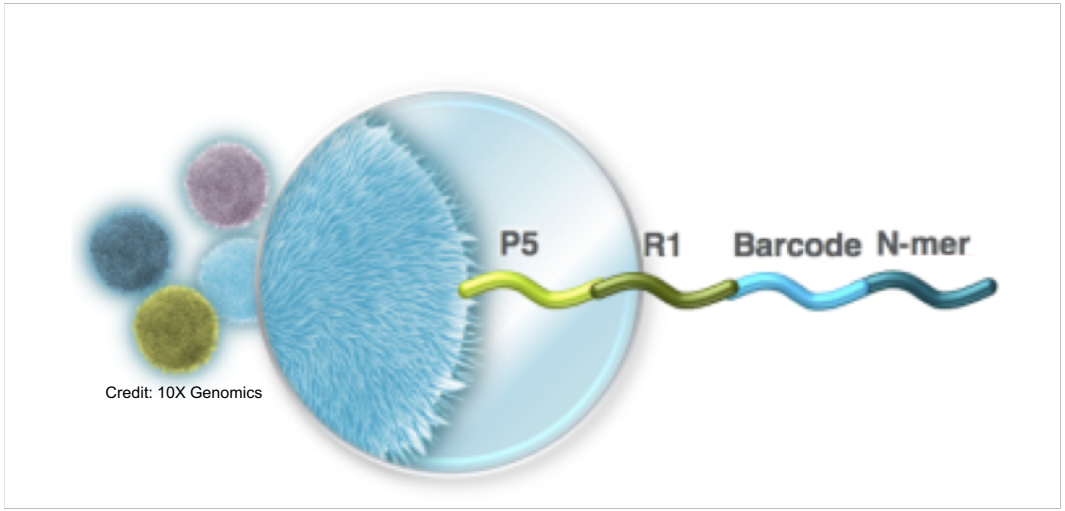
10X Genome Assembly Technology

and Single Cell CNV



Diana Burkart-Waco

DNA Technologies and Expression Analysis Cores

12-19-2018

10X Chromium Genome

linked read assembly

…providing *de novo* genome assembly, variant calling, and genome structure information…

- Upstream sample preparation
  - Sample QC guidelines
- 10X Chromium Genome
  - Technology
  - Applications
  - UC Davis projects
- NEW: Copy Number Variant kit

DNA Quality and Applications

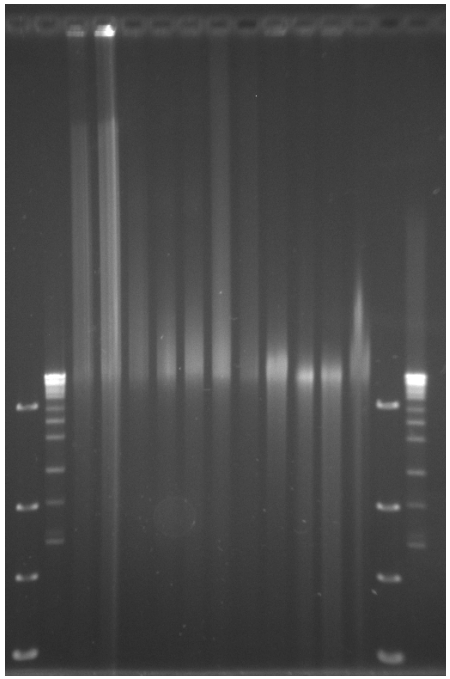
DNA Quality Level	DNA size (reported by 10x Chromium Genome pipeline)	Applications
5	>80 kb	De novo assembly with Supernova and Long Ranger analysis
4	60-80 kb	De novo assembly with Supernova and Long Ranger analysis
3	40-60 kb	Long Ranger analysis
2	20-40 kb	Long Ranger analysis ‡
1	<20 kb	Long Ranger analysis possible, performance not thoroughly characterized ‡

10X Technical note: “Single-stranded DNA Damage and its Effects on Chromium Genome Application Performance”

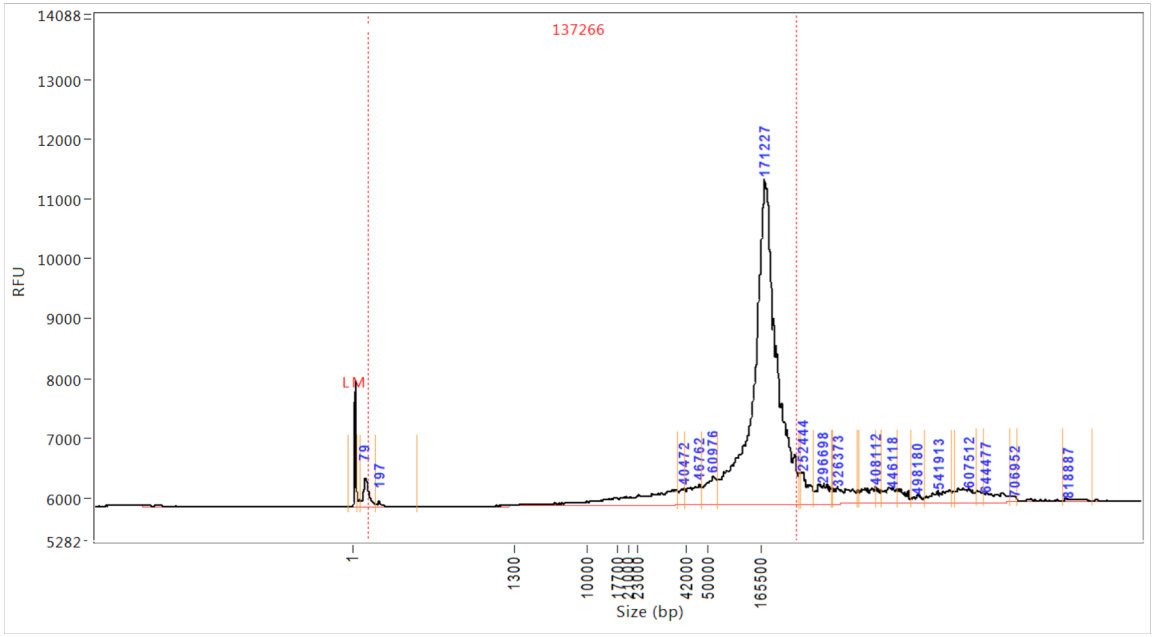
QC options

- Fragment analysis needed to determine size and degree of degradation.

