UCDAVIS Bioinformatics Core

Single-cell transcriptomics

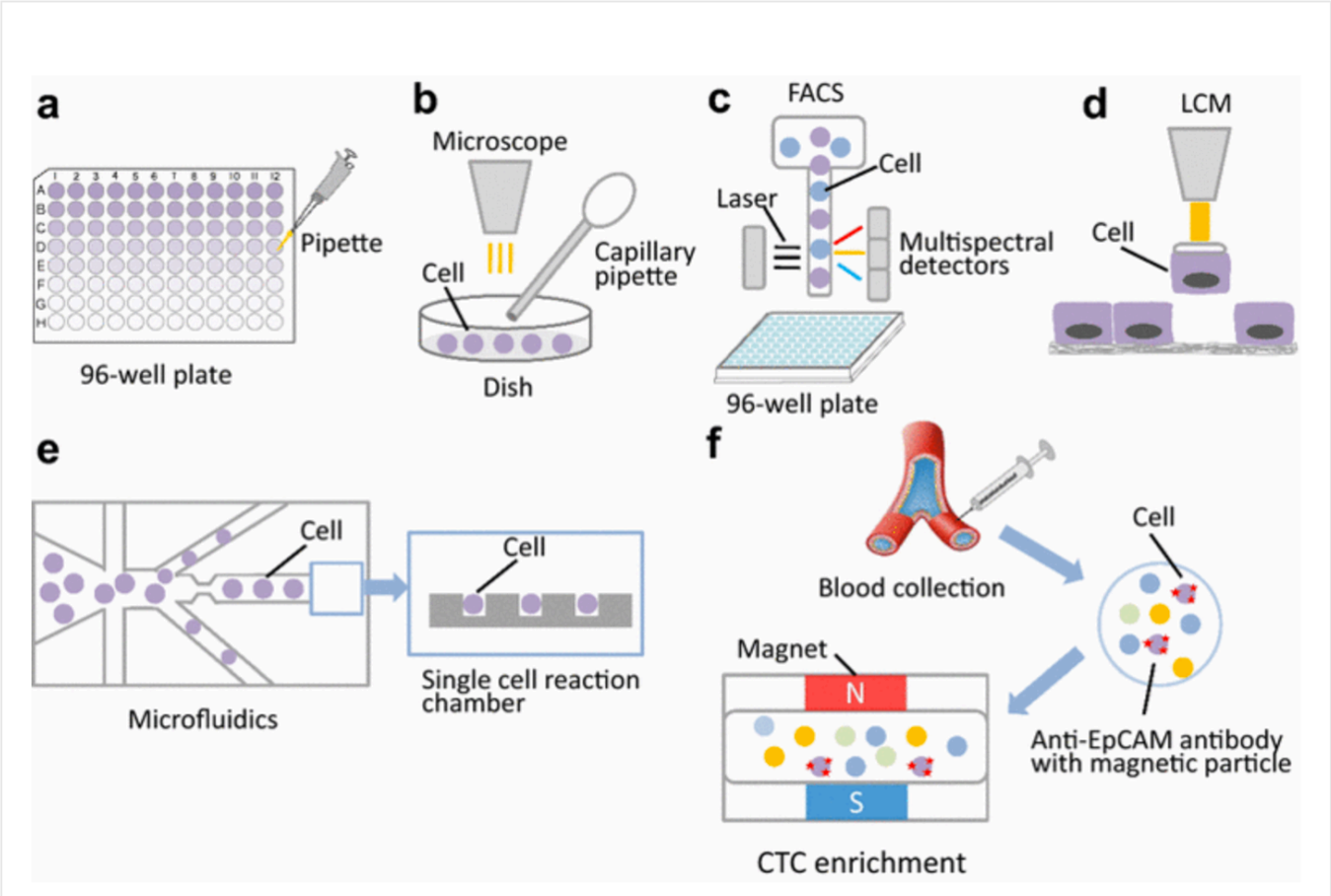
Dr. Matthew L. Settles

Genome Center
University of California, Davis
settles@ucdavis.edu

Sponsored by Lexogen

The sequencing of the transcriptomes of single-cells, or single-cell RNA-sequencing, has now become the dominant technology for the identification of novel cell types and for the study of stochastic gene expression.

Single-cell isolation methods



Single cell sequencing: a distinct new field
Jian Wang and Yuanlin Song

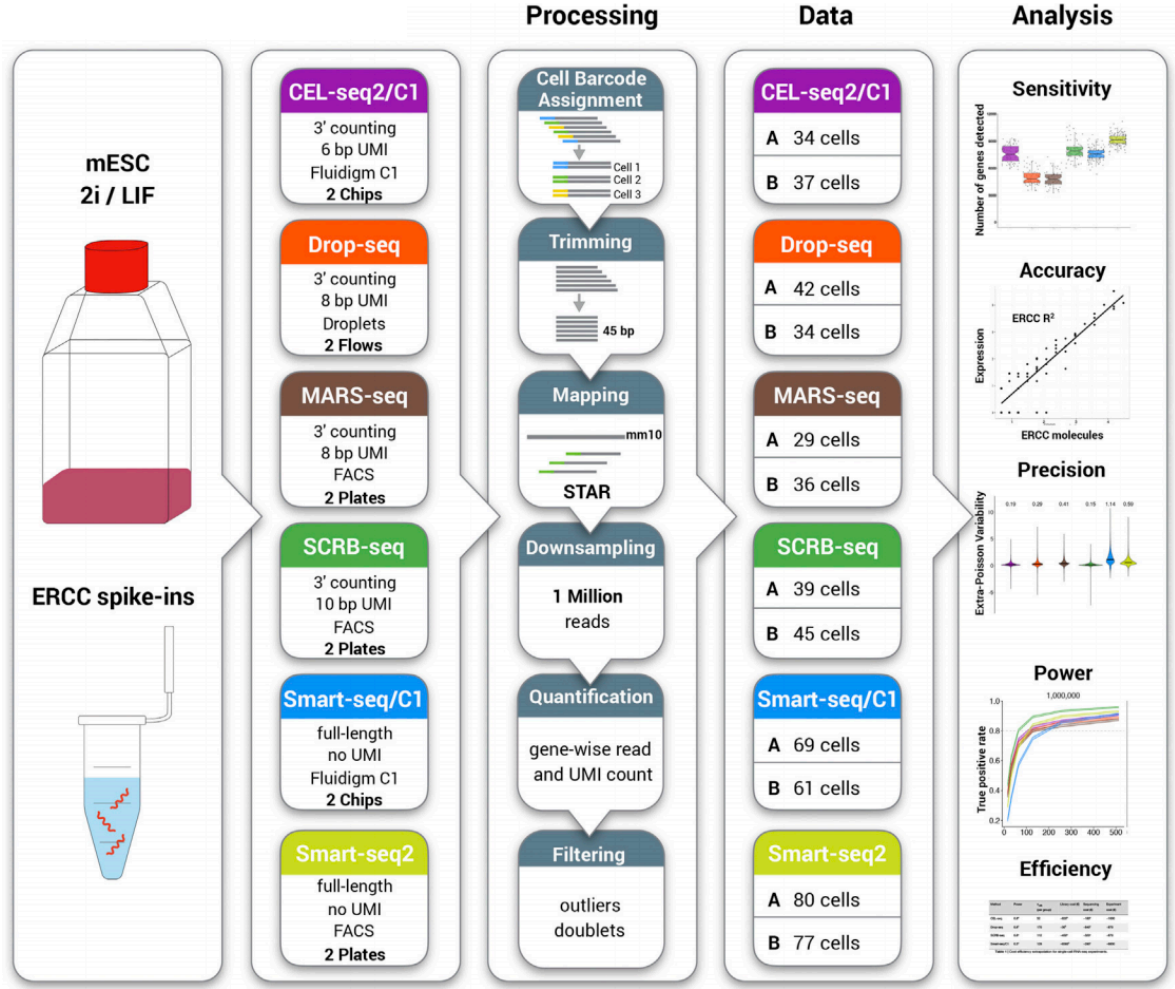
UCDAVIS Bioinformatics Core

Molecular Cell

Comparative Analysis of Single-Cell RNA Sequencing Methods

Authors

Christoph Ziegenhain, Beate Vieth,
Swati Parekh, ..., Holger Heyn,
Ines Hellmann, Wolfgang Enard



UCDAVIS Bioinformatics Core

Single-Cell with 10x genomics

Gene expression profiling at scale with single cell resolution

What 10x is currently used to do

- **Genome** — Genome Resequencing
  - Call the full spectrum of variants (particularly long INDELS/CNV and structural variants) and unlock previously inaccessible regions from ***a single library at equivalent coverage as standard genome resequencing projects***

  Exome – Subselect reads using capture techniques (Agilent)
  - Enable phasing of genes and detection of structural and copy number variation
  - Agilent SureSelect baits improve gene phasing by closing gaps, and recovering hard-to-map loci in the genome (future kits to include previously failed regions)
- **Assembly – de Novo genome assembly**
- **Single Cell 3' RNAseq**
  - High-throughput single cell RNA sequencing
  - Scalable transcriptional profiling of 1,000s to 10,000s of individual cells

10x Chromium Box

Basic Stats

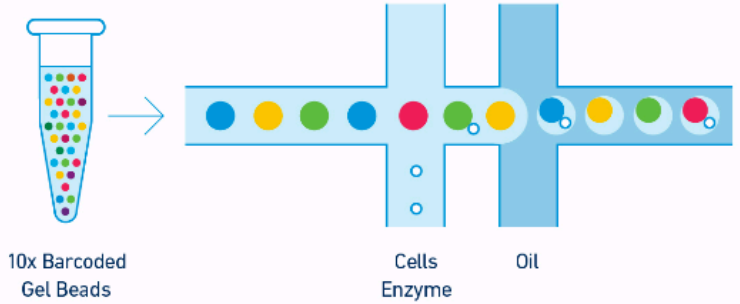
- Up to 8 channels processed in parallel
- 500 to 6,000 (V1) 10,000 (V2) cells per channel
- 10 minute run time per chip
- Up to 30 um cell diameter tested
- ~50 % cell processing efficiency

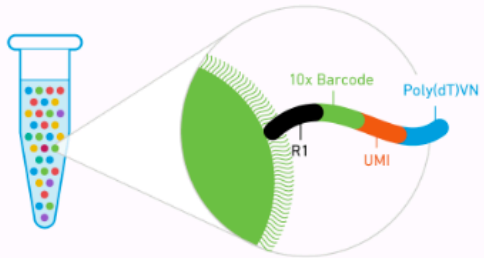
Number of cells	Expected Doublet Rate (%)
1,200	~1.2
3,000	~2.9
6,000	~5.7

