10X Chromium

Single Cell and Long Read Sequencing Applications and Workflow

Marie Adams, MS
Genomics Core Director
Van Andel Institute

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®

Introduction

- Overview of Chromium Instrument Workflow

- Chromium and Linked Read Sequencing

- Chromium and 3' scRNA Sequencing

- *Chromium sc V(D)J – human T-cells only*

10X Genomics

Gel bead scaffold → Functional oligo with barcode → High-diversity library

30 um

P5   R1   Barcode   N-mer

Allows for Genome sequencing on HiSeq X
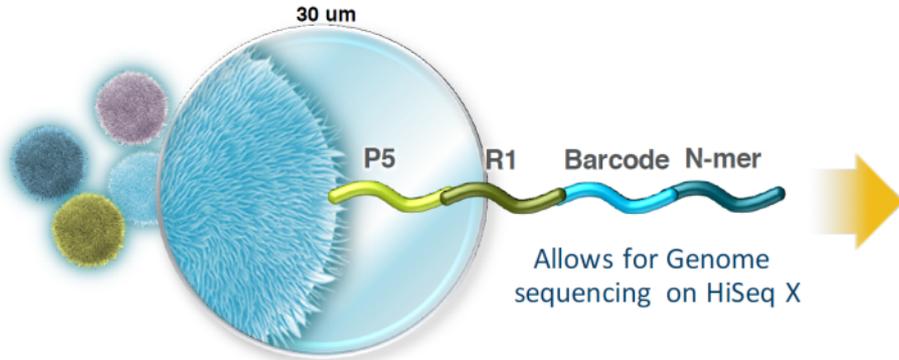
- 16bp barcode
- Defined sequence
- Highly uniform size and representation
- Built-in sequencing adapter and primer content

>4M discrete reactions in one tube!

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®

Reagents and Partitioning
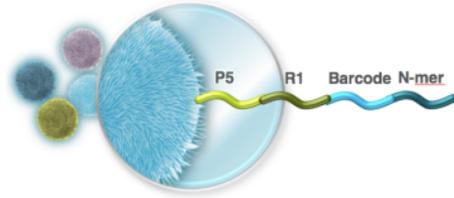
3 Components:

1. Gel Beads
2. Genomic DNA with Enzyme Mix
3. Partitioning Oil

GEMs
(Gel beads in EMulsion)

DNA

Barcoded Primer Library → Enzyme Oil → Collect

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®
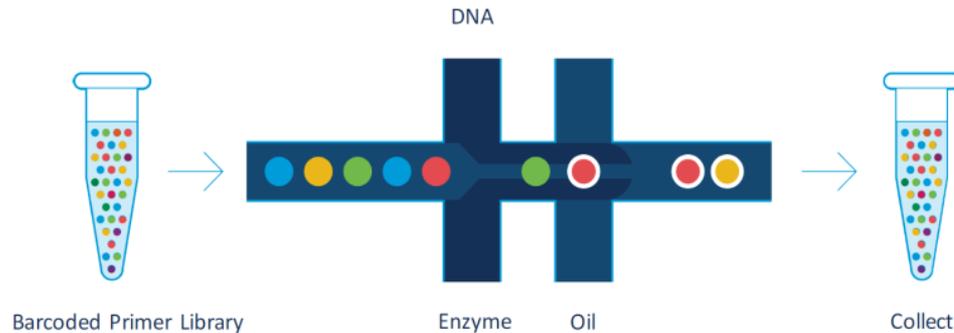
Inside the GEM

Linked Read
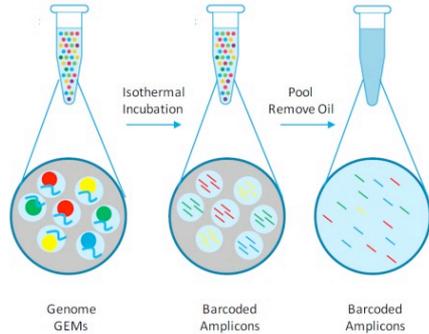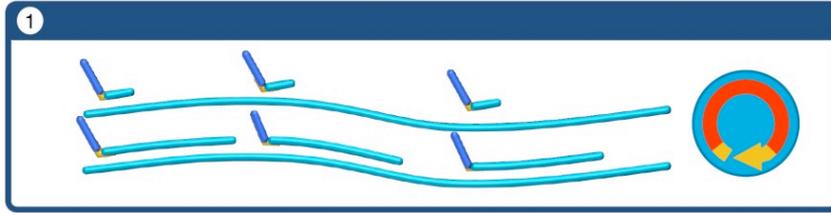
Single Cell
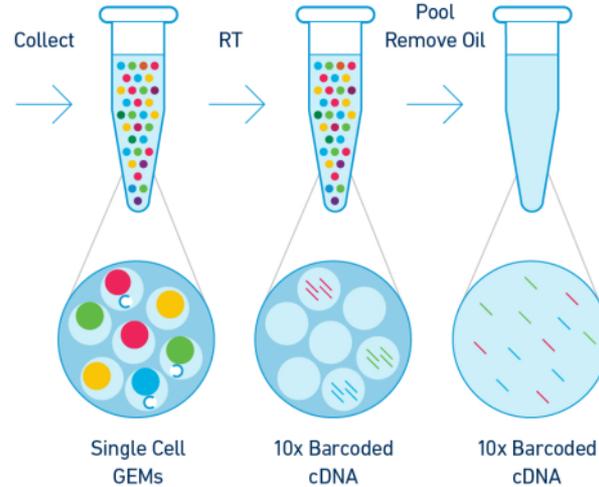
After emulsion is broken, move forward into Illumina library preparation and sequencing

VAN ANDEL INSTITUTE
100% TO RESEARCH, DISCOVERY & HOPE®
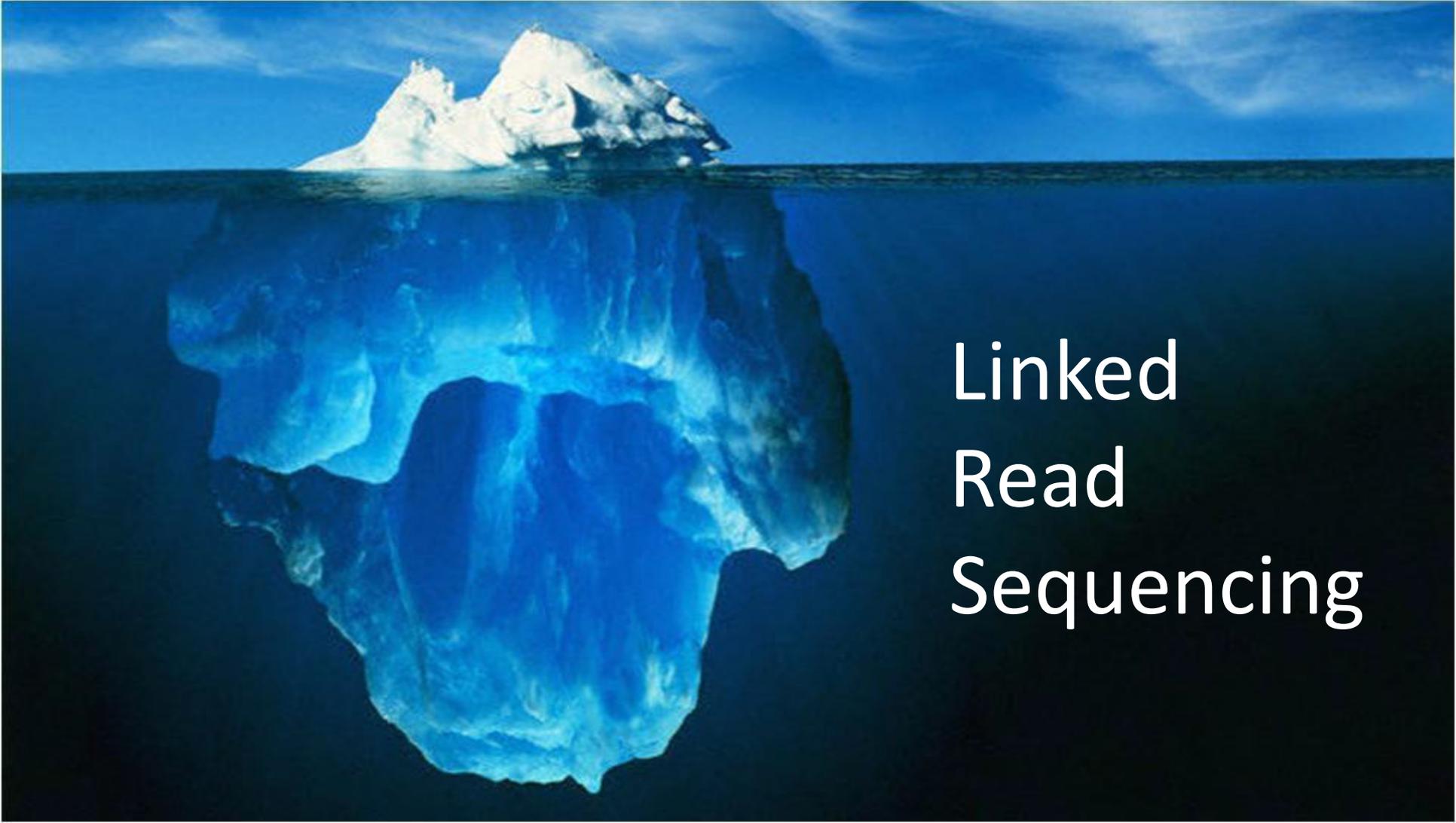
Linked Read Sequencing

