Unravelling the structural variability in chicken genomes by long-read sequencing

J. Geibel[1,2*], J. Schauer[3], A. Weigend[3], C. Reimer[1,2,3], D.-J. de Koning[4], H. Simianer[1,2] and S. Weigend[2,3]

[1] University of Goettingen, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany; [2] University of Goettingen, Center for Integrated Breeding Research, Carl-Sprengel-Weg 1, 37075 Göttingen, Germany; [3] Friedrich-Loeffler-Institut, Institute of Farm Animal Genetics, Höltystraße 10, 34535 Neustadt, Germany; [4] Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, Box 7023, SE-750 07 Uppsala, Sweden; [*]johannes.geibel@uni-goettingen.de

Abstract

Structural variants (SV) have gained increasing interest within the last years. However, calling SVs from array genotypes or short-read data is known to suffer from high false-positive rates and low specificity at the same time. In contrast, long-read technologies promise to derive accurate and sensitive SV callsets. We hereby present the first study that uses PacBio and Nanopore long-reads to derive quantitative estimates of the amount of SVs in chicken, based on four trios. We estimated the mean size of the reference genome affected by at least heterozygous SV calls to range from 0.2% for insertions to 0.7% for deletions. We further compared the results to a short-read based calling approach. This revealed more than 2/3 of short-read based initial SV calls not being backed by long-read based calls, while more than 50% of the long-read based calls would have been missed by the short-read approach.

Introduction

During the last years, some prominent phenotypes were shown to be causally affected by structural variants (SV), e.g. different comb phenotypes in chickens (Imsland *et al.*, 2012). Effects of SV may, however, be missed by analyses based on single nucleotide polymorphism (SNP) markers, as linkage disequilibrium between SVs and neighbouring marker SNPs is not necessarily high (Geibel *et al.*, 2021). This has driven the interest in identifying SVs along the genome. SV calling so far was mostly based on SNP-array data or short-read sequence data, and heavily relied on auxiliary information like probe intensity, coverage, insert size distributions, split reads, or local assembly. This led to high false positive rates and low specificity at the same time, especially when calling insertions (INS; Delage *et al.*, 2020), or SVs in repetitive regions (Belyeu *et al.*, 2021). Previous estimates of the SV content of chicken genomes therefore varied massively and reached estimates up to 12.84% of the autosomal genome size (Fernandes *et al.*, 2021).

Many problems of calling SVs from short reads can be overcome by using long reads from Oxford Nanopore Technologies (ONT) or Pacific Biosciences (PacBio) sequencing, which are nowadays considered to strongly improve SV identification (Ho *et al.*, 2019; Sedlazeck *et al.*, 2018b). However, studies identifying SVs based on long read technologies in chickens have not been published to this date.

The goal of this study was to provide a reliable quantitative estimate of the affectedness of the chicken genome by SVs using long-read sequencing of four chicken trios. We further verified how many of these structural variants would have been discovered by short-read sequencing, providing an insight into the false-positive and false-negative rates of the latter approach.

Materials & Methods

For the discovery of SVs, four chicken trios were sequenced. The trios were from a commercial White Leghorn line (WLA), a commercial Rhode Island Red line (BLA), a New Hampshire line (L68) and the intercross of an Araucana x WLA backcross experiment. Sequencing was done using Illumina short read sequencing (13 – 42X, 2*151bp), ONT sequencing (6 – 22X; 3 – 22kb mean read length) and PacBio HiFi sequencing (3 – 25X; 16 – 22kb mean read length). Note that for the BLA trio, a fullsib of the offspring chicken, which was sequenced by Illumina and PacBio, was sequenced with ONT due to a shortage of available DNA.

SV calling was performed separately for the different sequencing techniques. The short reads were mapped against GRGC6a by bwa-mem (Li, 2013) and SVs were called by a consensus calling approach that involved parallel SV calling by delly (Rausch *et al.*, 2012), manta (Chen *et al.*, 2016) and lumpy (Layer *et al.*, 2014), followed by feeding the union of the SV calls into svtyper (Chiang *et al.*, 2015) for genotyping. As svtyper is not able to genotype INS, only the calls of manta were considered for INS. Non-default filters for the short read data were not applied at this stage, as the main intention was to see whether short-read data would have provided support for SVs found by long reads. Long reads (PacBio and ONT) were mapped by ngmlr (Sedlazeck *et al.*, 2018a), followed by SV calling and genotyping through sniffles (Sedlazeck *et al.*, 2018a), requiring a threshold of at least five supporting reads. Long-read-based SV detection was performed for the technologies separately as well as based on merged bam alignment files (LRcombined). Finally, a merged VCF was produced for further analysis, using survivor (Jeffares *et al.*, 2017) together with custom awk scripts to account for incompatibilities between the different tools. To account for bad breakpoint resolutions, SV of the same type with breakpoint differences of 1kb or less were further merged into one SV. Analyses were thereby limited to autosomal deletions (DEL), duplications (DUP), insertions (INS), and inversions (INV) between 50bp and one Mb. Note that these are early stage results, which will require further validation, e.g. by assessing Mendelian errors in the trios.

