Bioinformatics

from **genotype** to **phenotype**

April 3rd. **Data Rave. @damiankao**

What is it

bioinformatics

computational biology

quantitative biology

biostatistics

What is it

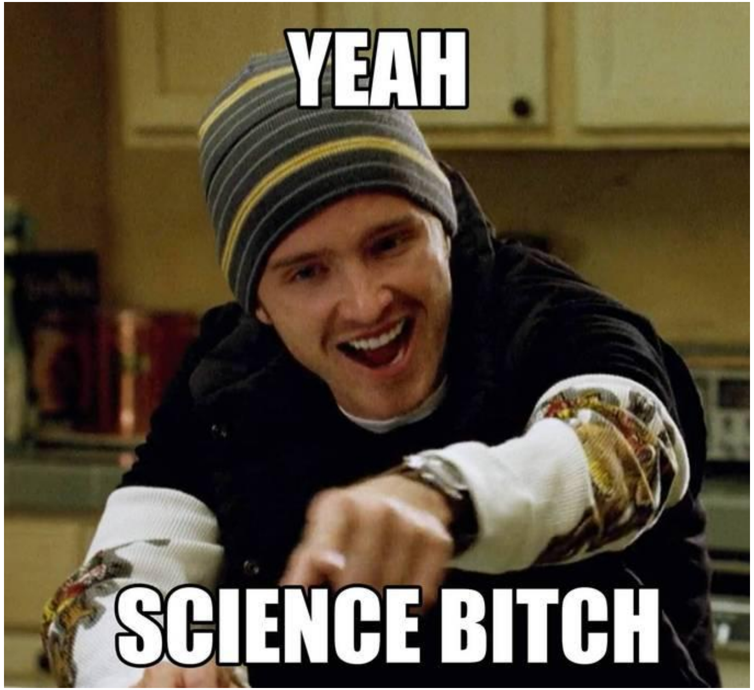
bioinformatics

computational biology

Science

quantitative biology

biostatistics



Genotype vs phenotype

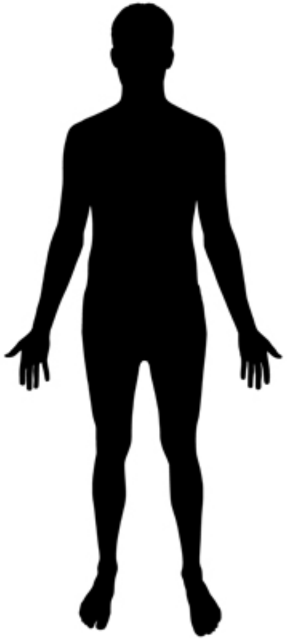
every cell in an organism contains a full set of genome

ATGAGATAGAGATAGCCCTA
TACTCTATCTCTATCGGGAT

GACTAGATAAAGACAGA..
CTGATCTATTTCTGTCT..

- 4 bases (A, G, T, C)
- human contains ~3 billion bases
- genotype is information

Genotype vs **phenotype**

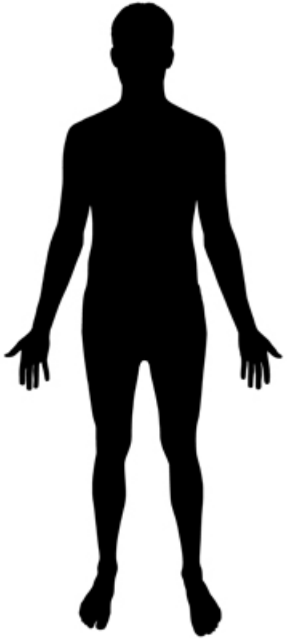


- Phenotype is a vague term
- It is any feature or trait that is NOT genotype
- Hair color, height, blood type, rate of production of a certain protein, fertility….etc
- Can be hard to quantify