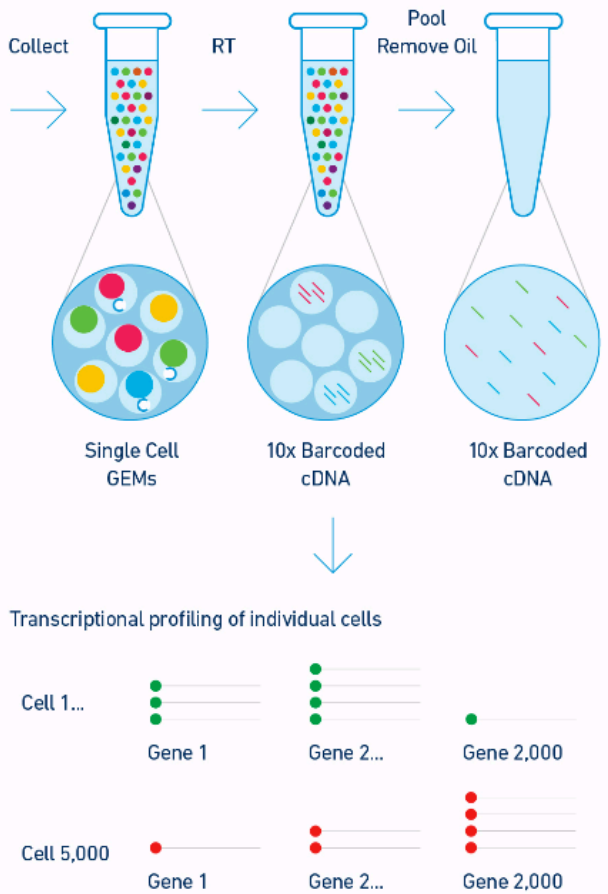
Number of cells	Expected Doublet Rate (%)
500	~0.4
1,000	~0.8
3,000	~2.3
5,000	~3.9
10,000	~7.6

User controlled trade off between cell numbers and doublet rate

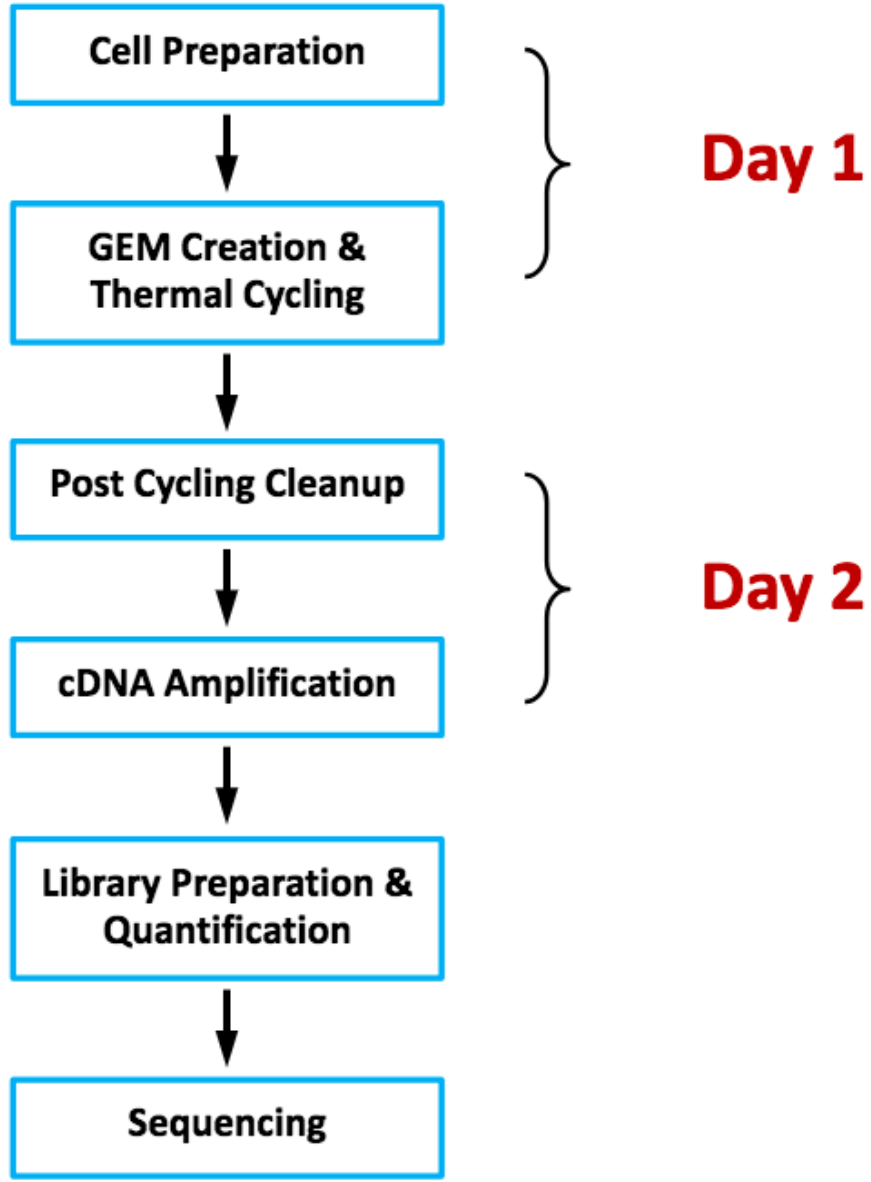
UCDAVIS Bioinformatics Core





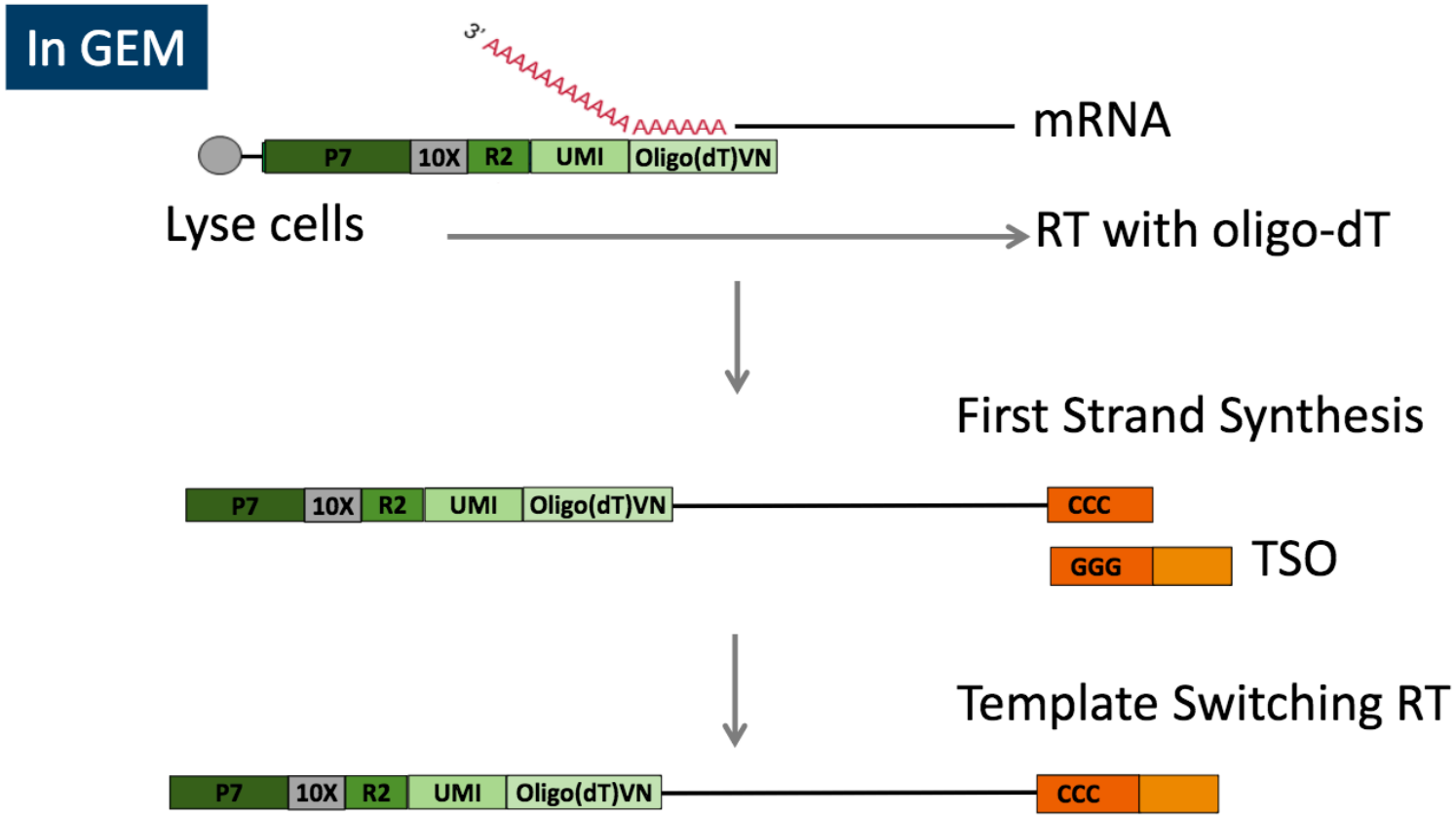


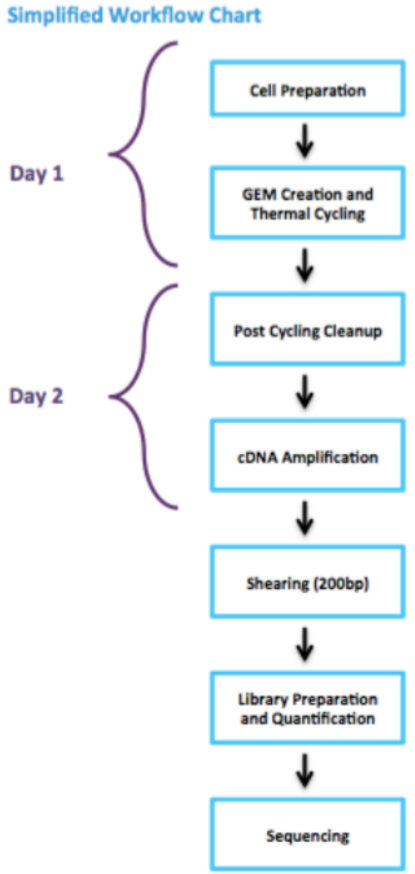
- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell



UCDAVIS Bioinformatics Core

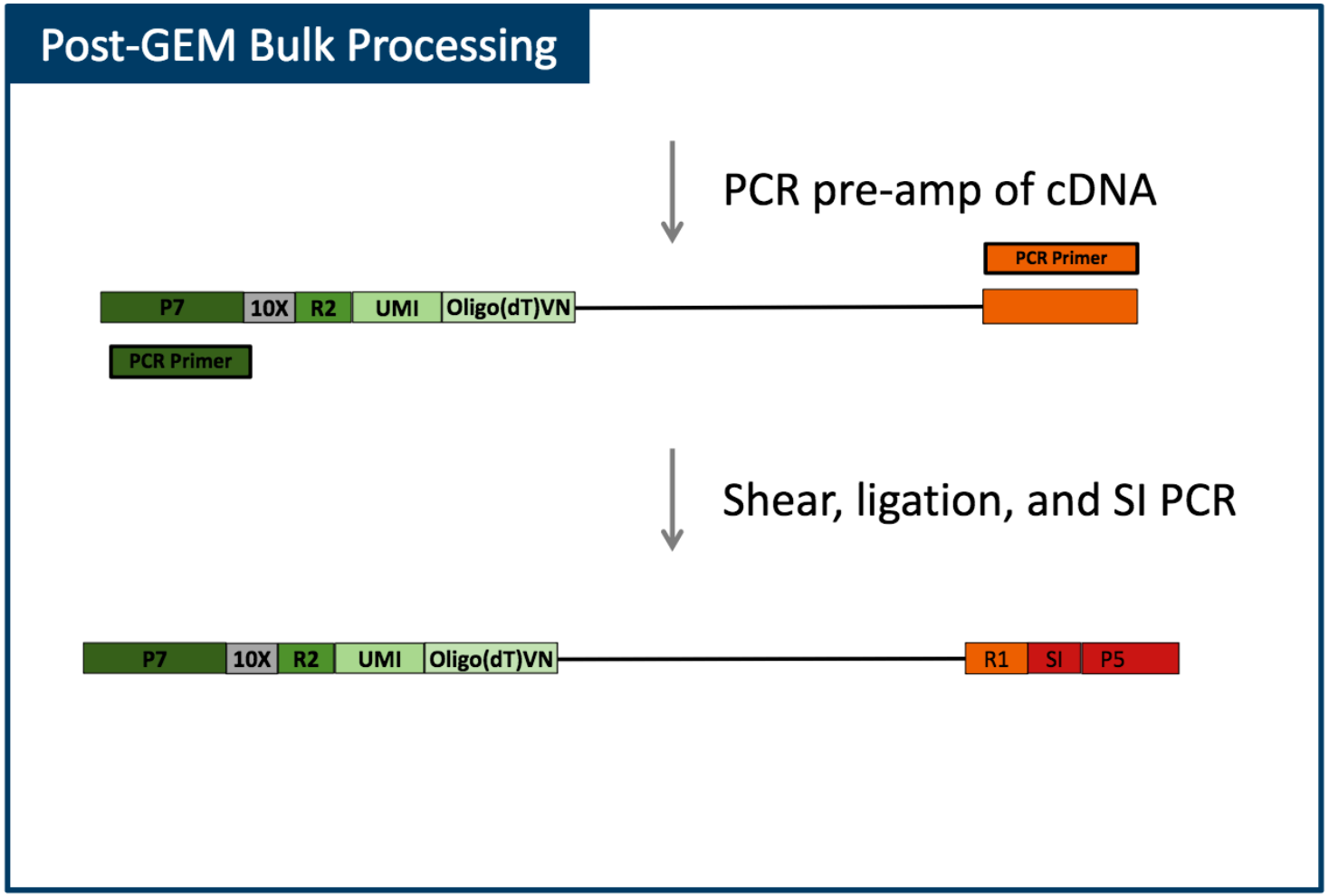
Basically a TAGseq protocol per cell 3’ expression

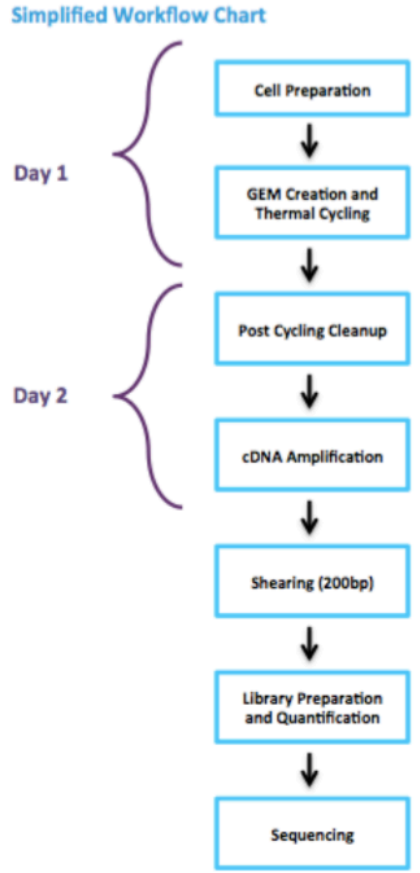




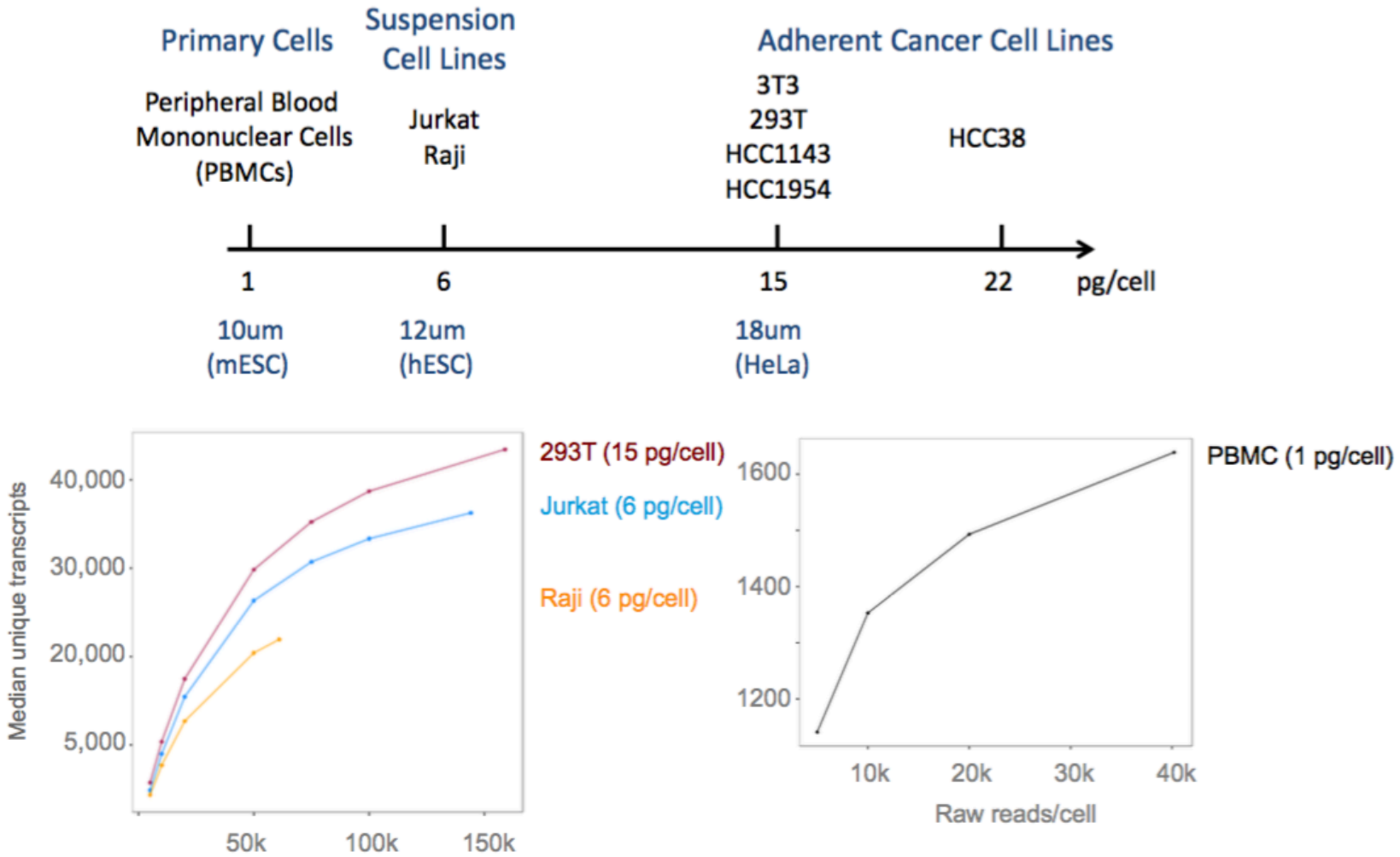
UCDAVIS Bioinformatics Core

Post-GEM Bulk Processing





Cells of differing sizes and complexity



Cell size requirement are minimal

Cells Tested	Species	Cell Source	Total RNA (pg/cell)*	Cell Size (µm)
PBMC	human	extracted from blood	~0.75	~5-10
E18 neuron	mouse	brain tissue	~ 2 - 3	~9
Jurkat	human	suspension	5.5	~12
Raji	human	suspension	7.3	~12
293T	human	adherent	14.2	~18
3T3	mouse	adherent	16.1	~18
HCC1954	human	adherent	15.7	~18
HCC38	human	adherent	21.6	~30

Wide window of RNA input

Cell Load	Total RNA Input*			
	PBMC's (RNA content per cell: 1 pg)	Jurkat Cells (RNA content per cell: 6 pg)	293T cells (RNA content per cell: 15 pg)	HCC38 cells (RNA content per cell: 22 pg)
500	0.5 ng	3 ng	7.5 ng	11 ng
1000	1 ng	6 ng	15 ng	22 ng
2000	2 ng	12 ng	30 ng	44 ng
3000	3 ng	18 ng	45 ng	66 ng
4000	4 ng	24 ng	60 ng	88 ng
5000	5 ng	30 ng	75 ng	110 ng
6000	6 ng	36 ng	90 ng	132 ng
7000	7 ng	42 ng	105 ng	154 ng
8000	8 ng	48 ng	120 ng	176 ng
9000	9 ng	54 ng	135 ng	198 ng
10000	10 ng	60 ng	150 ng	220 ng

Sequencing, V1

Recommendation

- 60,000 raw reads per cell is the recommended sequencing depth for typical samples.
- 30,000 raw reads per cell is sufficient for RNA-poor cell types such as PBMCs.
- Given variability in cell counting/loading, extra sequencing may be required if the cell count is higher than anticipated.

Validated on

- HiSeq 2500 Rapid Run
- NextSeq
- MiSeq

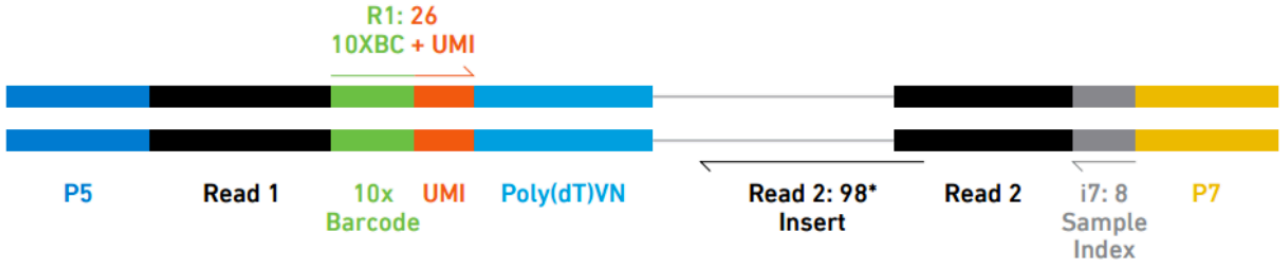
Custom sequencing run, with 4 reads, V1 kits

Sequence Read	Recommended Length	Read Description
Read 1	98bp	Transcript Tag
I7 Index	14bp	Cell Barcode Read
I5 Index	8bp	Sample Index Read
Read2	10bp	UMI Read

@ full capacity 6,000 cells per sample and 60K reads per cell = 360M reads or ~1 lane/sample

UCDAVIS Bioinformatics Core

Sequencing, V2



Recommendation

- 50,000 raw reads per cell is the recommended sequencing depth for typical samples.
- 30,000 raw reads per cell is sufficient for RNA-poor cell types such as PBMCs.
- Given variability in cell counting/loading, extra sequencing may be required if the cell count is higher than anticipated.

