

BUS format

Caltech Bioinformatics Symposium
February 14, 2019

The wild west of single-cell RNA-seq



- **The good (biology)**

- New insights into cell types and heterogeneous tissue
- Avoid Simpson's paradox



- **The bad (analysis)**

- Sampling is sparse and non-uniform
- Geometry and statistics in high dimension



- **The ugly (informatics)**

- Complex protocols -> complicated bioinformatics
- Technology is in flux, software is a mess



The informatics challenges

- Many technologies: 10x Genomics, Cel-seq2, Drop-seq, inDrops, SureCell, SCRB-seq, etc.
- Technologies are changing rapidly: 10x v1, v2, v3 chemistry, Cel-seq v1, v2, etc.
- Increasingly complex assays: cite-seq/REAP-seq, multiplexing, etc.
- Workflows require numerous software programs: CellRanger, STAR, Seurat, velocity, etc.
- Numerous languages involved: C/C++, R, python, etc.
- Datasets are large: one single-cell RNA-seq dataset is about 10 times larger than a bulk RNA-seq dataset

Existing workflows



barcode error correction

- correct sequencing errors in barcodes

read alignment

- align reads to a reference genome

UMI error correction

- correct sequencing errors in UMIs

cell assignment

- decide which reads are associated with which barcodes

cell filtering

- remove barcodes that correspond to failed cells

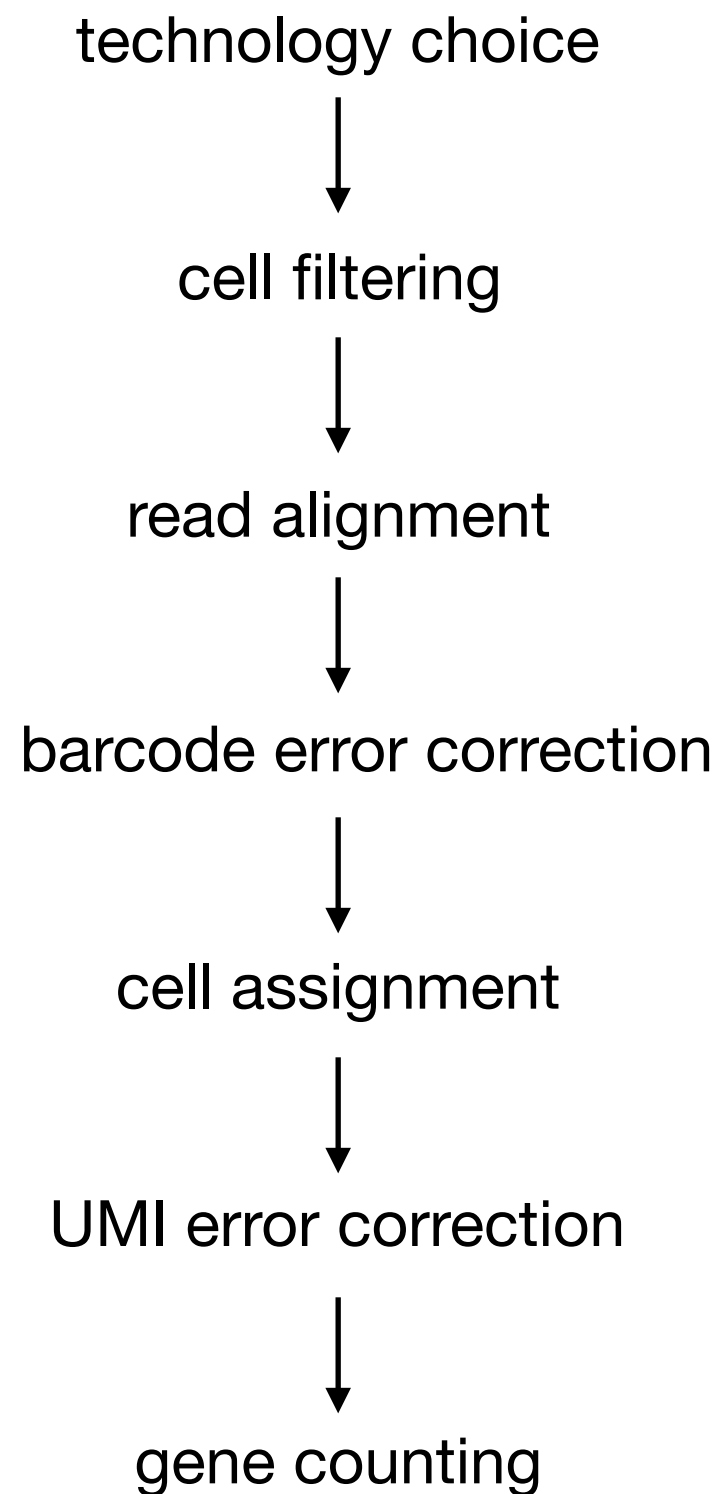
technology choice

- determine how to extract information from reads

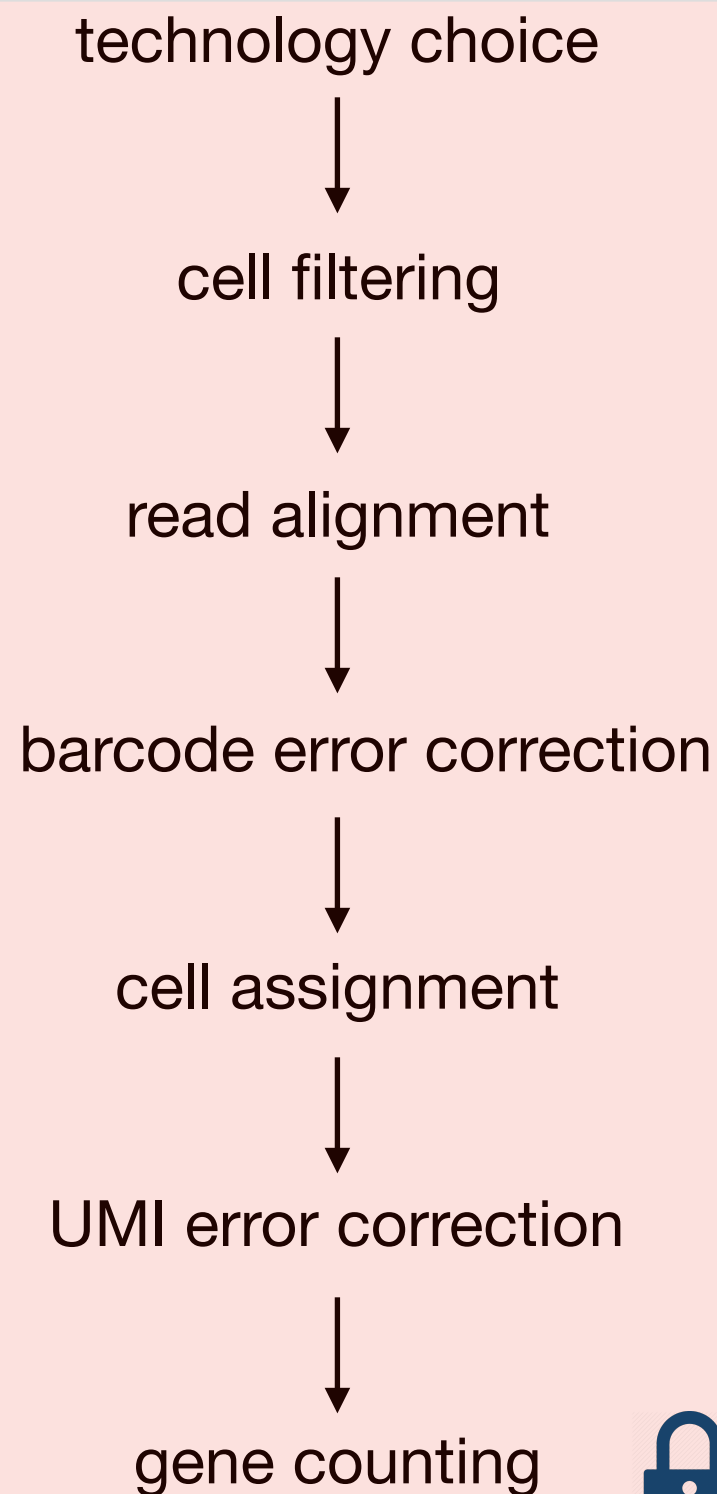
gene counting

- produce cell x gene matrix of read counts

Existing workflows



Existing workflows

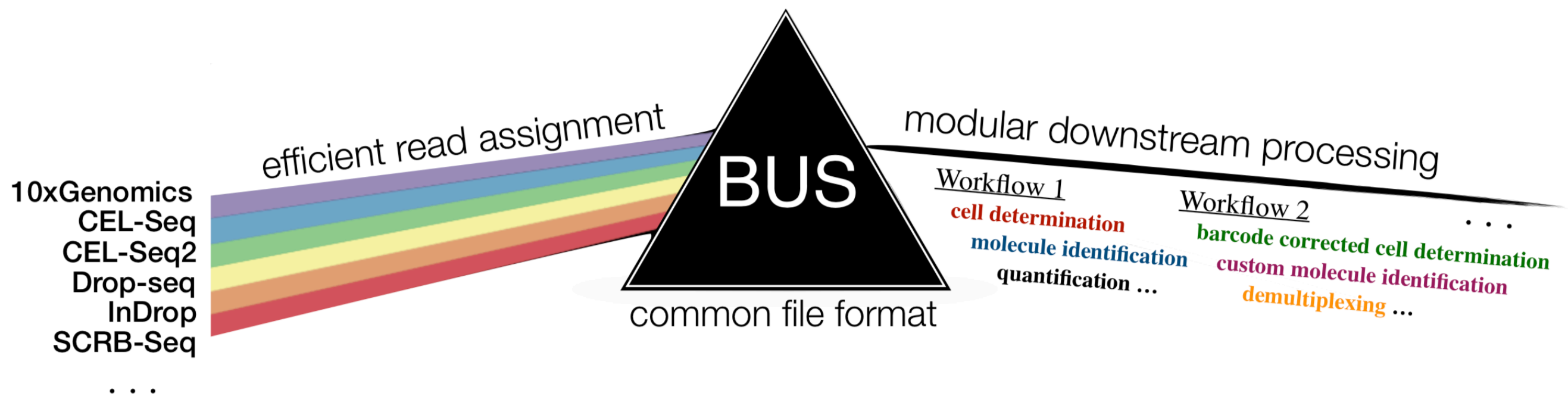


Godot: go out (and) dance our trash (takes forever)

- **Godot** requires a ton of memory
- **Godot** will take a day to run
- **Godot** requires a server
- Want to analyze a different kind of experiment? LOL!
- BUT....
- **Godot** is open source!!

Proposal

- A new format which decouples technology dependencies from algorithm choices.



- We call this format **B**arcode, **U**MI, **S**et (BUS) format.

Common structure to data

V2 library:

[illegible]

V3 library:

Diagram illustrating the structure of a sequencing read, showing the alignment of various components:

- 5' -** AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC (Illumina P5)
- 3' -** TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTACA (Illumina P7)
- Truseq Read 1** (16 bp)
- cell barcode** (16 bp)
- UMI** (12 bp)
- cDNA**
- Truseq Read 2** (8 bp)
- Sample Index** (8 bp)

10x v2 and v3

The diagram illustrates a sequencing read layout with the following components and labels:

- Read Sequence:** 5' - AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTACNNNNNNTAGCCATCGCATGCNNNNNNACCTCTGAGCTGAANNNNNNCGNNNNNNNNGA(dT)VXXX...XXXCGTCTCTTATACACATCTCCGAGCCACAGAGACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTGTTACTATGCCGCTGGTGGCTCTAGATGTGCGGACAGGCGCCTTCGTCAACCATAGTTGCGTCTCATGNNNNNNATCGGTAGCGTACGNNNNNNTTGGAGACTCGATTNNNNNNGCNNNNNNNNCT(pA)BXXX...XXXGCAGAGAATATGTGTAGAGGCTCGGGTGCTCTGNNNNNNNNTAGAGCATACGGCAGAACACGAAC - 5'
- Regions and Labels:**
 - Barcode 1:** Indicated by a pink box around the first 'NNNNNN' region.
 - Barcode 2:** Indicated by a pink box around the second 'NNNNNN' region.
 - Barcode 3:** Indicated by a pink box around the third 'NNNNNN' region.
 - UMI:** Indicated by a blue box around the 'CGNNNNNNNNGA' region.
 - cDNA:** Indicated by a black box around the '(dT)VXXX...XXXCGTCTCTTATACACATCT' region.
 - ME:** Indicated by a black box around the 'CCGAGCCACAGAGAC' region.
 - s7:** Indicated by a red box around the 'NNNNNNNNATCTCGTATGCCGTCTTCTGCTTG' region.
 - 8bp:** Indicated by a blue box around the 'TTGGAGACTCGATTNNNNNN' region.
 - sample index:** Indicated by a black box around the 'TAGAGCATACGGCAGAACACGAAC' region.
- Platform Labels:**
 - Illumina P5:** Located below the first part of the sequence.
 - Illumina P7:** Located below the last part of the sequence.

SureCell

5' - AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTACCTNNNNNNNNNNNNNNNNNNNN(dT)XXX...XXXTGTCTCTTATACACATCTCCGAGCCACGAGACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGCGGACAGGCGCC TTCGTACCATAGTTGCGTCTCATGANNNNNNNNNNNNNNNNNNNN(pA)XXX...XXGXACAGAGAATATGTGTAGAGGCTCGGGTGCTCTGNNNNNNNNNTAGAGCATACGGCAGAAGACGAAC -5'