➢Pulsed-Field Gel Electrophoresis

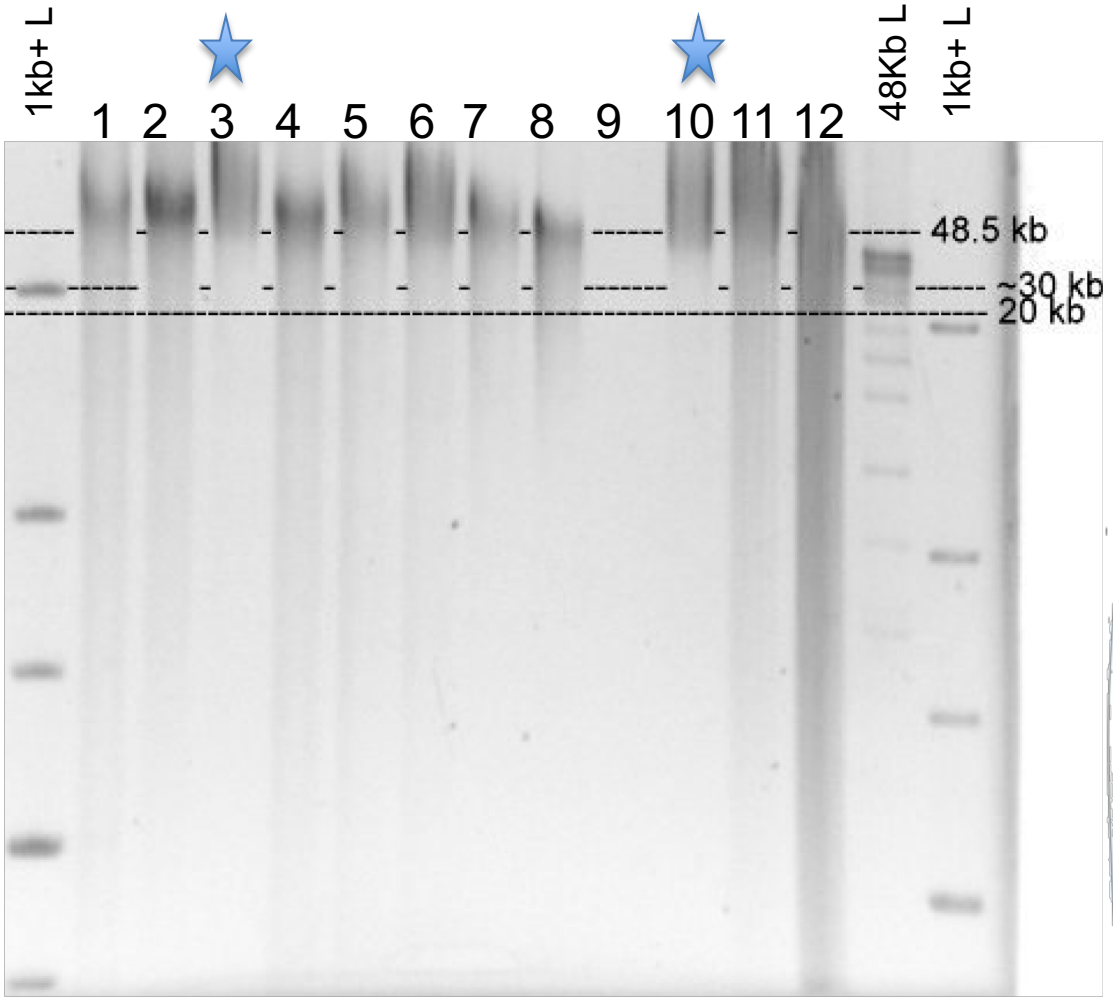


➢Femto pulse



HMW gDNA QC guidelines

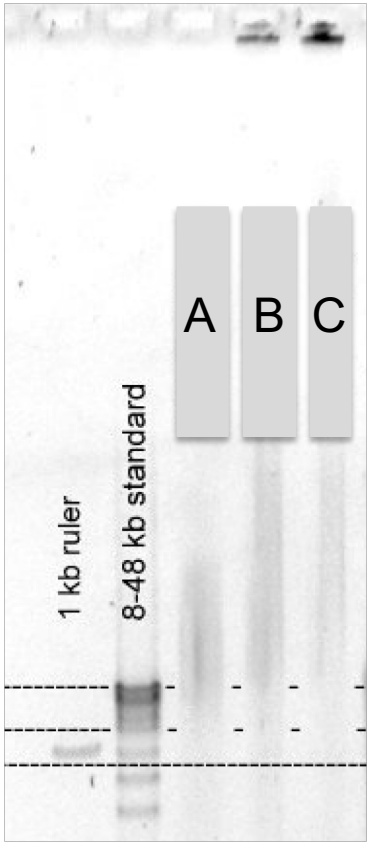
- Above 40kb!
- No smear below 20kb.
- Free of RNA, protein, and carbohydrates.
- Nanodrop ratio (2.0) for both 260/230 and 260/280.



