Coverage and Depth

- Often used interchangeably and differences are not clear
- Coverage is the number of times a base-pair is read (called) by the machine
- Coverage is represented as a **average** for a region/genome with letter X appended to it. It gives a rough estimate.

  Coverage = (Total Base pairs Sequenced)/(Sequenced Region length)
- 10X, 20X …
- Coverage per base has a distribution and needs mapping information

Coverage

Average Estimated Coverage

* Based on sequencing output
* Before mapping
* Start out estimate
* Total basepairs sequenced/(Total haploid DNA length)

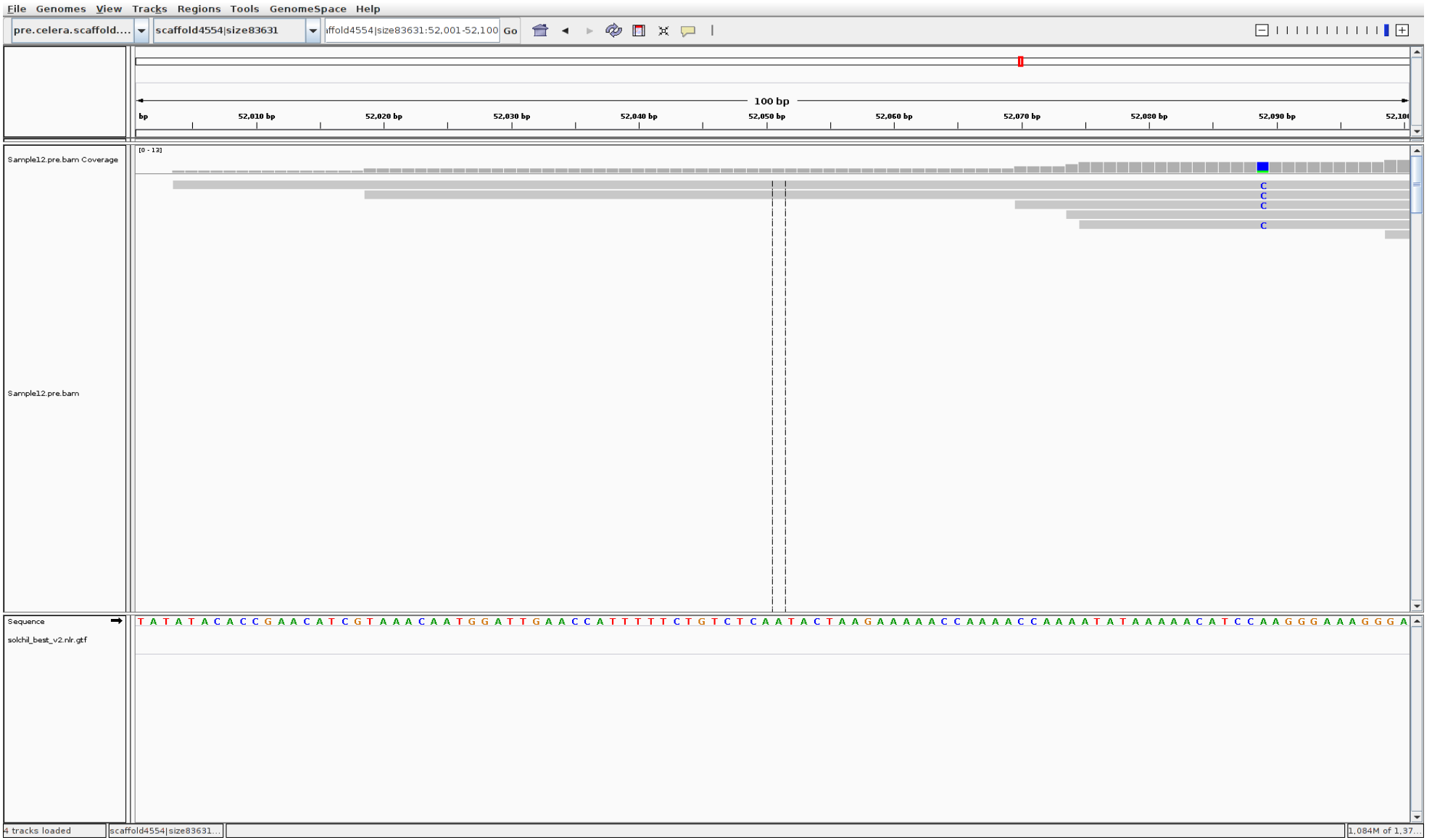
Per Site

* Based on mapping information (BAM)
* Forms a distribution for a genome/region
* Bedtools/Samtools