Genotype to phenotype

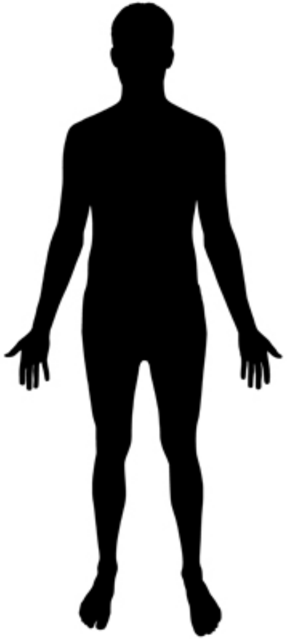
```
ATGAGATAGAGATAGCCCTA
TACTCTATCTCTATCGGGAT

GACTAGATAAAGACAGA..
CTGATCTATTTCTGTCT..
```

- ~ 3 billion bases for humans
- ~ 1 gigabyte of data so far



- How do you quantify all these dimensions?
- Probably a lot of data?

Genotype to phenotype

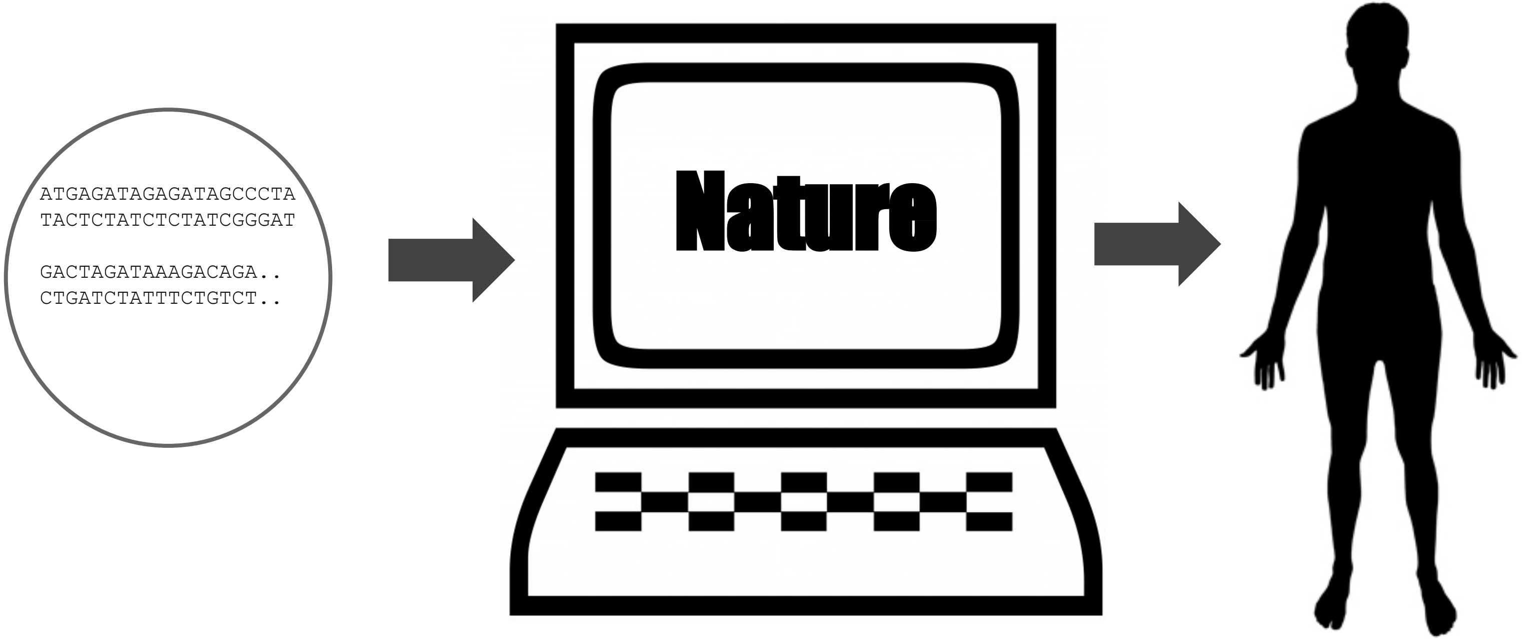


phenotype data
is algorithmically
compressed into
genotype data

greatest compression
algorithm ever??

```
ATGAGATAGAGATAGCCCTA
TACTCTATCTCTATCGGGAT

GACTAGATAAAGACAGA..
CTGATCTATTTCTGTCT..
```

