GENOME RESEARCH

CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2017 May; 27(5): 757–767.
doi: 10.1101/gr.214874.116

PMCID: PMC5411770

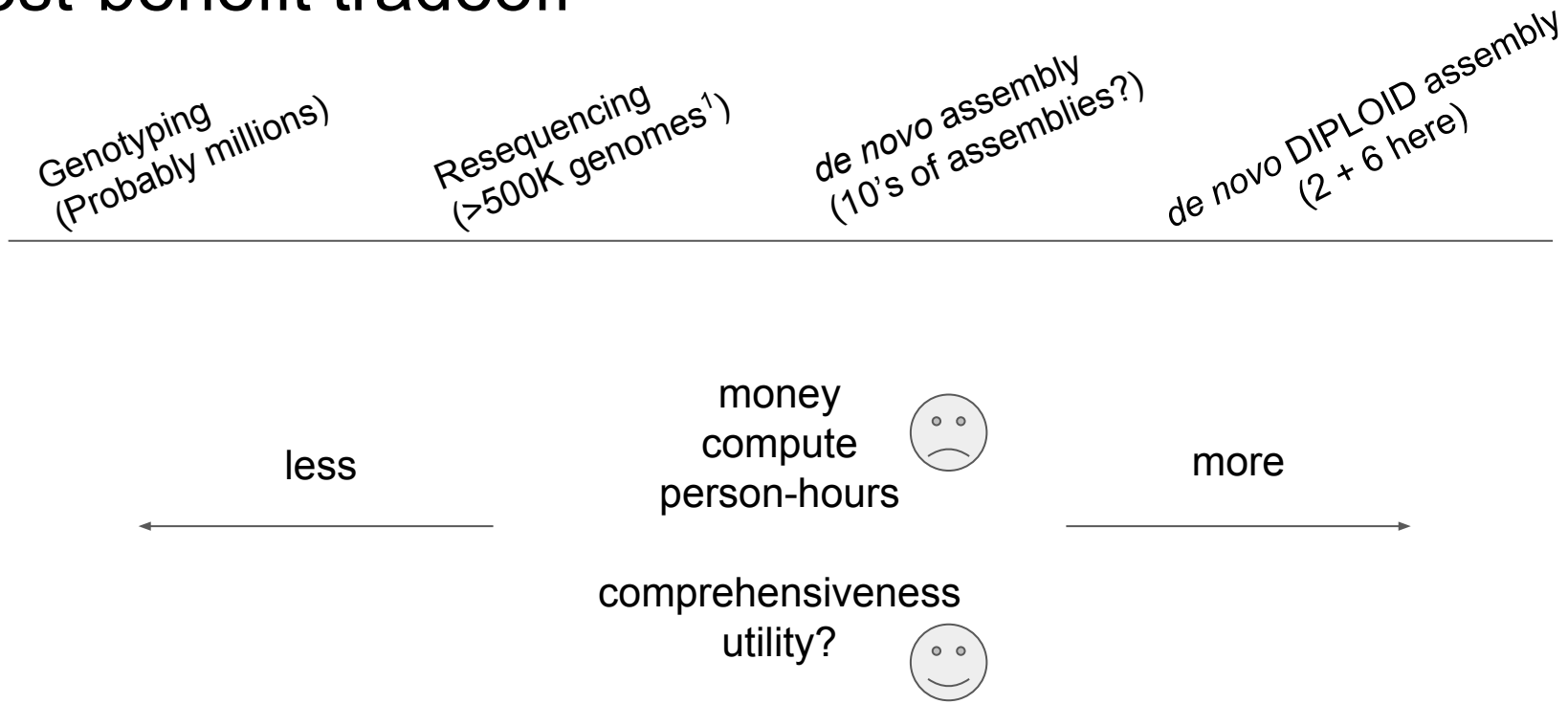
Direct determination of diploid genome sequences

