Welcome!

Titus Brown

Qingpeng Zhang

John Blischak

Goals!

- Drive-by introduction to:
  - Cloud computing
  - Basic Illumina sequence quality evaluation & control
  - De novo mRNAseq assembly
  - A (our) "protocol" for mRNAseq analysis => diff expr
  - Variant calling protocol, too

- This will let you explore other online resources to your heart's content, we hope!
  - Other protocols & tutorials
    - khmer-protocols
    - ged.msu.edu/angus/tutorials-2013

Our goals:

Answer your questions!

Help you figure out what questions to ask!

Point to further materials!

Structure of day

- Start by logging into cloud machines, grabbing data, running analyses.
- Coffee break at 10:30
- After lunch, check out variant calling.
- Coffee break at 2:30
- Some open time, if possible
  - Starting your own cloud machine (costs $$, but: freedom!).

Strategy

- Run stuff!
- Talk while it's running.
- Ask questions whenever!

Technology!

- Stickies
- Minute cards
- …Dropbox?

Etherpad?

Why... the cloud?

- Rental computers for small and BIG problems.
- Completely reproducible; independent of institution; so I can write tutorials!
- Once you get something working in the cloud, your local sysadmins can often help you get it running at your institution. If not, well, you can always pay $$.
- (How much? est $150 compute/$1000 mRNAseq sample)

The challenges of non-model transcriptomics

- Missing or low quality genome reference.
- Evolutionarily distant.
- Most extant computational tools focus on model organisms –
  - Assume low polymorphism (internal variation)
  - Assume reference genome
  - Assume somewhat reliable functional annotation
  - More significant compute infrastructure

…and cannot easily or directly be used on critters of interest.

The problem of lamprey...

- Diverged at base of vertebrates; evolutionarily distant from model organisms.
- Large, complicated genome (~2 GB)
- Relatively little existing sequence.
- We sequenced the liver genome…

Assembly

It was the best of times, it was the wor

, it was the worst of times, it was the

isdom, it was the age of foolishness

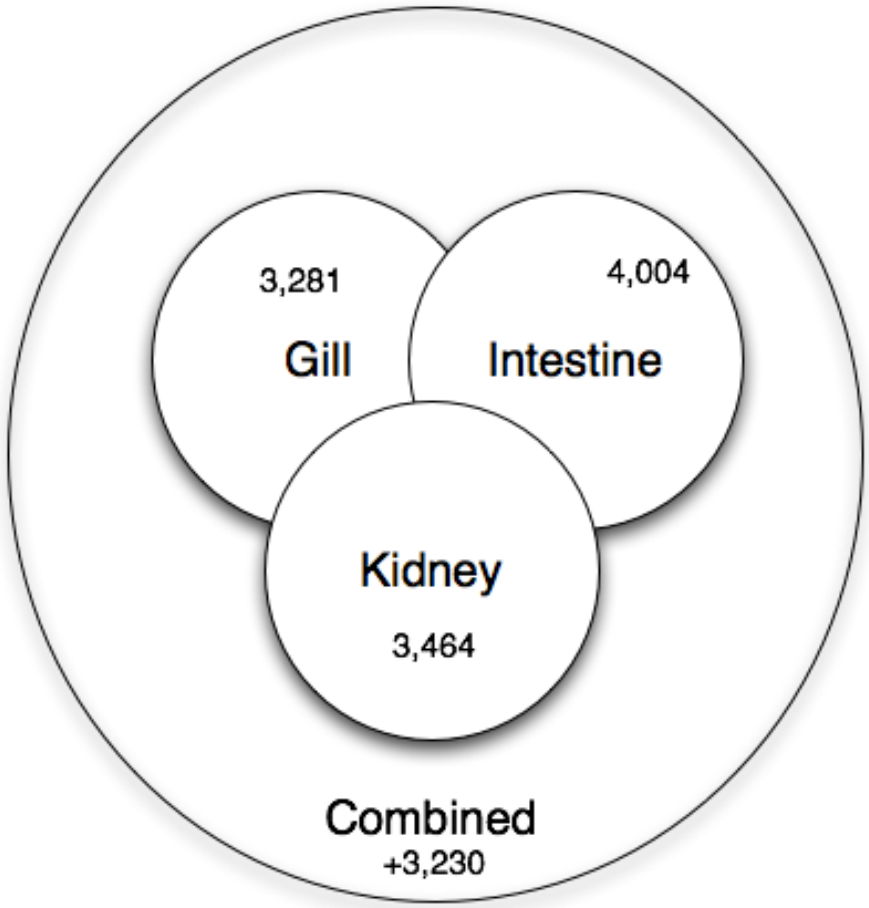
mes, it was the age of wisdom, it was th

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

…but for lots and lots of fragments!

Co-assembly is important for sensitivity

Shared low-level transcripts may not reach the threshold for assembly.