Genotype to phenotype



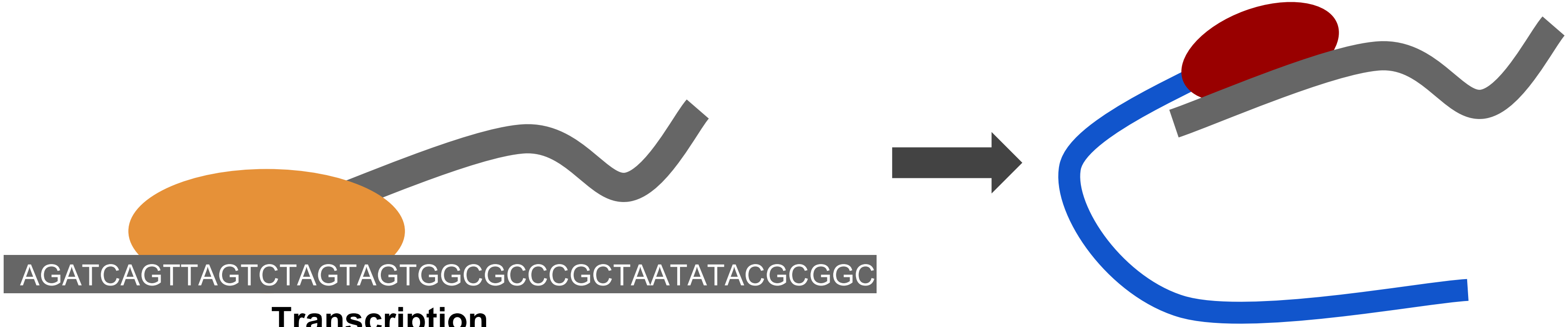
ATGAGATAGAGATAGCCCTA
TACTCTATCTCTATCGGGAT

GACTAGATAAAGACAGA..
CTGATCTATTTCTGTCT..

Genetics = How does this compression work?

Layers of complexity

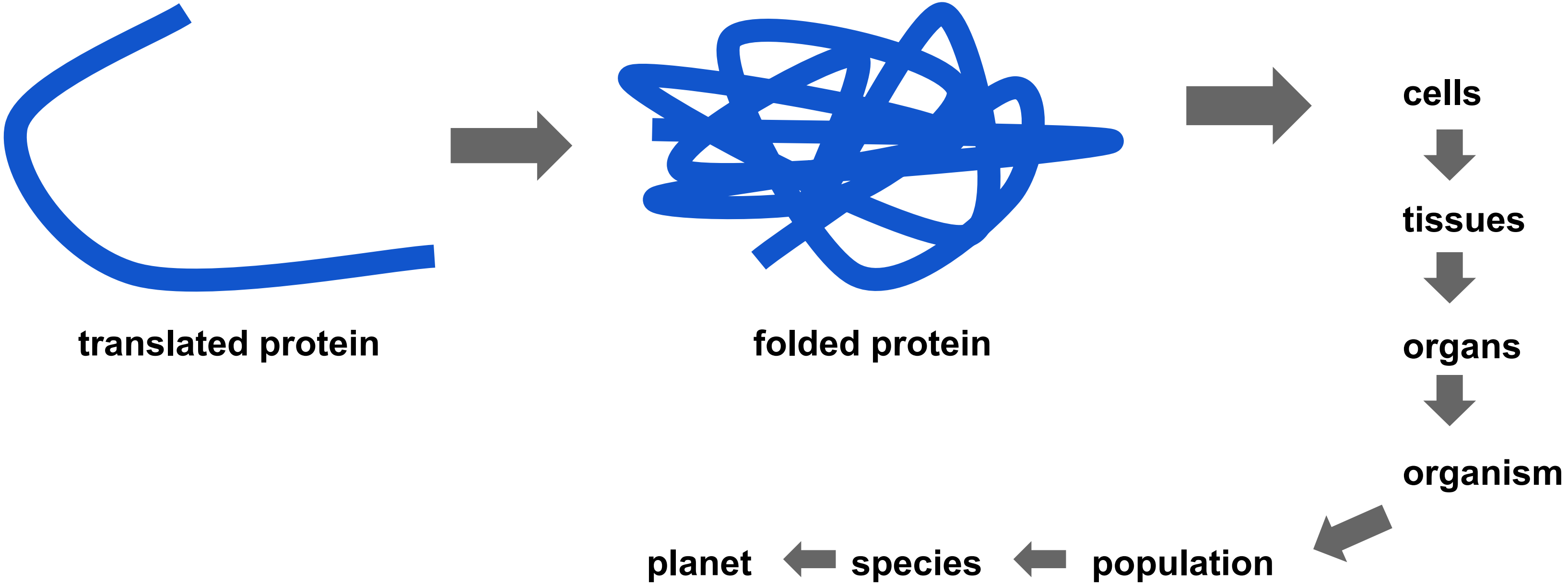
There are many layers of complexity above genotype that eventually leads to phenotype.