0.75% gel run for 16hrs – Pippin Pulse (5-150kb)

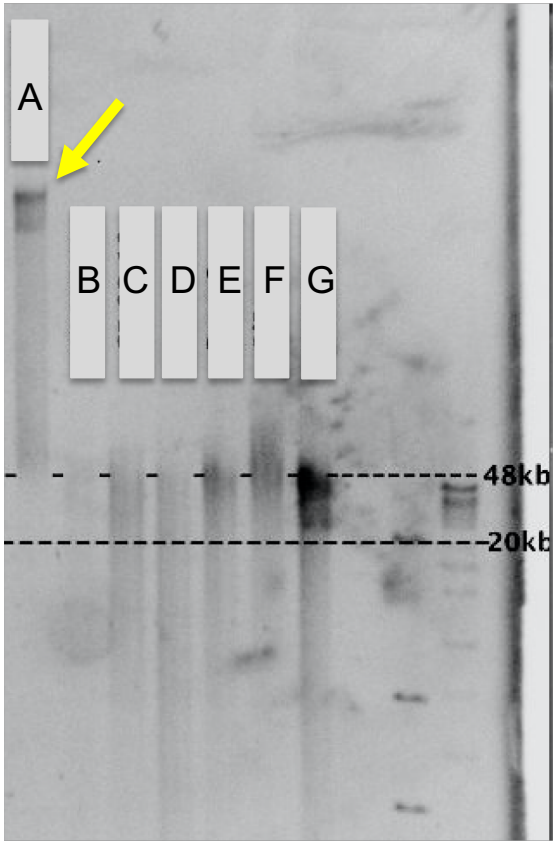
QC Examples

Example #1



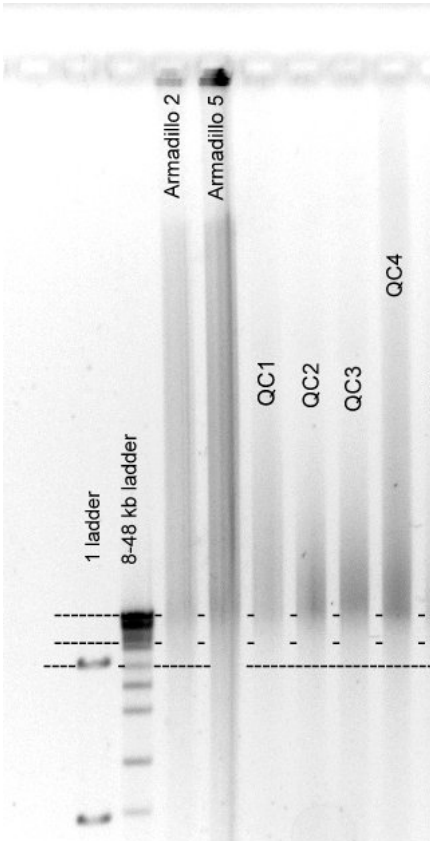
Look at loading wells.

Example #2



Bands are better than smear.

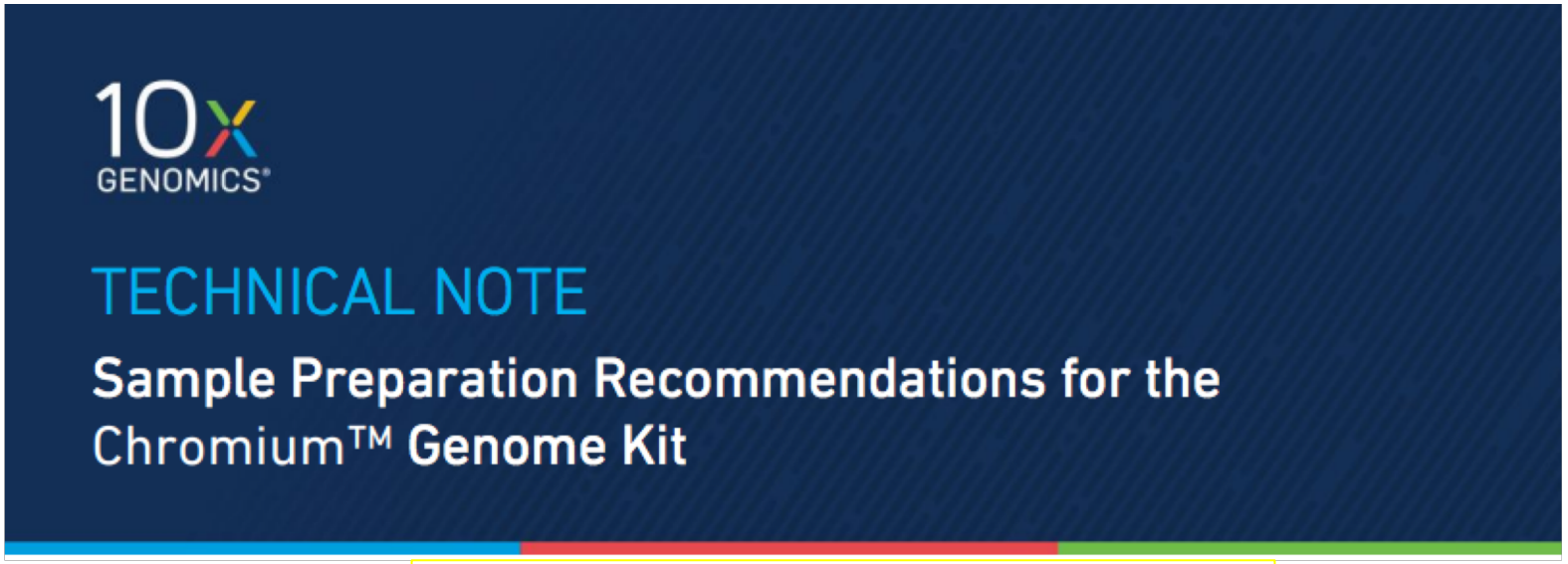
Example #3



Loading amount impacts QC.