Validated on

- HiSeq 4000
- HiSeq 2500 Rapid Run
- NextSeq
- MiSeq

Custom sequencing run, with 3 reads, V2 kits

Sequence Read	Recommended Length	Read Description
Read 1	100bp	10 barcode and UMI
I7 Index	8bp	Sample Index Read
Read2	100bp	Transcript Tag

@ full capacity 10,000 cells per sample and 60K reads per cell = 500M reads or ~1.25 lane/sample

Software – System requirements

Local

- Run on single, standalone Linux system
- CentOS/RedHat 5.2+ or Ubuntu 8.04+
- 8+ cores, 64GB RAM

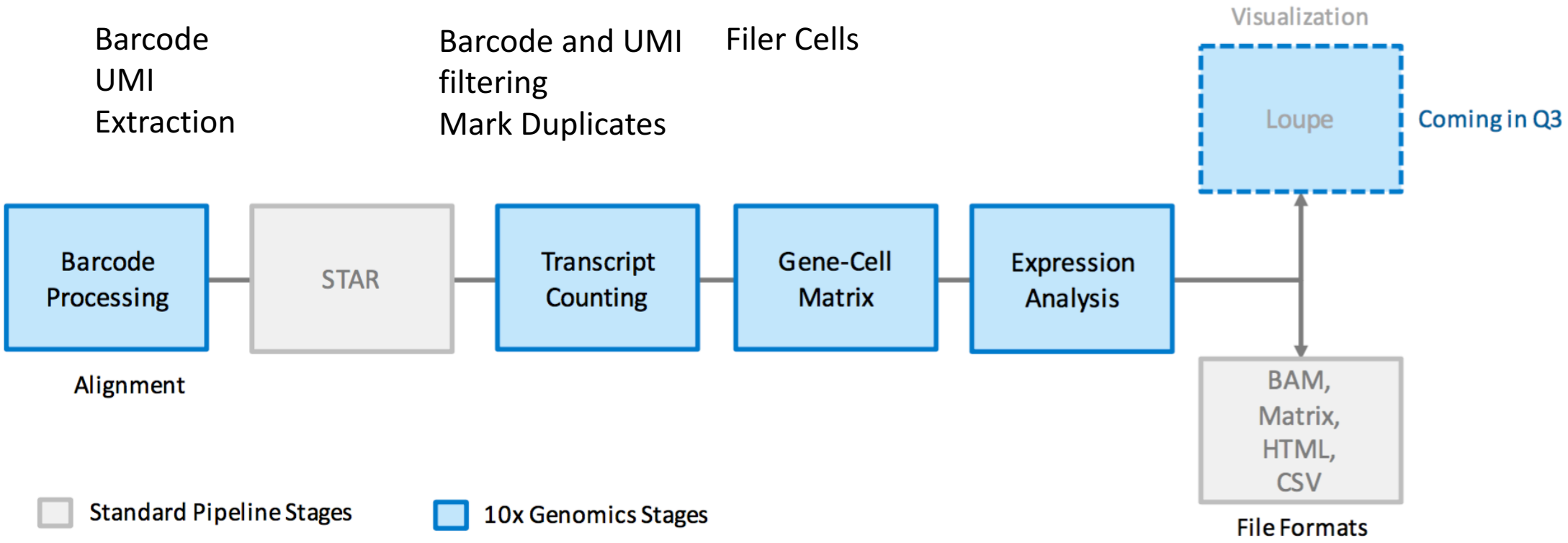
Cluster

- Run on SGE and LSF
- Each node must have 8+ cores and 8GB+ RAM/core
- Shared filesystem between nodes (e.g. NFS)

50 core-hours per 100M reads, 5000 cells, 40k reads/cell: 95 core-hours

UCDAVIS Bioinformatics Core

Analysis Workflow



Cell Barcode and UMI filtering

- Cell barcodes

– Must be on static list of known cell barcode sequences

– May be 1 mismatch away from the list if the mismatch occurs at a low-quality position (the barcode is then corrected).

- UMIs

– Must not be a homopolymer, e.g. AAAAAAAAAA

– Must not contain N

– Must not contain bases with base quality < 10

– UMIs that are 1 mismatch away from a higher-count UMI are corrected to that UMI if they share a cell barcode and gene.

Marking Duplicates

- Using only the confidently mapped reads with valid barcodes and UMIs,

– Correct the UMIs

  - UMIs are corrected to more abundant UMIs that are one mismatch away in sequence.

– Record which reads are duplicates of the same RNA molecule – Count only the unique UMIs as unique RNA molecules

– These UMI counts form an **unfiltered gene-barcode matrix.**

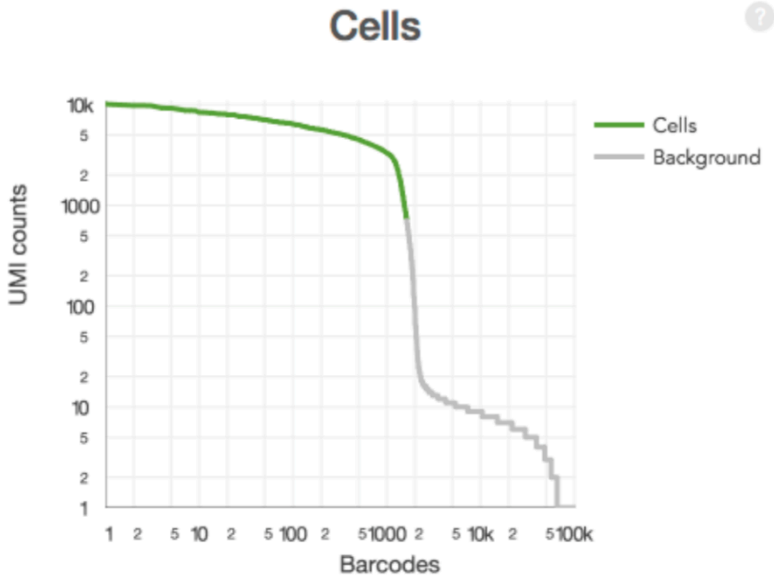
Filtering Cells

- Select GEMs that likely contain cells

– Sum UMI counts for each barcode

– Select barcodes with total UMI count >10% of the 99th percentile of the expected recovered cells.

- Produces a **filtered gene-barcode matrix**.

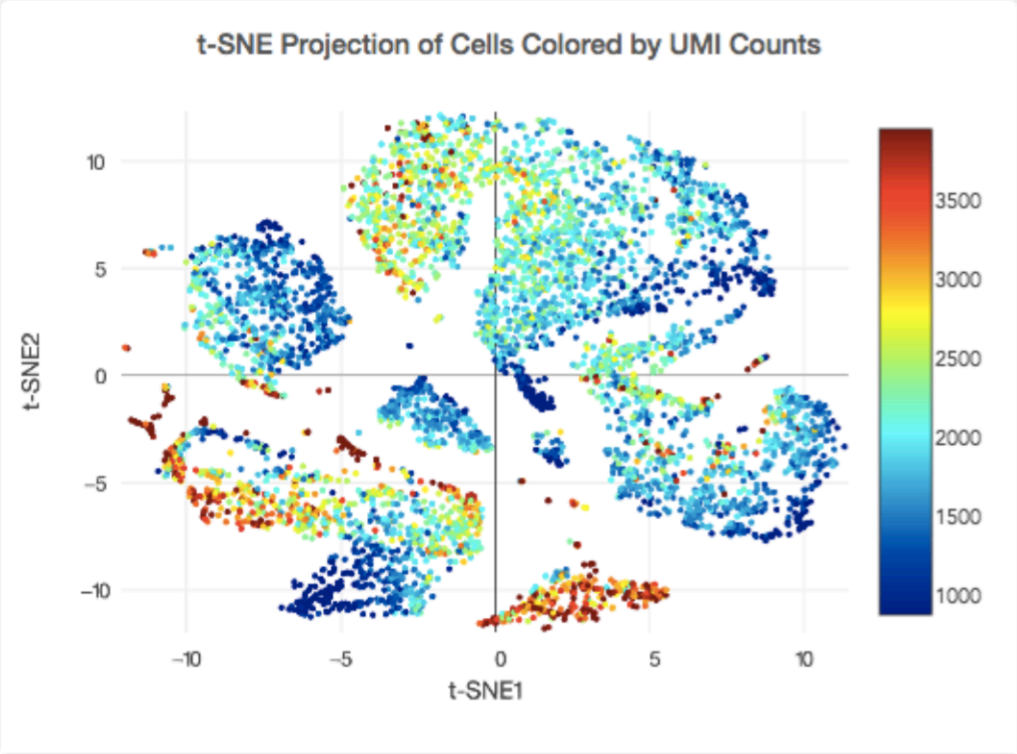


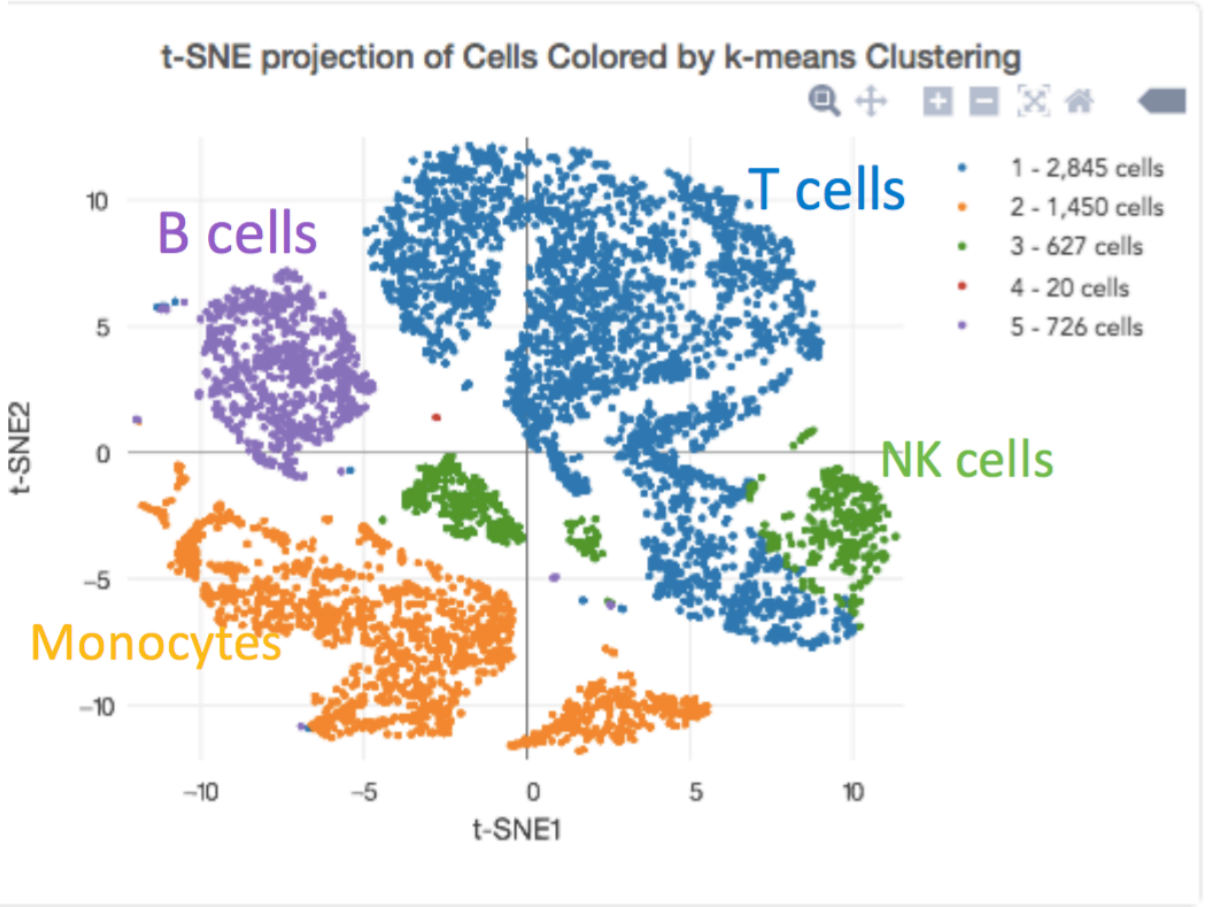
Downstream Analysis – offered by cell ranger