Average actual Coverage

* Mean of the persite distribution

Per Site Coverage



Biological Implications

- Coverage data has randomness and uncertainty. Due to random selection of DNA fragments
- **Coverage is proportional to the amount of DNA that comes from a region with some noise**. Duplication of the region in the sample would increase the coverage.
- Significant differences in WGS coverages could indicate
  - Duplications (CNVs)
  - Deletions
  - Transposon insertions
  - Disease associations and Adaptation
  - Examples
    - Anuoploidies XXY (Klinefelter syndrome), XYY, XXX
    - Durg resistance
    - Cancer

Tools for coverage calculations

- Samtools and bedtools can calculate site coverage for a particular site.
- Samtools idxstats <BAM> will give a the number of reads aligning to a particular chromosome.
  - Technically it does not give coverage
  - Its very fast

Problem

You just got resequencing data (fastq) for the organism A whose genome is 1Mb (10^6 bp). It contains 50 Million paired end reads of 100 base pairs. What is the average expected coverage ?

Problem

Your lab works on organism A whose genome is ~ one Megabase (10^6 bp). You are in-charge of resequencing a sample. A 10X mean coverage is desired. The NGS facility produces 100 base-pairs paired end reads. How many reads are needed. Would you sequence exactly whats needed ? If not why ?

Problem

You have enough money for resequencing at a 30X coverage largely for SNP discovery. But there are two samples from the same species but one is a wild line/strain and another is an inbred laboratory strain/line. Would you divide the coverage equally (15X,15X) ? If Not why?

Problem

In your home_directory/Lect_03/prob1 there are two bam files. The data is for human and both files have roughly the same number of reads. Which file has a female sample and which one has a male sample ?

HINT: Use idxstats, Human females are XX and males are XY

Problem

In your home_directory/Lect_03/prob2/ there are three bam files. The data is for human and all files have roughly the same number of reads. One sample has a disease. Can you -

- Identify the abnormality ?
- Which disease is it ?

HINT: Use idxstats

BAM and SAM

A Non Technical Analogy!

What is a BAM/SAM ?

- …. ?
- …. ?
- …. ?
- Analogy ?