# "Long read Sequencing": Unlocking Structural Variation

- Repeat/low complexity region resolution
- de novo genomes
- Phasing/haplotyping

VAN ANDEL INSTITUTE
100% TO RESEARCH, DISCOVERY & HOPE®
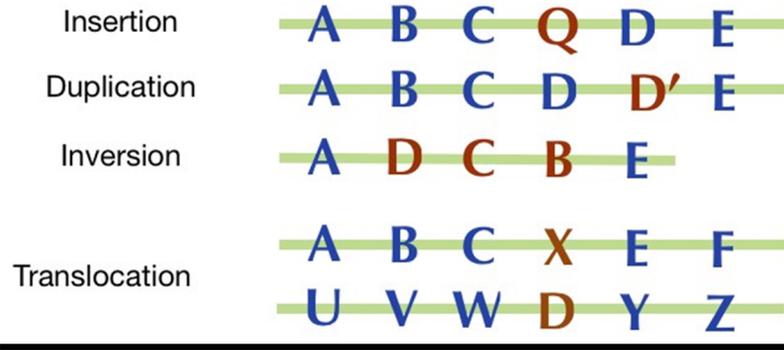
Repeats: Short vs Long Reads

Collapsed repeat consensus

Long reads make "cleaner" genomes
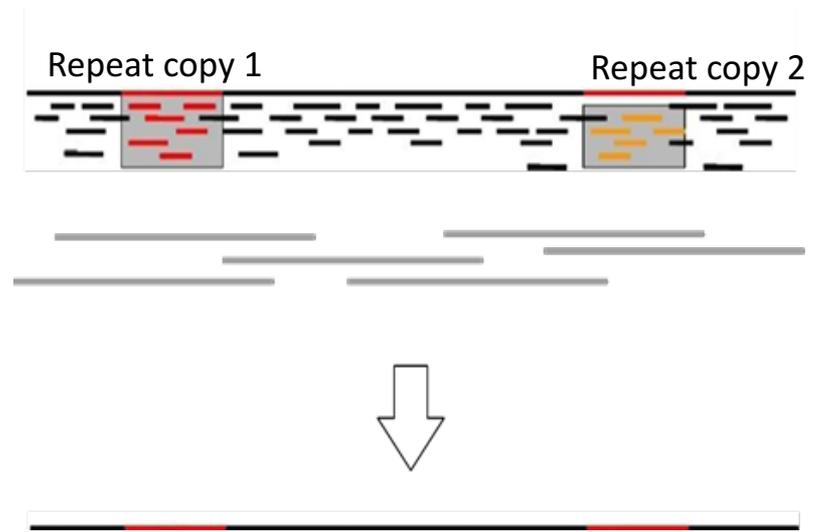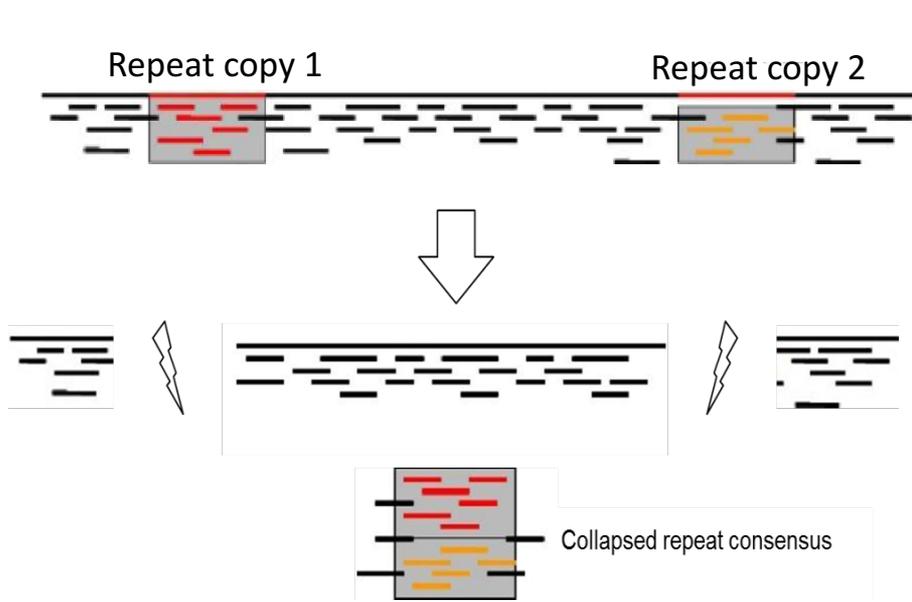
Long reads can phase haplotypes
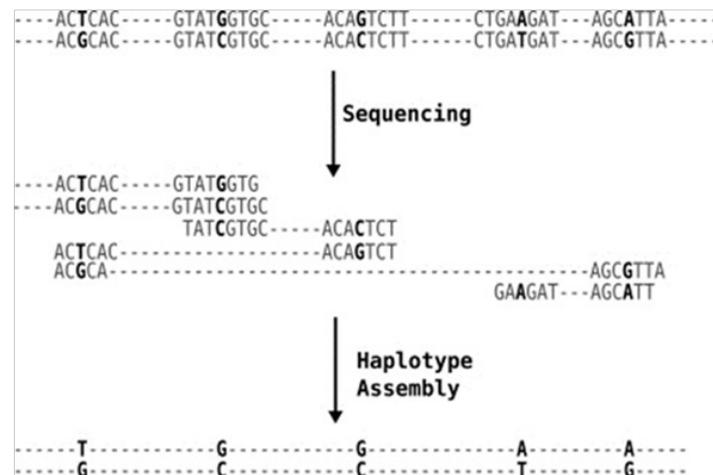
# 10X Genomics Linked-Reads

