General workflow single cell is similar to bulk RNASeq

Barcoding is more complex -> extra step in analysis
You have more fastq files -> loss of quality control steps in analysis
Library prep for single cell expression analysis

gel beads with bar codes + cells + oil

Dilution of cells => max 1 cell per droplet

bead + 1 cell in droplet oil

Cell lysis
Beads dissolve
Beads release bar codes
RT

https://www.10xgenomics.com/solutions/single-cell/
Barcoding is more complex

Each sample has a unique barcode

Each cell has a unique barcode

Each transcript has a unique barcode -> easy to detect duplicates
Barcodes for single cell sequencing

Barcode on the bead < sequencing adapter + primer

14 bp cell barcode
10 bp UMI
30 bp oligo(dT) -> initiate RT

Bias of fragments from 3'UTR of transcripts
Library prep for single cell expression analysis

PCR amplification with primers complementary to **barcode adapter**

Shearing

Ligation of adapters containing a **sample barcode**

Paired end sequencing with primers complementary to **adapters**
Library prep for single cell sequencing

Multiple files per sample

read1 and read2 are split into individual components

- read1
- read2

- Read < 98nt transcript sequences
- Index2 < 8nt sample barcodes

- Read < 10nt UMIs
- Index1 < 14nt cell barcodes

Recommended sequencing depth: 60,000 reads/cell
UMIs allow to detect technical duplicates
What coverage do you need?

Recommended sequencing depth: 60000 reads/cell

For cell-type discovery: more cells is better than higher coverage

Jaitin et al., 2014
What you get from your provider

Illumina sequencer's base call files .bcl

Binary representation of base calls and quality scores
Software for single cell RNASeq analysis

Cell ranger:  10xgenomics own pipeline

> cellranger mkfastq

Transform bcl into **multiple fastq files**:
Index file:
I1 sample indexes
Read files:
R1 cell barcodes + UMIs
R2 files contain reads

Generate quality metrics file
Parameters of cellranger mkfastq

--run path to folder with bcl files
--id name for output folder
--samplesheet path to Illumina file with sample information OR
--csv path to simple file with sample information
--localcores number of cores the command can use

16
Example of a simple sample sheet

<table>
<thead>
<tr>
<th>Lane</th>
<th>Sample</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D1</td>
<td>SI-3A-A1</td>
</tr>
<tr>
<td>1</td>
<td>D2</td>
<td>SI-3A-B1</td>
</tr>
<tr>
<td>1</td>
<td>M1</td>
<td>SI-3A-E1</td>
</tr>
<tr>
<td>1</td>
<td>M2</td>
<td>SI-3A-F1</td>
</tr>
<tr>
<td>1</td>
<td>M3</td>
<td>SI-3A-G1</td>
</tr>
<tr>
<td>1</td>
<td>M4</td>
<td>SI-3A-H1</td>
</tr>
<tr>
<td>2</td>
<td>D1</td>
<td>SI-3A-A1</td>
</tr>
<tr>
<td>2</td>
<td>D2</td>
<td>SI-3A-B1</td>
</tr>
<tr>
<td>2</td>
<td>M1</td>
<td>SI-3A-E1</td>
</tr>
</tbody>
</table>

...
Output of cellranger mkfastq

```
$ ls -l tiny-bcl/outs/fastq_path/
```

```
-rw-r--r-- 1 jdoe jdoe 20615106 Aug  9 12:26 Undetermined_S0_L001_I1_001.fastq.gz
-rw-r--r-- 1 jdoe jdoe 51499694 Aug  9 12:26 Undetermined_S0_L001_R1_001.fastq.gz
-rw-r--r-- 1 jdoe jdoe 152692701 Aug  9 12:26 Undetermined_S0_L001_R2_001.fastq.gz
```

```
$ tree tiny-bcl/outs/fastq_path/tiny_bcl/
```

```
Sample1
Sample1_S1_L001_I1_001.fastq.gz
Sample1_S1_L001_R1_001.fastq.gz
Sample1_S1_L001_R2_001.fastq.gz
```

- < sample indexes
- < cell barcodes + UMIs
- < reads
Output of cellranger mkfastq