Two problems:

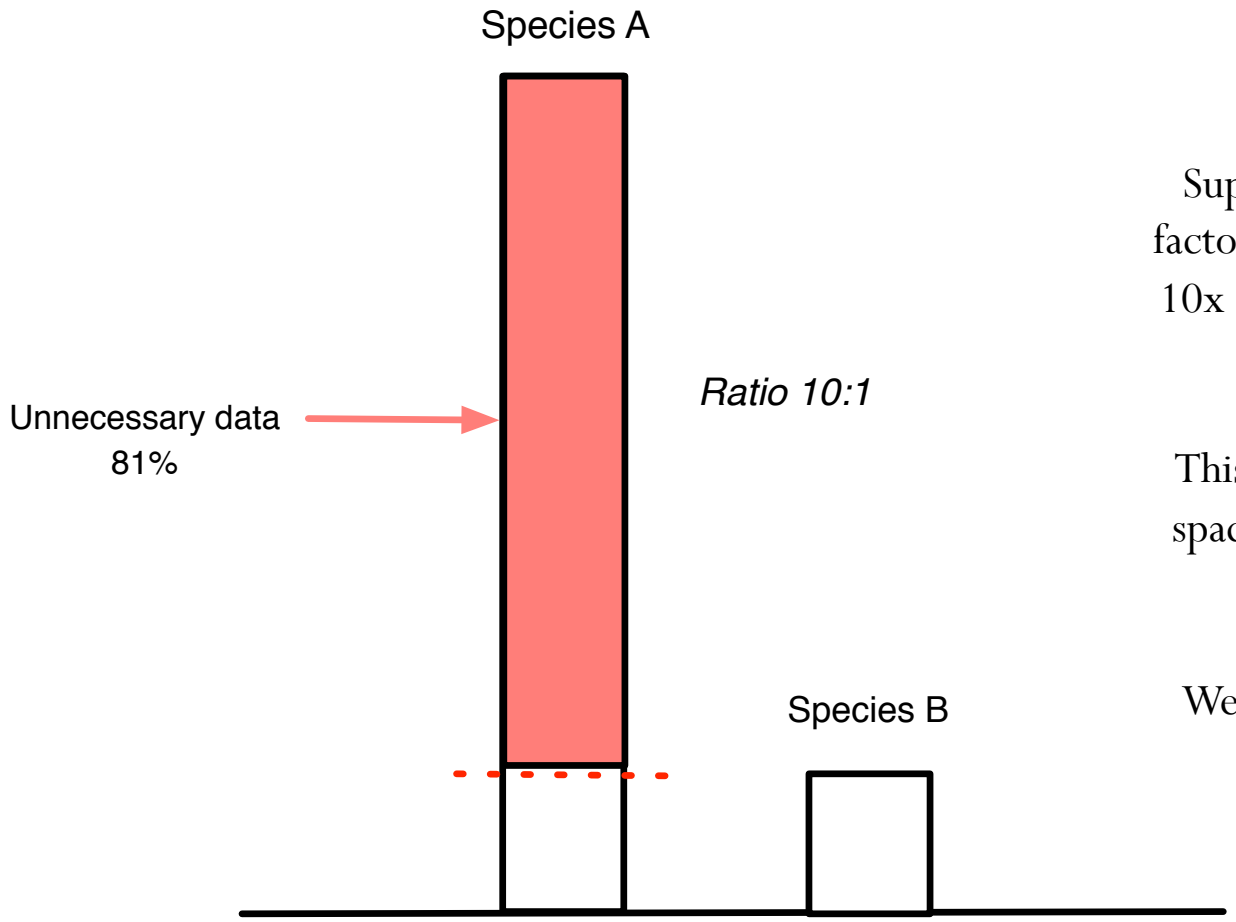
- We want to assemble a lot of stuff together.

- We need to construct transcript families (to collapse isoforms) without having a reference genome.

Diginorm

Solution: Digital normalization

(a computational version of library normalization)



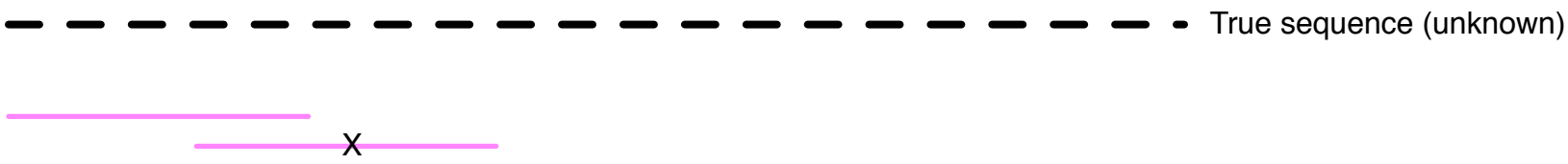
Suppose you have a dilution factor of A (10) to B(1). To get 10x of B you need to get 100x of A! Overkill!!

This 100x will consume disk space and, because of errors, **memory**.