Transcription

Translation

Layers of complexity



translated protein

folded protein

cells

tissues

organs

organism

population

species

planet

Sequencing the genome

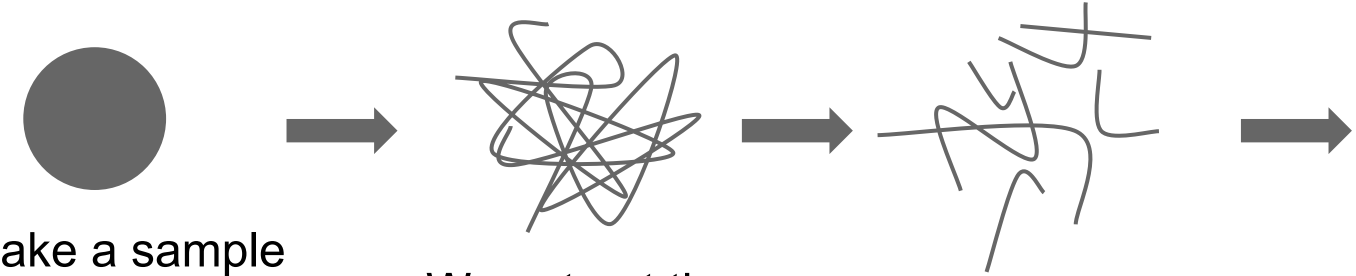
- There are billions of bases and bunch of chromosomes

- In a perfect world, we would get a really long string of bases representing each chromosome.

Sequencing the genome?

- Our sequencing technology isn’t there yet (maybe soon)
- We can sequence a lot of short fragments

Sequencing the genome



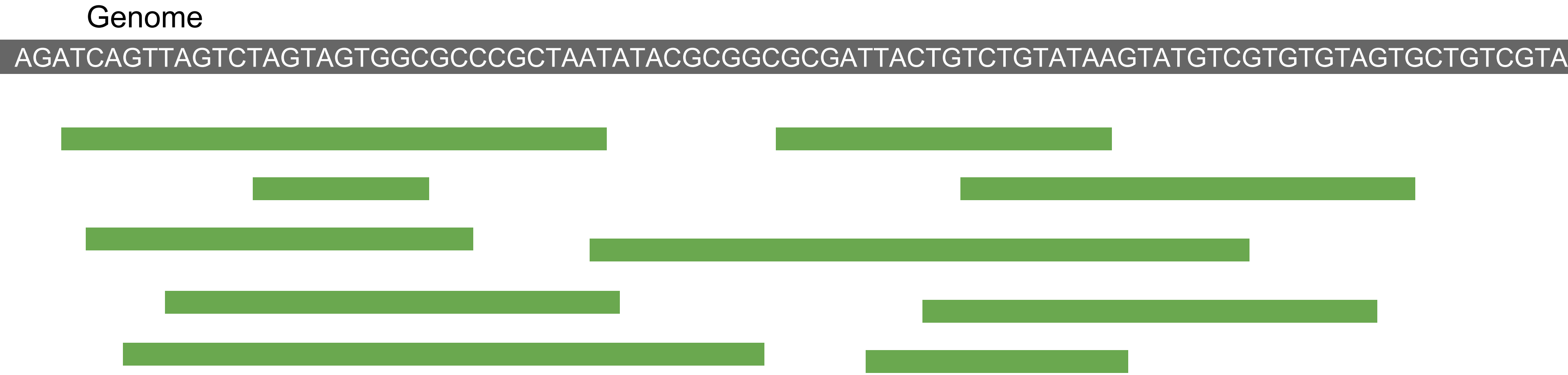
We take a sample from you (ie. cheek swab). This consists of millions of cells.