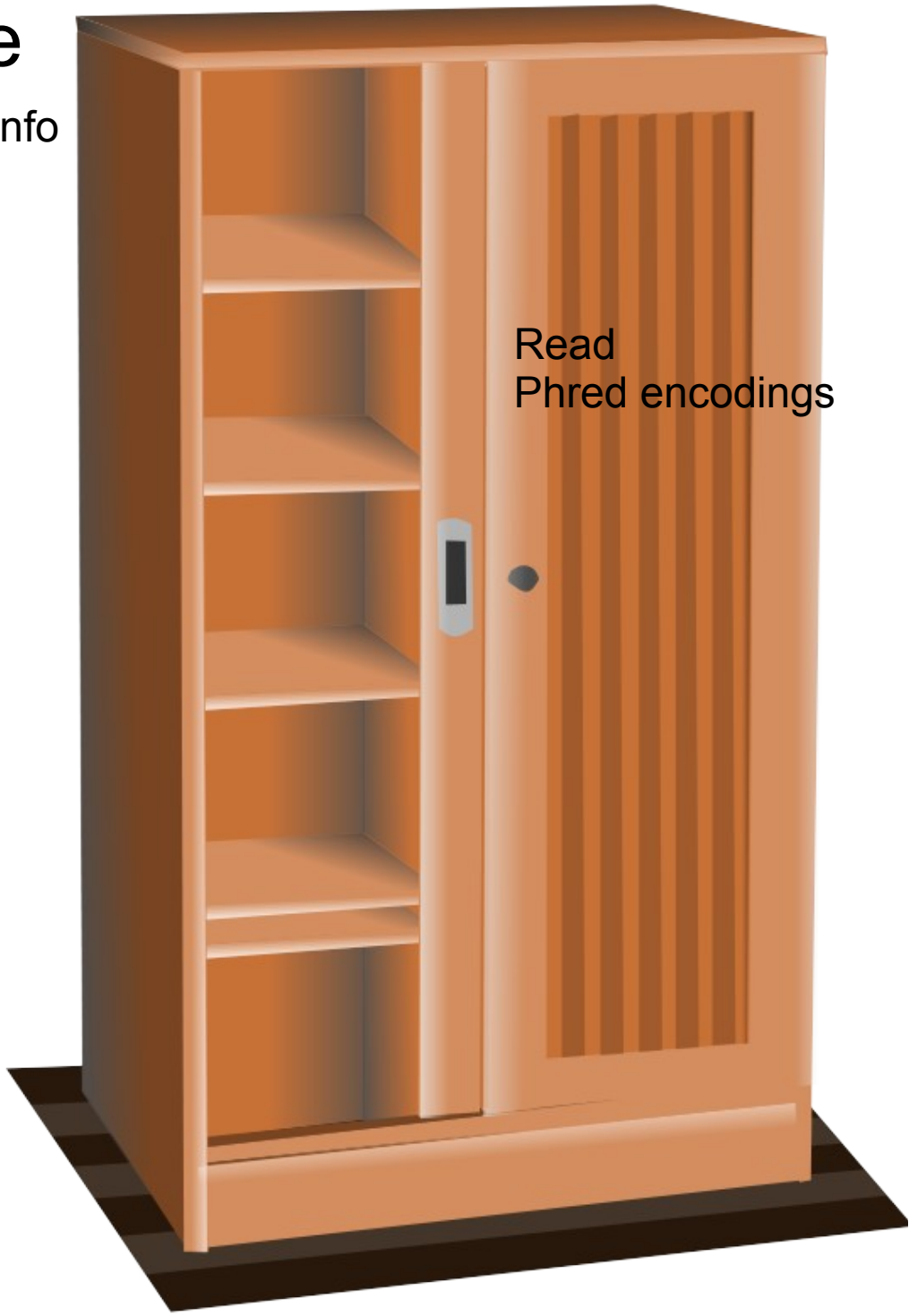
What is BAM/SAM

- Format/Specification
  - Unification/Common Language
  - Data Sharing
  - Downstream Processing
  - Flexibility/Evolvability
- A File
  - Human Readable?
  - Indexed (Random and FastData Access)
  - Data Visualization (Browsing, Zooming)
  - Compression (Storage)

Bam as Alignment Storage

@Reference Info
@Aligner Info

Mapping Info



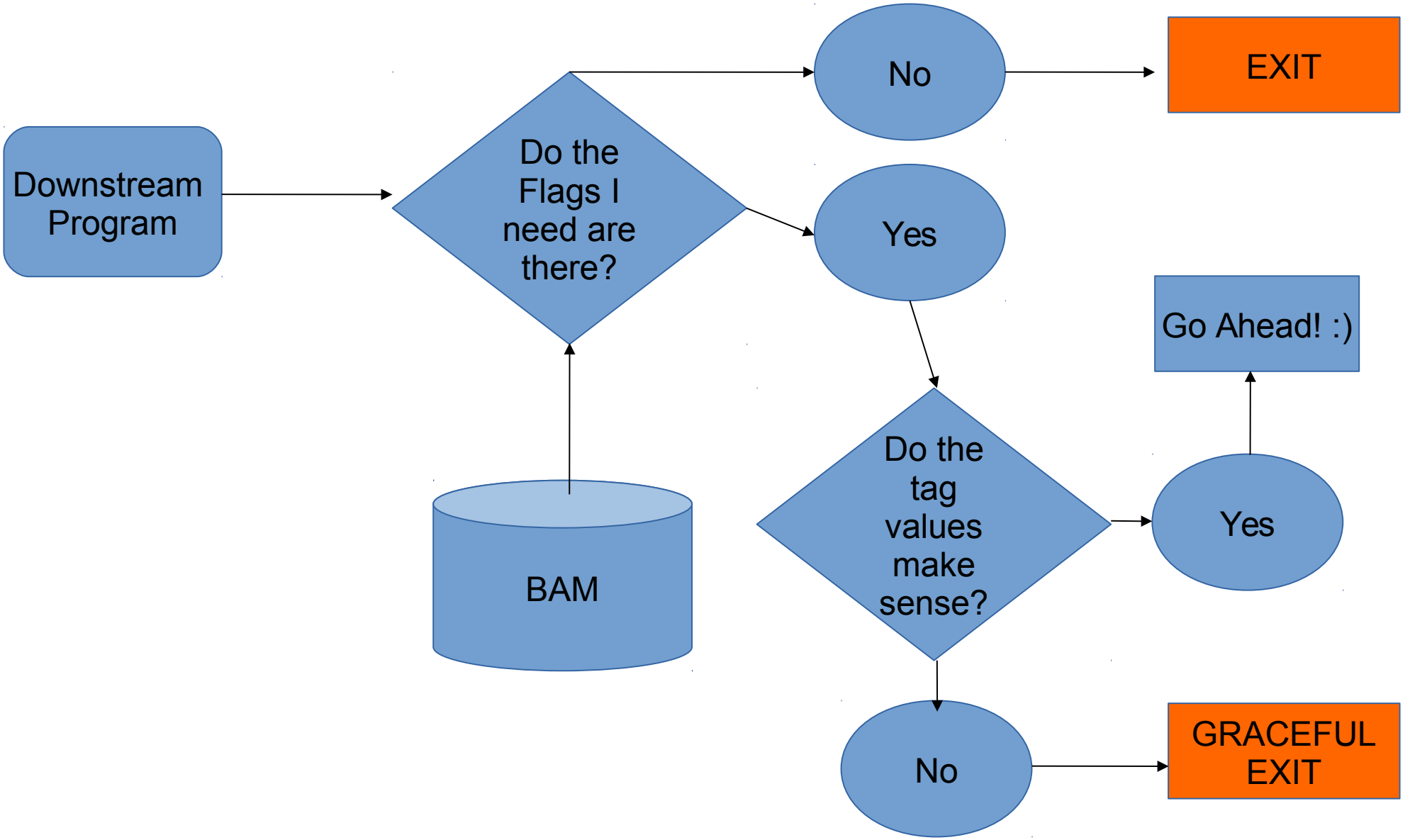
Who Fills the Closet?

