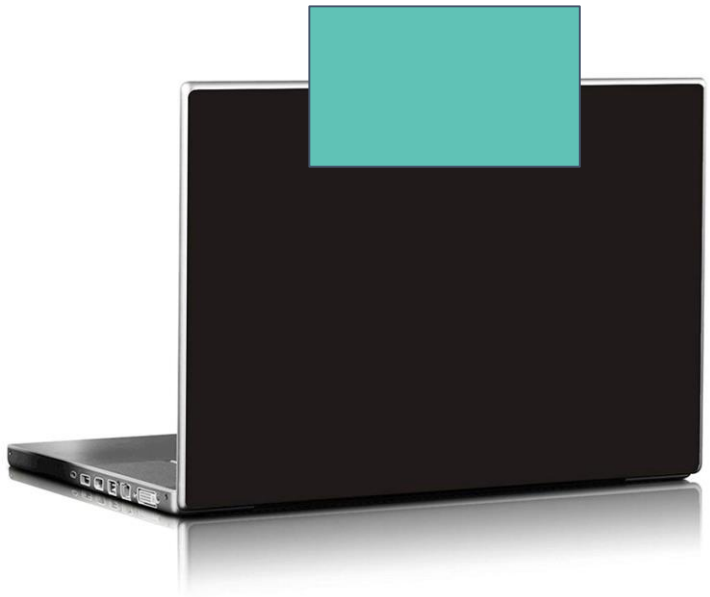
ChIP-seq introduction

Huitian Diao (Yolanda)

References:

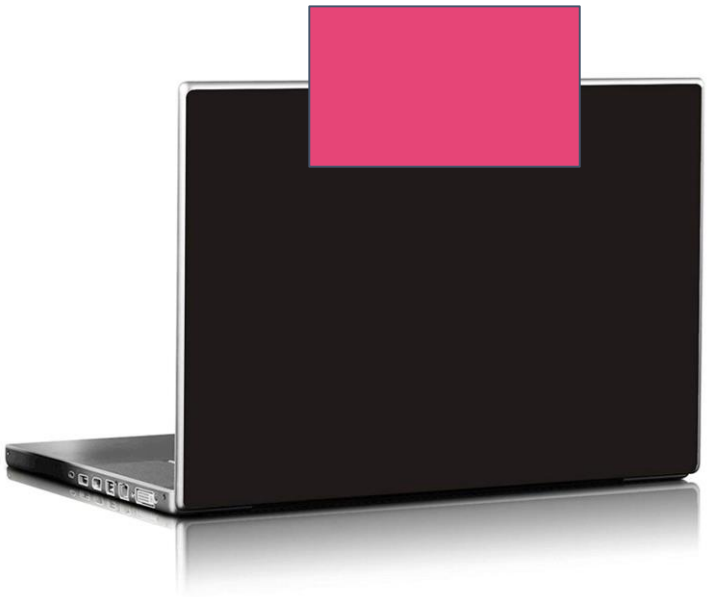
ENCODE ChIP-seq pipeline: https://www.encodeproject.org/pipelines/ENCPL138KID/