Sample requirements

- Input into library prep 0.6ng-1.25ng.
  - Input depends on genome size.
- Additional 200 ng for QC.
- 40kb minimum, but 60kb better.
  - Don’t size select (new reco from us), DNA damage repair optional.



https://support.10xgenomics.com/

10X Chromium Genome

linked read assembly

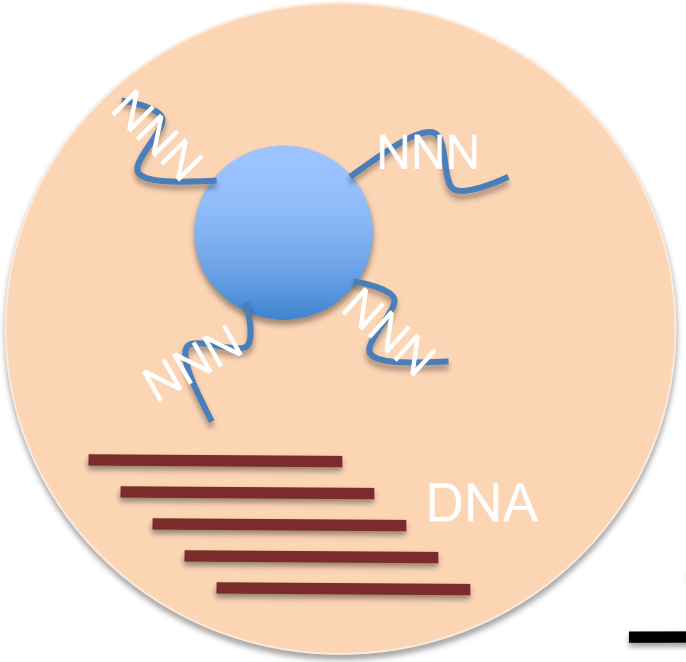
…providing *de novo* genome assembly, variant calling, and genome structure information…

- Upstream sample preparation
  - Sample QC guidelines
- 10X Chromium Genome
  - Technology
  - Applications
  - UC Davis projects
- NEW: Copy Number Variant kit

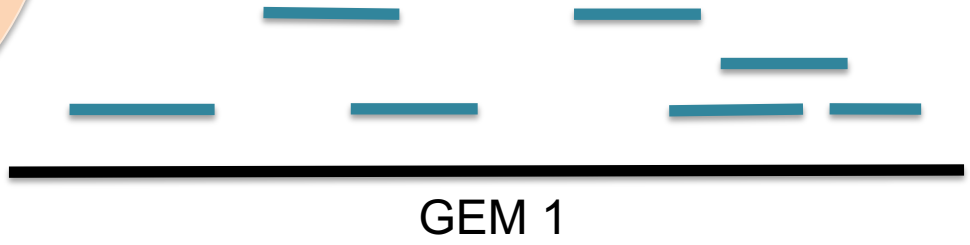
10X Genomics

(genomic DNA analysis, CNV, and SC)

GemCode technology



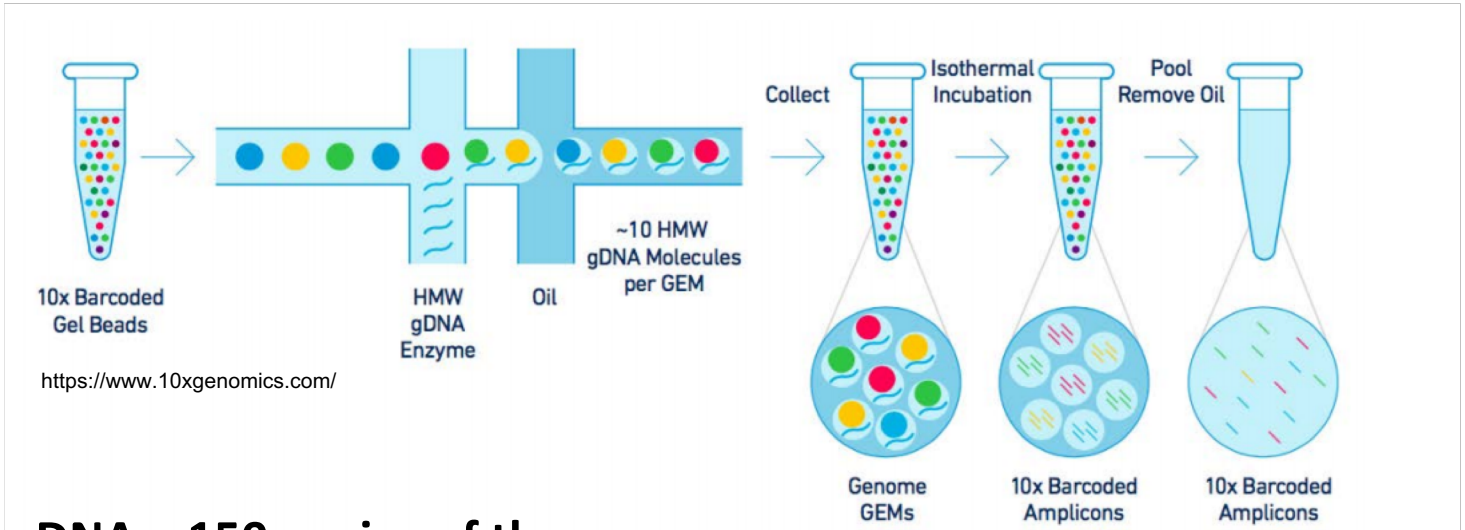
- Droplet-based technology. Subset of genome partitioned in oil droplets with beads with a millions of barcodes.



- Barcoded amplicons generated in gel beads provide building blocks of genome.

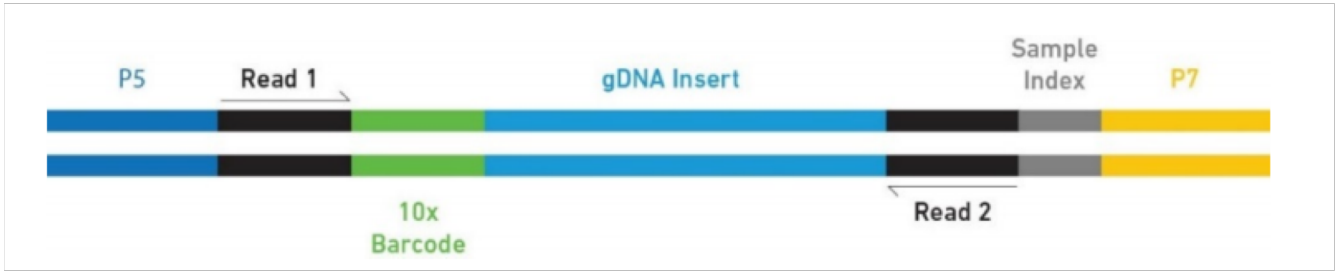
➢ **“Read clouds”: molecules inferred linked reads**