Who does what

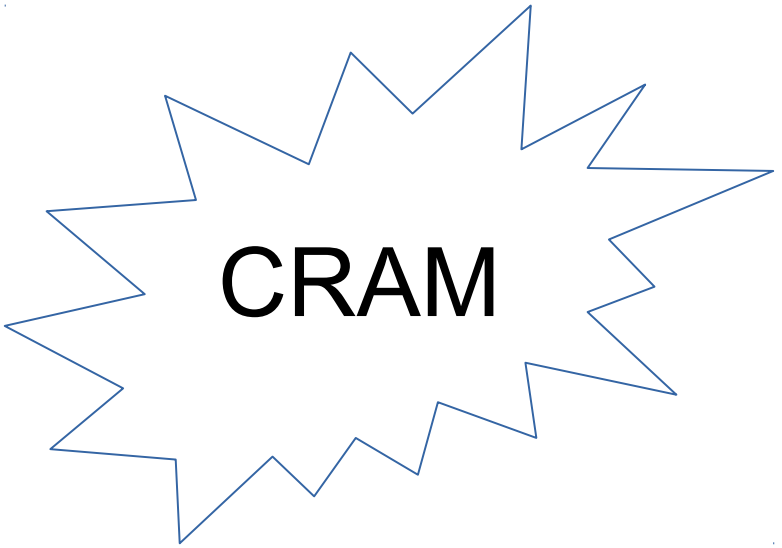
- It is the aligner who fills the bam
- Different aligners will produce different bams
- SAM/BAM specify Flagsets but ultimate values are put there by aligner/user
- Flags/Tags can modified in aligner settings
- Some Flags values can be set/added by downstream programs manipulating bams
- **It is You who are ultimately responsible for the Flag values**
- Knowing what the aligner is putting as a tag value is important
- Why?

Why ?

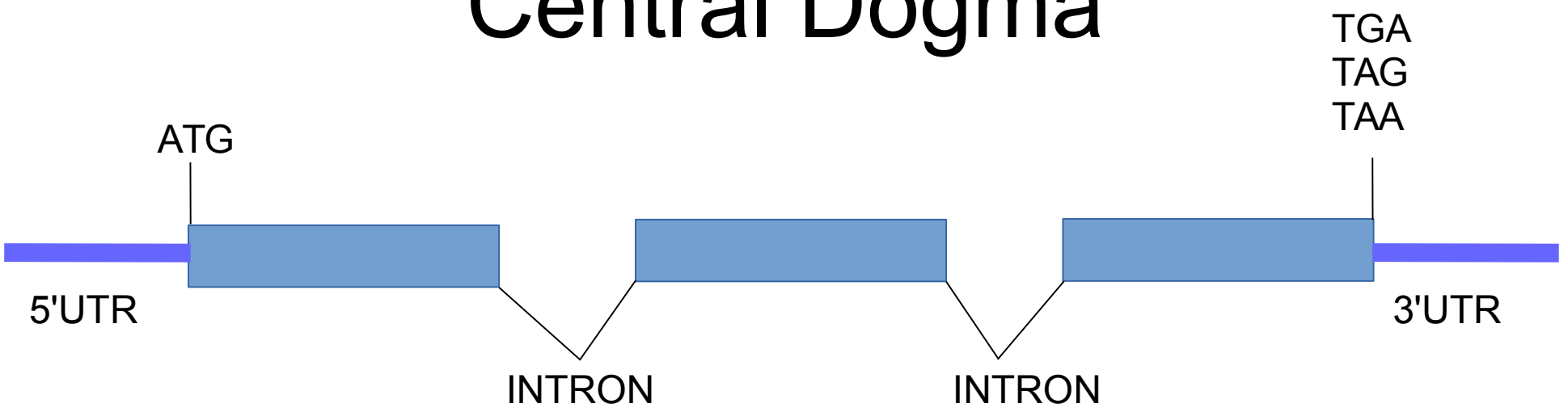
Downstream Program



New Developments



Central Dogma

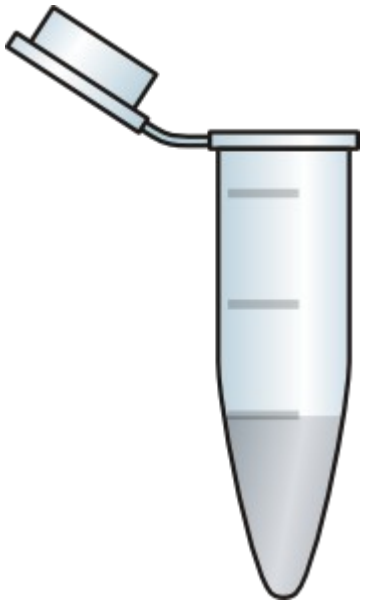


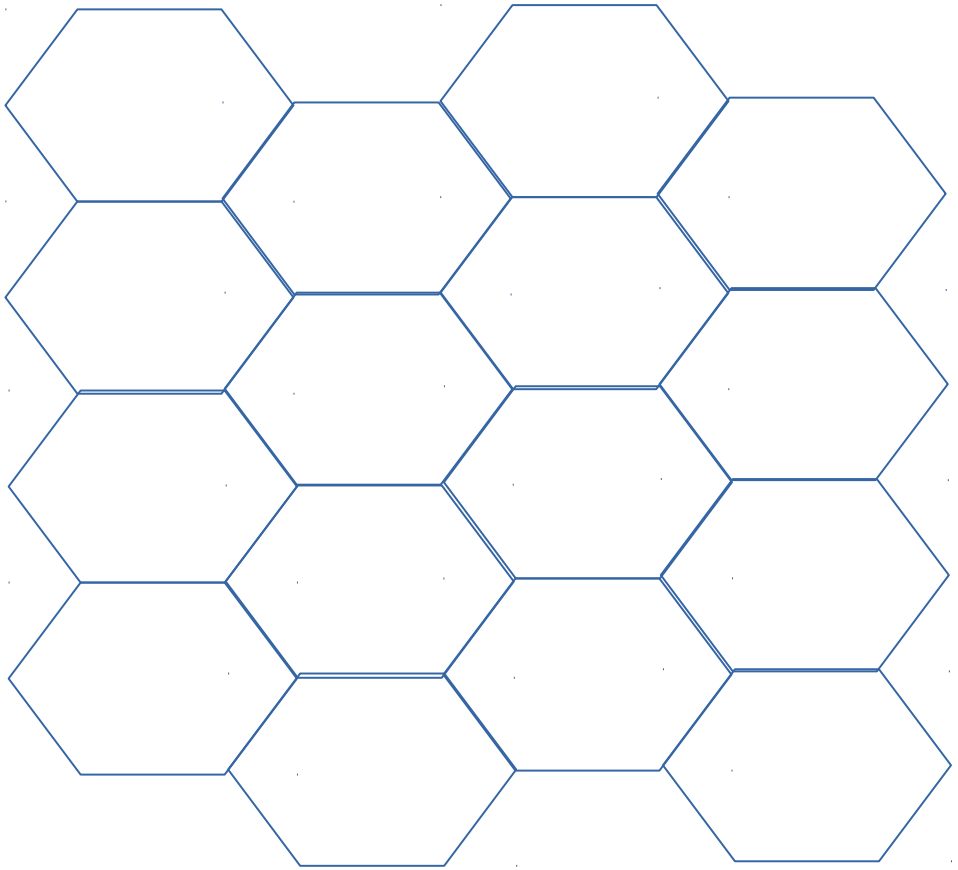
RNA-Seq

What can be done ?

- Unbiased Expression profiling
- Differential expression analysis
- Gene discovery
- Splice variant Profiling/Discovery
- Transcription start site discovery (CAGE)
- Single cell Genomics
- Strand specific RNA-Seq
- …..

Cells





Cell