Using sticky notes for feedback



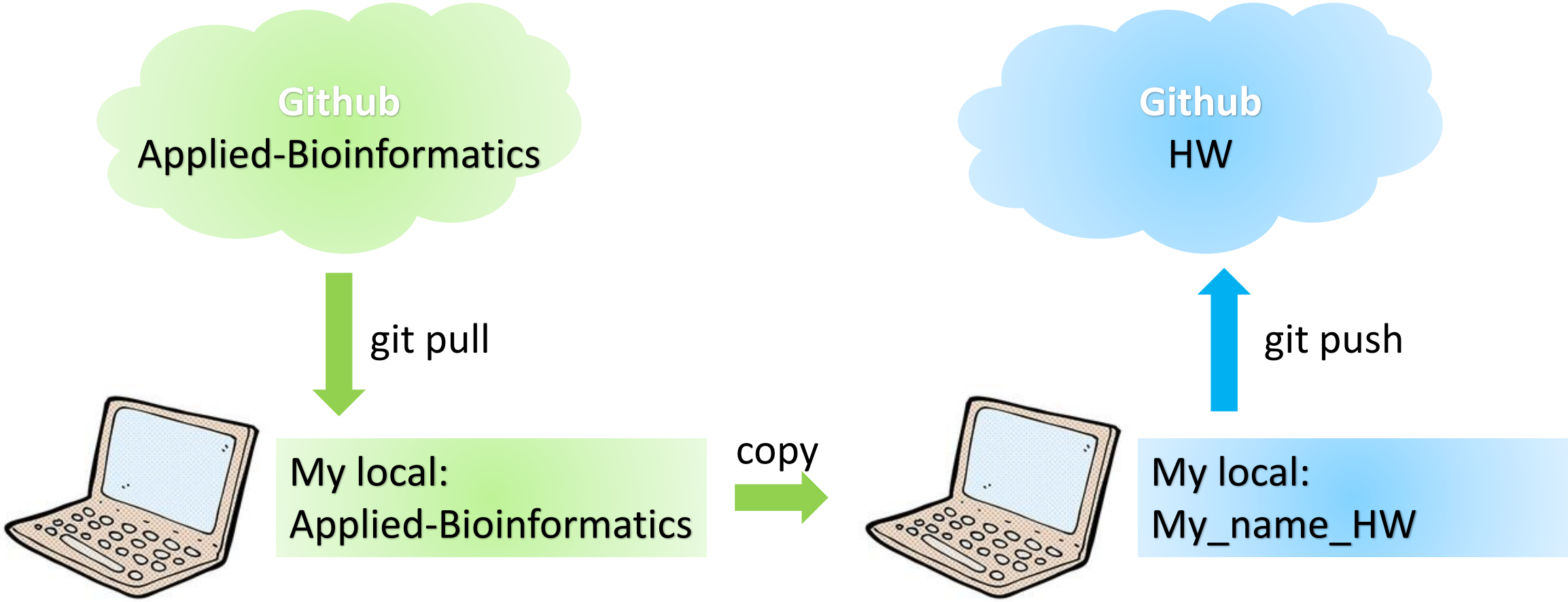
"I've got a good handle on things…"

"I think I understand but I'm still working through things…"



"I'm not understanding and I'm a little lost right now..."

Get course material



Q&A.1

- Why when using some packages we need to include the full path but for others we don't have to?

Step 1. Where are the executables?

```
%%bash
which grep
which fastqc
```

```
/usr/bin/grep
/Users/yolandatiao/anaconda/bin/fastqc
```

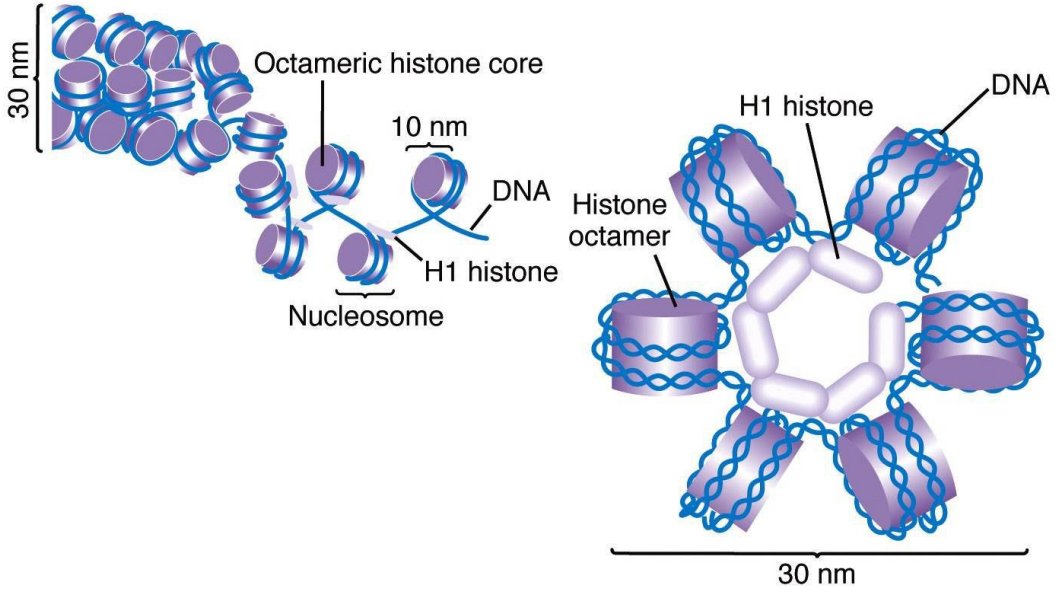
Step 1. How the computer find the executables?

```
%%bash
printenv | grep “PATH”
```