- Clustering analysis results
  - For each ‘K’ (number of clusters desired): – Which cells go into which clusters – Differentially expressed genes across cluster
- Principle Component Analysis (PCA) results
  - How much each gene contributes to the lower-dimensional space
  - PCA projection coordinates of cells
- t-SNE analysis results
  - The coordinates of each cell in 2-d space
- R packages
  - 10x genomics
  - Seurat https://github.com/satijalab/seurat

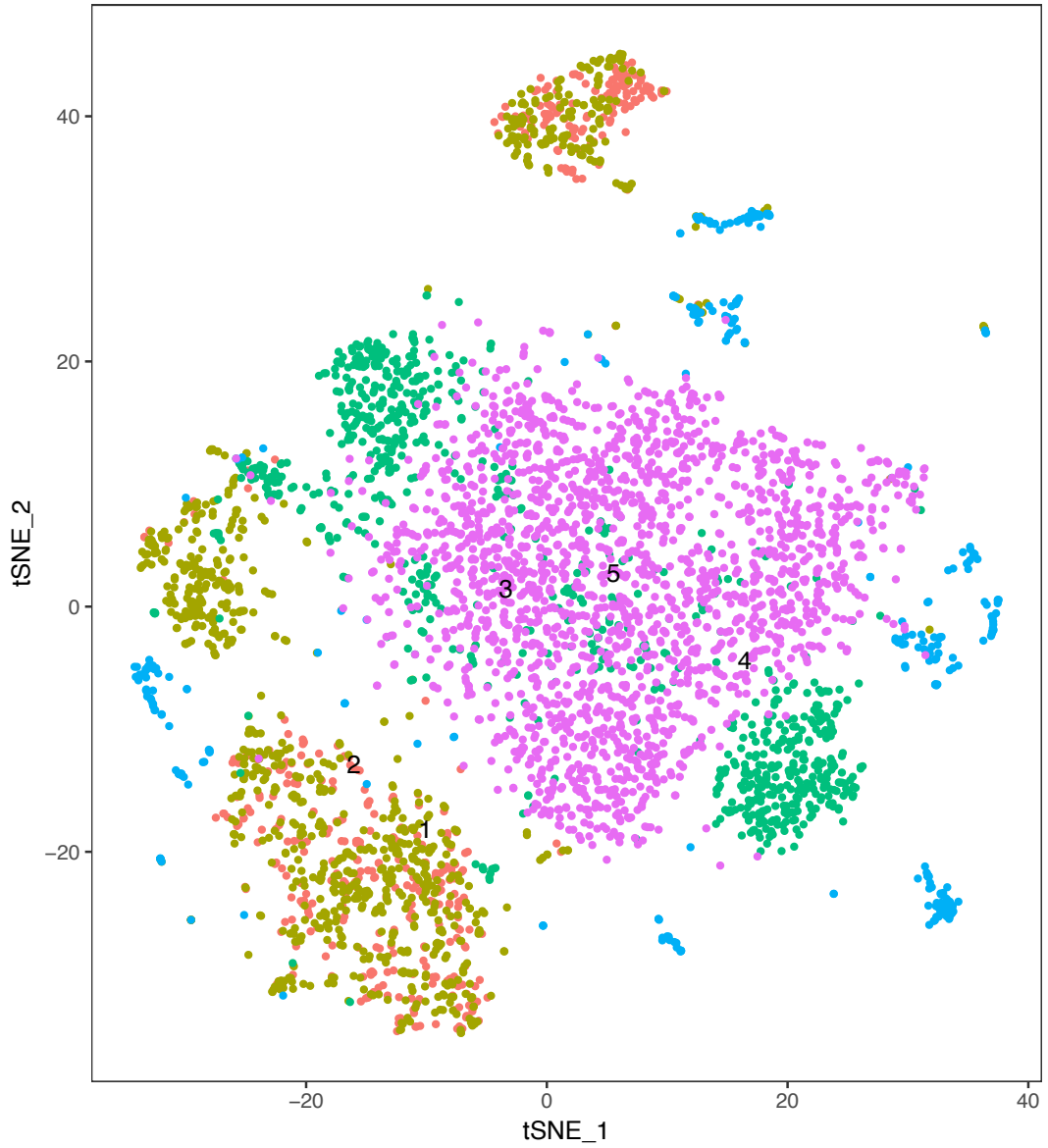
UCDAVIS Bioinformatics Core

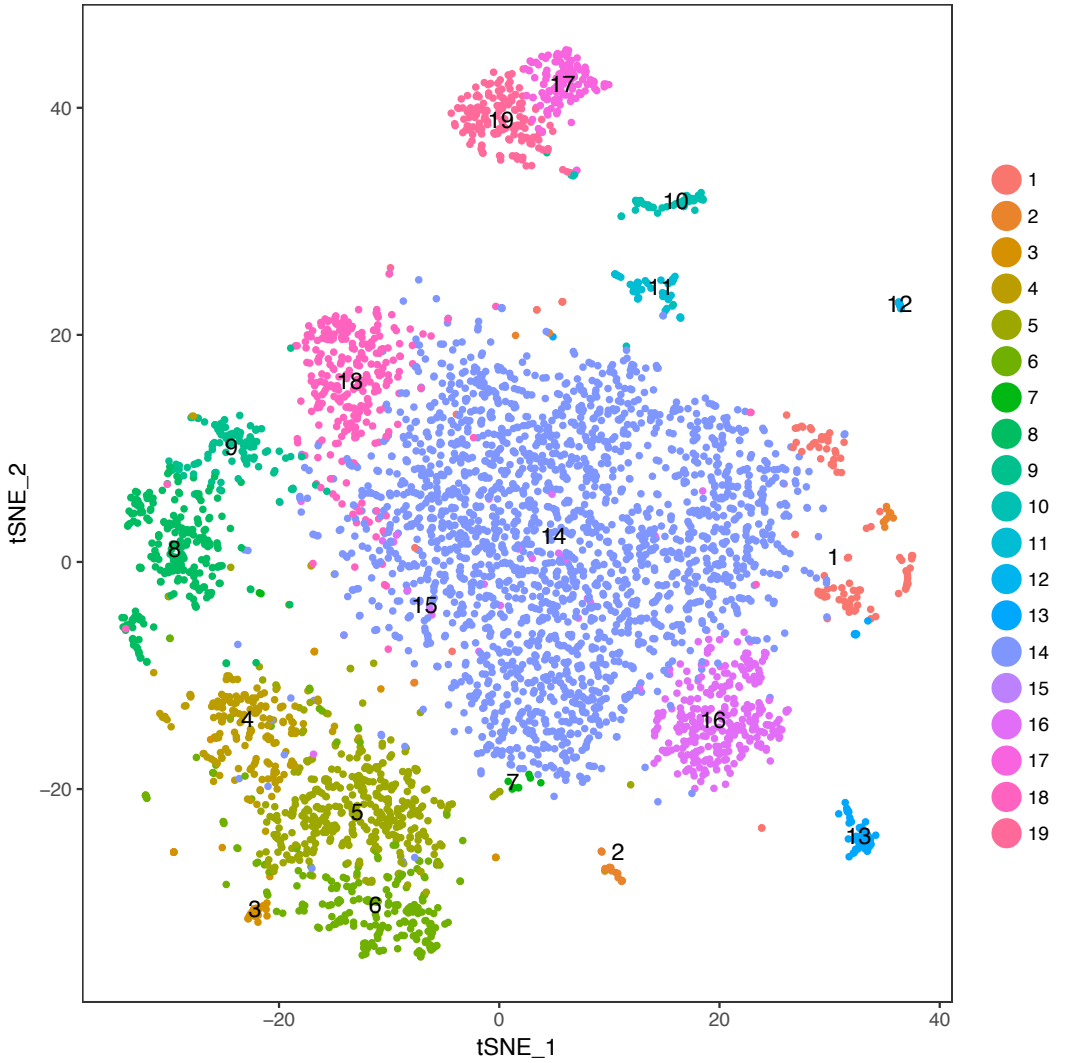