Long Range information from Short Length reads

1. Partition Long DNA fragments into individual GEMs
2. Fragment DNA once in the GEM
3. Append a GEM-specific barcode to each DNA fragment
4. Sequence short fragments
5. Use Barcode to link short reads back to larger input molecule
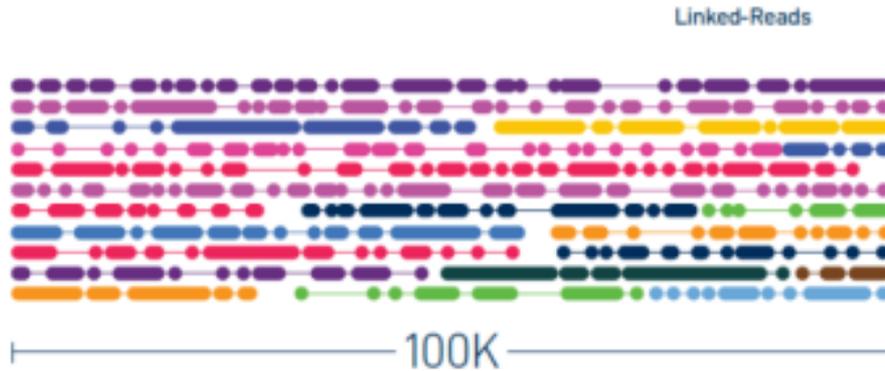
Physical vs molecule coverage

Each Long Fragment is sequenced sparsely (~0.2x coverage)

Long input molecule (50Kb)
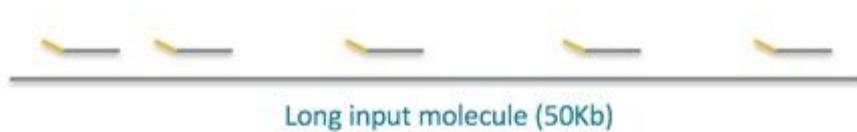
Trade - off between "Molecule Coverage" and "Read Coverage".   We may not see every base, but sparse coverage allows us to find overlap between different long read fragments. Lowers "dropout regions"

Sequencing cost ↑
Physical coverage ↓

Sequencing cost ≈
Physical coverage ↑

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®
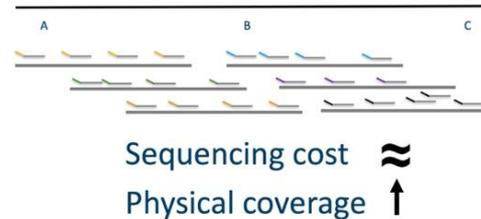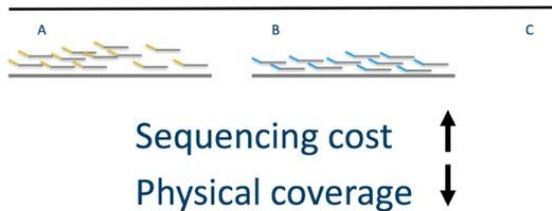
Physical vs molecule coverage
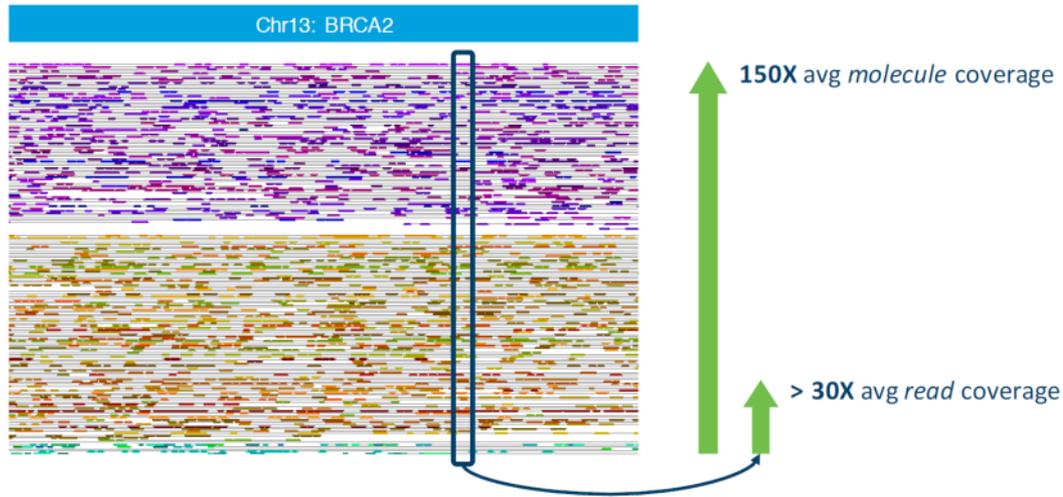
Example:
- 150 long molecules span region of interest (150x molecule coverage)
- Sequencing data available for 30 of those reads (30x physical coverage)

How do Linked Reads work?

- Locus Limiting dilution
  - We start with very dilute DNA, ~1ng
  - Each GEM has ~10 random input molecules
    - NOT 1:1!!!!
  - Goal: avoid getting >1 molecule covering the same locus to get the same barcode

Inside a GEM

150 genomes went into 1M partitions

Primers with Same Barcode
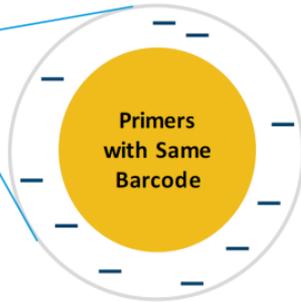
Each GEM contains:

- 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%

VAN ANDEL INSTITUTE

_ _0% TO RESEARCH, DISCOVERY & HOPE®

Practical input considerations

We want LONG DNA molecules – 50 KB min, 80-100 kb ideal

Phasing:template length
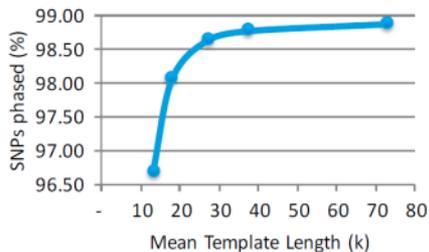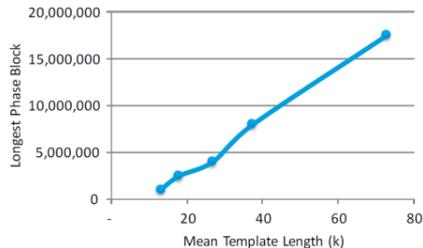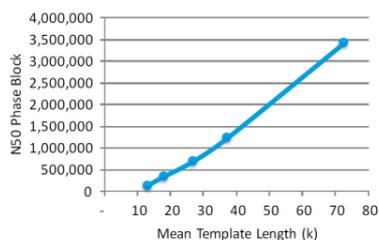
Assembly: template length
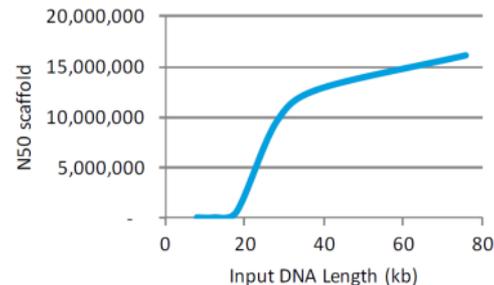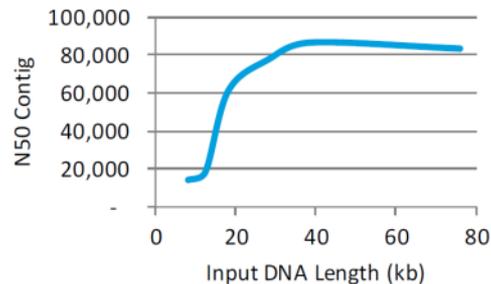
Input Considerations

- Recommendations:
  - Slow pipetting and genomic tips, minimal handling
  - Choose an extraction method meant to maintain long DNA fragments
  - Think about tissue/cell storage
    - Snap freezing, cryo-storage