Illumina P5 ISPCR/TSO 12bp cell barcode 8bp UMI cDNA ME s7 i7 Illumina P7

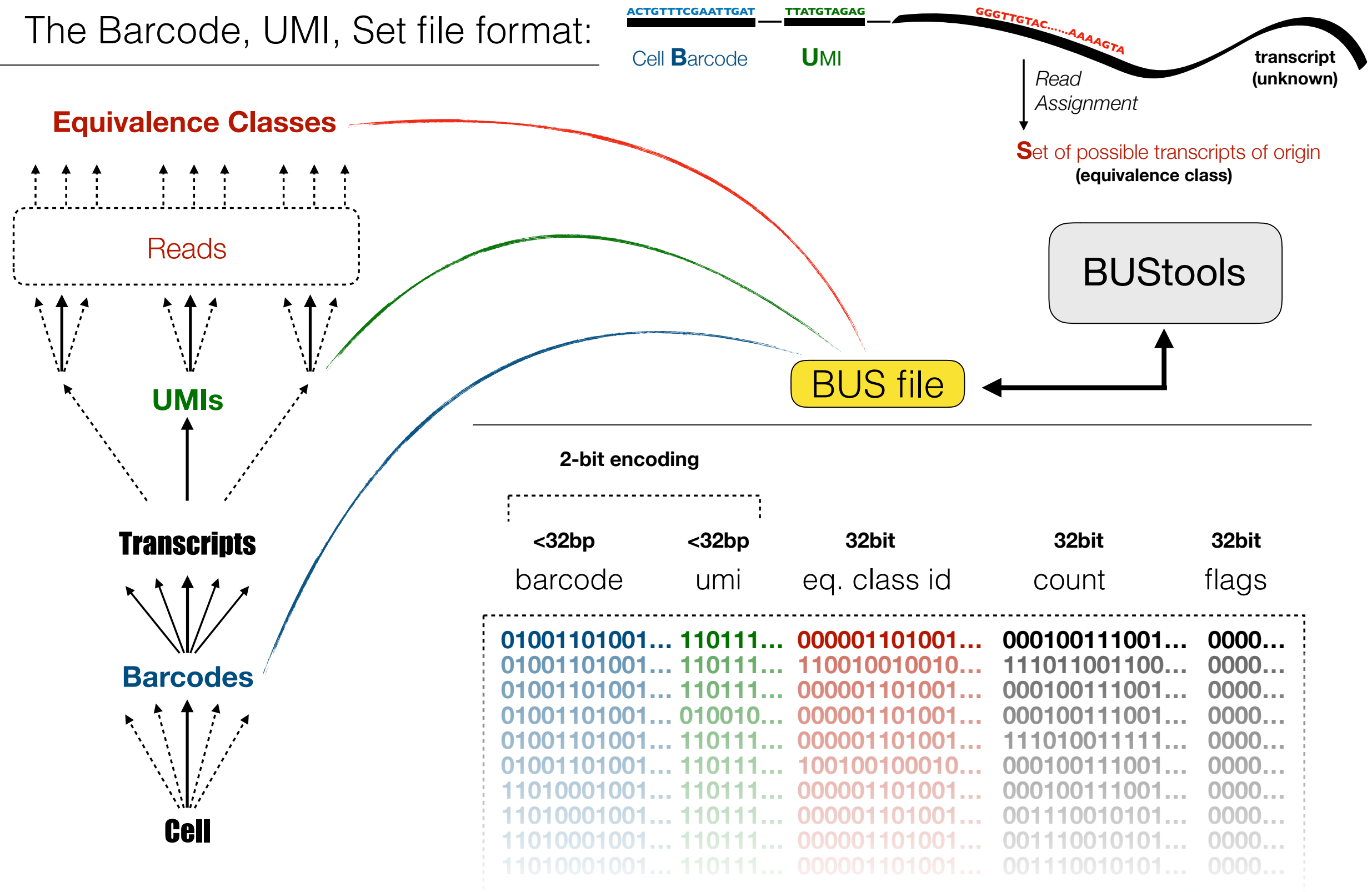
Drop-seq

5' – AATGATACGGCGACCACCGAGATCTACACGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCTXXX...XXXV(pA)NNNNNNNNNNNNNNNGASTGATTGCTTGTGACGCCTTNN...NNGATCGGAAGAGCGTCGTGTAGGGAAAGAGNNNNNNATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGCGAGAGCCGTAAGGACGACTTGGCGAGAAGGCTAGAXXX...XXXV(dT)NNNNNNNNNNNNNNNCTACTAACGAACACTGCGGAANN...NNCTAGCCTTCTCGCAGCACATCCCTTTCTNNNNNNNTAGAGCATACGGCAGAAGACGAAC –5'

Illumina P5 PE2 cDNA 6bp UMI 8bp barcode2 W1 barcode1 PE1 6bp sample index Illumina P7

InDrops

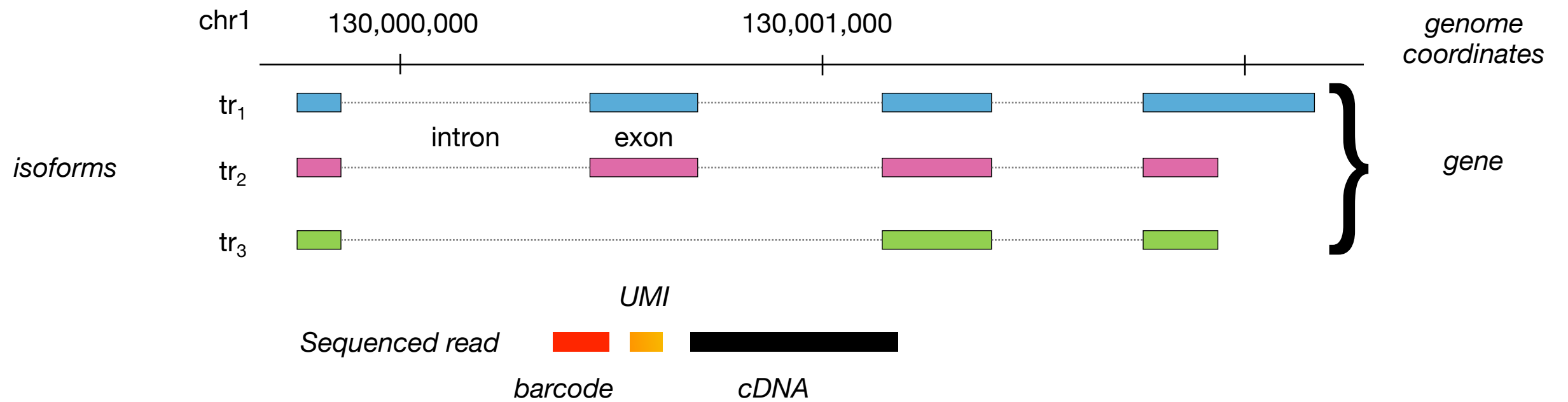
The Barcode, UMI, Set file format:



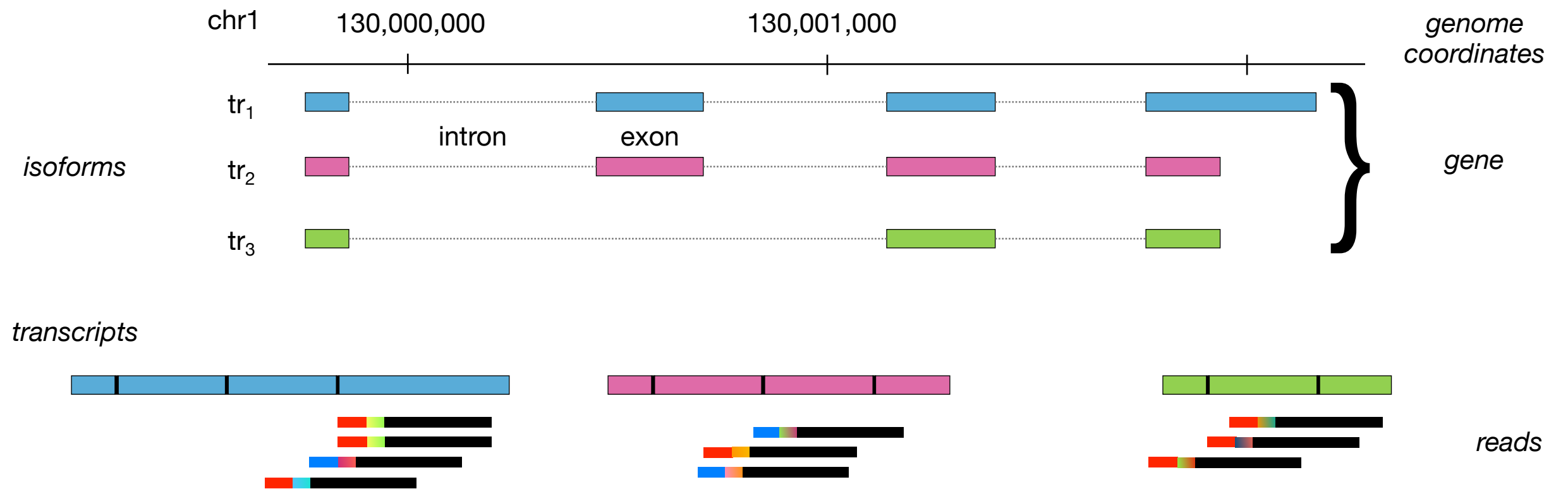
BUS centered workflow

- BUS can be generated with **kallisto** (Bray et al. 2016)
 - kallisto is fast: no sorting or alignment is required
 - kallisto streams bus records directly to disk, no memory overhead
 - Easy to process all technologies. kallisto already supports 10x v1,v2 and v3 chemistry, Drop-seq, inDrops, SureCell, etc.
- **BUStools** can be used for generic processing of BUS files
- Downstream processing **notebooks** in Python and R

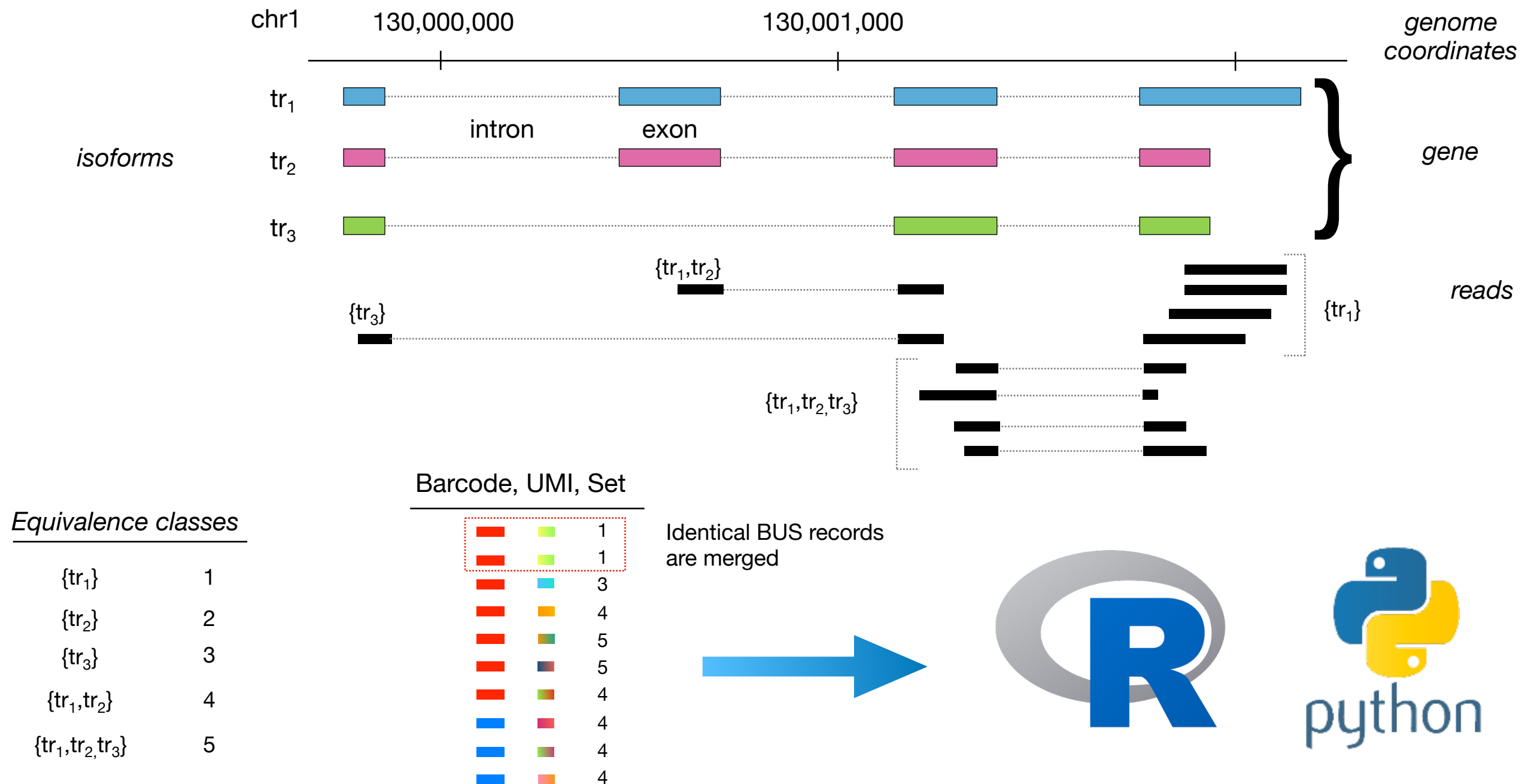
Example



Example



Example



BUS notebook review

- Download data
- Download reference transcriptome
- Build kallisto index
- Run kallisto bus
- Sort the bus file and convert to text
- Parse bus file in python
- Collate counts to make cell x gene counts matrix
- Analyze data...

BUS notebook review

- Run kallisto bus

```
kallisto bus -i index -o output R1.fastq R2.fastq
```

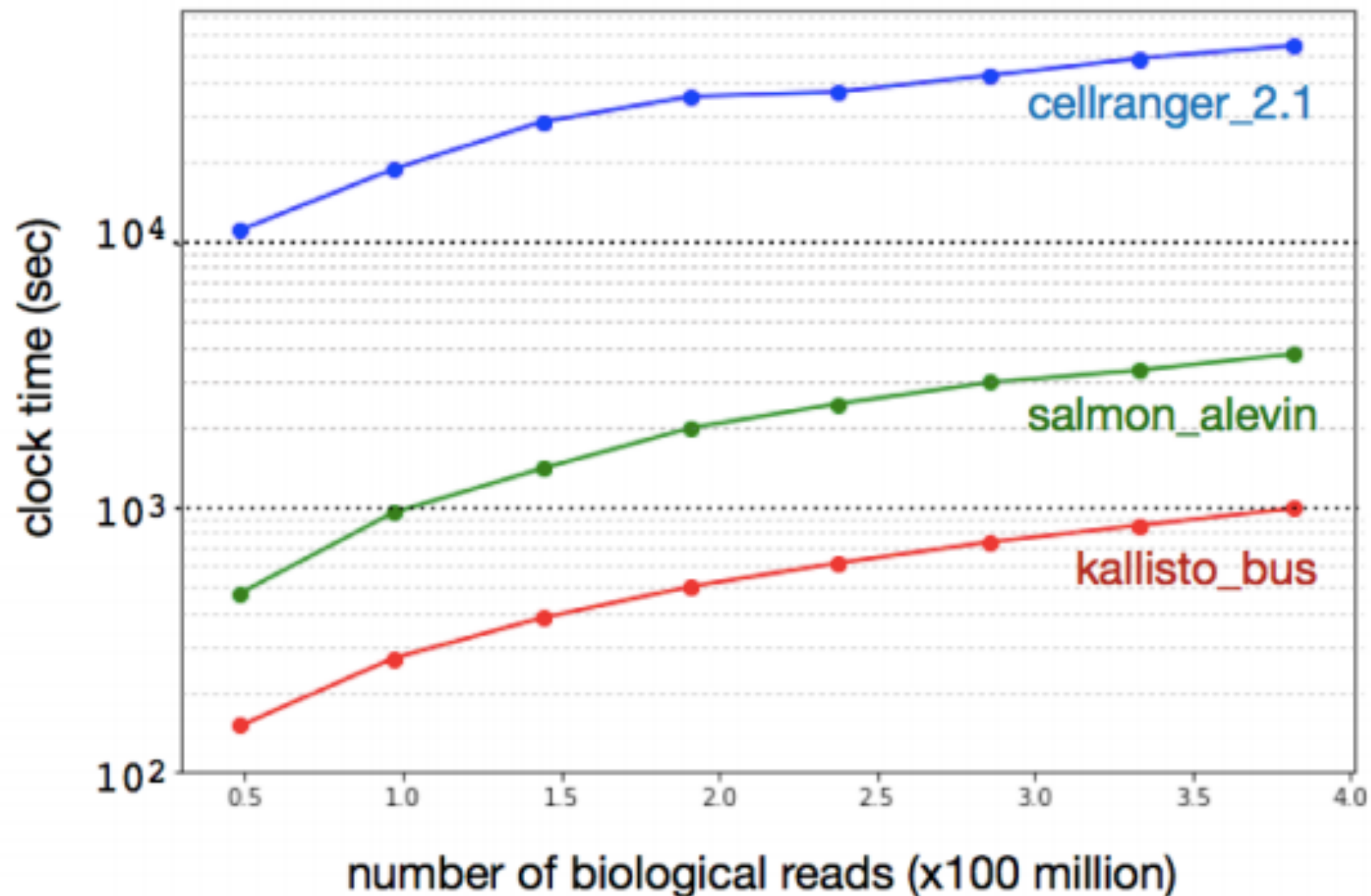
- Sort the bus file and convert to text

```
bustools sort -o output.sorted.bus output/output.bus  
bustools text -o output.sorted.txt output.sorted.bus
```

- Parse bus file in python
- Collate counts to make cell x gene counts matrix
- Analyze data...
 - Provided in notebooks

In practice...

- Running time for 350M reads, 8 threads: 10 minutes of **kallisto**



Current and future work

- Better algorithms for barcode and UMI correction
- Standardized workflows for popular technologies and assays
- RNA velocity workflows
- Compression of BUS format
- Large-scale processing of publicly available single-cell RNA-seq