From gDNA to library

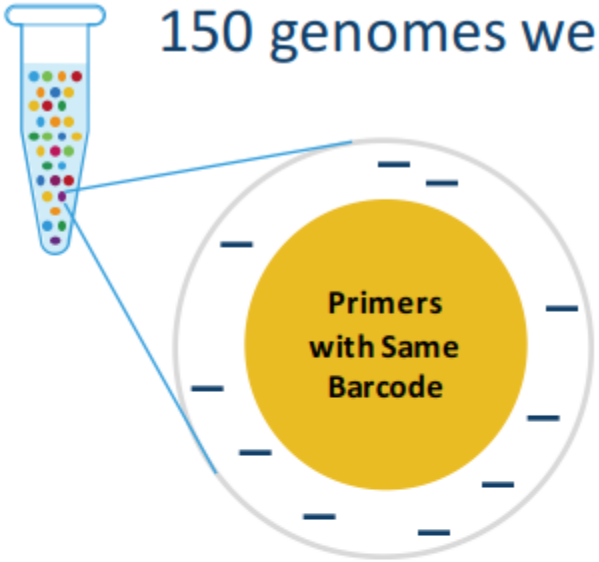


https://www.10xgenomics.com/

0.5ng DNA = 150 copies of the genome partitioned into ~1M GEMs.



Molecule partitioning – human



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:

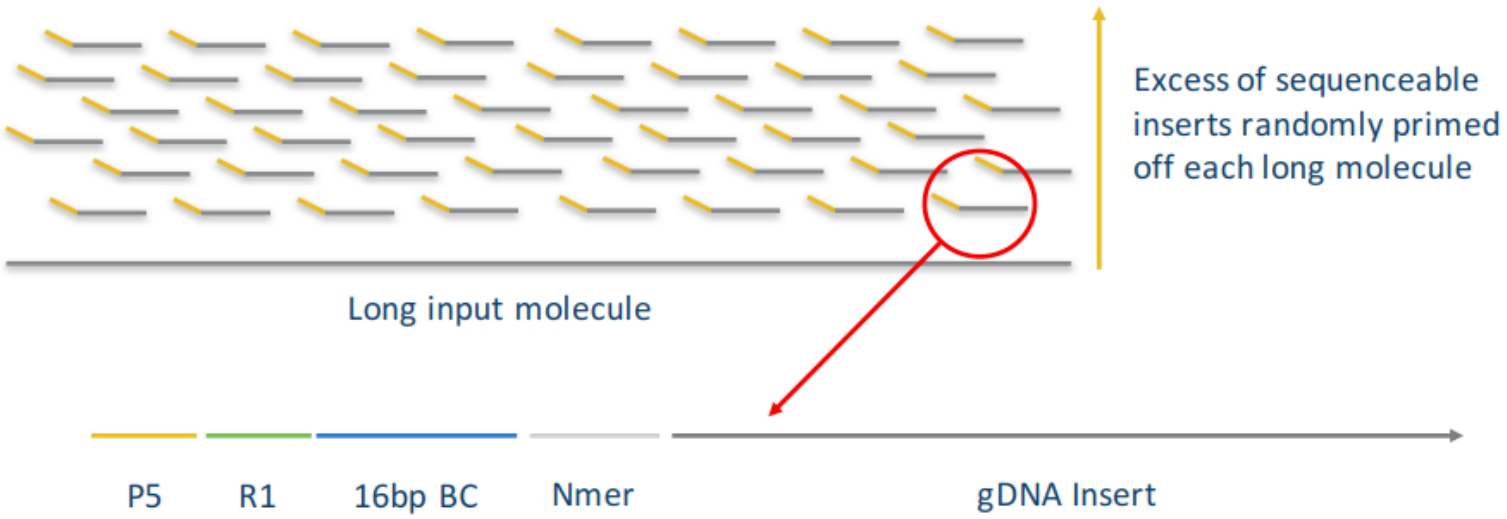
1 in 6000

Percent unique barcodes at any genomic locus:

99.98%

All graphics from 10X Genomics

Molecule coverage



At 30X read coverage, ~35 library fragments
will end up sequenced from each 50Kb input molecule

35 *x* 2 *x* 150bp ≈ 10Kb, or 0.2X read coverage per molecule

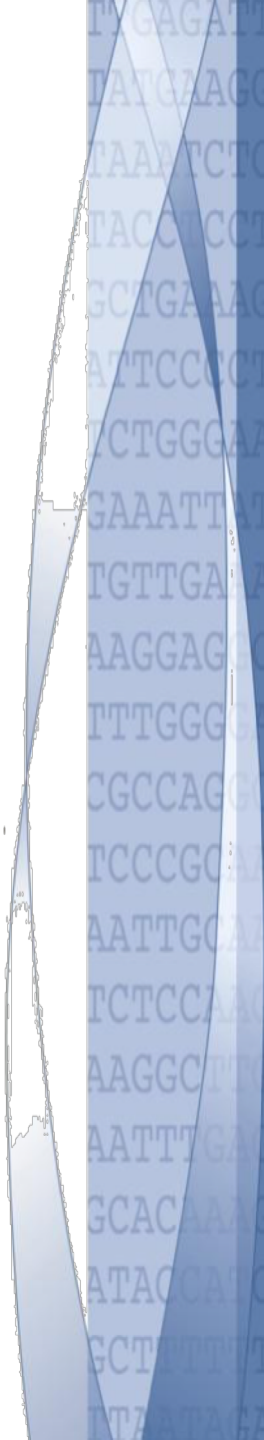
Reads from the same input molecule are called "Linked-Reads"

- Very little gDNA loaded into GEMs (some lost).
- Because so little gDNA added, unlikely that two haplotypes will have same barcode.

Read coverage recommendations

- Genome assembly: 60X coverage
- Structural variants: 25X coverage

- Too many reads doesn't improve assembly.
  - Worth running multiple assemblies with subsets of reads.



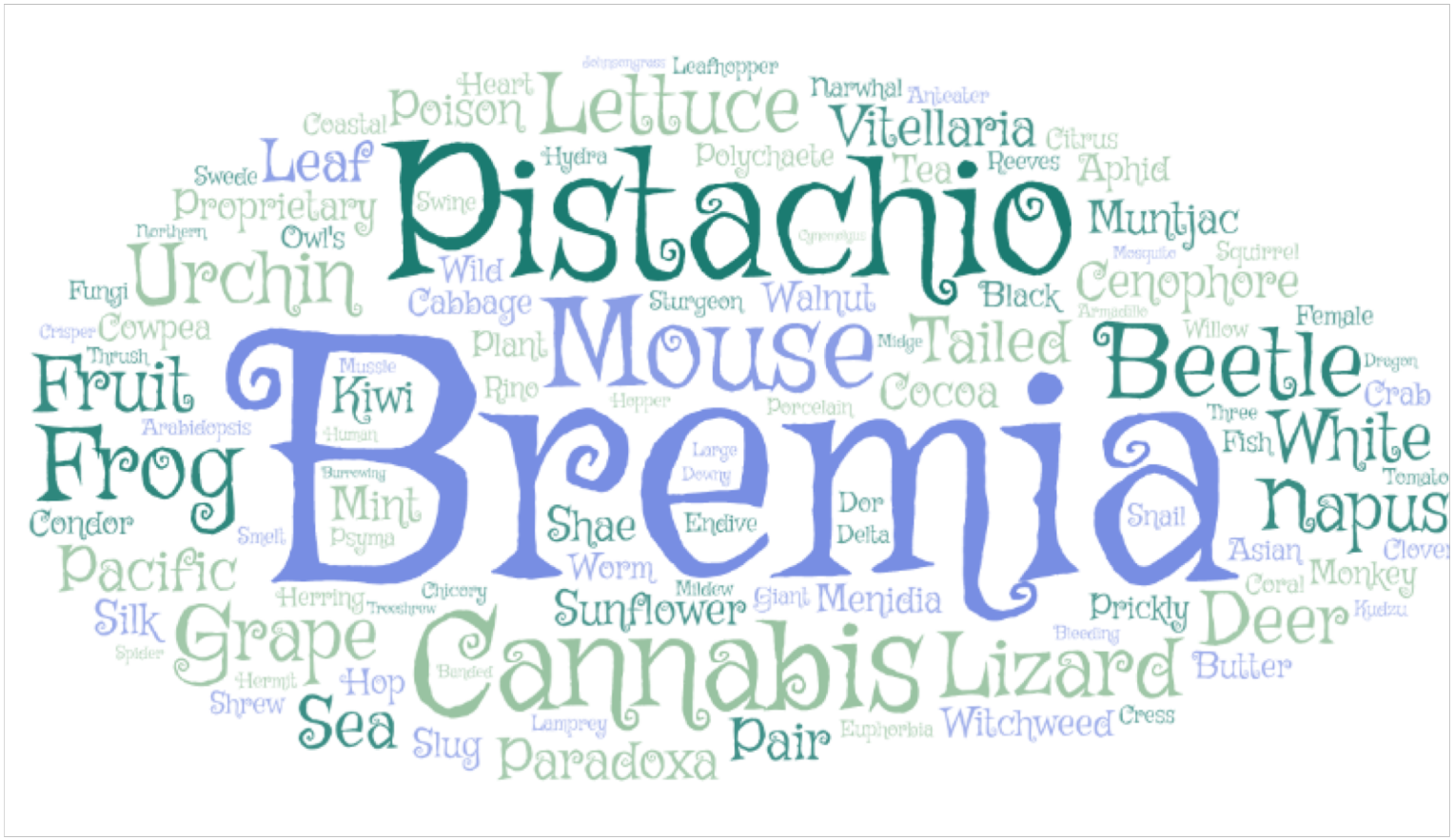
Structural variant detection

https://www.10xgenomics.com/

- Each colored line represents linked read.
- Linked reads used to infer alleles.

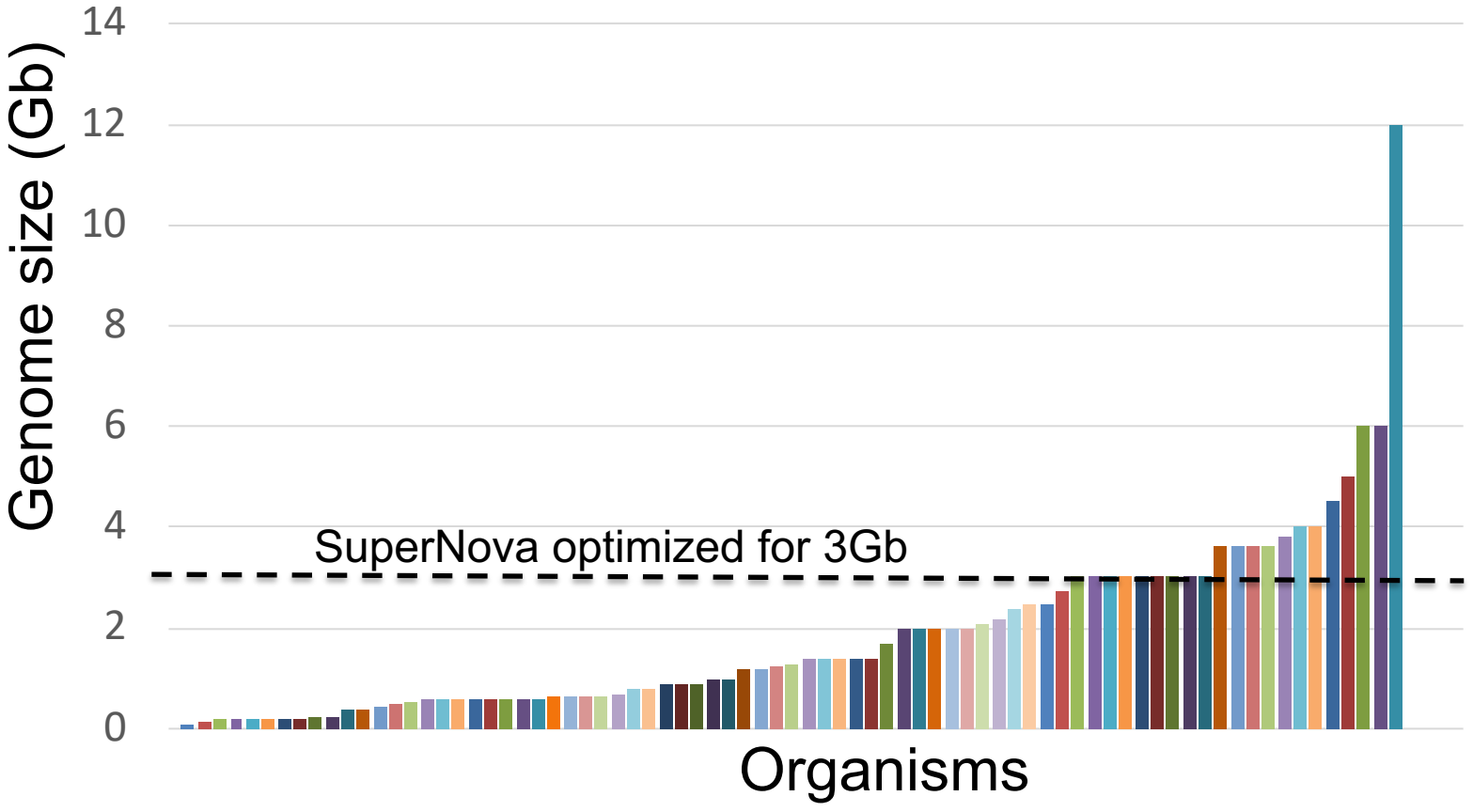
➢ 60 Kb deletion visible.

DNA Tech 10X Genome Assemblies



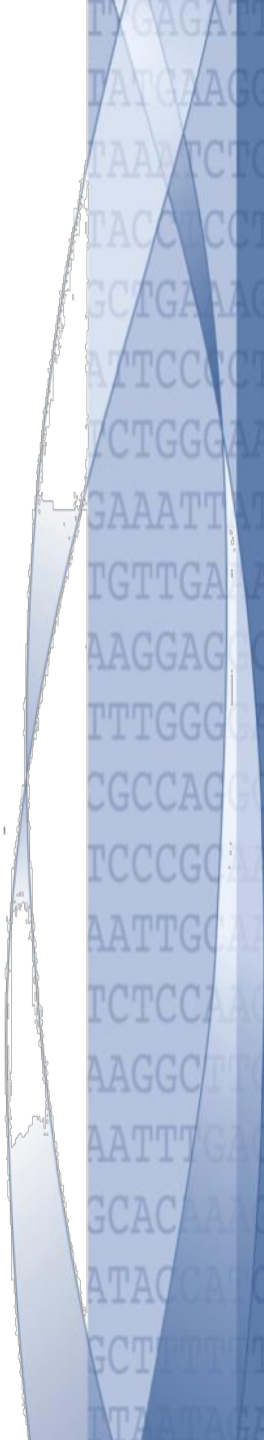
De novo genome assembly

- 120 genomes to date.
- Smallest genome: 78Mb (Oomycete)
- Largest genome: 12Gb (frog, way too big!)



Assembly Stats - Best

- Mammals, birds, and reptiles.
- Example #1 (3.01 Gb genome)
  - Assembly size: 2.49 Gb
  - Molecule length: 174.31 Kb
  - Contig N50: 334.53 Kb
  - Scaffold N50: 38.80 Mb (entire chromosome arms)
- Example #2 (3.00 Gb genome)
  - Assembly size: 2.3 Gb
  - Molecule length: 118.08 Kb
  - Contig N50: 87.32 Kb
  - Scaffold N50: 7.41 Mb



Assembly Stats - Suboptimal

- Insects, marine life, plants (variable)
  - Depends on genome architecture, gut contents, metabolites, heterozygosity / variant density, ploidy.
- Example #1 (400 Mb genome)
  - Assembly size: 200 Mb
  - Molecule length: 13.42 Kb
  - Contig N50: 13.86 Kb
  - Scaffold N50: 40 Kb
- Example #2 (790 Mb genome)
  - Assembly size: 369.98 Mb
  - Molecule length: 64.70 Kb
  - Contig N50: 16.60 Kb
  - Scaffold N50: 90.45 Kb

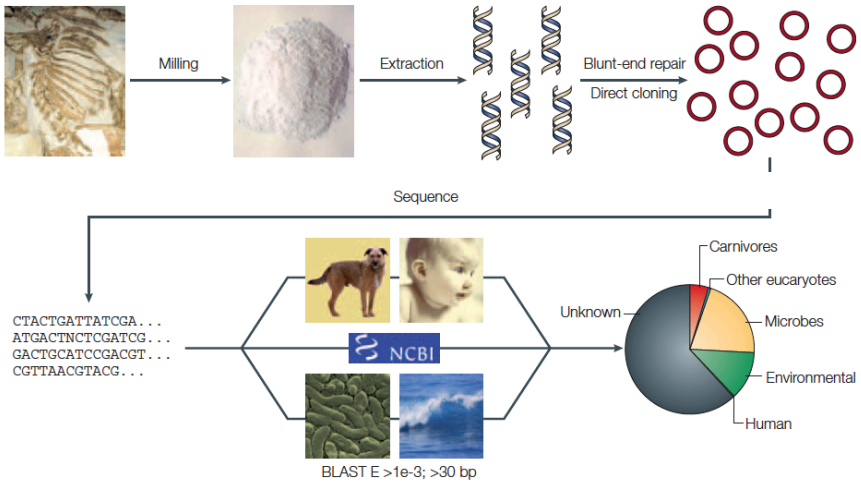
10X Chromium Genome

linked read assembly

…providing *de novo* genome assembly, variant calling, and genome structure information…

- Upstream sample preparation
  - Sample QC guidelines
- 10X Chromium Genome
  - Technology
  - Applications
  - UC Davis projects
- NEW: Copy Number Variant kit