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®

Inside GEMs – isothermal amplfication turns into short reads
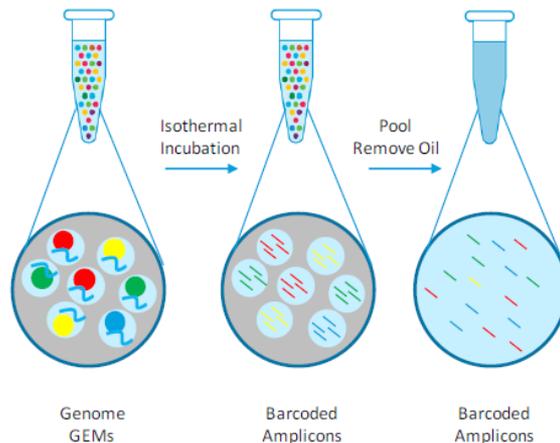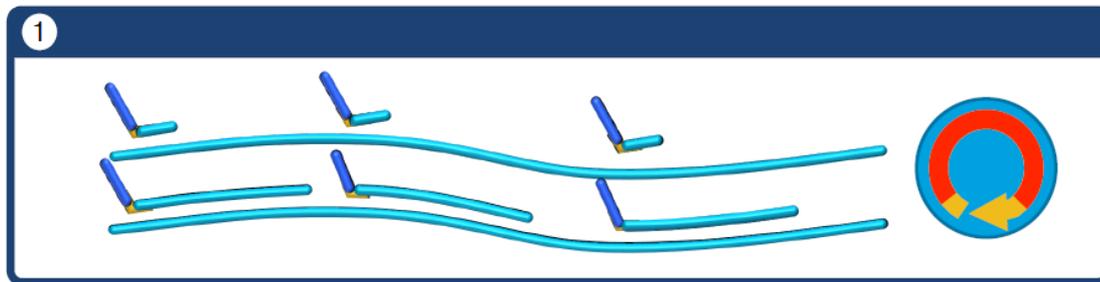
PostGEM library construction
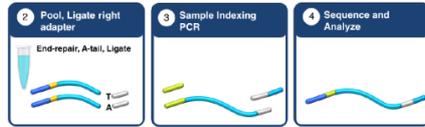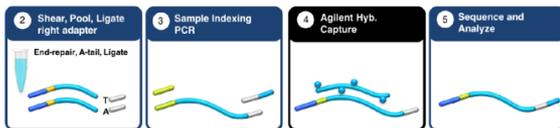
Recommend 38-56x raw coverage:
NextSeq: 2 flowcells, 2x150 bp
HiSeq4000: 3 lanes, 2x100 bp
NovaSeq: S2FC 1.5 lanes, S4 FC: ½ lane
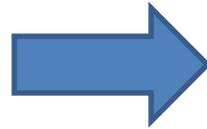
Chromium Software Suite

Powerful Linked-Reads analysis software
for a wide range of applications

BCL

Long Ranger
Analysis Pipelines

BAM

VCF

BEDPE

Standard Informatics
samtools, bcftools, bedtools

LOUPE

Loupe
Genome Browser

Supernova
De novo Assembly

FASTA
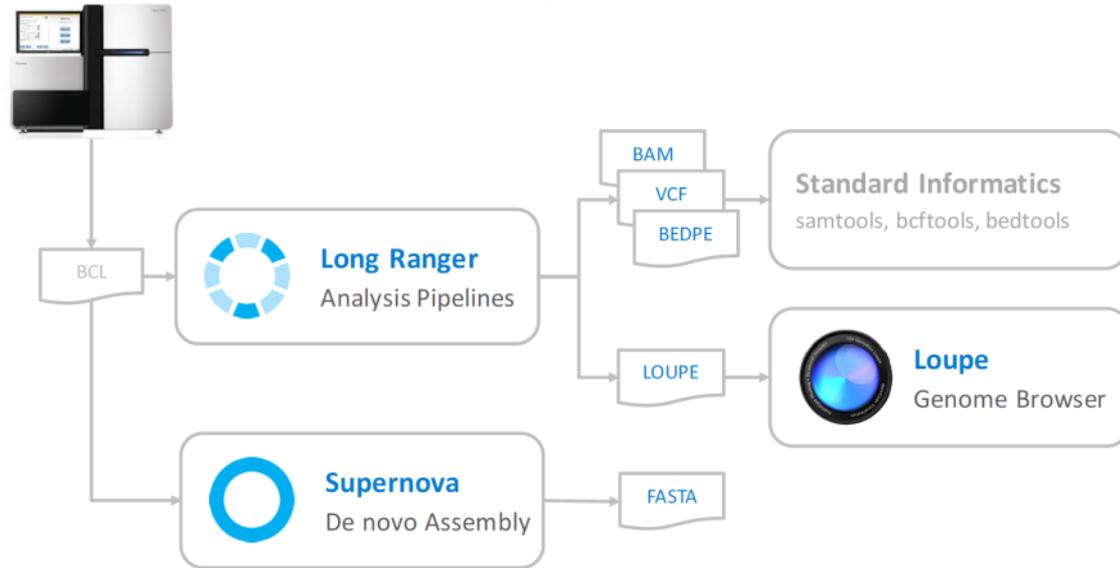
Analysis resource comparison

Single Box
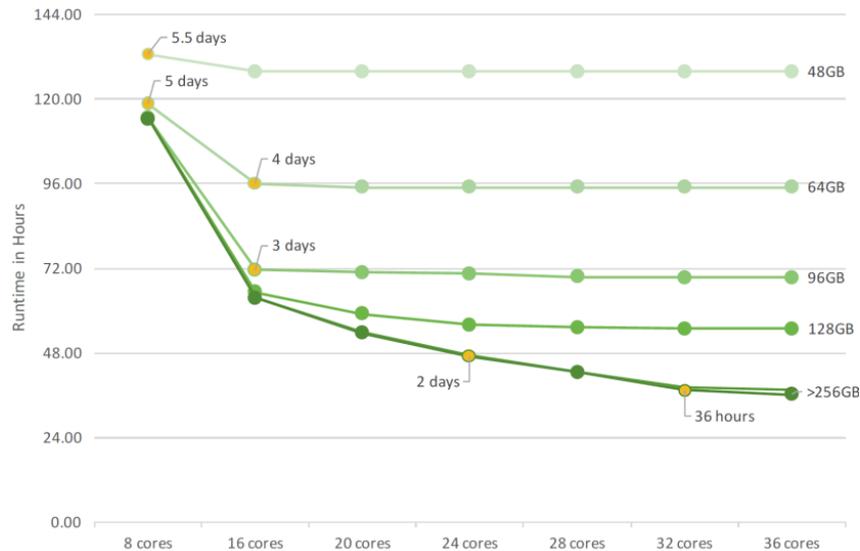
High Performance Cluster

L INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®
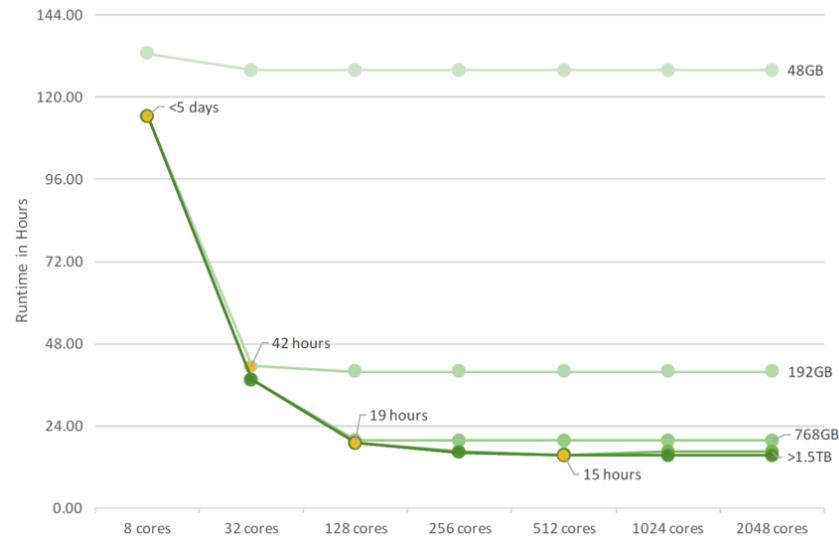
10X - Structural variation and haplotyping in psychiatric disorders

Psychiatric Disorder Sequencing Project Adds 10x Genomics Tech for Phasing, Structural Variant ID

Mar 03, 2017 | Monica Heger

⭐ *Premium*

NEW YORK (GenomeWeb) – The InPSYght project is about halfway towards its goal of sequencing 10,000 whole genomes in order to study schizophrenia and bipolar disorder, as well as to increase the amount of genomic data available from individuals of African ancestry. As part of the project, 500 genomes will also include linked read data generated using 10x Genomics' Chromium platform.