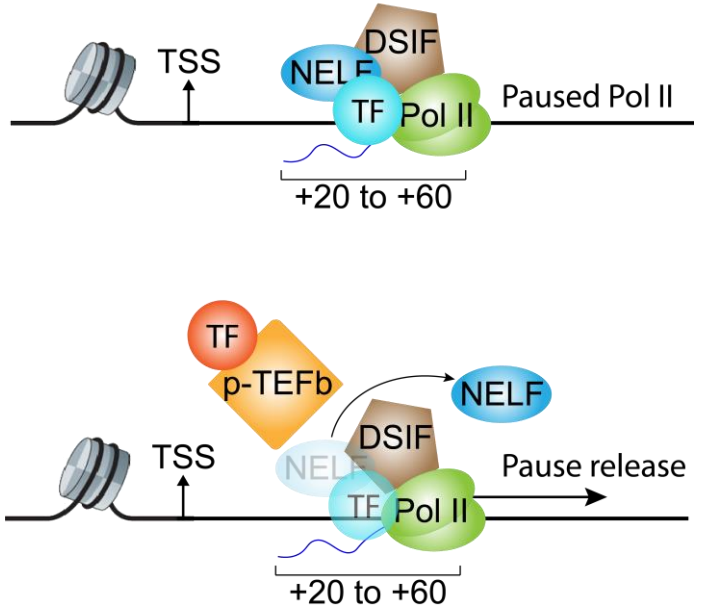
```
PATH=/Users/yolandatiao/anaconda/envs/HDpython3/bin:/Users/yolandatiao/anaconda/envs/HDpython3/bin:/Users/yolandatiao/Documents/0_Bioinformatics2017/201804_Cousera_Unix/Code/newCmd:/Users/yolandatiao/Documents/0_Bioinformatics2017/201804_Cousera_Unix/Code/Commands:/Users/yolandatiao/anaconda/bin:/Users/yolandatiao/anaconda/bin:/Users/yolandatiao/anaconda2/bin:/usr/local/bin:/usr/bin:/bin:/usr/sbin:/sbin:/Users/yolandatiao/.local/bin:/Library/Frameworks/Mono.framework/Versions/Current/Commands
```

Why ChIP-seq?

DNA in the nucleus is highly condensed



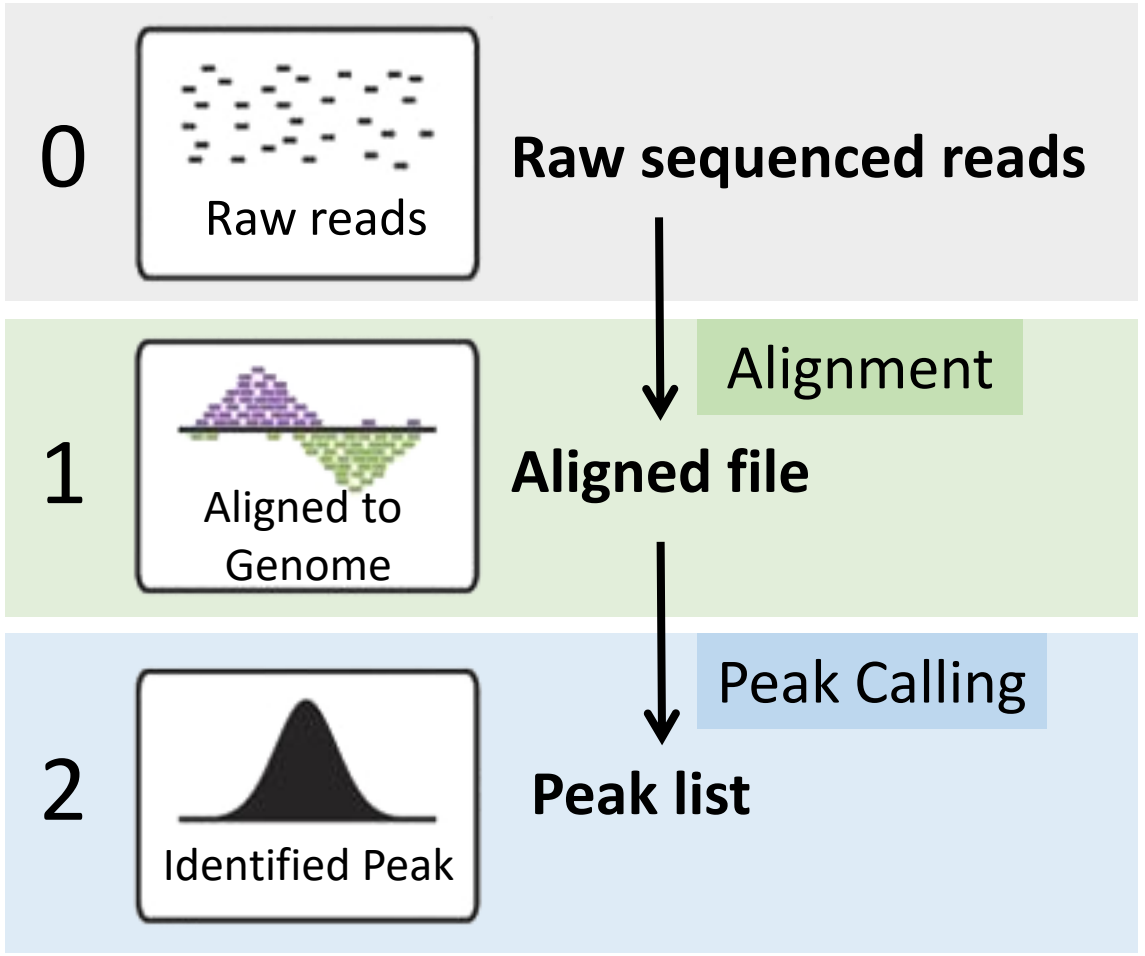
Protein-DNA interaction is crucial
For transcription regulation