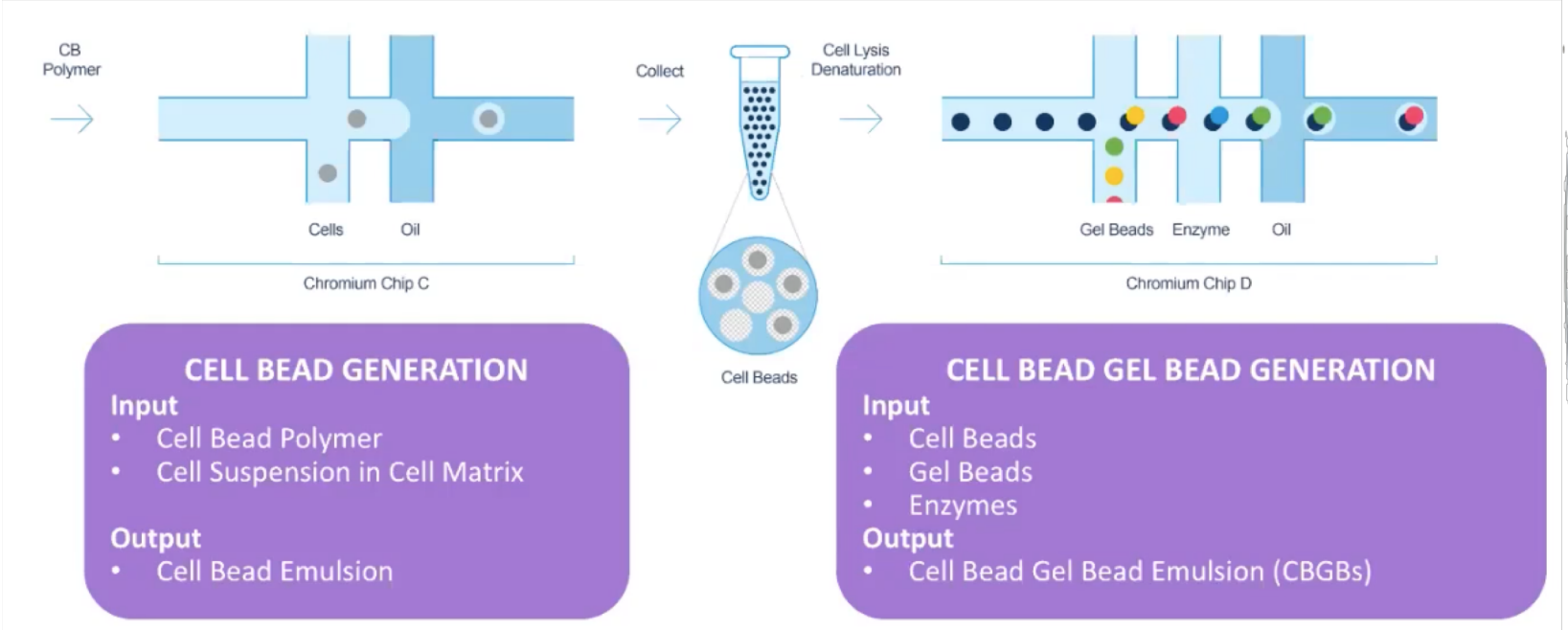
Summary of 10X Genome



- 10X great option if you are a human, bird, lizard, or diploid.
- Max genome size = 7.5 Gb / 2.14 B reads.
- 120 *de novo* genomes in core with linked reads.
  - High N50 = >300 Kb. Low N50 = 8 Kb (DNA damage).
- Plants are risky, but can still provide better assemblies.

Copy Number Variation

- Capture 100-1000s of single cell → copy number information.
- Calls single cell (or nuclei) CNV at 2 Mb resolution.
- Important tool to study dosage imbalances → changes in traits.
  - CNVs determine phenotypes more than SNPs.



PACIFIC BIOSCIENCES™

http://pacificbiosciences.com

- Read long molecules in real-time with polymerase.
- Very long reads.
  - Subread N50: up to 35kb.
  - Polymerase read length: up to 100Kb for CCS.
  - Yield: up to 50 Gb for CCS.
  - High error rate for raw data (~13%), but random (unlike Nanopore).

Iso-Seq Pacbio

- Sequence full length transcripts
  - Using TeloPrime protocol for mostly full length transcripts.
  - No assembly required.
- High accuracy – CCS data.
- More than 95% of genes show alternate splicing.
- On average more than 5 isoforms/gene.
- Precise delineation of transcript isoforms.
  (PCR artifacts? chimeras?).
- Ideal for gene annotation.

Please contact Oanh Nguyen (ohnguyen@ucdavis.edu)

Post Short Read Assemblies

- The future of sequencing is longer and longer reads.
  - Price dropping significantly.
  - Do 10X first because cheap?
- If 10X alone doesn't work, use combined assemblies (PacBio + 10X + Hi-C).
- Even suboptimal 10X data can be used for scaffolding with ARKS.
- Focusing on high molecular weight DNA can help obtain longer read lengths.
  - Junk in is junk out.
- But now we have to figure out how to use these data!

Price List – UC Rate

custom projects

- 10X Genome
  - Library prep: $918.
  - Sequencing: $1,500 for each 1.5Gb genome (NovaSeq, PE150).
- HMW gDNA extraction
  - Labor: $792 (plants, 1-4 samples)
  - Reagents: $100 per sample.
- 10X Single Cell CNV
  - TBD. But currently $$$$, but looking for testers.
- PromethION
  - $2,880 per experiment (library prep and sequencing).
- Hi-C
  - $1,690 (library prep only) + 100 million reads per 1.0 Gb genome (HiSeq4000 PE150).

Thank you!

"Safety first"

"Davis smog days"

From left to right:

Lutz – *Core Director*
Oanh – *PacBio*
Siranoosh – *HiSeq4000, MiSeq, and smallRNA*
Vanessa – *MiSeq, Genotyping*
Emily – *Library prep*
Ruta – *Nanopore, HMW gDNA extraction, Hi-C*