Validation of synthetic long reads for use in constructing variant graphs for dairy cattle breeding

M Keehan and C Couldrey

Research and Development, Livestock Improvment Corporation, Hamilton, New Zealand 3240
mkeehan@lic.co.nz (Corresponding Author)

Summary

Recently 10x Genomics have released 10x GemCode Technology which makes cost effective synthetic long read data available to animal breeders. In this paper we report quality metrics from a validation study of 10x Genomics data obtained from seven New Zealand dairy bulls. We use two separate 10x Genomics pipelines with the aim of determining which approach is best for use in the future for variant graph construction. The mapping/variant calling pipeline (Longranger) is tested as well as the parameter free, diploid Supernova de novo assembly pipeline to individually assemble the seven genomes. Longranger provides excellent Single Nucleotide Polymorphism (SNP) discovery, phasing and genotyping. The Supernova pipeline easily generates unbiased contigs of a reasonable quality. The 10x Genomics platform together with Longranger and Supernova assembly pipelines provide an excellent value for money approach for both variant detection and de novo assembly.

Keywords: sequencing, Sequence – gene annotation, assembly.

Introduction

Animal breeding has benefited from the use of whole genome sequencing to directly discover and genotype almost all variants within a population. However, with current short read technology, identification of variation longer than the read length and the accurate assignment of phase are challenging. Longer reads, such as those generated by Pacific Biosciences (PacBio), ease the task of recovering phase and enhance the discovery of structural variants (Sedlazeck et al., 2017). However, long read sequencing suffers not only from poor sequence accuracy, but also sequencing of multiple individuals in a species, let alone breed, is currently cost prohibitive. Recently 10x Genomics (Weisenfeld et al., 2017) has produced a system that combines the cost effectiveness and high accuracy of Illumina sequencing with the read length from PacBio to produce long synthetic reads.

While the promise of these synthetic reads is high, bioinformatic methods have been optimised for short read sequencing and the current linear reference model of read alignment and variant calling relies crucially on the reference genome. If the reference genome is not representative of the sample or has a genomic region that is highly divergent from the study sample the linear reference method is likely to fail in favour of the reference i.e. “reference bias” (Eggertsson et al., 2017). Here we begin to explore the potential of 10x Genomics platform, not only for effective calling of variants, but also de novo assemblies from which variant graphs (Paten et al., 2017) may be generated in the future.

Material and methods

Animals

Four Holstein Friesian (HF) bulls: Beamer, Esteem, Hammer and Hothouse, two Jersey (J): Speedway and Integrity and one HF x J crossbreed, Solaris were selected from the LIC commercial bull teams for whole genome sequencing using the 10x Genomics platform.

DNA sequencing and analysis

DNA was extracted from Whole Blood. DNA sequencing libraries were prepared as per 10x Genomics workflow protocols and sequenced on an Illumina HiSeq X Ten DNA sequencer by the Garvan Institute of Medical Research. One lane of Illumina 150 bp paired end reads was used for sequencing six bulls and the remaining sample (Esteem) was sequenced on two lanes.

10x Genomics Longranger 2.1.3 pipeline was initially tested on two animals using bovine UMD 3.1 as the reference genome. However, long run times were noted. As per the Longranger documentation, for SV calling, UMD3.1 contigs were filtered to remove contigs with a median genotype imputation allelic $R^2 < 0.97$ (unpublished data) and the 5% of contigs with the greatest mendel error rate were removed. Longranger (2.1.5) and the filtered contig set were used to call variants on all samples using the NZ National E-Science Infrastructure (NeSI).