The main goals of including 10x Genomics data are to enable phasing of the genomes as well as to improve structural variant detection, Chris Whelan, a computational biologist at the Broad Institute, said in an interview. Whelan also discussed a 50-whole-genome pilot project last month during a presentation at the Advances in Genome Biology and Technology conference in Hollywood Beach, Florida.

The National Institute of Mental Health funded the InPSYght project in 2014 with a $16 million grant to researchers from the Broad Institute, the University

Project Goals:
- WGS for 10k schizophrenia, bipolar and control samples.
- Identify variants associated with disease
- Reference panel of structural events

Differential Coverage

- 322 protein coding genes in regions covered by linked reads in the majority of our samples but not in standard libraries

MAPQ zero alignments in standard data

10X Data

Standard PCR-free Data

STRC Gene

Example deletion

2033bp deletion

10X Data

Haplotype 1

Haplotype 2

Unphased

PCR-free Standard Data

Mapping Quality 0 Reads

Segmental Duplication containing GOLGA8B
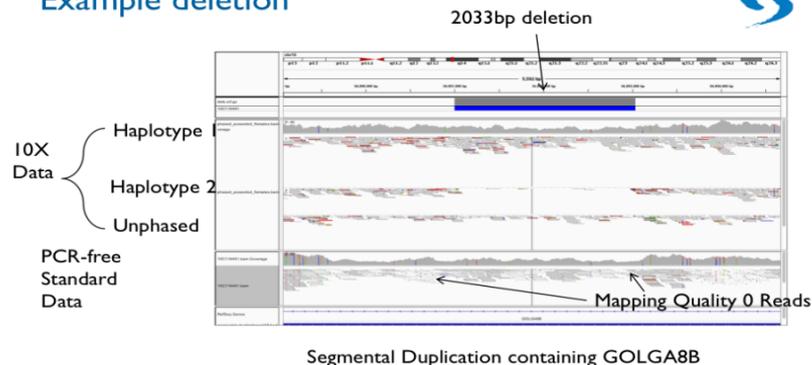
3' single cell sequencing
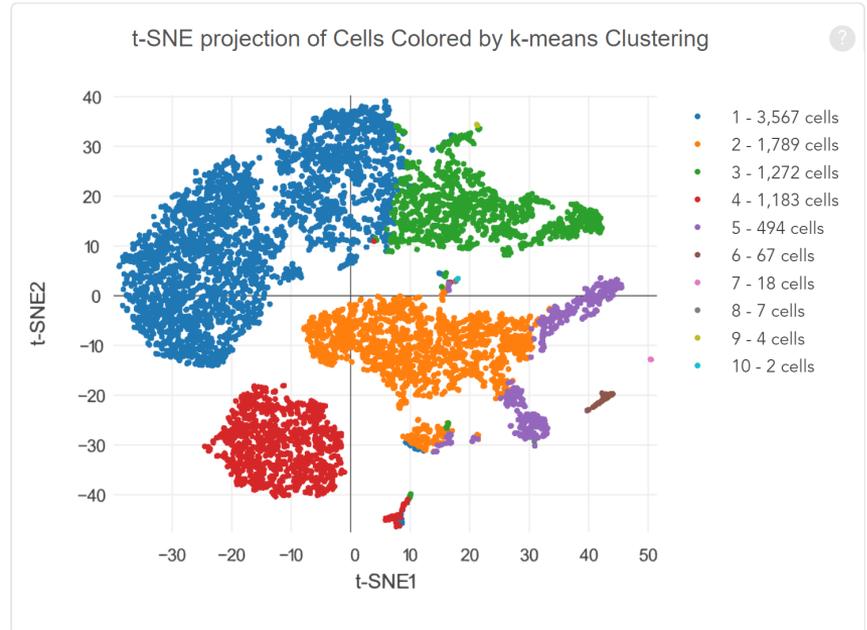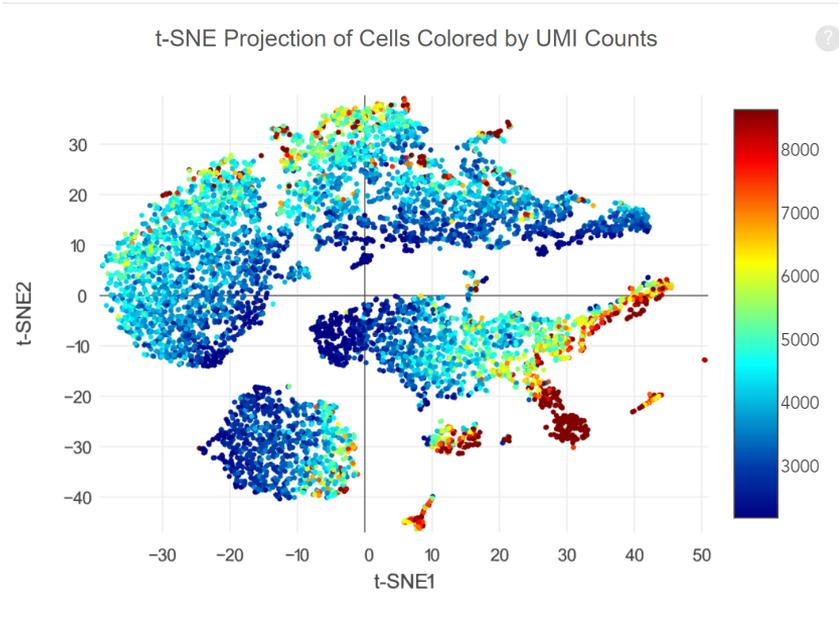
Why use single cell?

a) Deconvolving heterogeneous cell populations

(heterogeneous tissue or tumor)

dimensionality reduction (e.g. PCA)

Component 2 / Component 1
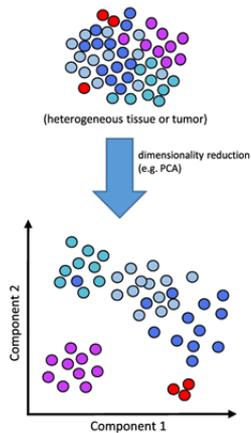
b) Trajectory analysis of cell state transitions

lineage A
lineage B

t_0 t_1 t_2

(cell differentiation, or response to stimulus)

trajectory analysis pipeline (e.g. Monocle, Wanderlust)

trajectory B
trajectory A

Component 2 / Component 1
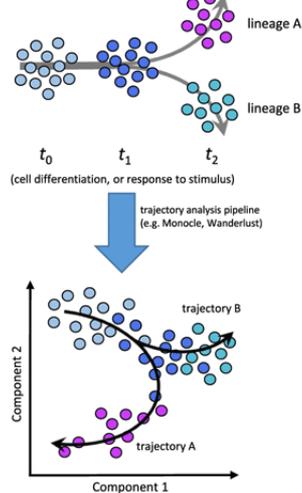
c) Dissecting transcription mechanics