We extract the DNA from all of the cells.

We chop them up into random smaller sized fragments

Sequencing machine

Sequencing the genome



Genome

AGATCAGTTAGTCTAGTAGTGGCGCCCGCTAATATACGCGGCGCGATTACTGTCTGTATAAGTATGTCGTGTGTAGTGCTGTCGTA

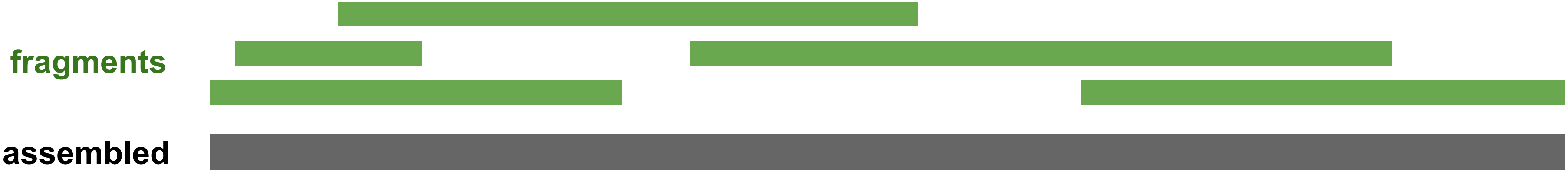
As a result of sequencing, we get back small fragments of the genome

Result file (.fasta)

```
>Fragment01
AGGTTAGGTTTTAGCTTGATGCTTAGCTTGATGCAGTATTATGTATCGTATCGTATATGTCGA
>Fragment02
GGTTTTAGCTTGATGCTTAGCTTGATGCAAGTTAGTCGTTAGTCGTTAGTCGTAGTGATG
>Fragment03
TGCTTAGCTGCTTAGCTTGATGCAAGTTAGTCGTTGCTTAGCTTGCTTAGCTTCGTTAG
…
…
```

Computational problems

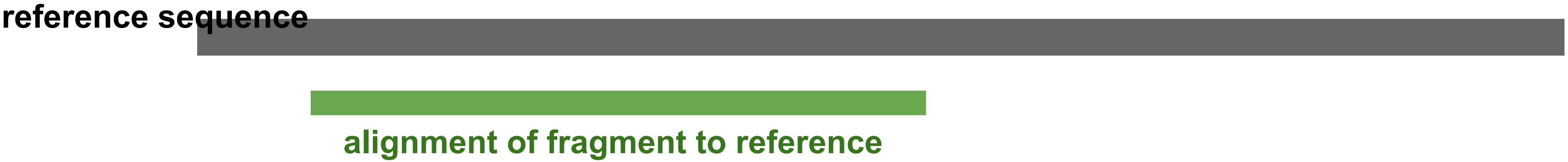
Assembly of fragments



fragments

assembled

Alignment of fragments



reference sequence

alignment of fragment to reference

Analysis problems

Variant calling

Bob AGTC**C**GTTATT**G**ACTTCGTAGTC

Tom G

Mary AGTC**C**GTTATT**G**ACTTCGTAGTC

Sam G

AGTC**A**GTTATT**T**ACTTCGTAGTC

Correlation to phenotype

G

AGTC**A**GTTATT**G**ACTTCGTAGTC

G

Data engineering problems

- **Compression** of raw data.
- **Querying** the data.
- **Standardization** of formats.
- **Accessibility** of the data.