Results and Discussion

The short-read based consensus calling approach (SRconsensus) identified a total of 32,111 SVs (15.6k DEL, 6.9k DUP, 6.6k INS, 2.9k INV), while 18,487 SVs were identified by the combined set of PacBio and ONT reads (LRcombined; 8k DEL, 1.2k DUP, 8.8k INS, 445 INV). Figure 1 shows the overlap between the different callsets. It is remarkable that more than 50% of the LRcombined SVs (10.5k) did not show even modest support in the SRconsensus set, while 2/3 of the SRconsensus SVs (21.6k) were not supported by long-read based calls. This indicates the known limitations of relatively non-sensitive and unreliable short-read calling algorithms requiring high manual curation effort (Belyeu *et al.*, 2021). However, against expectation (Delage *et al.*, 2020), more than 50% of the INS calls by manta were supported by long reads. Further, only very few DUP (1,244) were supported by the LRcombined set with close to no overlap to the SRconsensus set. This might partly be due to the fact that a high share of raw DUP calls in short-read data sets is expected to stem from mapping errors in badly assembled repetitive regions (Geibel *et al.*, 2021), which might be better resolved by long reads and called as INS rather than DUP then.

Generally, increasing the depth by combining the long-read sets improved the confidence in variant calls and 4k SVs were only discovered in the LRcombined set. Interestingly, slightly fewer SVs (3.3k) were called when only the PacBio set was used, but no longer when combining the sets. This might be a calling artefact, as sniffles has a special HiFi mode that trusts mappings with more confidence, but which could not be used for LRcombined, as LRcombined also included less accurate ONT data. Anyway, this needs further investigation.

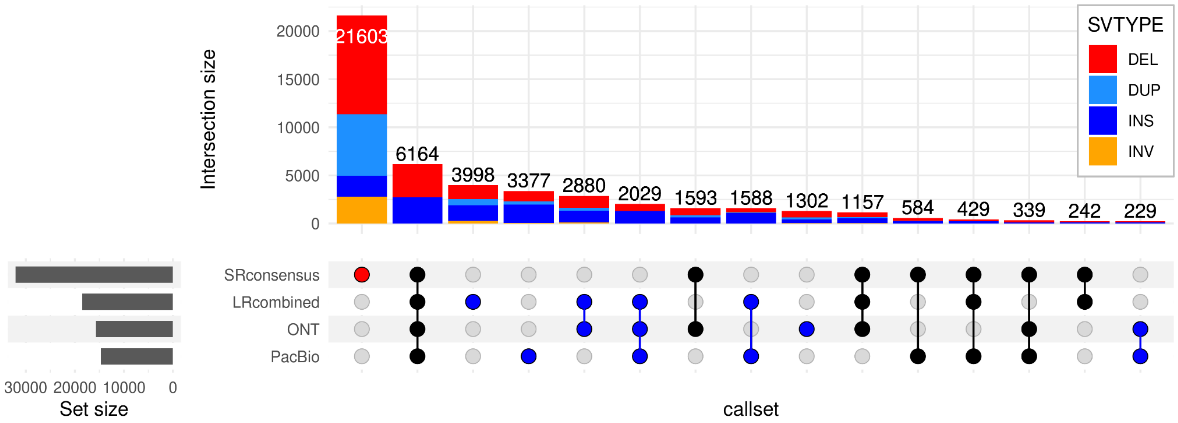

Figure 1: Upset plot showing the overlap between the callsets. Colours in the top plot indicate the proportion of SV types in the intersection. Colours in the intersection matrix instead indicate whether the SVs were found by long reads solely (blue), short reads solely (red) or by both techniques (black).

The length of the called SVs was only investigated for the most reliable set, LRcombined. Length distributions of the SVs (Figure 2 A) spanned the complete investigated range between 50bp and one Mb, with lengths >100kb (DUP and INV) respectively >10kb (DEL and INS) being rare events. Median sizes of 913bp (DUP), 379bp (INV), 153bp (DEL), and 132bp (INS), however, showed that the vast majority of called SVs was even smaller than 1kb.
We finally calculated the fraction of the autosomal reference genome that was affected by SV per individual (Figure 2 B). This number varied on average between 0.2% for INS and 0.7% for DEL. This is more than what was discovered previously by highly curated short-read based callsets (e.g. Geibel *et al.*, 2021), but by far less than what array-based CNV studies regularly estimate (e.g. Fernandes *et al.*, 2021).

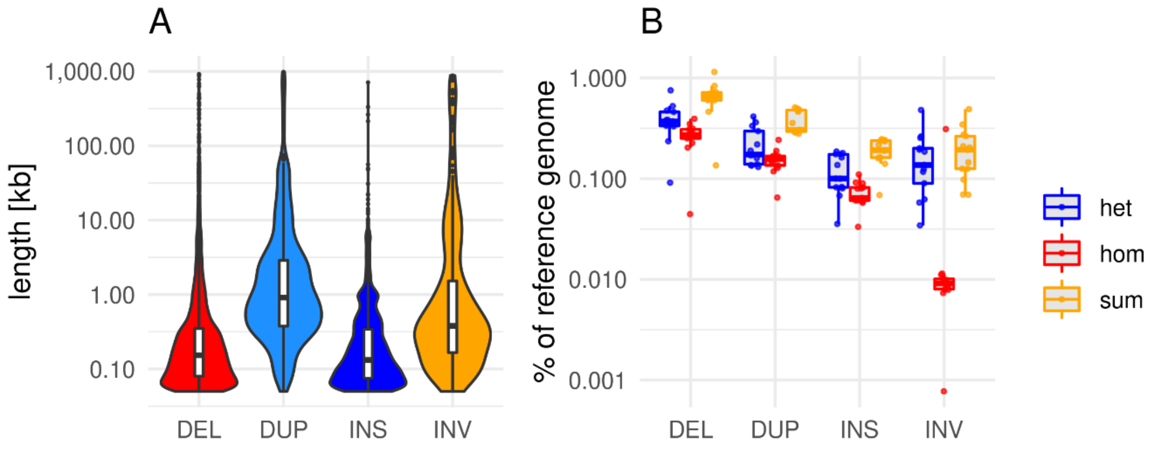

Figure 2: Distribution of SV lengths in the LRcombined callset (A). Percent of autosomal reference genome affected by heterozygous (het), homozygous (hom), or any (sum) SV genotypes by SV type in the LRcombined callset for the different chicken samples (B). Note the log-scaled y-axes.

Acknowledgements

The sequencing of the chicken was funded within the scope of the project IMAGE (www.imageh2020.eu; European Union's Horizon 2020 research and innovation programme under grant agreement No 677353). The Illumina sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, Sweden. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. Further, the PacBio sequencing was performed at NGI/ Uppsala Genome Center that has been funded by RFI/ VR and Science for Life Laboratory, Sweden. It was further supported by the National Genomics Infrastructure (NGI)/ Uppsala Genome Center and UPPMAX by providing assistance in massive parallel sequencing and computational infrastructure.

References

Belyeu, J.R., Brand, H., Wang, H., Zhao, X., and Pedersen, B.S., et al. (2021) Am J Hum Genet 108(4): 597–607. https://doi.org/10.1016/j.ajhg.2021.02.012

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., and Schlesinger, F., et al. (2016) Bioinformatics (Oxford, England) 32(8): 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., and Rose, D.B., et al. (2015) Nature methods 12(10): 966–968. https://doi.org/10.1038/nmeth.3505

Delage, W.J., Thevenon, J., and Lemaitre, C. (2020) BMC genomics 21(1): 762. https://doi.org/10.1186/s12864-020-07125-5

Fernandes, A.C., Da Silva, V.H., Goes, C.P., Moreira, G.C.M., and Godoy, T.F., et al. (2021) BMC genomics 22(1): 354. https://doi.org/10.1186/s12864-021-07676-1

Geibel, J., Praefke, N.P., Weigend, S., Simianer, H., and Reimer, C. (2021) Research Square (preprint). https://doi.org/10.21203/rs.3.rs-861830/v1

Ho, S.S., Urban, A.E., and Mills, R.E. (2019) Nature Reviews Genetics. https://doi.org/10.1038/s41576-019-0180-9

Imsland, F., Feng, C., Boije, H., Bed'hom, B., and Fillon, V., et al. (2012) PLoS genetics 8(6): e1002775. https://doi.org/10.1371/journal.pgen.1002775

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., and Shaw, L., et al. (2017) Nature Communications 8 14061. https://doi.org/10.1038/ncomms14061

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014) Genome Biology 15(6): R84. https://doi.org/10.1186/gb-2014-15-6-r84

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://arxiv.org/pdf/1303.3997

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., and Benes, V., et al. (2012) Bioinformatics (Oxford, England) 28(18): i333-i339. https://doi.org/10.1093/bioinformatics/bts378

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., and Nattestad, M., et al. (2018a) Nature methods 15(6): 461–468. https://doi.org/10.1038/s41592-018-0001-7

Sedlazeck, F.J., Lee, H., Darby, C.A., and Schatz, M.C. (2018b) Nature Reviews Genetics 19(6): 329–346. https://doi.org/10.1038/s41576-018-0003-4