ChIP-seq: identify DNA binding sites for proteins

https://www.mun.ca/biology/scarr/Histone_Protein_Structure.html
H. Diao Candidacy Proposal, 2016

ChIP-seq analysis steps



0.0 | Acquire raw reads | fastq-dump

Bardet, Anaïs F., et al. *Nature protocols* 7.1 (2012): 45.

Practice 4.1 (Dump-fastqc)

Tips:

- SRA run selector: https://www.ncbi.nlm.nih.gov/Traces/study/?go=home
- fastq-dump
- fastqc

Objectives:

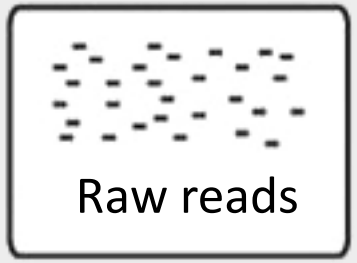
1. **Find Run numbers for ChIP-seq data from this paper:**
   B. H., Immunity, 2016
   1.1 Search for GEO accession number from paper
   1.2 Find SubSerie for ChIP-seq data
   1.3 Search for ChIP-seq data accession number in SRA Run Selector
   1.4 Save **RunInfo Table**

2. **For the first file in the RunInfoTable:**
   Save the first 50000 spots into a file named **test_50000.fastq**

3. **Run fastqc for test_50000.fastq**
   Check the quality of the sequcnce

ChIP-seq analysis steps

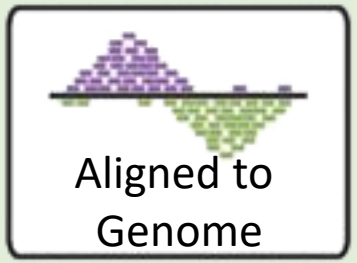
0



Raw sequenced reads

Alignment

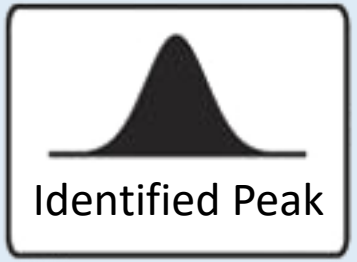
1



Aligned file

Peak Calling

2



Peak list

0.0	Acquire raw reads	fastq-dump	√
0.1	Fastq quality control	fastqc	√
0.2	Trimming	TrimGlore	

Bardet, Anaïs F., et al. *Nature protocols* 7.1 (2012): 45.