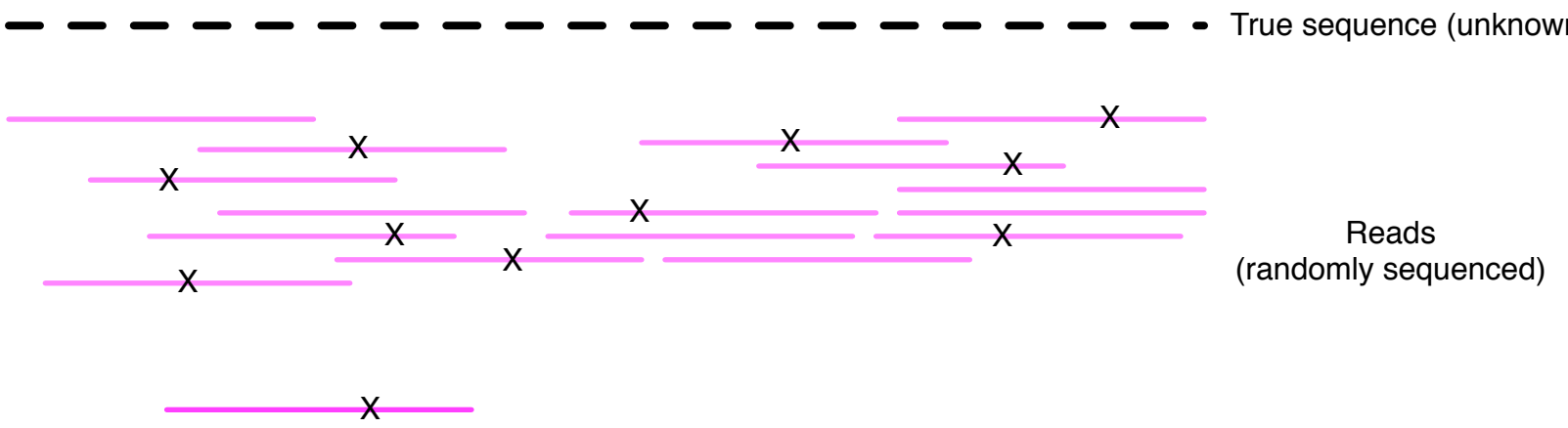
We can discard it for you…

Digital normalization

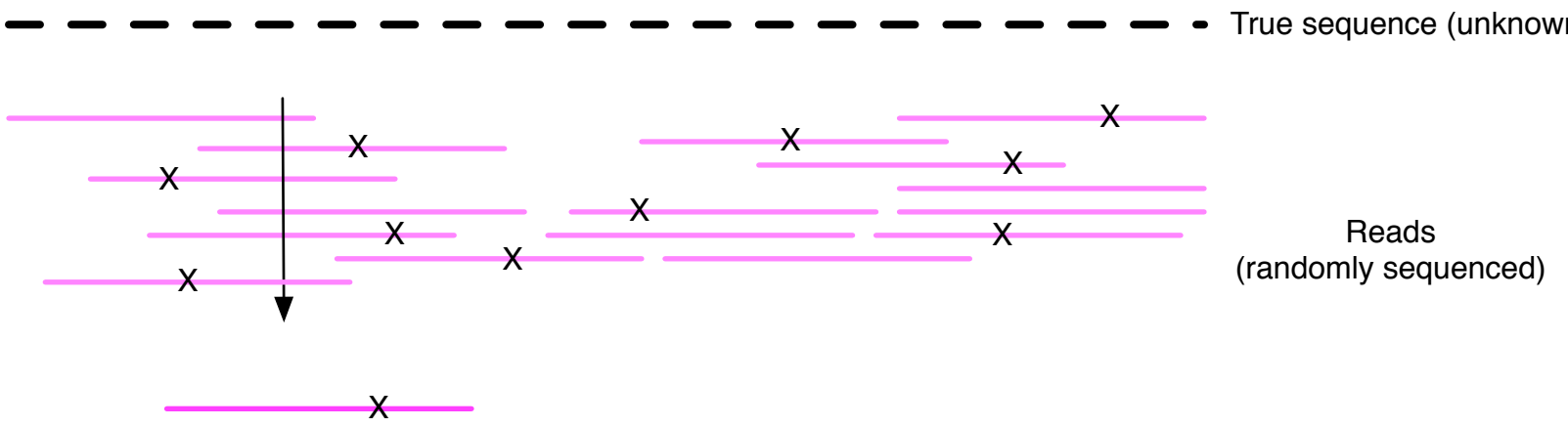
Digital normalization



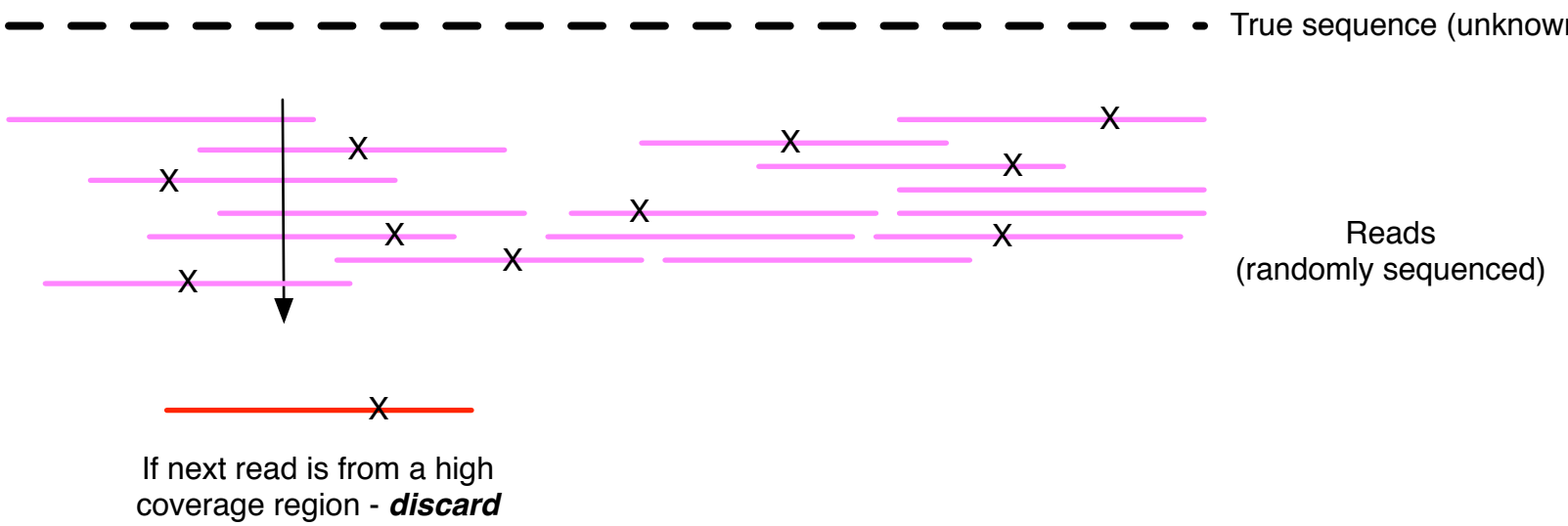
Digital normalization



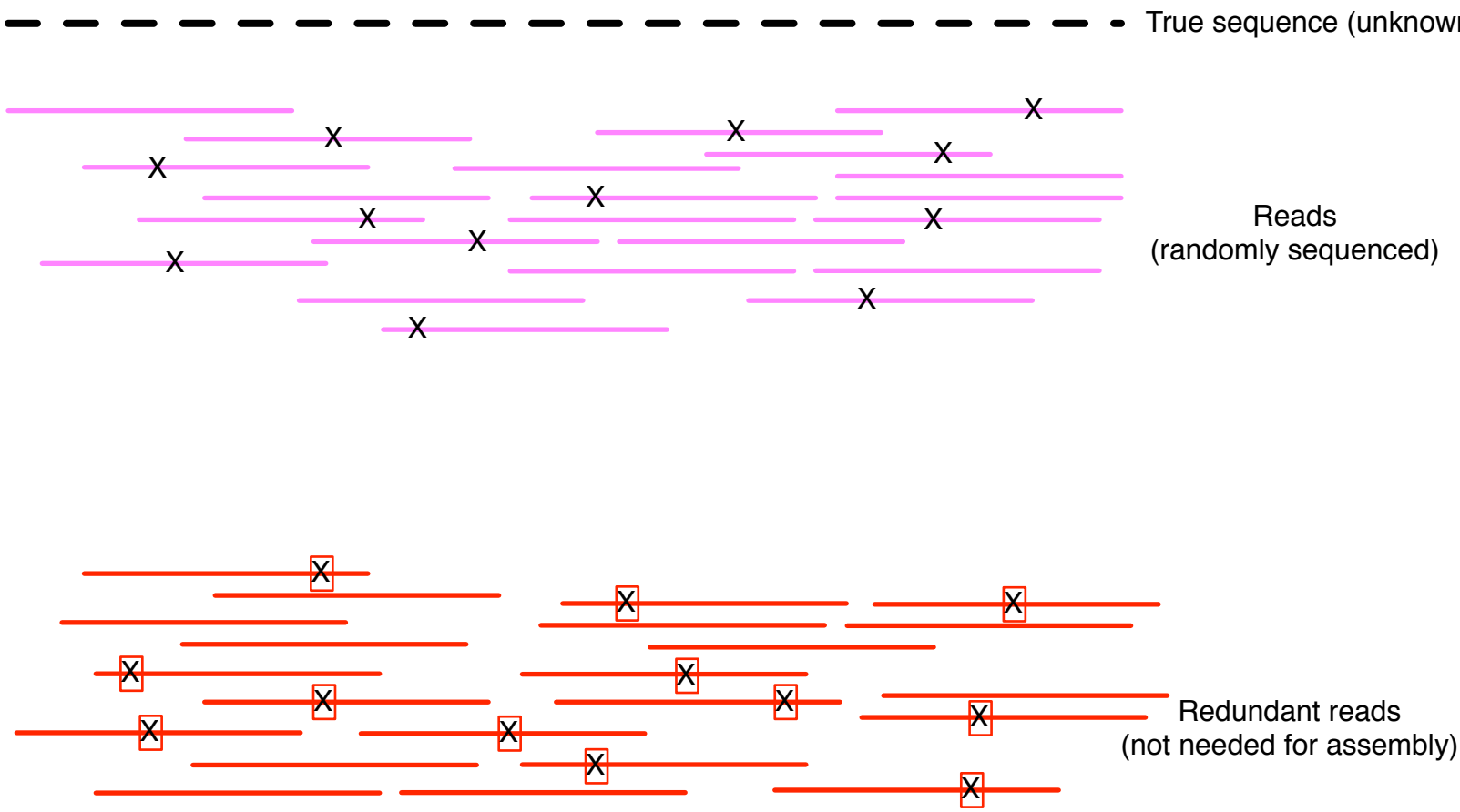