Gene transcription "off"

RNA polymerase disassociated from gene

RNA polymerase bound and transcribing gene

Gene transcription "on"

(transcriptional bursting and stochastic gene expression)

d) Network inference

Cells
Genes

module 1
module 2
module 3

Low   High

(identifying modules of co-regulated genes)

network inference

(inference of gene regulatory networks/subnetworks)
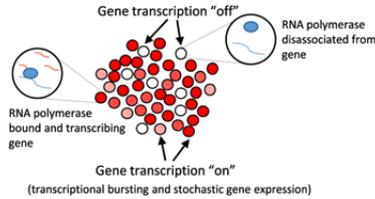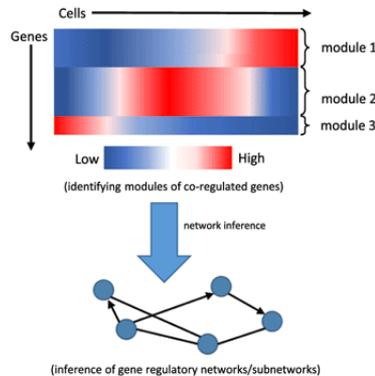
- Transcript heterogenetity
- Novel cell type discovery
- New markers for known cells
- Cell type separation
  - Not all cell types/species have defining flow cytometry markers!
- Developmental pathway & cellular trajectory analysis
- Cell signalling and Cellular networks

Liu S and Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges [version 1]. F1000Research 2016, 5:182 (doi: 10.12688/f1000research.7223.1)

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®

Single cell prep considerations

- We've got live cells, people!! Treat them gently, but not too gently
  - Single cell preparation is vital to avoid clogs– trypsinize tissues, strain to remove debris
  - Conversely non-cell mRNA can contaiminate GEMs. Wash cells with PBS+BSA
- Quantitate accurately – should be the last thing you do before loading
  - 65% of cells recovered in GEMs
  - Under/over loading = non optimal sequencing depth
  - Count cells at rough concentration of 700-1400 cells/mL for best accuracy
- Cell Viability - >85% live cells
  - Miltenyi live/dead kit to increase live cell %
  - Dead cells take up GEMs too, and can confound results
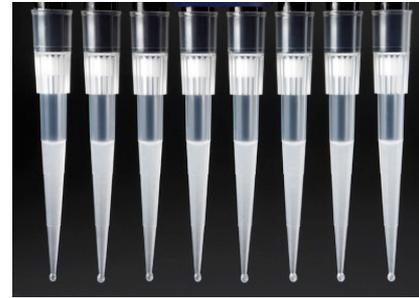
Successful emulsions
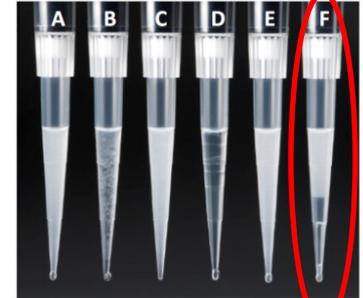
Emulsions from poor cell prep

Clog ☹

Live/Dead Staining of Final Suspension

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®
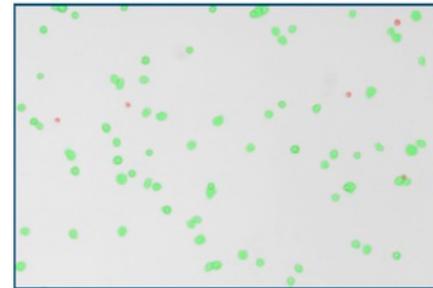
Single cell prep considerations

- Cell Size
  - 5-30 uM, preferably round
  - No cell walls. Sorry plant folks.
- What types of cells have been validated?
  - PBMCs, neurons, Jurkat, Raji, 293T cells, *young* cardiomyocytes
  - Nuclei (brand new protocol!)
- Cell Ranges
  - Wide dynamic range: 0.5ng-220 ng RNA
  - Equivalent to:
    - 870-17,800 cells input;
    - 500-10,000 cells output
- How do I get my cells to VAI?
  - Want to keep on ice, and load ASAP – ideally 30 min, not more than 2hrs
  - Some types of cells are stickier than others – we're looking at you PBMCs!!!
  - Cryopreservation protocols are available

PBMC E18 neuron 293T HCC38

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®
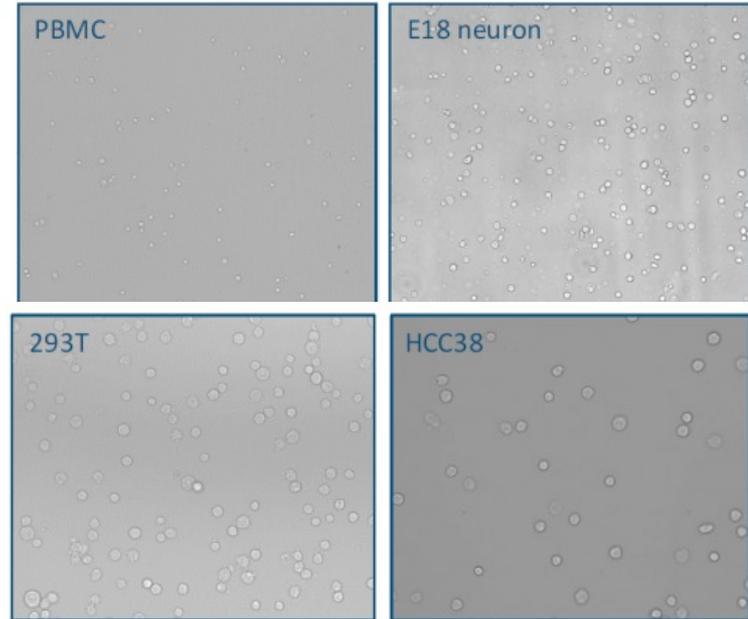
scRNA is not always single cell

User controlled trade-off between cell numbers and doublet rate

| Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~870	~500
~0.8%	~1700	~1000
~2.3%	~5300	~3000
~3.9%	~8700	~5000
~7.6%	~17400	~10000
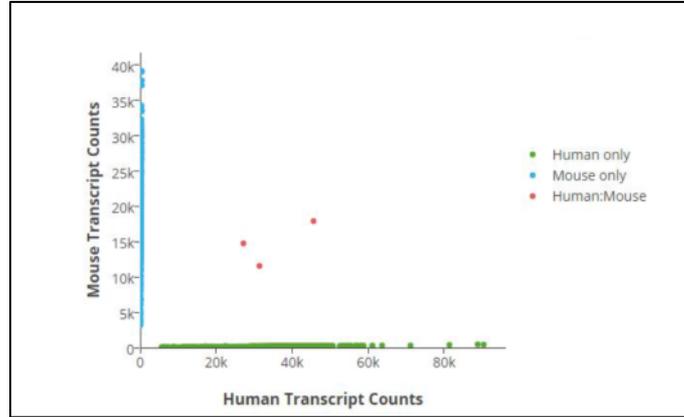
Small proportion of cells will have doublets
- Shown here as mouse:human transcript mixes
   - 8 mouse:human cells out of 1500
- Cell number:doublet rate is linear
   - Multiplet rate 1700 cells = 0.8%
   - Multiplet rate 17000 cells = 7.6%