Neil I. Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe

10X GENOMICS®

https://vimeo.com/120429438

Data acquisition for genome analysis is a constant cost-benefit tradeoff



1. https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet

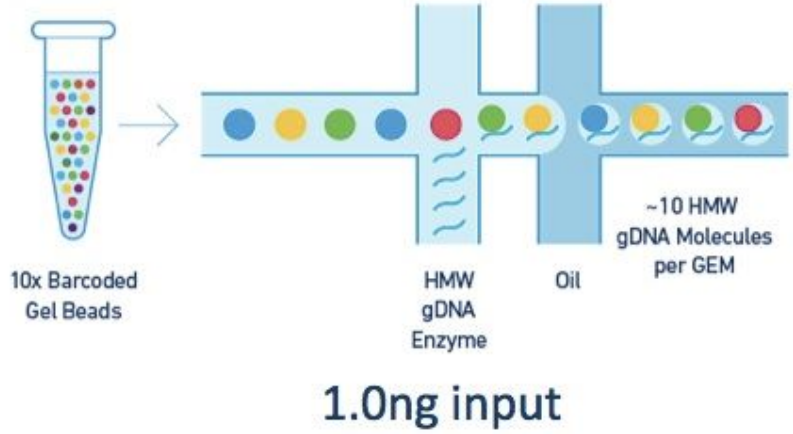
Thesis

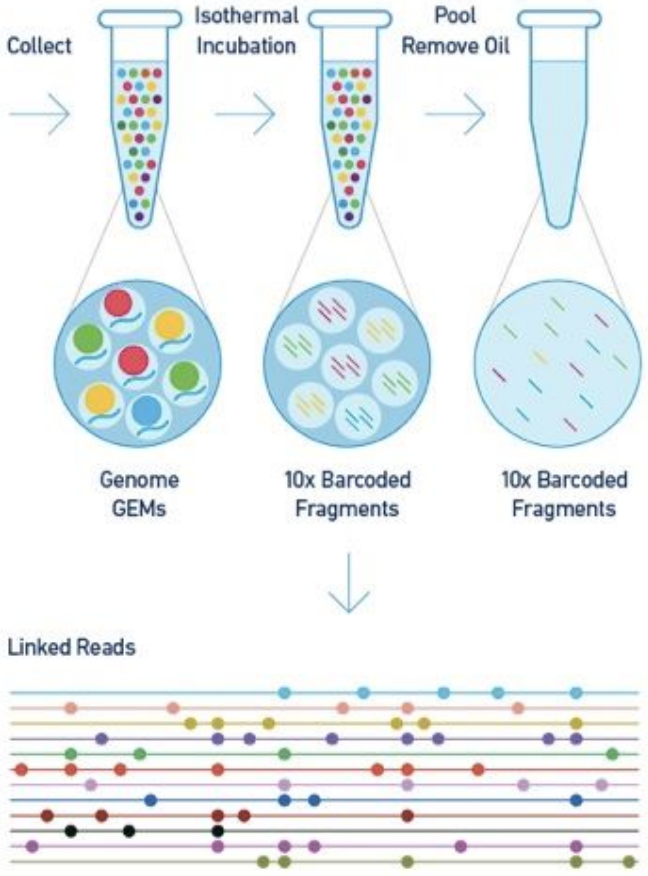
“Our method provides a scalable capability for determining the actual diploid genome sequence in a sample, opening the door to new approaches in genomic biology and medicine.”

It’s not yet clear exactly how much more you can do with a diploid genome, than say variant calls or even phased variants, but now that diploid *de novo* assemblies are possible, this can be studied

Decreased cost of a *de novo* assembly with competitive quality metrics should not be underestimated either - although 10X not ideal for non-human, progress is possible on that front

10X Genomics' linked-reads may be an effective compromise





An automated microfluidic system allows for functionalized gel beads to be combined with high molecular weight DNA (HMW gDNA) and oil to form a 'Gel Bead in Emulsion (GEM)'. Each GEM contains ~10 molecules of HMW gDNA and primers with unique barcodes. Isothermal incubation allows for the addition of a unique barcode to all DNA within the GEM, and barcoded fragments (typically ~350 bp) are generated. These fragments are then sequenced on an Illumina instrument, yielding groups of reads, organized by barcode, called Linked-Reads.