Digital normalization



Digital normalization



Digital normalization



Digital normalization approach

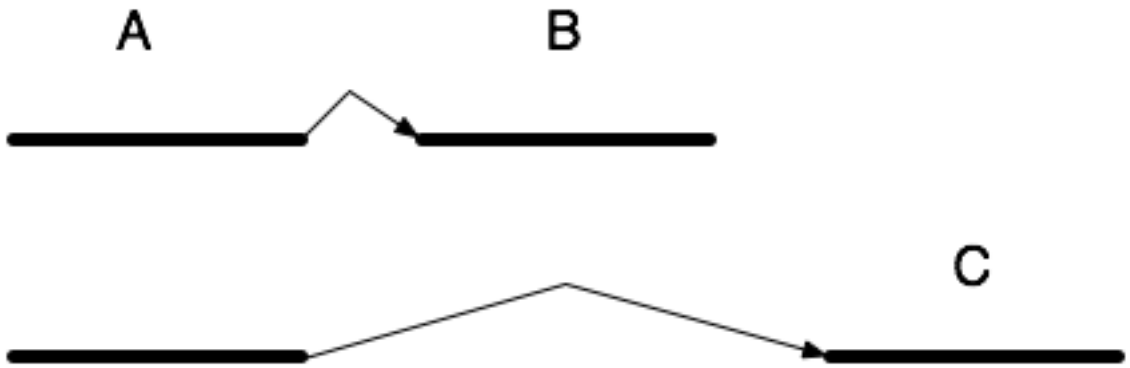
A *digital* analog to cDNA library normalization, diginorm:

- Is single pass: looks at each read only once;
- Does not "collect" the majority of errors;
- Keeps all low-coverage reads;
- **Smooths out coverage of regions.**

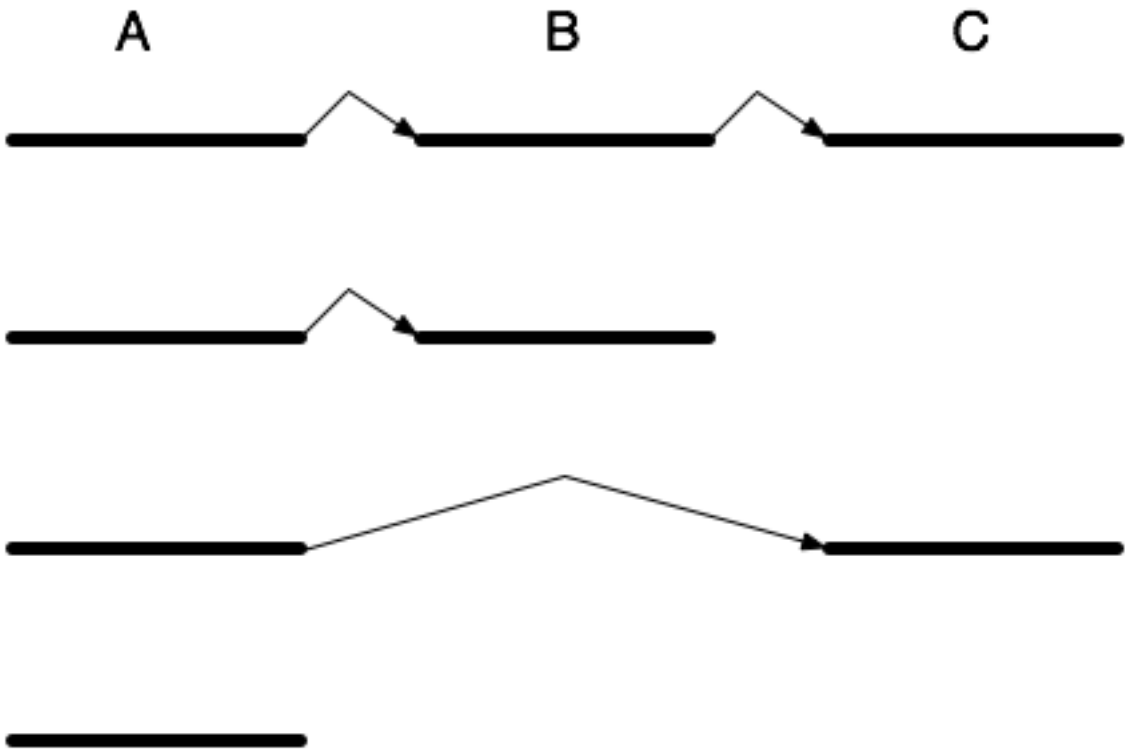
=> Enables analyses that are otherwise completely impossible.