The classical clustering plot

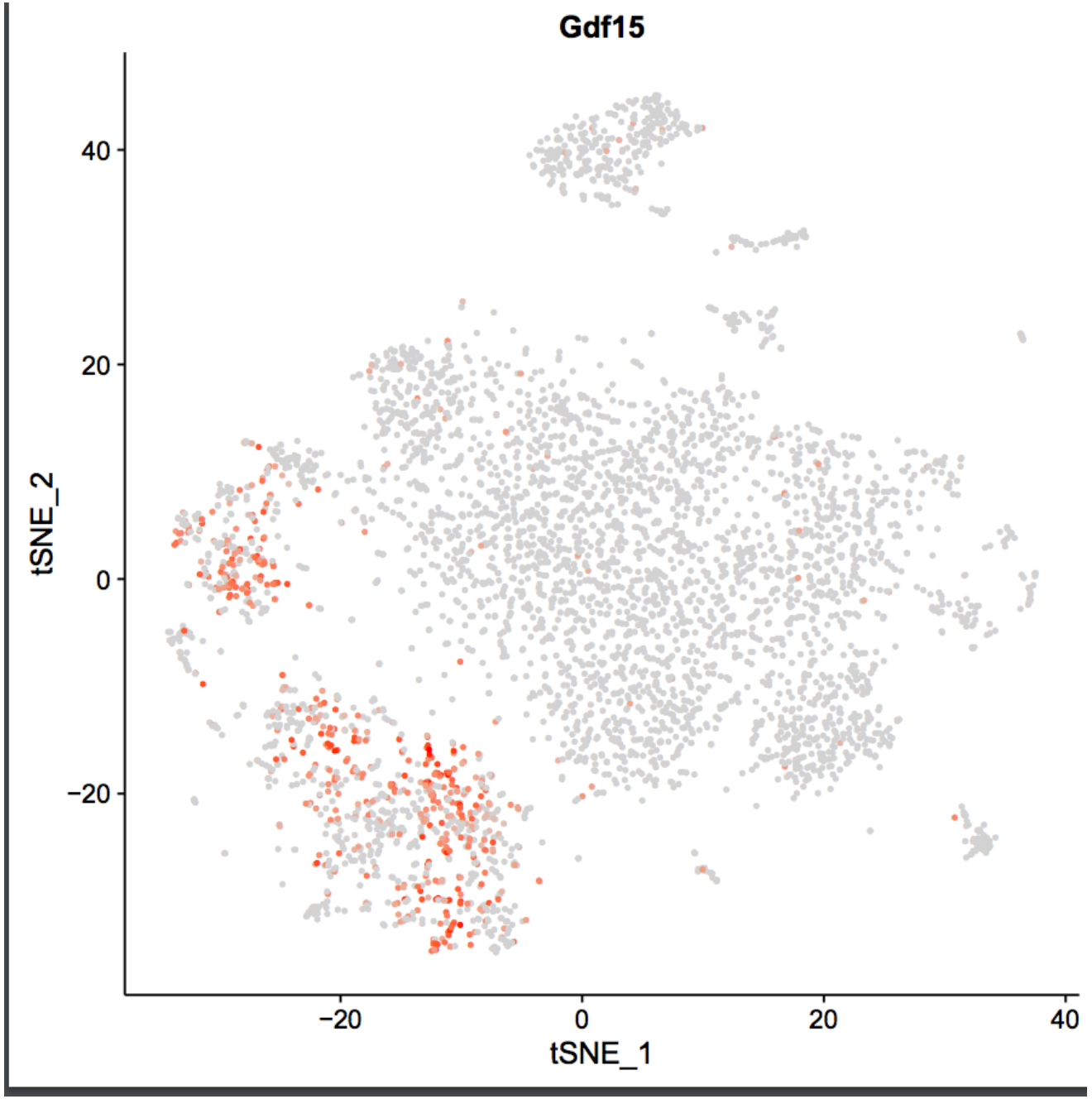




UCDAVIS Bioinformatics Core





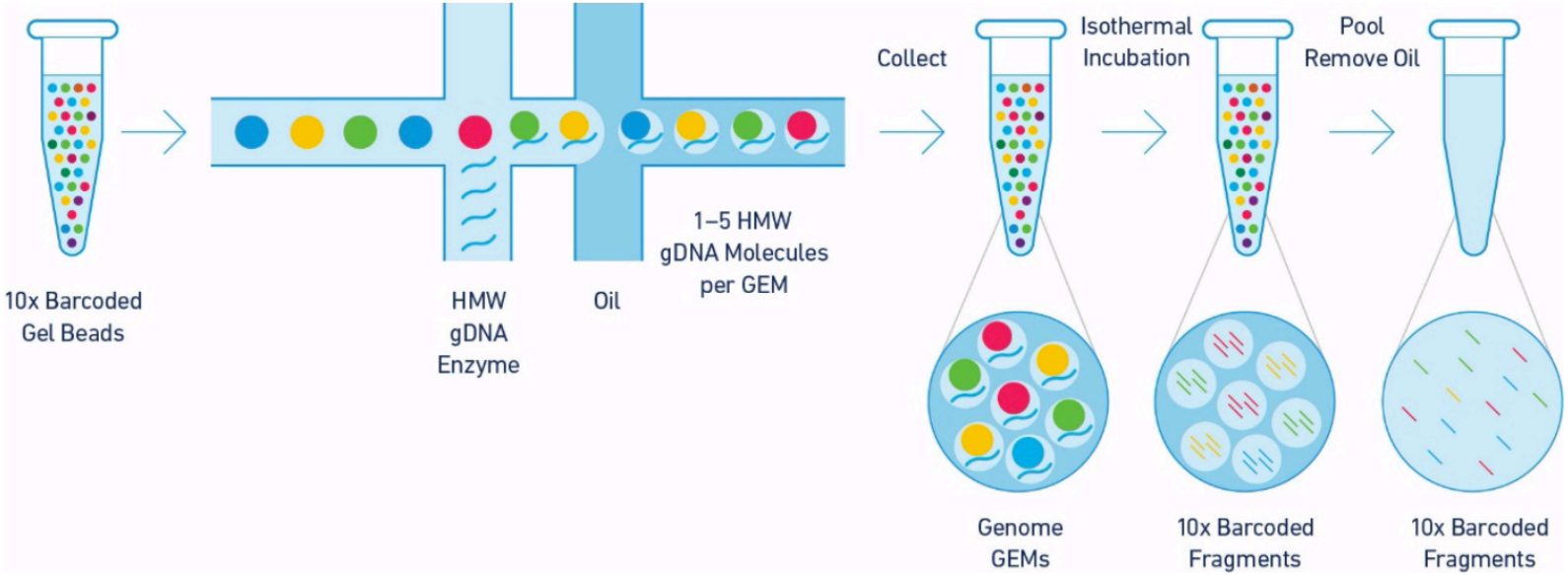


UCDAVIS Bioinformatics Core

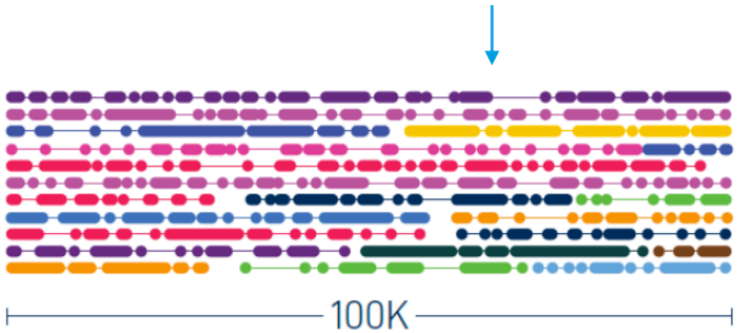
Genome and Exome Analysis

Long-range analysis and phasing of SNVs, indels, and structure variants

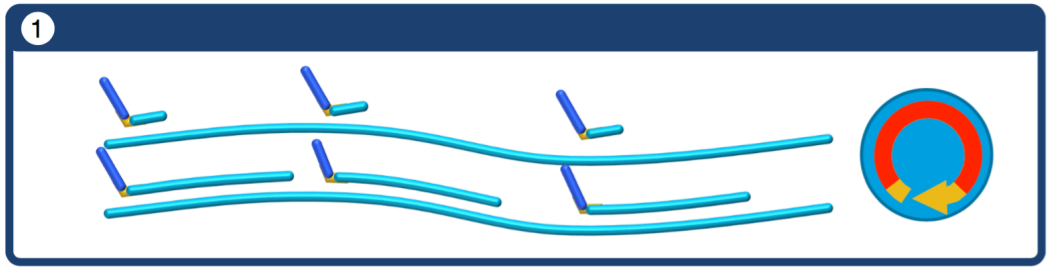
In a nut shell

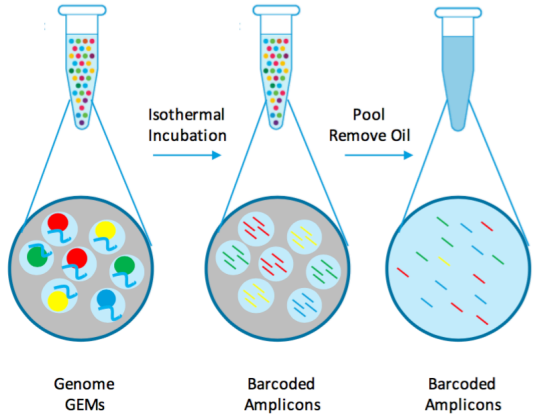


Linked-Reads



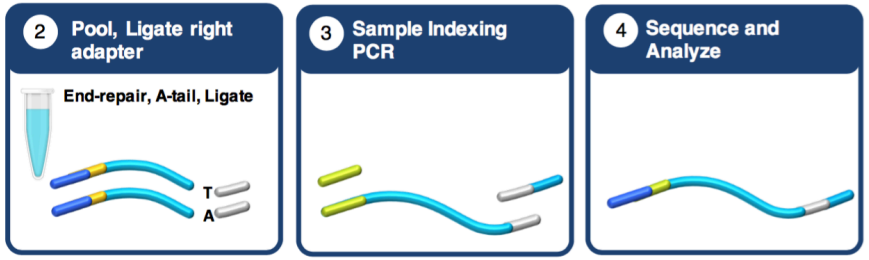
Laboratory Workflow



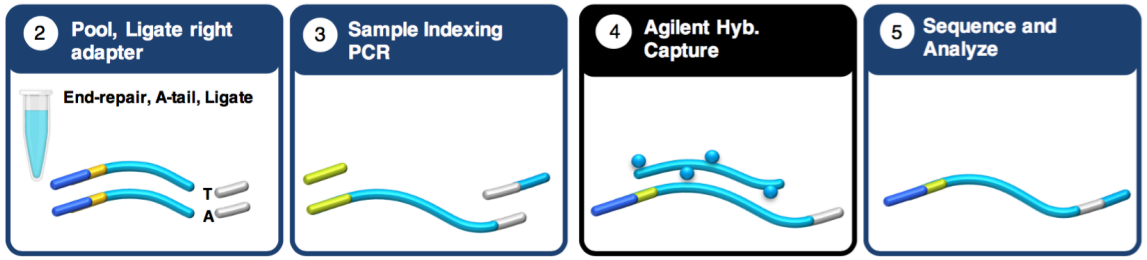


Shear?