scRNA: inside the GEM

Lib Prep details

Sequencing details

R1: 26
10XBC + UMI

P5 | Read 1 | 10x Barcode | UMI | Poly(dT)VN | Read 2: 98* Insert | Read 2 | i7: 8 Sample Index | P7
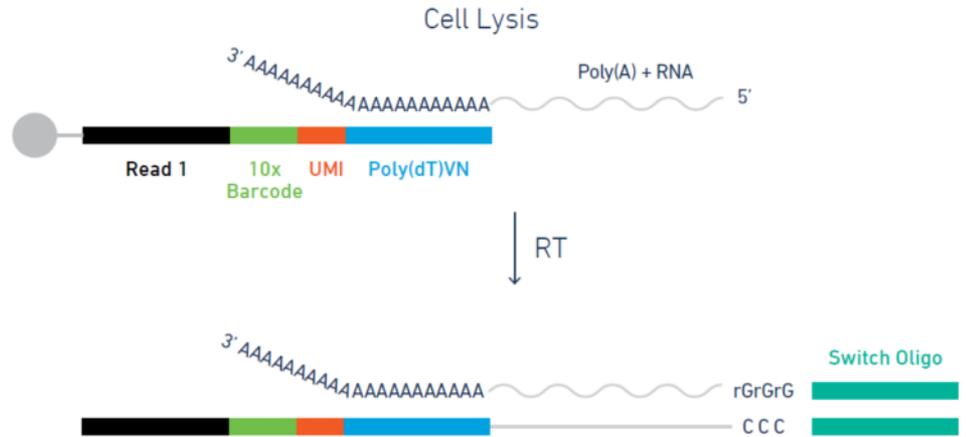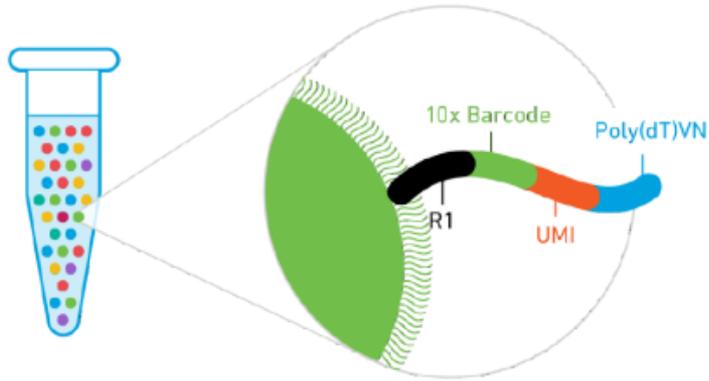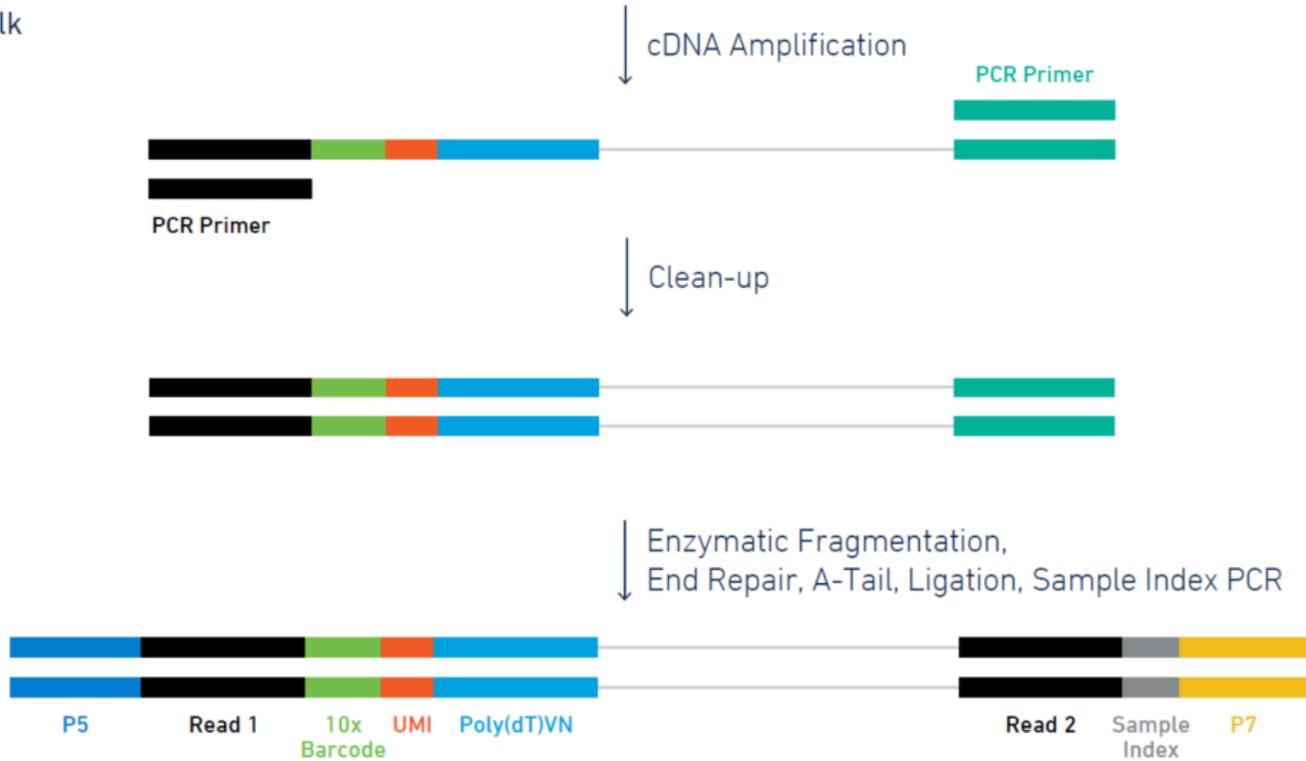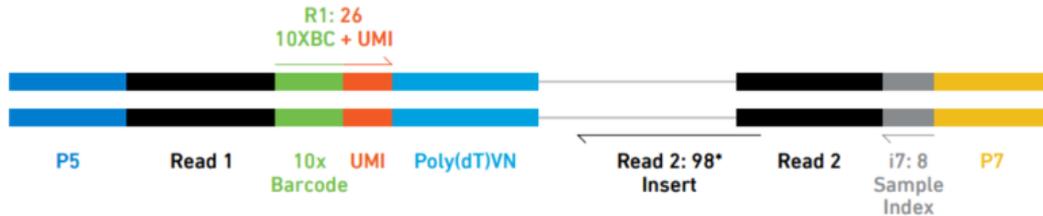
Sequencing Read	Description	Recommended Number of Cycles	Notes
Read 1	10x Barcode Read (Cell) and Randomer Read (UMI)	26 cycles	It cannot be shorter than 26 bp If longer than 26 bp it will be ignored by CellRanger
i7 Index	Sample Index Read	8 cycles	If longer than 8 bp it will need to be trimmed during basecalling
i5 Index	N/A	0 cycles	N/A
Read 2	Insert Read (Transcript)	98 cycles	Can be adjusted*

*User controlled trade-off between read length and mapping rate

Sequencers:
NextSeq/MiSeq:
Compatible

HiSeq/NovaSeq:
TALK WITH YOUR CORE!
- Due to weird read length configuration hard to mix on flowcell w/ "standard libraries"

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®

Sequencing depth

- 50K raw reads/cell is minimum recommendation
- Cell type dependent - 90K raw reads/cell tends to be start of "limiting returns" for more complex cells
- Note – this is still a LOT of sequencing! 50k raw reads x 4000 cells = 200M reads, 90K x 4000 = 360M reads

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®
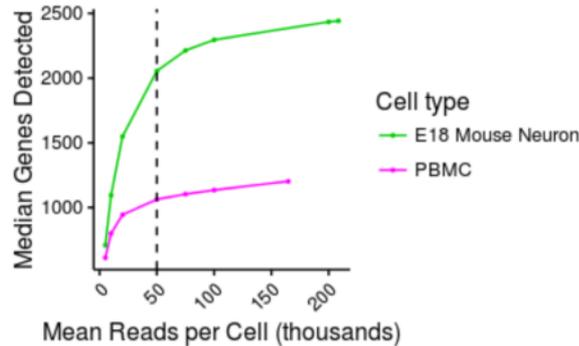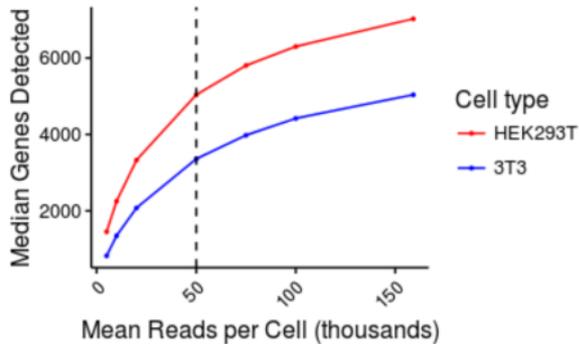
scRNA analysis- Cell Ranger Pipeline

Runs on Linux, builds on existing algorithms – eg STAR alignment pipeline
Processing intensive – 50 core hours/100M reads, one project can take 90-100 core hours
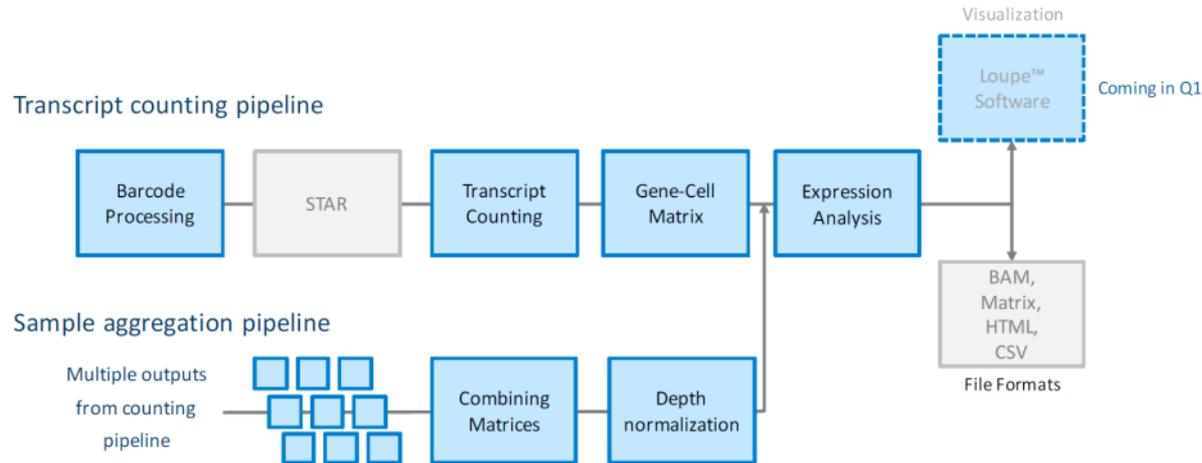
scRNA pipeline – defining a "cell"

- Sum UMI counts for each barcode
  - Barcode: must be on a static list, 1 mismatch allowed
  - UMI: must be non-homopolymer, non N, base Q scores >10
- Retain only barcodes with total UMI count >10% of 99th percentile of total expected cells

VAN ANDEL
INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®
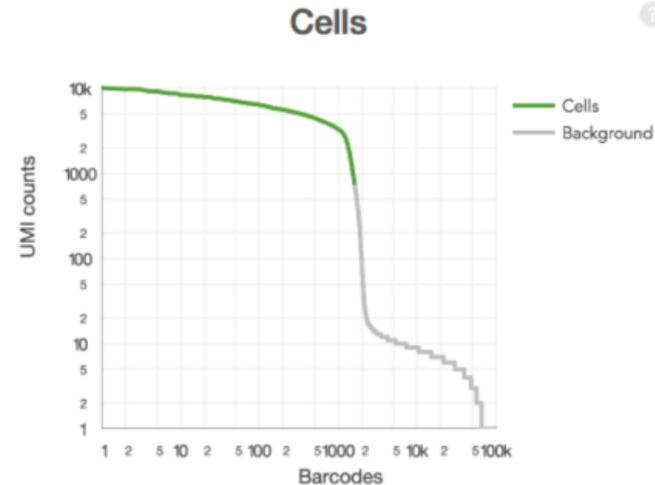
scRNA defined pre/post transplant cell populations

"we developed a computational method to distinguish donor from host cells in bone marrow transplant samples by genotype.....to compare subpopulation changes in acute myeloid lekemia (AML) patients" before and after transplant.

Massively parallel digital transcriptional profiling of single cells

Grace X. Y. Zheng, Jessica M. Terry […] Jason H. Bielas ✉

Altmetric: 168   Citations: 27   More detail »

Article | OPEN
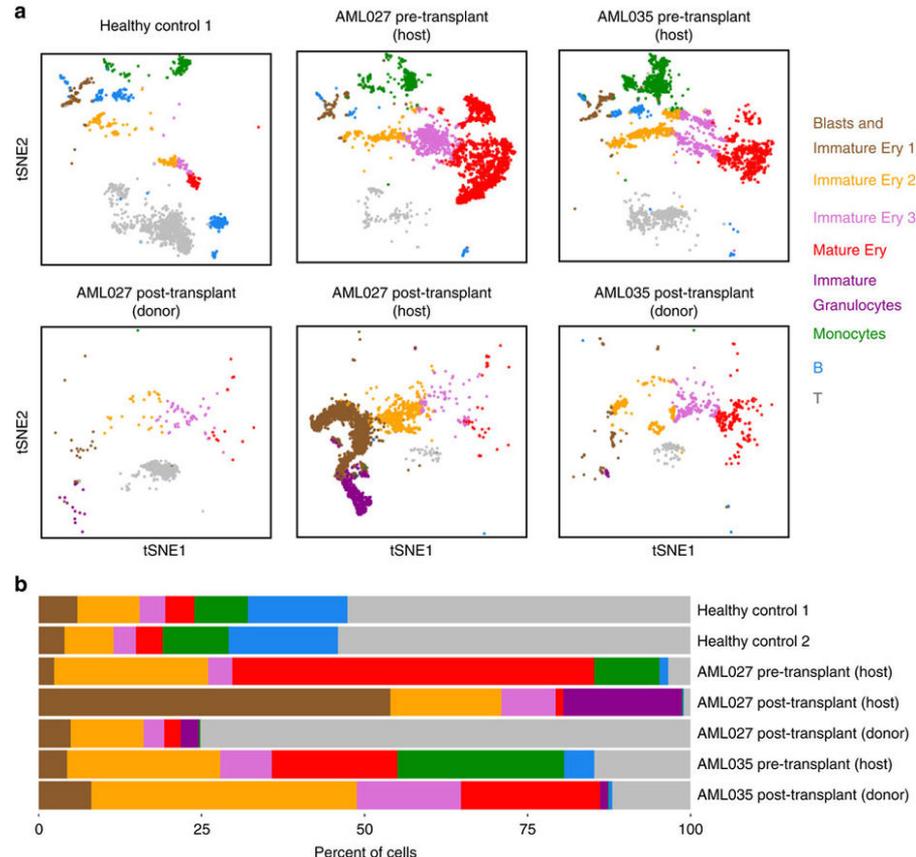
VAI Cores and Services

Genomics Core

https://genomicscore.vai.org/

- NGS Sequencing
- Illumina iScan arrays
- Cytogenetics

Other Cores:

- Bioinformatics and Biostatistics
- Confocal Microscopy & Quantitative Imaging
- Cryo-EM
- Flow Cytometry
- Pathology and Biorepository
- Small Animal Imaging
- Vivarium and Transgenics Imaging

Julie Koeman, ASCP

Mary Rhodes

VAN ANDEL INSTITUTE

100% TO RESEARCH, DISCOVERY & HOPE®