Partitioning transcripts into families based on overlap

Isoform analysis – some easy...

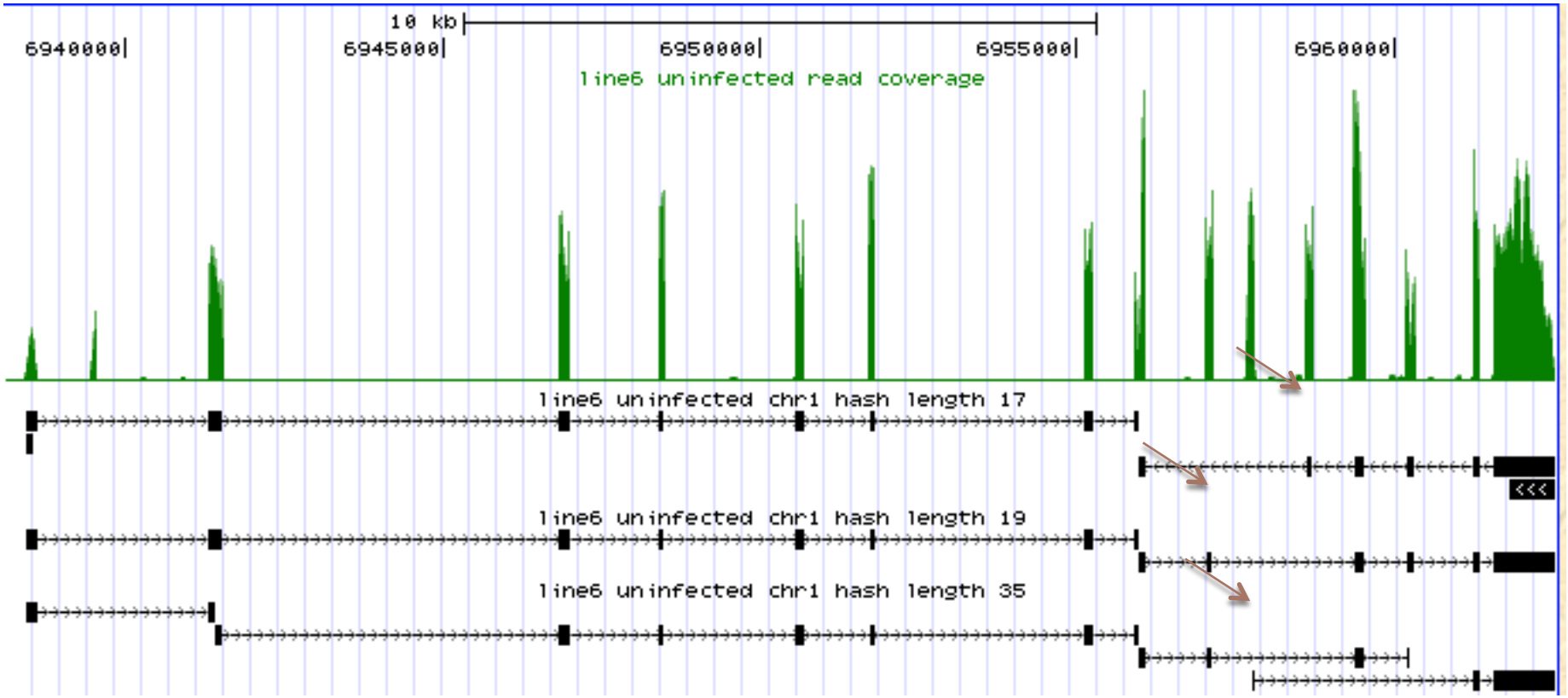


Isoform analysis – some hard

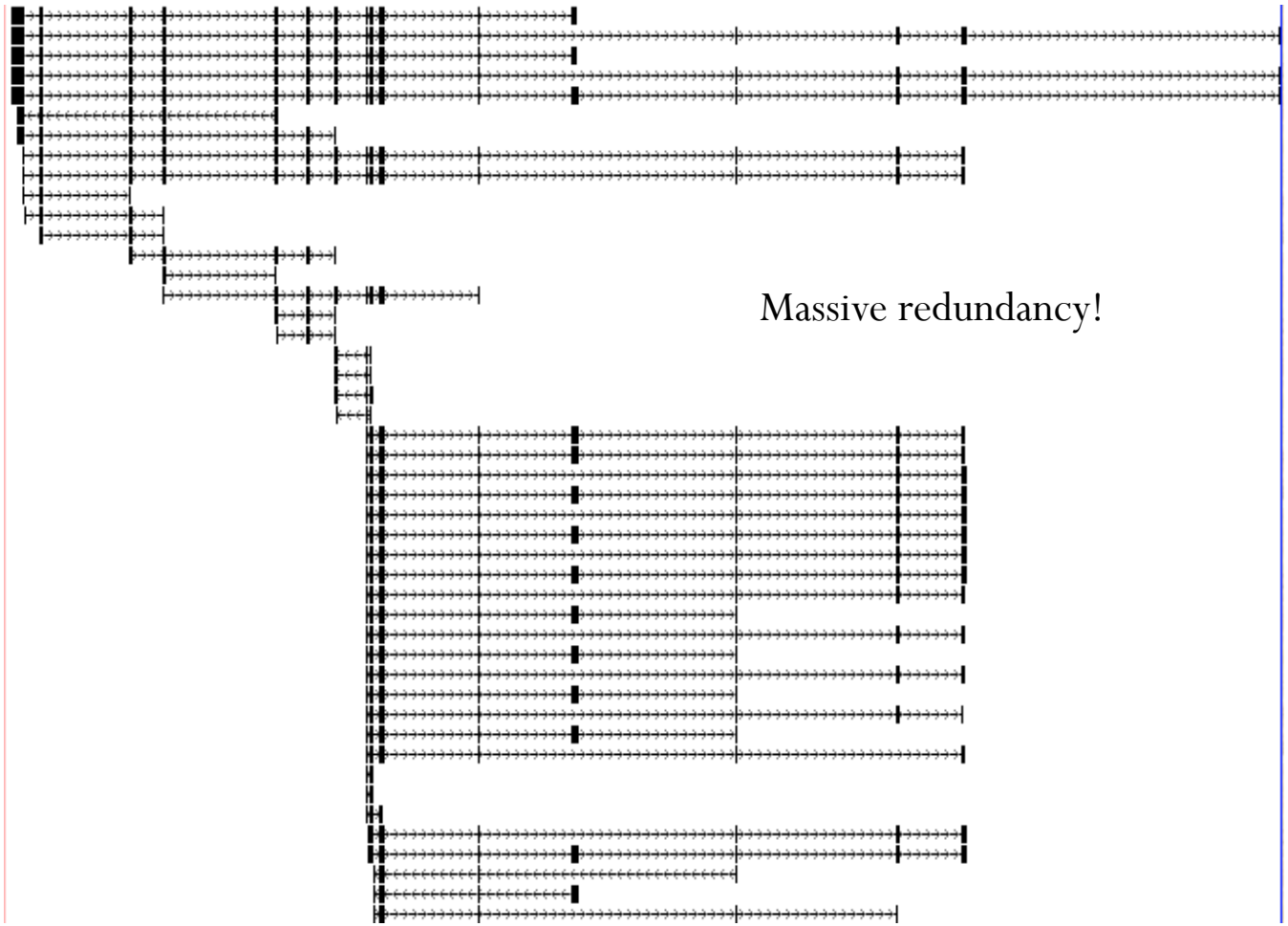