Whole **Genome** Sequence

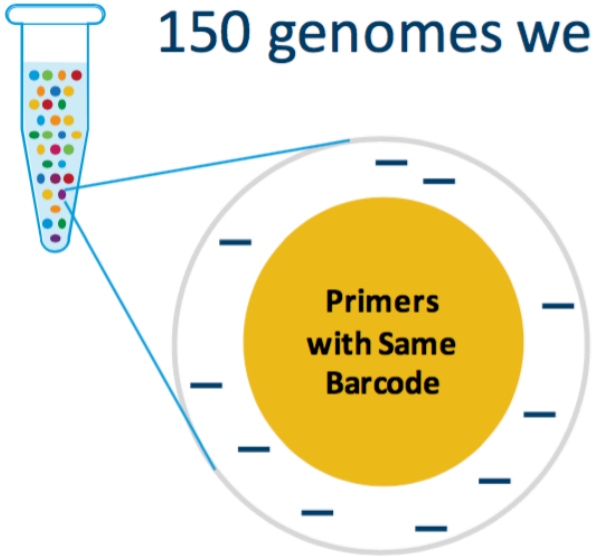


Whole **Exome** Sequence



The Math

1 ng Input DNA
= 300 genomes
copies of the
genome



150 genomes went into 1M partitions

Each GEM contains:

- One barcode (many copies)
- 1/6000 of the genome (500 Kb)
- At 50Kb length, 10 molecules

Chance that 2 molecules covering a locus are in same GEM:

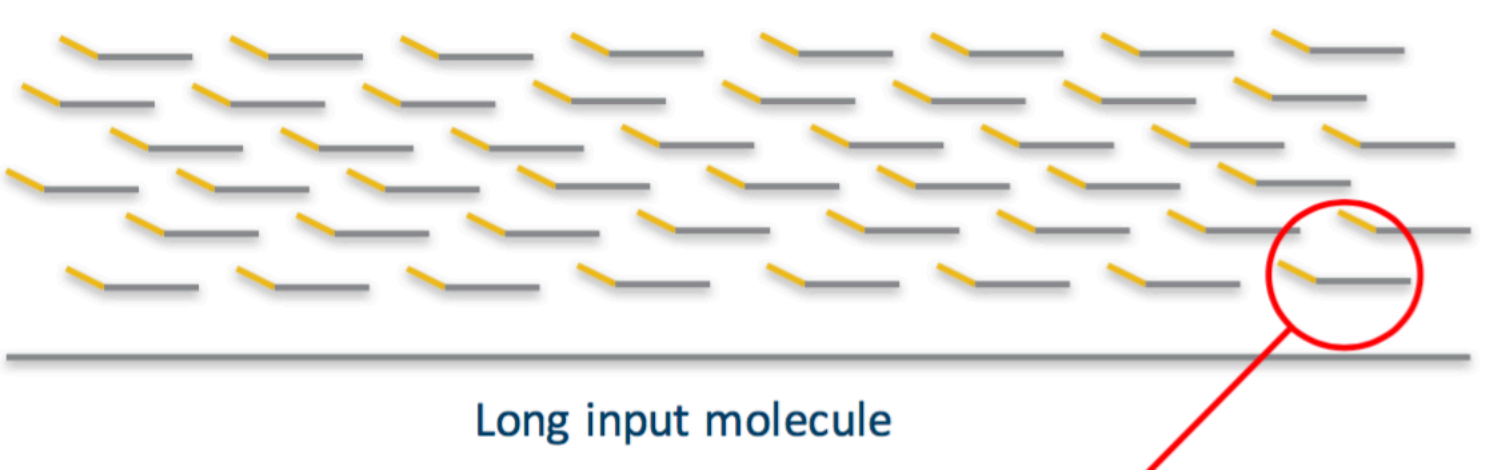
1 in 6000

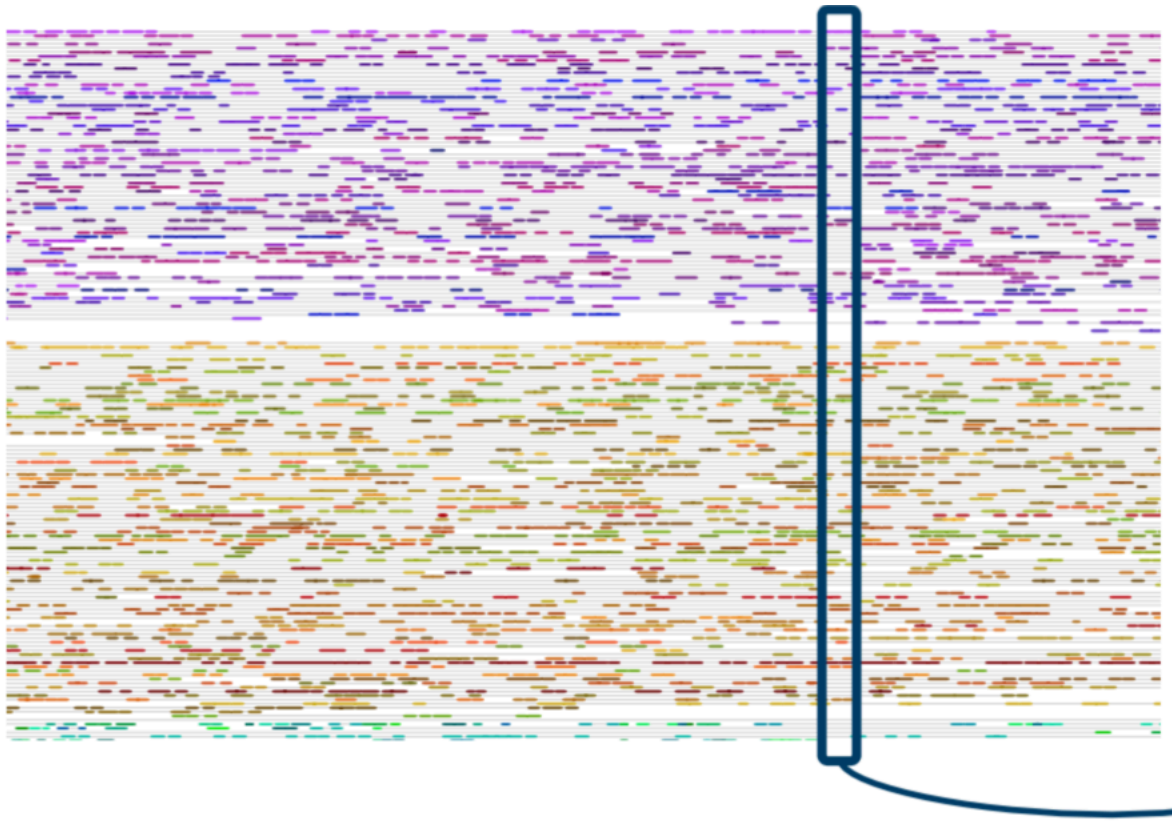
Percent unique barcodes at any genomic locus:

99.98%

Calculations imply that
about 50% of all possible
fragments end up in a bead

UCDAVIS Bioinformatics Core





150X avg *molecule* coverage

@ Recommended Loading

- Each locus will have 150 molecules
- Each locus will have 30x read depth

- ~35 fragments per molecule
  @50Kb molecules = 0.2x/molecule

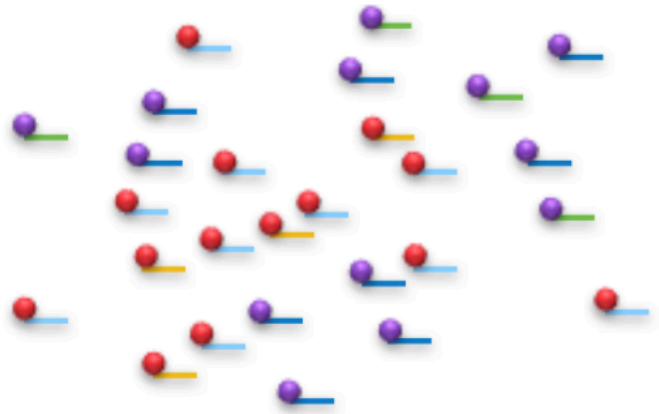
> 30X avg *read* coverage

Analysis – Biological Questions

- At recommended specs (for human genome)
  - Get ~30x coverage, adequate for standard variant analysis SNPs, small INDELS
  - Increased mapability to difficult regions [multi-mapped reads can be resolved by considering linked reads information], variants previously undetermined
  - Detect large SV and CNV
  - Phased information

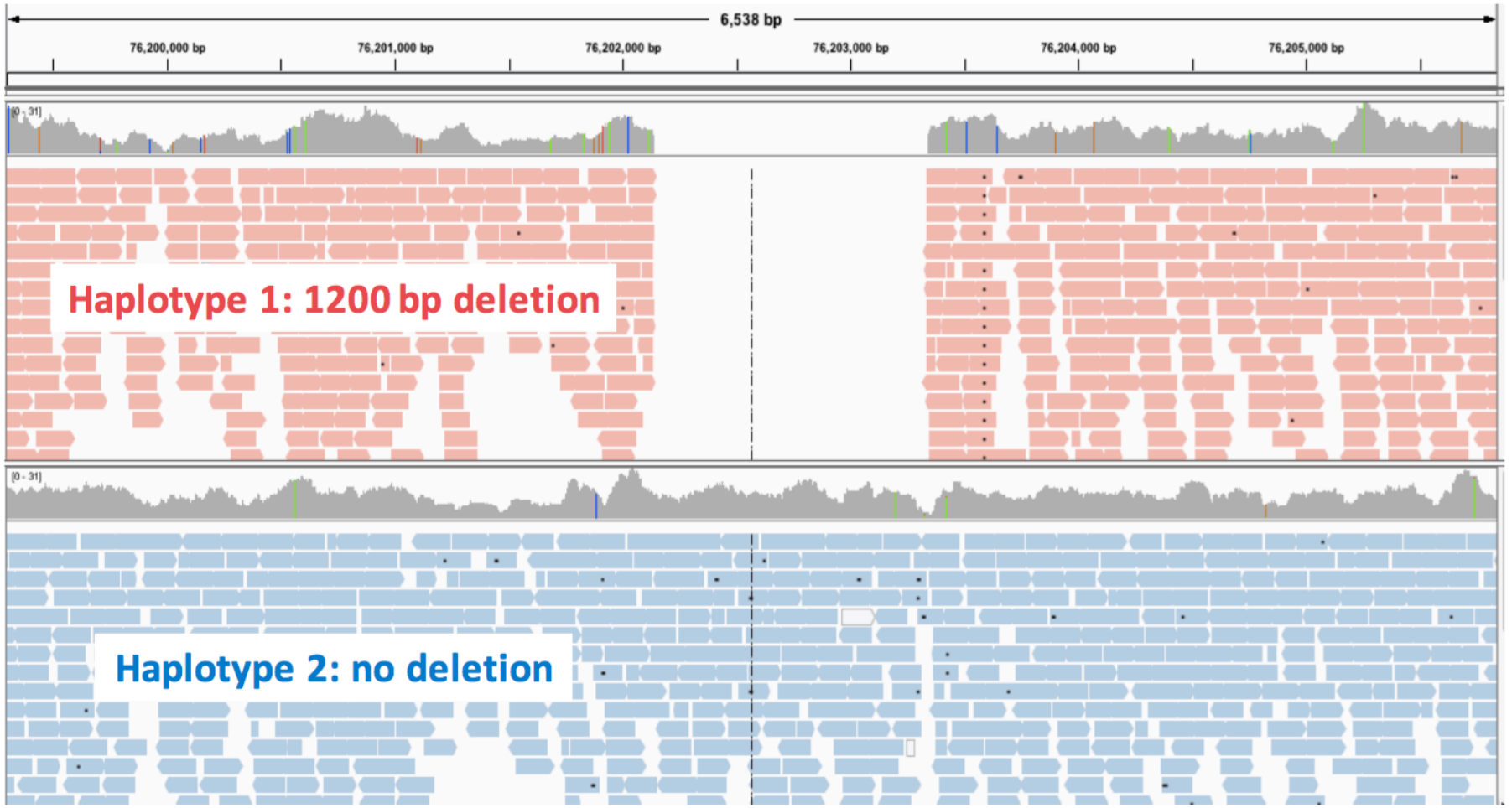
Detection of SV and CNV requires advanced computational techniques
Phasing has been used extensively in GWAS (usually imputed) to enhance analysis and inferences, slow to get to sequence based data
Potential applications for the technology are likely still yet to come