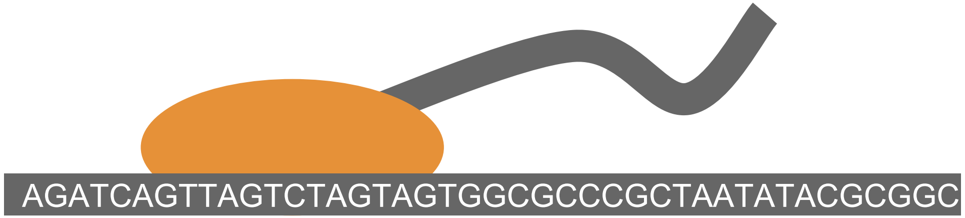
DNA-seq

DNA sequencing (DNA-seq) is NOT the only type of sequencing that is being done.

(RNA-seq, ChIP-seq, DNase-seq, RAD-seq, BisChIP-seq…..)

RNA-seq

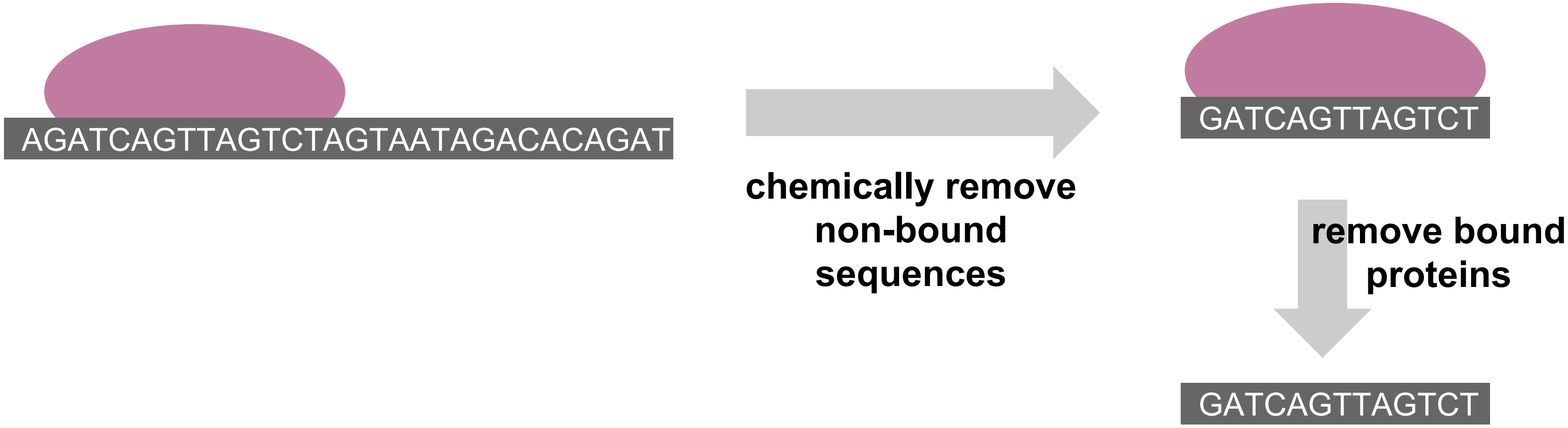
Sequencing transcripts can give us quantitative data on gene expression



Transcription

ChIP-seq

Sequencing bound DNA regions tells us where on the genome activity is happening



AGATCAGTTAGTCTAGTAATAGACACAGAT

chemically remove non-bound sequences

GATCAGTTAGTCT

remove bound proteins

GATCAGTTAGTCT

Future

- Better sequencing technologies will make many problems non-issues

- Maybe we can co-opt quantify self data as phenotype data

- Open source data

- Questions??