Counting methods mostly rely on presence of *unique sequence* to which to map.

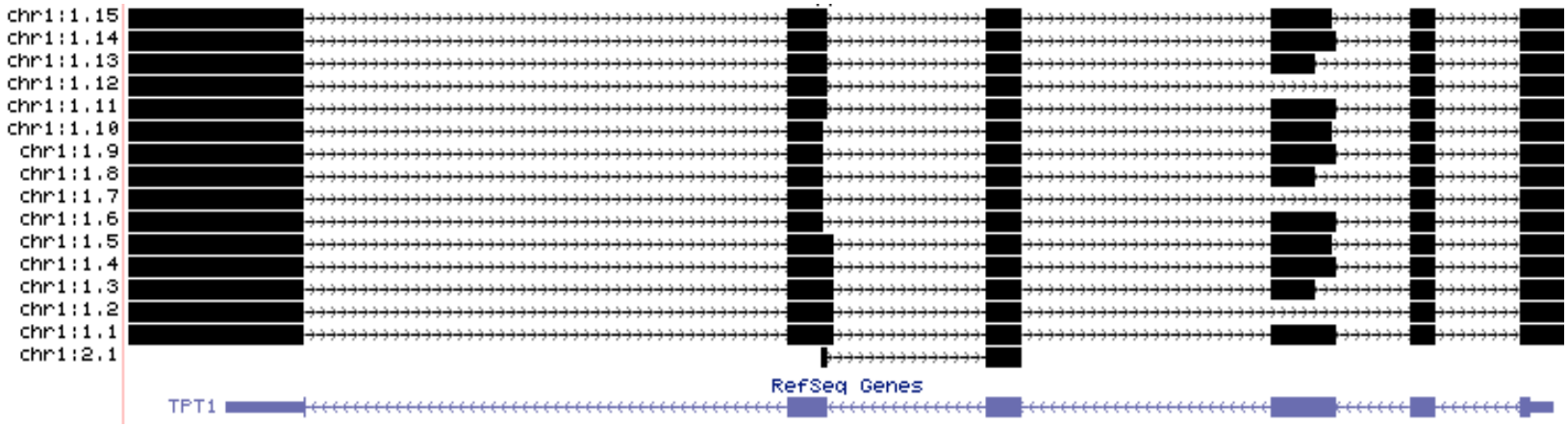
Exons are easy to locate, given genomic sequence