Supernova

DNA sequence from two bulls, Esteem and Speedway, were assembled using Supernova 1.1.5. The remaining bulls were assembled by Supernova 1.2.0 using NeSI. Phased scaffolds were generated using the pseudohap2 option of Supernova. Runs of contiguous N bases, used by Supernova to space contigs within scaffolds, were removed, and fasta files split at their location. These fragmented scaffolds were then aligned to UMD 3.1 using Minimap2 (Li, 2017) v2.0-r295-dirty and variants called using bcftools 1.5 and htsbox r340

Genotype Comparison

Bcftools gtcheck and Realtime Genomics (RTG) vcfeval (Cleary et al., 2015), which accounts for differing but equivalent allelic representations in VCF files, were then used to compare the genotypes from Longranger and htsbox to Illumina BovineHD Genotyping BeadChip (BovineHD) genotypes of the seven bulls.

Results and discussion

Runtimes

Longranger took approximately 1920 cpu hours (5 days) wall clock time using up to 80 GB RAM per sample. Esteem with two lanes of coverage required 8 days wall clock time. De novo assemblies using Supernova required approximately ~5500 cpu hours using up to 240 GB RAM. Supernova runtimes compare favourably with Canu (Koren et al., 2016) estimates of 10,000-25,000 cpu hours for a mammalian genome. With these times it is feasible to call

variants and undertake de novo assemblies on large numbers of animals on a standard high performance computing cluster.

Longranger

The 10x Genomics Longranger pipeline aligns sequence reads, calls, and phases structural and non-structural variants (analysis of structural variants is described by Couldrey et al. and Reynolds et al in this publication). A summary of Longranger output is shown in Table 1. The two lanes of sequence generated from Esteem results in approximately twice as much read depth and a greater N50 linked reads per molecule but the extra sequence did not improve phasing. Given the similarity in sequence coverage for the remaining six animals there is no clear correlation between sequence coverage and quality of data produced. Longest phased SNP block is linked to molecule length and dependent on quality of DNA extracted for library preparation. SNP call sets from all samples had a Ti/Tv ratio very close to 2.1, indicating that the default settings are well calibrated in Longranger.

Table 1: Summary of Longranger results. N50 is per molecule, Longest phase block is Mbp

Statistic	*Beamer*	*Esteem*	*Hammer*	*Hothouse*	*Integrity*	*Solaris*	*Speedway*
Mean read depth	35.4	69.5	36.3	38.0	36.4	29.4	36.1
N50 linked reads	20	48	10	23	14	19	17
Longest phase block	8.57	8.39	3.98	7.64	6.99	7.66	5.60
SNPs phased (%)	0.988	0.991	0.988	0.988	0.987	0.982	0.988

Summary of supernova de novo assemblies

Table 2 shows a summary of the de novo assemblies generated by Supernova and comparable UMD3.1 statistics where available. The average Supernova assembly length is 86% of the UMD 3.1 assembly length. The higher sequence coverage obtained for Esteem as recommended by 10x Genomics, resulted in the largest assembly size and best contig N50. Esteem and Hothouse report a scaffold N50 greater than UMD 3.1. These results suggest assemblies close to the quality of UMD3.1 (in terms of contigN50) could be cost effectively generated for large numbers of animals.

Table 2. Summary statistics from multiple de novo assemblies.

Statistic	*Beamer*	*Esteem*	*Hammer*	*Hothouse*	*Integrity*	*Solaris*	*Speedway*	*UMD3.1*
Assembly size	2.35	2.47	2.34	2.37	2.36	2.02	2.36	2.67
Contig N50 (bp)	59312	74891	58692	65223	61841	35480	59101	96955
Scaffold N50 (Mbp)	5.15	7.56	0.76	8.42	4.46	1.87	3.36	6.38
Scaffolds_10kb_plu	3068	1658	8139	2794	3306	7843	3567	NA
Scaffolds_1kb_plus	58760	36961	65874	55169	55426	125432	56230	NA

Comparative statistics with Illumina BovineHD

Supplementary Table 1 reports genotype concordances from the Longranger and Supernova + htsbox variant calling strategies to the BovineHD genotypes. The Supernova + Bcftools variant calling strategy was tested, however, variant output terminated prematurely (possibly due to the extreme length of some alignments) and was therefore not pursued further. In Table

3 we report genotype comparisons derived from 10x Genomics and Illumina BovineHD using RTG vcfeval.

Table 3:RTG vcfeval comparision of Longranger and SuperNova+htsbox to BovineHD

Statistic	*Variant Caller*	*Beamer*	*Esteem*	*Hammer*	*Hothouse*	*Integrity*	*Solaris*	*Speedway*
Precision	Htsbox	0.888	0.924	0.873	0.898	0.894	0.816	0.893
	Longranger	0.943	0.944	0.946	0.943	0.943	0.943	0.943
Sensitivity	Htsbox	0.857	0.932	0.817	0.878	0.869	0.717	0.866
	Longranger	0.995	0.997	0.995	0.996	0.996	0.988	0.995

Variant calling after alignment of de novo assembly contigs to a linear reference gave disappointing results compared to optimised Longranger variant calling. This is perhaps in part due to the tools for analysing contigs being experimental. Furthermore missassemblies present in the UMD3.1 reference genome exacerbate these problems.

Conclusions

10x Genomics sequencing represents good value for money in terms of enabling high quality variant calling (Longranger) as well as enabling reasonably high quality de novo assemblies of multiple animals. However, the latter does require high sequence coverage. While more work is required to determine a truth set for phase comparison, to date 10x Genomics appears to be an excellent platform for beginning to develop and explore variant graphs.

Acknowledgements

Funding for this project was provided by the Ministry for Primary Industries as a Primary Growth Partnership

List of References

Cleary, J.G., R. Braithwaite, K. Gaastra, B.S. Hilbush, S. Inglis, S.A. Irvine, A. Jackson, R. Littin, M. Rathod, D. Ware, J.M. Zook, L. Trigg, and F.M.M. De La Vega. 2015. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv*. 23754. doi:10.1101/023754.

Eggertsson, H.P., H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, A. Jonasdottir, A. Jonasdottir, I. Jonsdottir, D.F. Gudbjartsson, P. Melsted, K. Stefansson, and B. V Halldorsson. 2017. Graphtyper: Population-scale genotyping using pangenome graphs. *bioRxiv*. doi:10.1101/148403.

Koren, S., B.P. Walenz, K. Berlin, J.R. Miller, and A.M. Phillippy. 2016. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv*. 71282. doi:10.1101/071282.

Li, H. 2017. Minimap2: fast pairwise alignment for long DNA sequences. 1–3. doi:10.1101/169557.

Paten, B., A.M. Novak, J.M. Eizenga, and E. Garrison. 2017. Genome Graphs and the Evolution of Genome Inference. *bioRxiv*. 665–676. doi:10.1101/101816.

Sedlazeck, F.J., P. Rescheneder, M. Smolka, H. Fang, and M. Nattestad. 2017. Accurate

detection of complex structural variations using single molecule sequencing. *bioRxiv*. 1–24. doi:10.1016/j.ajhg.2017.06.005.

Weisenfeld, N.I., V. Kumar, P. Shah, D.M. Church, and D.B. Jaffe. 2017. Direct determination of diploid genome sequences. *Genome Res.* 27:757–767. doi:10.1101/gr.214874.116.

Supplementary Table1: bcftools gtcheck genotype concordance with Illumina BovineHD on Chr1.

Statistic	*Variant Caller*	*Beamer*	*Esteem*	*Hammer*	*Hothouse*	*Integrity*	*Solaris*	*Speedway*
Genotype concordance	Htsbox	0.922	0.969	0.913	0.941	0.935	0.823	0.939
	Longranger	0.996	0.997	0.995	0.996	0.996	0.995	0.995
Number Of variants	Htsbox	27109	28851	26797	27608	26436	24613	26603
	Longranger	30038	30203	30470	29929	28885	30392	29086