<table>
<thead>
<tr>
<th>Example of I1 file: read-I1-si-ACGCGGAA-lane-001-chunk-001.fastq</th>
</tr>
</thead>
<tbody>
<tr>
<td>@D00547:633:HKMNVCXX:1:1101:1183:2064 2:N:0:0</td>
</tr>
<tr>
<td>AATCTCTGCTTTAC</td>
</tr>
<tr>
<td>@D00547:633:HKMNVCXX:1:1101:1121:2103 2:N:0:0</td>
</tr>
<tr>
<td>ACGCGAAGGCTTTGCTTT</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>AAAAGGGG.AAGGG</td>
</tr>
<tr>
<td>&lt;A&lt;&lt;GAAGG&lt;AGG</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example of I2 file: read-I2-si-ACGCGGAA-lane-001-chunk-001.fastq</th>
</tr>
</thead>
<tbody>
<tr>
<td>@D00547:633:HKMNVCXX:1:1101:1183:2064 3:N:0:0</td>
</tr>
<tr>
<td>ACGCGGAA</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>GGGGGGIIII</td>
</tr>
<tr>
<td>@D00547:633:HKMNVCXX:1:1101:1121:2103 3:N:0:0</td>
</tr>
<tr>
<td>ACGCGGAA</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>GAGGAGGG</td>
</tr>
</tbody>
</table>

ID  sequence cell barcode  quality
ID  sequence sample barcode  quality

All identical sequences because sorted per sample
Output of cellranger mkfastq

Example of R file: read-RA-si-ACGCGGAA-lane-001-chunk-001.fastq

Transcript sequence

UMI sequence
Sequencing errors in barcodes are corrected

Barcodes contain sequencing errors just like reads:
Barcodes in reads are compared to list of known 10x barcodes
Errors are corrected
Hamming distance to determine how much error is allowed
    = number of mismatches in end-to-end alignment
Cell ranger:    = 1

One mismatch allowed
Quality metrics of cellranger mkfastq

For every sample:

barcode_exact_match_ratio: 0.93  barcodes exactly matching known 10x barcode
barcode_q30_base_ratio: 0.96   barcode bases with quality > 30
bc_on_whitelist: 0.94          barcodes matching 10x barcode after error correction
mean_barcode_qscore: 37.77     mean quality score of barcode bases
number_reads: 2748155          # reads matching the sample's sample index
read1_q30_base_ratio: 0.89     read1 bases with quality > 30
read2_q30_base_ratio: 0.78     read2 bases with quality > 30
Software for single cell RNASeq analysis

> cellranger count

```
Align reads using STAR
Tag reads with gene, transcript hits
Correct UMI sequences
Count UMI(s) by (cell, gene)
Select cell-associated barcodes
```

- Link genes – reads
- Uniquely mapped reads only
- 1 Hamming distance
- Link genes - # transcripts
- Link cells -
- Per cell: #transcripts for each gene
Parameters of cellranger count

Run on each sample (I1, R1, R2)

--id   name of the directory where the output will be written

--fastqs   path to folder with fastq files

--sample   sample name (= name of fastqs without _S1_L001_R1_001.fastq)

--reference   path to folder that contains the reference

    /usr/summer/refs/sc/Homo...

--expect-cells   number of cell you think were analyzed

--localcores   number of cores the command can use

    16
### Output of cellranger count

#### Outputs:

- **Run summary HTML:** `/opt/sample345/outs/web_summary.html`
- **Run summary CSV:** `/opt/sample345/outs/metrics_summary.csv`
- **BAM:** `/opt/sample345/outs/possorted_genome_bam.bam`
- **BAM index:** `/opt/sample345/outs/possorted_genome_bam.bam.bai`
- **Filtered gene-barcode matrices MEX:** `/opt/sample345/outs/filtered_gene_bc_matrices`
- **Filtered gene-barcode matrices HDF5:** `/opt/sample345/outs/filtered_gene_bc_matrices_h5.h5`
- **Unfiltered gene-barcode matrices MEX:** `/opt/sample345/outs/raw_gene_bc_matrices`
- **Unfiltered gene-barcode matrices HDF5:** `/opt/sample345/outs/raw_gene_bc_matrices_h5.h5`
- **Secondary analysis output CSV:** `/opt/sample345/outs/analysis`
- **Per-molecule read information:** `/opt/sample345/outs/molecule_info.h5`
- **Loupe Cell Browser file:** `/opt/sample345/outs/loupe.loupe`