Contents

Practice 4.2 (Dump-multiqc)

Tips:

Q2.2: Self-help

Google: generate array of number in shell

Google: shell seq output scientific notation

Google: shell convert stdout to array

Q3: Self-help

Google: **multiqc** documentation

Questions:

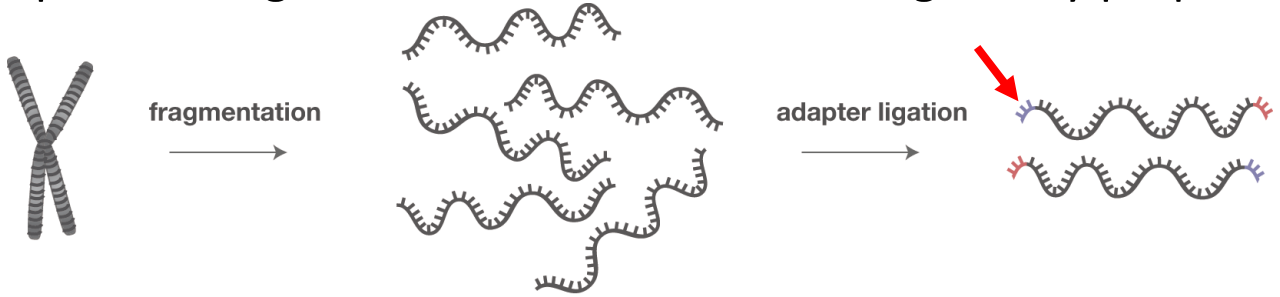
1. **Find Run numbers for ChIP-seq data from this paper:**
   B. H., Immunity, 2016
   1.1 Search for GEO accession number from paper
   1.2 Find SubSerie for ChIP-seq data
   1.3 Search for ChIP-seq data accession number in SRA Run Selector
   1.4 Save **RunInfo Table**

2. **Build a folder named Sample_fastq**
   2.1 Redirect to the new folder
   2.2 Save the first 500 spots of each file into **SRRXXXXX.500.fastq** with a **for loop**
   2.3 Run **fastqc** on all the files

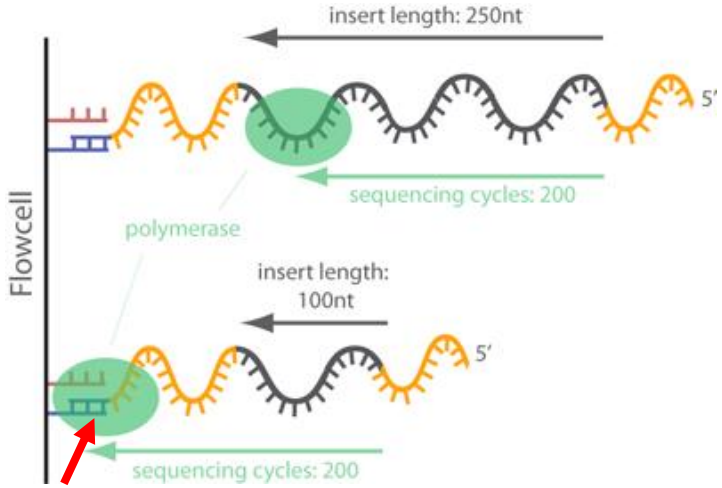
3. **Run multiqc on the fastqc outputs**
   Read fastqc and multiqc outputs. What did you find?

