The promise and challenges of structural variant discovery: A conservation case study in the critically endangered kākāpō (*Strigops habroptilus*)

Jana R Wold[1*], Joseph G Guhlin[2], Peter K Dearden[2], Anna W Santure[3], and Tammy E Steeves[1]

[1] University of Canterbury, Christchurch New Zealand

[2] Genomics Aotearoa and Biochemistry Department, University of Otago, Dunedin, New Zealand

[3] University of Auckland, Auckland, New Zealand

[*]Corresponding author: jana.wold@canterbury.ac.nz

GitHub Repository

Abstract

There is growing interest in the role of structural variants (SVs) as drivers of local adaptation and speciation. From a conservation genomics perspective, the characterisation of SVs in threatened species provides an exciting opportunity to complement existing approaches that use single nucleotide polymorphisms (SNPs) to detect adaptive variation, identify conservation units, guide pairing decisions and inform conservation translocations. However, little is known about whole-genome SV frequency and size distributions, especially for small populations. To explore the impacts that SV discovery and genotyping strategies may have on characterisation of SV diversity in non-model organisms, we explore a near whole-species resequence dataset, and long-read sequence data for a subset of highly represented individuals in the critically endangered kākāpō (*Strigops habroptilus*). We demonstrate that even when using a highly contiguous reference genome, different discovery and genotyping strategies can significantly impact the type, size and location of SVs characterised, which indicates researchers should exercise caution when drawing conclusions at the individual-scale. Further, we find that genotyping SVs discovered with long-read data at the population-scale with short-read data remains challenging. Despite this, we found that all six strategies used to characterise SVs in kākāpō reflected similar trends at the population-scale including the identification of population structure. We are optimistic that increased accessibility to long-read sequencing and advancements in bioinformatic approaches (e.g., multi-reference approaches like genome graphs) will alleviate challenges associated with resolving SV characteristics below the species level and facilitate the characterisation of population- and individual-level SVs in threatened species around the globe.

Keywords: structural variation, conservation genomics, population genomics, small population paradigm, Illumina, Oxford Nanopore Technologies

Introduction

The increased accessibility of whole-genome sequencing (WGS) technology has revolutionised population genetic/genomic studies in non-model organisms, and continues to provide valuable insights into the mechanisms underpinning genome divergence during speciation as well as the interplay between mutation, genetic drift, selection, and gene flow in the context of population demography (Cruickshank and Hahn 2014; Campbell *et al.* 2018; Lado *et al.* 2020; Chueca *et al.* 2021; Mathur and DeWoody 2021; Formenti *et al.* 2022). To date, the vast majority of these studies use single nucleotide polymorphisms (SNPs) to investigate these processes, yet there is a growing interest in the evolutionary and adaptive significance of structural variants (SVs), which are genomic rearrangements that include deletions, duplications, insertions, translocations, and inversions (Wellenreuther and Bernatchez 2018; Mérot *et al.* 2020). SVs have been shown to influence the evolutionary trajectory of populations by determining traits associated with reproductive strategies (Huynh *et al.* 2011; Küpper *et al.* 2016), local adaptation and adaptive potential (Dorant *et al.* 2020; Huang *et al.* 2020; Cayuela *et al.* 2021; Kess *et al.* 2021; Tigano *et al.* 2021; Berdan *et al.* 2021). There is also growing evidence that SVs may lead to speciation (Davey *et al.* 2016; Todesco *et al.* 2020; Funk *et al.* 2021).

Previous studies exploring SV diversity in natural populations have generally combined multiple sequencing technologies (e.g., short- and long-read sequencing, optical mapping) and large sample sizes (reviewed in Wold *et al.* 2021). Further, many studies to date have aimed to identify SVs in close association with specific traits of interest and subsequently validate them with more traditional approaches (e.g., vonHoldt et al., 2017). There is ample opportunity to develop 'good' practice to reliably investigate population-level differences in SV frequency, location or size distributions in non-model species. However, agricultural and human genomics studies have identified caveats to consider before using short-read sequence data to call SVs. For example, we expect to observe a high false-positive rate and biases in the type and size range of SVs detected (English *et al.* 2015; Cameron *et al.* 2019; Mahmoud *et al.* 2019; Ho *et al.* 2020). This is in part because SV discovery tools commonly use discordant reads (i.e., those that are improperly aligned and/or depart from expected and observed insert lengths) and read depth to identify putative variants (Alkan *et al.* 2011; Rausch *et al.* 2012; Layer *et al.* 2014; Chen *et al.* 2016; Cameron *et al.* 2017). Although discordant reads do occur as a result of 'true' SVs, they may also arise as the result of mapping/sequencing error or reference error (Hurgobin and Edwards 2017; Bayer *et al.* 2020).

Distinguishing between the underlying sources of discordant read mapping generally requires independent data, such as extensive long-read sequencing, PCR amplification and Sanger sequencing, or Optical mapping (Ho *et al.* 2020). Such resource intensive approaches may not be feasible for many non-model species, especially those of conservation concern. Given that long-read sequences have been shown to outperform short-read data for SV discovery (Alkan *et al.* 2011; Mahmoud *et al.* 2019; Chaisson *et al.* 2019; Mérot *et al.* 2022), researchers may choose to use a strategic approach that combines long-read sequencing for SV discovery and short-read sequencing for population-scale genotyping (e.g., Huddleston *et al.* 2017; Chander *et al.* 2019; Jun *et al.* 2021). Guidelines around the application of genotyping SVs with short-read data in non-model species remain somewhat unclear (e.g., target sequence depth, ideal read insert size distribution, considerations for polyploids). This is in large part due to the lack of datasets–excluding human genomic datasets–suitable for benchmarking SV discovery and genotyping strategies (e.g., Cameron *et al.* 2019; Kosugi *et al.* 2019).

The critically endangered kākāpō is a nocturnal ground parrot endemic to Aotearoa New Zealand. Once widely distributed throughout the North and South Islands of Aotearoa, kākāpō populations rapidly declined as a result of anthropogenic disturbances and introduced mammalian predators (Williams 1956; