Direct determination of diploid genome sequences

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

Supplemental Notes: contents

Supplemental Note 1. Comparison of performance on HiSeq 4000 versus HiSeq X10

Supplemental Note 2. Supernova performance as DNA length changes

Supplemental Note 3. Supernova performance as coverage is reduced

Supplemental Note 4. Sources for comparison assemblies

Supplemental Note 5. Supernova gap size estimation

Supplemental Note 6. Assembly accuracy assessment (N50 perfect stretch)

Supplemental Note 7. Phasing accuracy assessment sources and methodology

Supplemental Note 8. Measurement of completeness relative to the reference sequence

Supplemental Note 9. Analysis of contamination in novel sequences

Supplemental references

Supplemental Note 1. Comparison of performance on HiSeq 4000 versus HiSeq X10

We created a single Linked-Read library from NA12878 DNA, sequenced 800M reads of length 150 bases at six sites, then assembled using the same (but older) version of Supernova, and compared results. Contig lengths for assemblies of HiSeq X10 data were ~90% longer than those generated on the HiSeq 4000.

class	site	% of read one bases that had quality ≥ 20	% of read two bases that had quality ≥ 20	N50 perfect stretch (kb)	N50 contig (kb)	N50 phase block (Mb)	N50 scaffold (Mb)
X10	Hudson-Alpha	91.6	89.8	11.8	92.8	1.46	13.3
X10	Genewiz	91.1	89.4	9.9	89.0	1.49	14.9
X10	Macrogen	94.8	93.3	12.7	102.6	1.62	14.1
4000	10x	93.1	91.5	6.8	50.0	1.30	11.5
4000	OMRF	90.5	88.5	6.8	44.5	1.30	12.2
4000	UNC	93.0	91.8	7.8	56.2	1.43	12.8

The quality scores assigned to read bases do not appear to explain assembly quality differences. However, we note that HiSeq 4000 performance on homopolymers is worse than HiSeq X: the mean coverage of 80-mers containing an A or T homopolymer of length exactly 20 at positions 30-50 and having overall GC content 35-40% is 0.1 for HiSeq 4000, as compared to 0.7 for HiSeq X. Both are very low, but the HiSeq 4000 coverage is seven times lower. Very low coverage at homopolymers would tend to cause contigs to be shorter.

Supplemental Note 2. Supernova performance as DNA length changes

The data below show assembly performance on data from four libraries (not shown), constructed from NA12878 DNA of various lengths, and sequenced to 38x coverage. An earlier version of Supernova was used for these assemblies.

DNA length estimates (kb)		output statistics			
from gel measurement	length weighted mean, from Supernova	N50 perfect stretch (kb)	N50 contig (kb)	N50 phase block (Mb)	N50 scaffold (Mb)
13	14	2.0	14.2	0.0	0.0
20	21	7.9	61.4	0.1	0.6
30	35	9.0	78.1	0.3	9.6
68	48	8.3	86.9	0.5	12.8

Effect of DNA length reduction. DNA of four different lengths was prepared, and one library was made from each, which was sequenced to 38x (800M reads). DNA size was estimated both from a gel and by Supernova. Output statistics: see **Table 1**.

Supplemental Note 3. Supernova performance as coverage is reduced

The data below show assembly performance at three coverage levels, subsampled from the HGP dataset for assembly F (**Table 1**). An earlier version of Supernova was used for these assemblies.

input statistics		output statistics			
coverage (x)	**number of reads (M)**	**N50 perfect stretch (kb)**	**N50 contig (kb)**	**N50 phase block (Mb)**	**N50 scaffold (Mb)**
56	1200	16.9	95.2	4.1	16.9
47	1000	15.1	95.0	3.3	15.1
38	800	11.8	89.8	2.3	11.8

Effect of coverage reduction. HGP sample was assembled at three levels of coverage. Output statistics: see **Table 1**. See also **Supplemental Table 3**.

Supplemental Note 4. Sources for comparison assemblies

Assembly H was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_001013985.1_ASM101398v1/GCA_001013985.1_ASM101398v1_genomic.fna.gz.

Assembly I was downloaded from http://kwoklab.ucsf.edu/resources/nmeth_201604_NA12878_hybrid_assembly.fasta.gz. The following data were used to create the assembly (not including phasing data):

library type	number of libraries	coverage (x)
fragment	2	39
jumping	1	24
10x Gemcode	2	97
Bionano Genomics	1	N/A

Assembly J is 'MERAC PE + HiRise 1.0 + L3' as described in (Putnam et al. 2016). It is Bioproject PRJNA305315, and was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_001483605.1_hirise_NA12878_merac_L1_L2_L3/GCA_001483605.1_hirise_NA12878_merac_L1_L2_L3_genomic.fna.gz. Data description, based on (Putnam et al. 2016; Nik Putnam personal communication; Gnerre et al. 2011):

library type	# of libraries	coverage (x)	data source
fragment	2	84.0	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20101201_cg_NA12878/NA12878.hiseq.wgs.bwa.raw.bam; number of libraries inferred from BAM header
jumping	2	45.9	Gnerre et al. 2011
Fosmid	2	5.3	Gnerre et al. 2011
HiRise	3	15.3	Putnam et al. 2016

Assembly K [unpublished, from cell line NA24385] was downloaded from ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/UMD_PacBio_Assembly_CA8.3_08252015/trio2.quiver.fasta. See (Zook et al. 2016) for data description.