Increased Mapability



1. *Confident* mapping provides anchors
2. *Barcodes* recruit short reads into paralogous loci

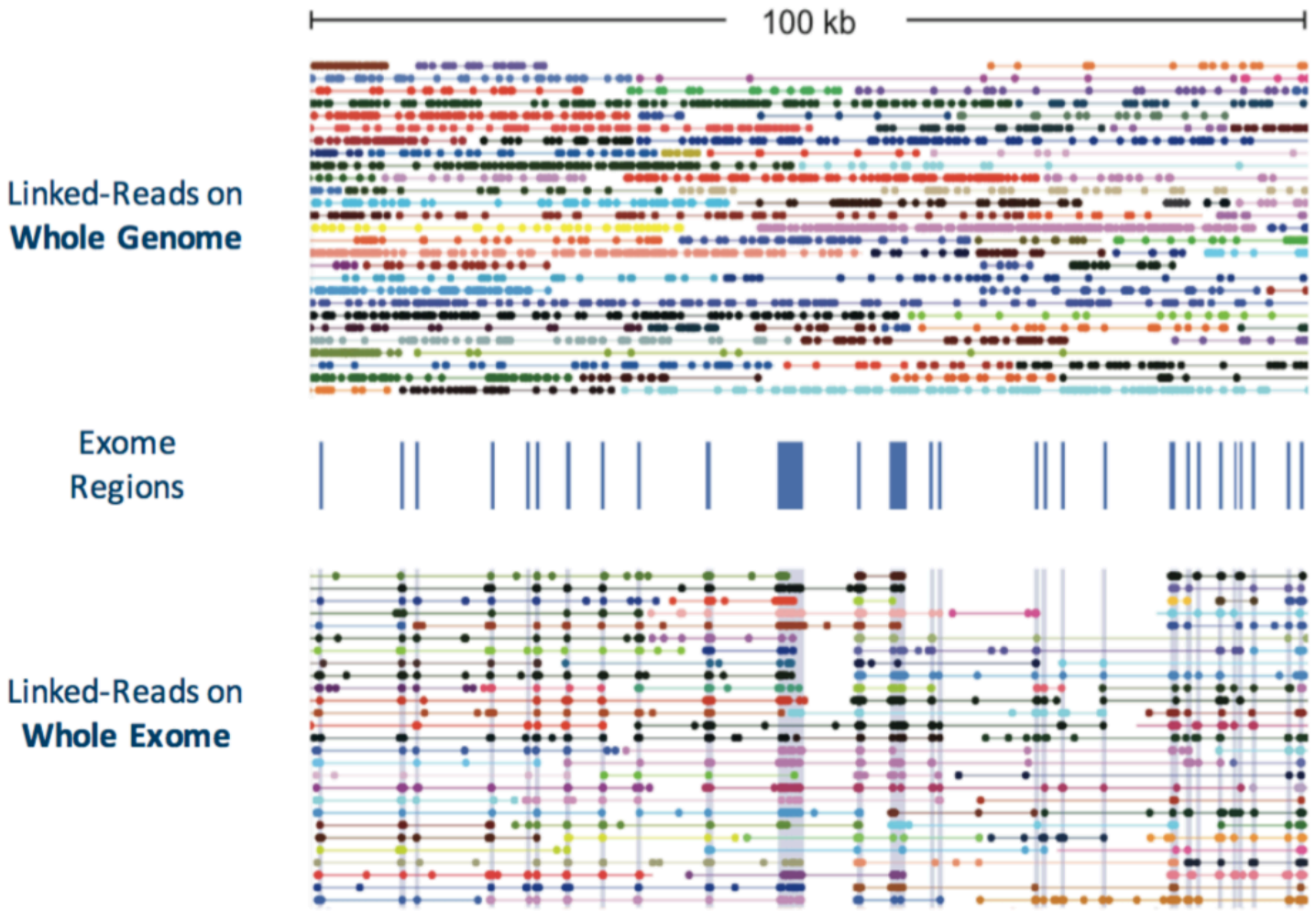
Linked Reads



Capture – linked genes

Enrich reads of interest instead of random selection

Depending on size of capture, can pool more samples/lane



UCDAVIS Bioinformatics Core

Assembly

de novo assemblies

Sample Requirements - Supernova

- **Genome size**: Supernova has been tested on genomes in the size range 1.0-3.2 Gb.
- **Other genome characteristics**: Supernova has not been tested on genomes having repeat content far greater than human, nor on genomes having extreme GC content.
- **Clonality**: strongly recommend that DNA be obtained from an individual organism or clonal population.
- **DNA size**: Recommend that this value be at least 50 kb, and preferably 100 kb. DNA length is highly correlated with several assembly statistics, including contig length, phase block length and scaffold length.

Sequencing - Supernova

Instrument	Configuration	Result	Lanes
HiSeq X	Standard	Excellent	2
HiSeq 2500	Rapid run	Excellent	4
HiSeq 2500	High Output	Not tested	
HiSeq 4000	Standard	Useable, but observed contig length half as long as those from HiSeq X	2
Miseq	standard	Not tested	

- **Read length**: Supernova requires as input 2x150 base reads.
- **Sequencing depth**: Recommends sequencing to depth between 38x and 56x. For highly polymorphic organisms, we recommend 56x. Coverage higher than 56x may not improve results. Sequence twice as much as mapping application

System Requirements - Supernova

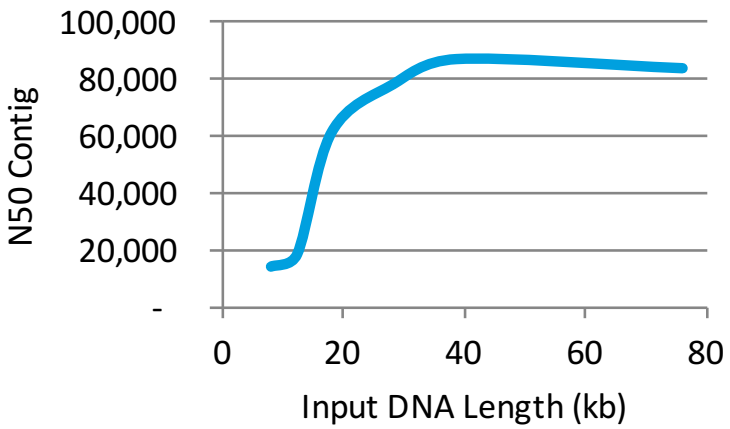
- 16-core (or greater) Intel or AMD processor
- **384 GB RAM**
- 2 TB free disk space
- 64-bit CentOS/RedHat 5.2+ or Ubuntu 8+
- Bcl2fastq 2.17
- No other large processes running on the system **

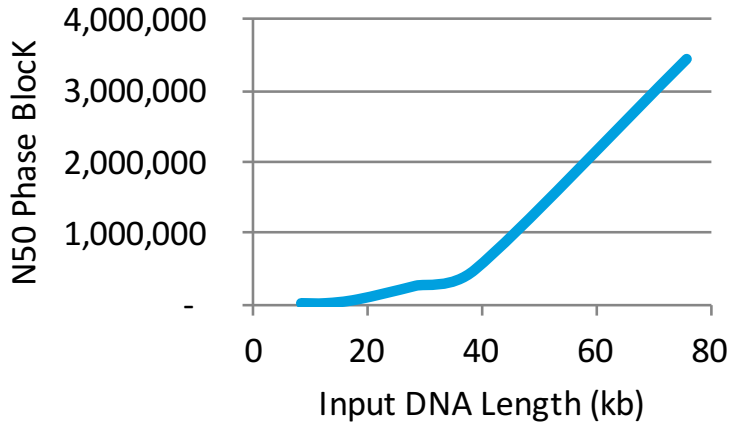
Supernova should be run with at most 1.2 billion reads (single reads), and at 38-56x coverage of the genome.

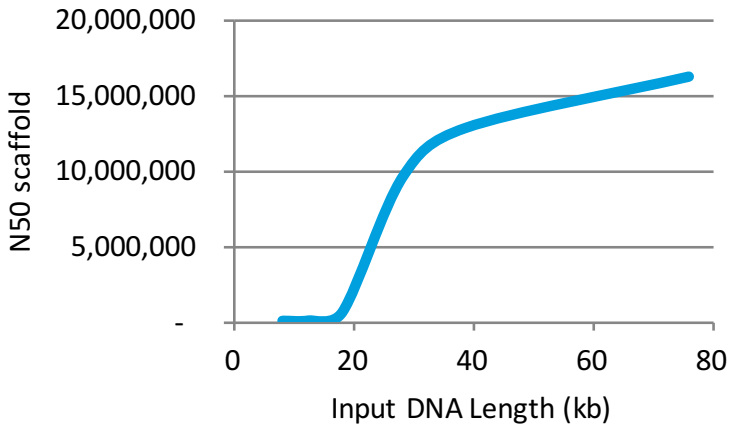
Assembly Performance vs. DNA length (Supernova)

10X GENOMICS™

- Increased Assembly performance with increased input DNA length
  - Recommended: >50Kb length, longer lengths may be required for some organisms







UCDAVIS Bioinformatics Core

Genome Stats

Genome	Size (Gb)	DNA size(Kb)	N50 contig(Kb)	N50 scaffold(Mb)	N50 phase block (Mb)
NA12878	3.2	95.5	85.0	12.8	2.8
NA24385	3.2	111.3	90.0	10.4	3.9
HGP	3.2	138.8	104.9	19.4	4.6
Yoruban	3.2	126.9	100.5	16.1	11.4
Komodo dragon	1.8	85.4	95.3	10.2	0.4
Spotted owl	1.5	72.2	118.3	10.1	0.2
Hummingbird	1.0	86.2	87.6	12.5	10.1
Monk seal	2.6	92.3	93.8	14.8	0.6
Chili pepper	3.5	53.3	84.7	4.0	2.1

Genome Quality - Pac Bio vs 10x

- Qualitatively
  - Pac Bio will produce >> N50 contig lengths (contiguous sequence) on the order of 10 – 50x larger contig N50
  - 10x will will produce >> N50 scaffold length (ordered and arranged contigs with gaps) on the order of 2-5x larger scaffold N50
- Costs
  - Human Genome sequenced at ~60x coverage

    (recommended depth for both Pac Bio and 10x)
    - $70,000 Pac bio
    - $8,500 (4 lanes, HiSeq 2500 2x150, rapid mode) [~$5000 on the X platform]
- Input DNA
  - 10x requires ~ 1ng input DNA relative to ~10ug of input DNA for Pac Bio

In Conclusion

- As with many of these newest technologies, sample/cell preparation is crucial to success
  - Expect to purchase more equipment related to sample/cell preparation
- Expect some time spent on optimization, especially with single cell RNA
- As with every new technology there are still detail and kinks to be worked, ex. HiSeq 4000 is inevitable
- Ahead of its time for phased analysis/experiments (Diplomics)
- Expect other applications, methodologies to grow out of the technology, as people play new “seqs” will arise