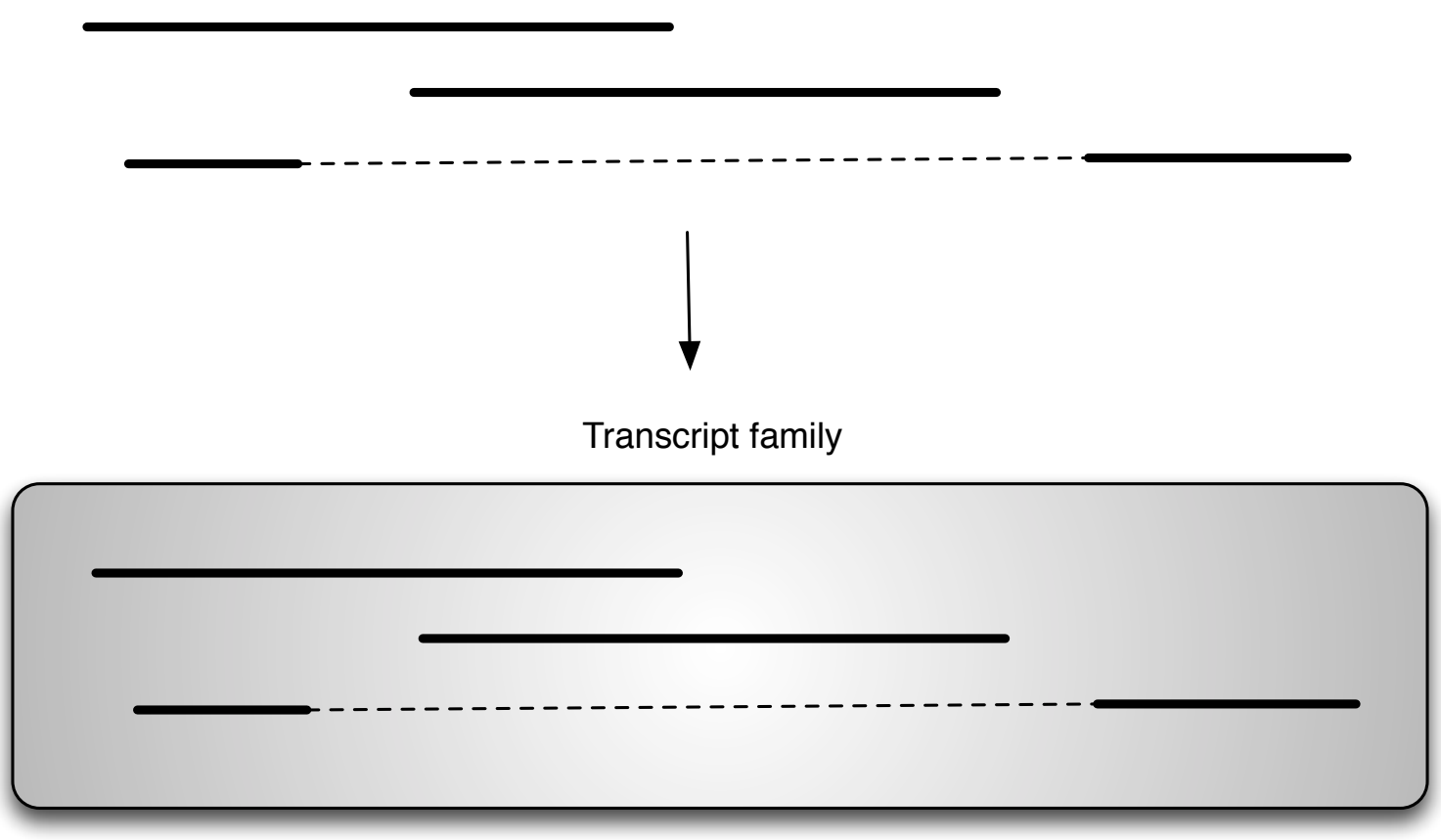
Genome-reference-free assembly leads to many isoforms.



Gene models can be “collapsed” given genomic sequence… But don’t always have.



Solution: Partitioning transcripts into "transcript families"



Pell et al., 2012, PNAS