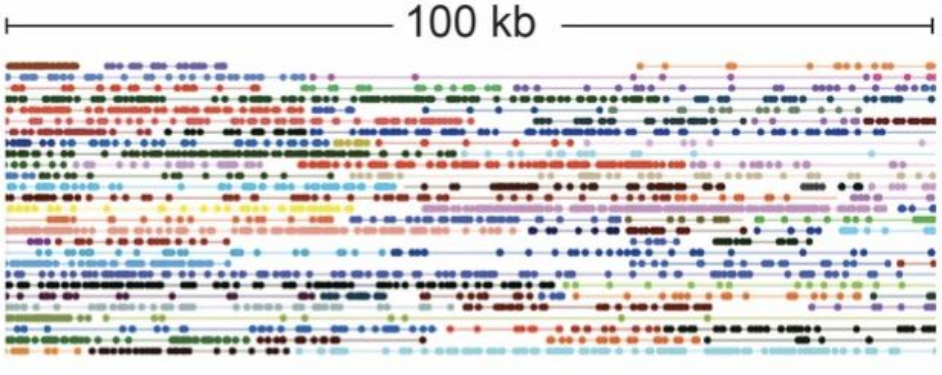
Rough calculations of the experimental parameters

1.25 ng DNA

10^6 partitions (GEMs)

50kb molecules *

10 molecules per partition



56X genome-wide coverage (1200M reads)

60 reads per molecule = **0.36X linked-read coverage**

* They used >90kb, and 20-50kb was much worse

AFAIK, Supernova is the only 10X / linked-read assembly algorithm available

48-mer de Bruijn assembly graph from single reads (Discovar)

Retain heterozygous loci as “small bubbles” and gaps

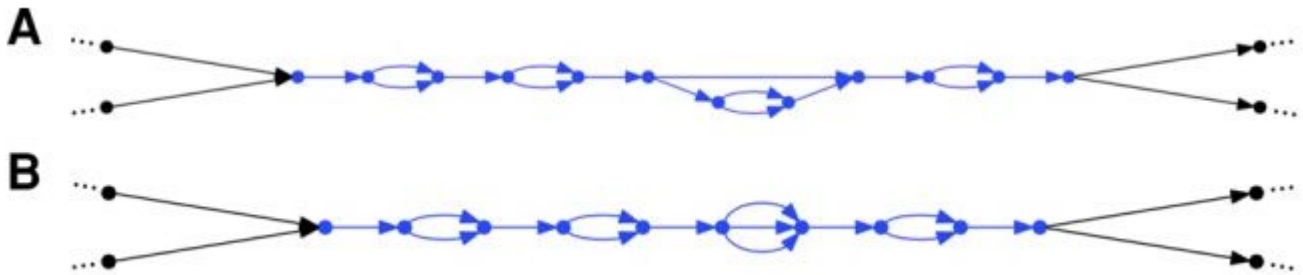
Use read pairs to scaffold “lines” from the dBG

Use barcode groups to scaffold further into “megabubbles”

Local assembly of all reads from molecules that originated from a region

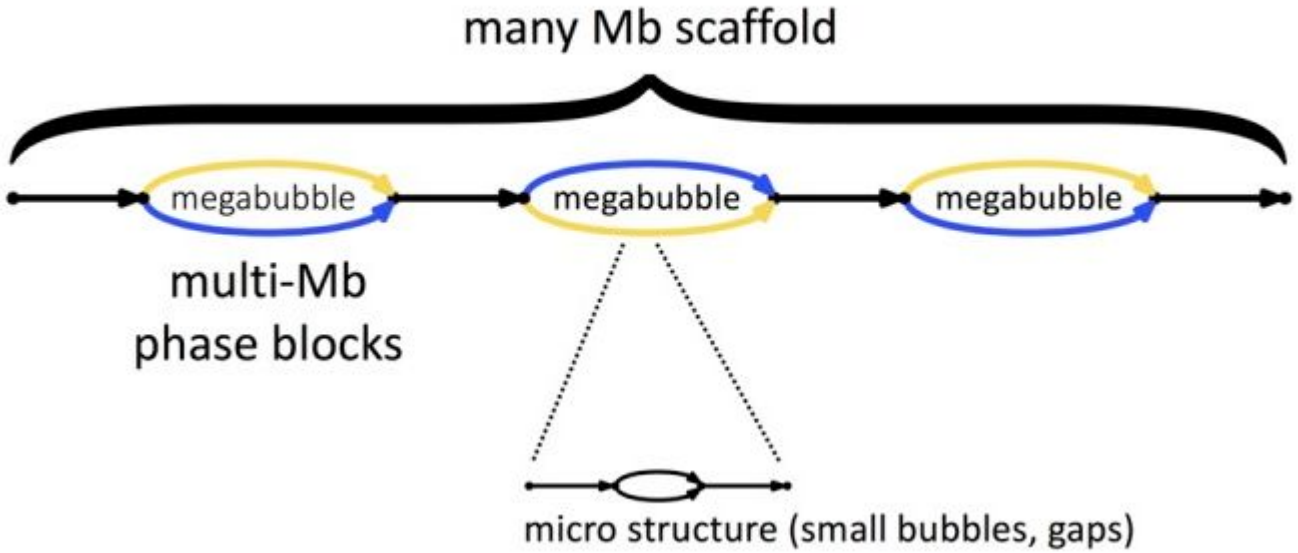
Phase

“Lines” in the short-read assembly graph preserve heterozygous variants



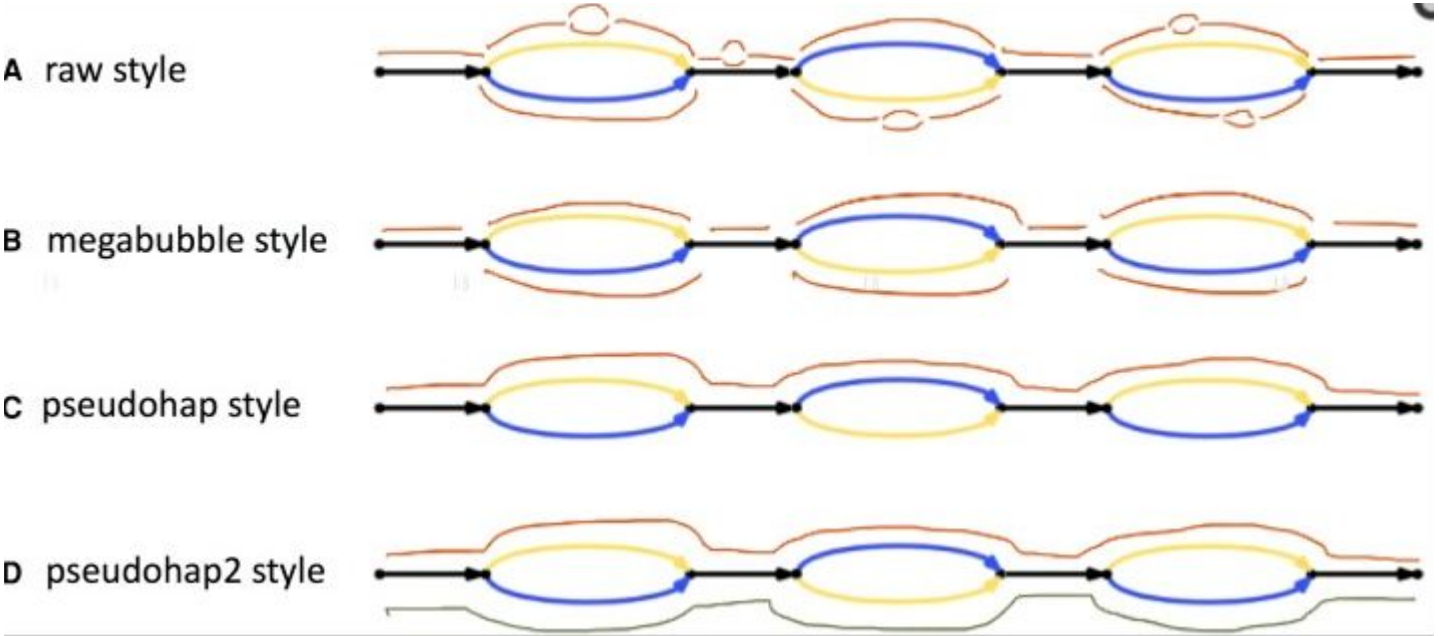
Lines in an assembly graph. Each edge represents a DNA sequence. (A) Blue portion describes a line in an assembly graph, which is an acyclic graph part bounded on both ends by single edges. The line alternates between five common segments and four bubbles, three of which have two branches. The third bubble is more complicated. The entire graph may be partitioned so that each of its edges lies in a unique line (allowing for degenerate cases, including single edge lines, and circles). (B) The same line, but now each bubble has been replaced by a bubble consisting of all its paths. After this change, **each bubble consists only of parallel edges**.

Read pairs and barcodes enable scaffolding into “megabubbles”



Supernova assemblies encode diploid genome architecture. Each edge represents a sequence. Blue represents one parental allele, and gold represents the other. Megabubble arms represent alternative parental alleles at a given locus, whereas sequences between megabubbles are homozygous (or appear so to Supernova). **Successive megabubbles are not phased relative to each other.** Smaller scale features appear as gaps and bubbles.

Assembly reporting modes trade contiguity for accuracy



A. Every edge is a contig. "Bubbles and gaps generally appear once per 10–20 kb; consequently, FASTA records from A are much shorter (~100 times) than those from B, C, and D"
B. Collapse megabubble arms into a single contig. Report both arms.
C. Choose an arbitrary megabubble arm and make a haploid assembly
D. Choose an arbitrary megabubble arm for each haplotype of diploid assembly

Supernova assemblies are graphs that can be represented as FASTA files. Here we see a toy example:

Each edge represents a DNA sequence, that is, a portion of a single strand of DNA. Each edge has left and right vertices. For example, edge 10 has left vertex 1 and right vertex 2. Edges that abut each other (like edges 10 and 20) overlap at their junction by K-1 bases. (Currently K = 48.) Any path through a Supernova graph is a list of edges, for example 10,20,40 is a path in the above graph. A sequence may be associated to it by concatenating successive edges and deleting the shared K-1 bases at each junction. Such a path might be encoded in a FASTA record as:

```
>55 edges=10,20,40 left=1 right=4 ver=1.3 style=2
ACTTTAGACGGGGACCCTAGACTTACTTGAGAAAACGTTTTTACACTTACCA
```

We've augmented the information normally encoded in a FASTA file by adding tags after the sequence identifier (e.g. >55) for each FASTA record. Note that we use serial numbers that count upward from zero for the sequence identifiers themselves, which otherwise have no specific meaning. The tags used are defined as follows:

Field	Sample Value	Meaning
edges	10,20,40	path of edges in the assembly that the sequence describes
left	1	identifier of vertex at left end of the path
right	4	identifier of vertex at right end of the path
ver	1.3	Supernova output format version number
style	2	output style identifier (see below)

https://support.10xgenomics.com/de-novo-assembly/software/pipelines/latest/output/graphs

These 6 new *de novo* assemblies appear to hold their own against the “state of the art”

A	NA19238	Yoruban	F	One 10× library
B	NA19240	Yoruban	F	One 10× library
C	HG00733	Puerto Rican	F	One 10× library
D	HG00512	Chinese	M	One 10× library
E	NA24385	Ashkenazi	M	One 10× library
F	HGP	European	M	One 10× library
G	NA12878	European	F	One 10× library
H	NA12878	European	F	Unknown number of PacBio libraries plus BioNano Genomics data
I	NA12878	European	F	Six libraries (fragment, jumping, 10×)
J	NA12878	European	F	Nine libraries (fragment, jumping, Fosmid, Chicago)
K	NA24385	Ashkenazi	M	Seven PacBio libraries
L	NA24143	Ashkenazi	F	Two PacBio libraries
M	YH	Chinese	M	~18,000 Fosmid pools and six fragment and jumping libraries, Illumina sequenced, plus Complete Genomics data

Contiguity
Completeness
Correctness

Only M is phased diploid

Contiguity

N50 : Half of the base pairs in the assembly are in a contig/scaffold/phase block of this size or larger.

Contig N50 (coverage-tech) = 120 kb

71-PacBio > 46-PacBio + BioNano > 30-PacBio > **56-10X** > 702-fosmid + others > 150-jump/fosmid > 160-jump/10X

Scaffold N50 = 18 Mb

150-jump/fosmid > 160-jump/10X > 46-PacBio + BioNano > **56-10X** > 702K-fosmid + others > 71-PacBio > 30-PacBio

coverage

One 10× library	56
One 10× library	56
One 10× library	56
One 10× library	56
One 10× library	56
One 10× library	56
One 10× library	56
Unknown number of PacBio libraries plus BioNano Genomics data	46
Six libraries (fragment, jumping, 10×)	160
Nine libraries (fragment, jumping, Fosmid, Chicago)	150
Seven PacBio libraries	71
Two PacBio libraries	30
~18,000 Fosmid pools and six fragment and jumping libraries, Illumina sequenced, plus Complete Genomics data	702

Completeness

2-3% “N” ‘s - other assemblies (that use N’s) have 0.6 - 10%

Discover 6-8 Mb of non-reference sequence per individual

12-15% 100-mers missing in haploid (9-10% diploid)

Expect 10% from real sample-reference differences

Correctness

Compare to assembly of same individual:

N50 “perfect stretch” = 16-20 kb

Compare to known haplotype:

Phasing error rate (counting every difference as a mismatch) = very small

Compare to reference genome:

1Mb inconsistency: 1-2% (Others 0.5-15)

10Mb inconsistency: 0.5-2% (Others 1.6-18)

“Of k-mer pairs at the given distance in the assembly, and for which both are uniquely placed on the reference, fraction for which either the reference chromosome, orientation, order, or separation (±10%) are inconsistent (includes bona fide sample/reference differences).”

Improvements of a diploid assembly

4% more 100-mers are represented compared to Supernova haploid

Phase block N50: 3-8 Mb (compared to 0.5 Mb in YH genome)

Computational performance

~2 days per genome; v1.2 uses 2TB disk space, 244GB RAM

Overview of *De Novo* Assembly Software

Supernova is a software package for *de novo* assembly from Chromium Linked-Reads that are made from a single whole-genome library from an individual DNA source. A key feature of Supernova is that it creates diploid assemblies, thus separately representing maternal and paternal chromosomes over very long distances. Almost all other methods instead merge homologous chromosomes into single incorrect 'consensus' sequences. Supernova is the only practical method for creating diploid assemblies of large genomes.

The Supernova software package includes two processing pipelines:

- `supernova mkfastq` wraps Illumina's bcl2fastq to correctly demultiplex Chromium-prepared sequencing samples and to convert barcode and read data to FASTQ files.
- `supernova run` takes FASTQ files containing barcoded reads from `supernova mkfastq` and builds a graph-based assembly. The approach is to first build an assembly using read kmers (K = 48), then resolve this assembly using read pairs (to K = 200), then use barcodes to effectively resolve this assembly to K ≈ 100,000. The final step pulls apart homologous chromosomes into phase blocks, which are typically multi-megabase for human genomes.

and for post-processing:

- `supernova mkoutput` takes Supernova's graph-based assemblies and produces several styles of FASTA suitable for downstream processing and analysis.

https://support.10xgenomics.com/de-novo-assembly/software/overview/welcome

Supernova Guidance

Sample Requirements and Library Considerations

Technical Note, Last Modified on May 24, 2017, Permalink

Supernova is a *de novo* assembly program that has been designed to assemble **germline human** genomes, from data generated in a precise fashion that we outline below. For users who are interested in using Supernova to assemble human and non-human genomes, this guidance document outlines some important considerations.

- For Supernova assemblies of nonhuman genomes, we have seen both successes and failures, depending on the characteristics of the genome and the quality of the data. Due to the diversity of genome and data characteristics, it is impossible to know a priori how well a specific dataset will assemble. For that reason, using Supernova to assemble non-human genomes remains *experimental* and may lead to unexpected failures or limited assembly quality.

- Supernova requires long, undamaged DNA. We provide protocols for DNA extraction from blood and cell lines, both of which have been validated for human samples. Getting good material from other sample types and organisms can be challenging.

Please Note: Even following all of the guidance in this document by no means guarantees success in assembling non-human genomes.

https://support.10xgenomics.com/de-novo-assembly/sequencing/doc/technical-note-supernova-guidance

Related work: findings from simulation

Human genome:

100-200kb molecules are best: longer, and **linked-read coverage** is too low.

Optimal coverage is ~50X and 1.5M GEMs, which are close to defaults

Smaller genome (Arabidopsis):

Better results with 200K GEMs: far less than machine is made to use

Optimal coverage is ~32X: lower than optimal for human

Here: authors recommend (no data) less input DNA, higher coverage, subsample barcodes before assembly, which is basically the same thing just working within the confines of the machine as-is

http://www.biorxiv.org/content/early/2017/01/26/103549 (LRSim)

Cut through the self-promo for some takeaways

“The ~100 kb molecules underlying Linked-Reads enable the long phase blocks that are difficult to achieve with other technologies.”

I agree from experience! This is/will probably be the main draw of 10X.

“Our diploid assemblies open the door to new analytical approaches, including alignment of assemblies to a reference sequence to call variants.”

If reference bias is so bad as to preclude even variant calling with 50X coverage…but why not assemble if you can

“The aggregate experiment burden [and cost] of our approach is dramatically lower than that for all the human assemblies that we compared to.”

Definitely, especially YH (Chinese diploid genome project)