---

**Run summary metrics and charts in HTML or CSV format**

**Alignments + index file**

**Filtered and unfiltered gene-barcode matrices in MEX or HDF5 format**

*unfiltered: includes background and non-cellular barcodes*

**Info used by cellranger aggr to aggregate samples into larger datasets**

**Analysis data: dimensionality reduction, cell clustering, differential expression**

**Data for visualization and analysis in Loupe**
Aggregation of multiple samples with cellranger aggr

--id       name of the directory where the output will be written
--csv      CSV file with a header line containing the following columns:
            library_id: unique identifier for each sample
            molecule_h5: path to the molecule_info.h5 file produced by cellranger count

```
library_id,molecule_h5
LV123,/opt/runs/LV123/outs/molecule_info.h5
LB456,/opt/runs/LB456/outs/molecule_info.h5
LP789,/opt/runs/LP789/outs/molecule_info.h5
```

--normalize how to normalize depth across samples
          mapped: (default) subsample reads from higher-depth samples
          until they all have an equal number of mapped reads/cell
          raw:    subsample reads from higher-depth samples
          until they all have an equal number of total reads/cell
          none:   do not normalize
### Output of cellranger aggr

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregation metrics summary HTML</td>
<td><code>/home/jdoe/runs/AGG123/outs/web_summary.html</code></td>
</tr>
<tr>
<td>Aggregation metrics summary JSON</td>
<td><code>/home/jdoe/runs/AGG123/outs/summary.json</code></td>
</tr>
<tr>
<td>Secondary analysis output CSV</td>
<td><code>/home/jdoe/runs/AGG123/outs/analysis_csv</code></td>
</tr>
<tr>
<td>Filtered gene-barcode matrices HDF5</td>
<td><code>/home/jdoe/runs/AGG123/outs/filtered_gene_bc_matrices_h5.h5</code></td>
</tr>
<tr>
<td>Filtered gene-barcode matrices MEX</td>
<td><code>/home/jdoe/runs/AGG123/outs/filtered_gene_bc_matrices_mex</code></td>
</tr>
<tr>
<td>Filtered molecule-level info</td>
<td><code>/home/jdoe/runs/AGG123/outs/filtered_molecules.h5</code></td>
</tr>
<tr>
<td>Unfiltered gene-barcode matrices HDF5</td>
<td><code>/home/jdoe/runs/AGG123/outs/raw_gene_bc_matrices_h5.h5</code></td>
</tr>
<tr>
<td>Unfiltered gene-barcode matrices MEX</td>
<td><code>/home/jdoe/runs/AGG123/outs/raw_gene_bc_matrices_mex</code></td>
</tr>
<tr>
<td>Unfiltered molecule-level info</td>
<td><code>/home/jdoe/runs/AGG123/outs/raw_molecules.h5</code></td>
</tr>
<tr>
<td>Barcodes of cell-containing partitions</td>
<td><code>/home/jdoe/runs/AGG123/outs/cell_barcodes.csv</code></td>
</tr>
<tr>
<td>Copy of the input CSV</td>
<td><code>/home/jdoe/runs/AGG123/outs/aggregation_csv.csv</code></td>
</tr>
<tr>
<td>Loupe Cell Browser file</td>
<td><code>/home/jdoe/runs/AGG123/outs/cloupe.cloupe</code></td>
</tr>
</tbody>
</table>

**Aggregation summary metrics and charts in HTML or JSON format**

**Alignments + index file**

**Filtered and unfiltered gene-barcode matrices in MEX or HDF5 format**

  * unfiltered: includes background and non-cellular barcodes

**Analysis data: dimensionality reduction, cell clustering, differential expression**

**Data for visualization and analysis in Loupe**