
  - Multiple versions
  - only useful for feature counting if annotated with a transcriptome!