Assembly L [unpublished, from cell line NA24143] was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_001549595.1_GIAB_Ashkenazim_Mother_HG004_NA24143_hu8E87A9_PacBio_Assembly_with_PBcR/GCA_001549595.1_GIAB_Ashkenazim_Mother_HG004_NA24143_hu8E87A9_PacBio_Assembly_with_PBcR_genomic.fna.gz. See (Zook et al. 2016) for data description.

Assembly M was downloaded from ftp://public.genomics.org.cn/BGI/yanhuang/assemble/YH.haplotype-resolved.fa.gz.

Supplemental Note 5. Supernova gap size estimation

The size of assembly gaps within a scaffold that were not bridged by read pairs ('barcode-only gaps') were estimated using the following approach. First we consider 'fake' gaps of sizes 0, 5, 10, 15, …, 100 kb in a given assembly. To do this we find many loci in scaffolds where there is sequence of the form ABC, where A and C have size 10 kb, B has the given gap size, and there are no barcode-only gaps in ABC. We consider the set of all barcodes that have at least one read aligned to A or C. Then we determine the fraction of these barcodes, that contain both a read aligned to A and a read aligned to C. This is the 'bridge fraction'. We compute the mean value over all loci under consideration for a given gap size. Next for each barcode-only gap, we determine if there is 10 kb on both sides that is free of other barcode-only gaps. If not, we set the size of the gap to 3000. Otherwise, we compute the bridge fraction for the given gap (as above), and then assign the gap the gap size corresponding to the fake gap whose bridge fraction is closest to the observed value. For the HGP assembly, in comparison to GRCh37, we observed a mean error rate of 2.2 kb for these estimates and a median error rate of 0.4 kb.

Supplemental Note 6. Assembly accuracy assessment (N50 perfect stretch)

To compute the N50 perfect stretch, we chose the best alignment of a given finished clone to the assembly, consistent with the graph structure. This alignment corresponded to a single path through the assembly graph, and thus by definition could not traverse both parental alleles in a megabubble. We also excluded from consideration all scaffolds less than 10 kb. We then partitioned the clone sequence into nonoverlapping intervals. These consisted of the perfectly matching intervals, plus single base intervals, to fill in the gaps between them (and thus penalizing for lack of coverage and errors). We pooled these intervals across clones, and computed their N50 size, after a statistical correction to compensate for the limited length of the clones, as follows. We chose an ordering for the clones, and combined the interval at the end of clone i with the interval at the beginning of clone i+1. For a given ordering of the clones, we computed the N50 perfect stretch length, and we repeated this process for 1000 random orderings of the clones, then took the mean of these N50 values.

Supplemental Note 7. Phasing accuracy assessment sources and methodology. We used publicly available phased VCF files for four samples, as follows:

NA12878:
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/analysis/RTG_Illumina_Segregation_Phasing_05122016/sp_v37.7.0.NA12878.vcf.gz.

HG00733:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20160704_whatshap_strandseq_10X_phased_SNPs/PUR/PR05.wgs.whatshap.strandseq-10X.20160704.phased-genotypes.vcf.gz
This is for GRCh38. We translated to GRCh37 for the analysis in this work.

NA24385:
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/Rutgers_IlluminaHiSeq300X_rtg_11052015/rtg_allCallsV2.vcf.gz

NA19240:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20160704_whatshap_strandseq_10X_phased_SNPs/YRI/Y117.wgs.whatshap.strandseq-10X.20160704.phased-genotypes.vcf.gz
This is for GRCh38. We translated to GRCh37 for the analysis in this work.

Supplemental Note 8. Measurement of completeness relative to the reference sequence. Here we provide technical notes on our completeness computations. We first excluded kmers in GRCh37 containing any ambiguous base. We used only the records for chr1-22, X and Y. By a kmer, we mean a kmer in these records (K = 100), regardless of how many times it or its reverse complement appears. Kmers were divided into duplicate kmers and nonduplicate kmers. To do this, for each kmer, we considered all instances of it or its reverse complement. One instance (chosen arbitrarily) were placed in the nonduplicate kmer bucket. All other instances were placed in the duplicate kmer bucket.

Supplemental Note 9. Analysis of contamination in novel sequences.

All novel sequences (see text) were aligned to NT using blast 2.5.0 (blastn with default parameters), finding the top ten hits. Sequences having hits to bacterial genomes or genuses *Capsicum*/*Solanum* (~50 kb, assembly E only) were excluded from the analysis. For all but 25 of the remaining sequences, the top 10 hits were to primate species.

Analysis of the 25 remaining sequences. There were 3 with no hits. All others hit primate. Some also hit the nonprimate species *Spirometra* and *Ovis*. For all 25, we computed (number of occurrences of CATTC or its reverse complement), divided by the number of bases. For random sequences, the expected value of this is about $1/(4^5)$ = 0.1%. However for the 25 sequences, the values were (in percent): 2.4, 2.4, 3.7, 3.7, 4.7, 4.7, 4.7, 4.8, 5.8, 5.9, 5.9, 6.1, 6.2, 6.3, 6.3, 6.3, 7.2, 7.2, 7.3, 7.7, 7.8, 7.8, 8.7, 8.9 and 9.7. This would be consistent with the hypothesis that these sequences consisted at least in part of human heterochromatic CATTC repeats.

Supplemental references

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* **108:** 1513-1518.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025.