Why trim Adapter -- Adapter contamination leads to lower alignment rate

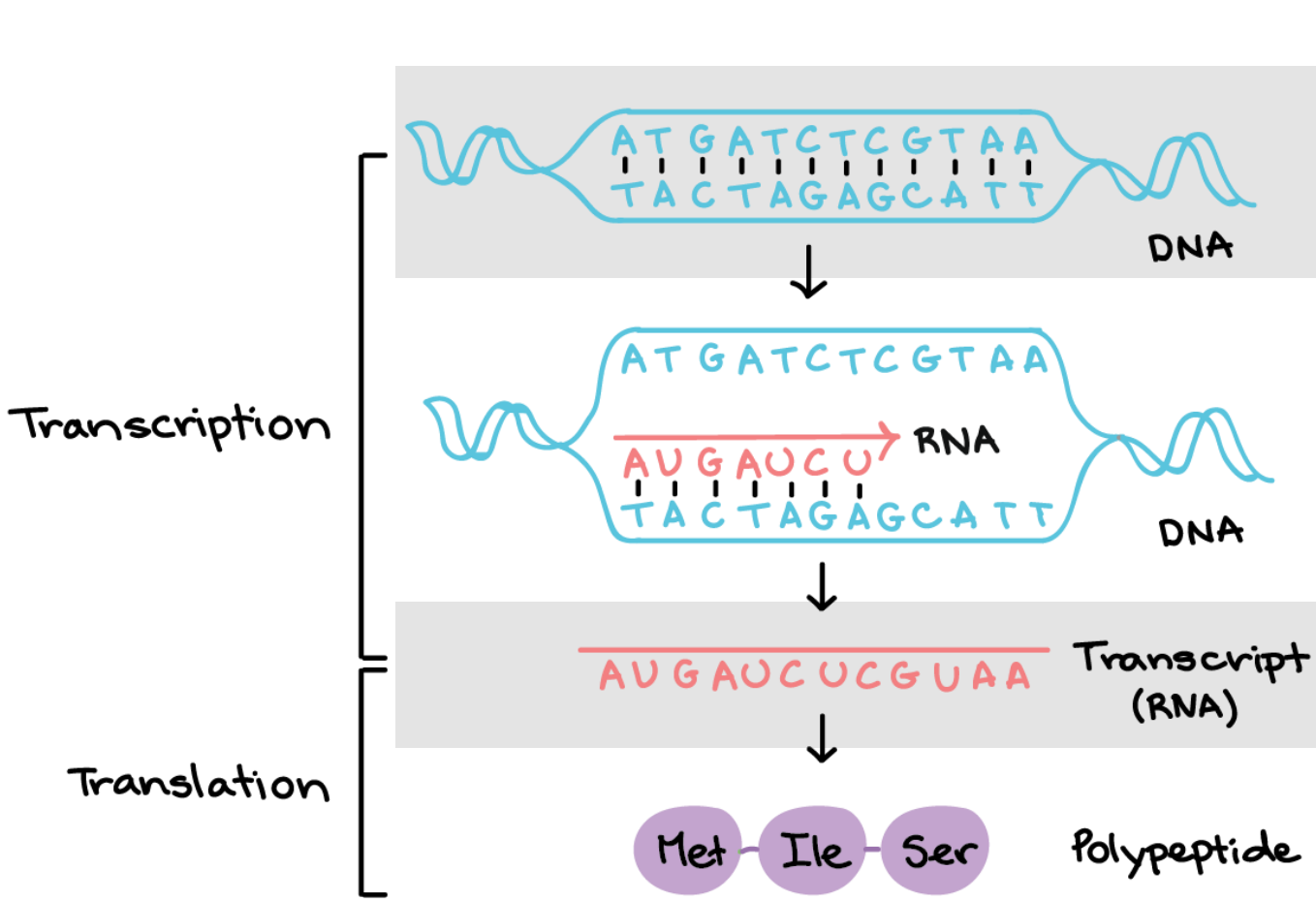
Adapters are ligated to DNA molecules during library preparation



In Illumina sequencing, 3’ end adapter would be sequenced if DNA insert is shorter than number of sequencing cycles



Alignment: ChIP-seq v.s. RNA-seq



Genome sequencing

ChIP-seq
ATAC-seq
Hi-C ...

Transcriptome sequencing

RNA-seq

https://www.khanacademy.org/science/biology/gene-expression-central-dogma

File formats – fastq, sam

Fastq File

Sequence identifier
Sequence
Quality score identifier (+)
Quality score

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<????#=
```

Sam File

Header

```
@HD   VN:1.3   SO:coordinate
@SQ   SN:conticA   LN:443
@SQ   SN:contigB   LN:1493
@SQ   SN:contigC   LN:328
```

Alignment info

```
readID43GYAX15:7:1:1202:19894/1   256   contig43   613960   1   65M   *   0   0
CCAGCGCGAACGAAATCCGCATGCGTCTGGTCGTTGCACGGAACGGCGGCGGTGTGATGCAC
GGC   EDDEEDEE=EE?DE??DDDBADEBEFFFDBEFFEBCBC=?BEEEE@=:?::?7?:8-
6?7?@??#   AS:i:0   XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:65  YT:Z:UU
```

Name	Sam Flag	Ref Name	Position	MAPQ quality	CIGAR string	Mate name	Mate Pos	Temp length	Read Sequence	Read Quality	Add. Info

* Columns separated by Tab (/t)

All you want to know about file formats is here: https://genome.ucsc.edu/FAQ/FAQformat.html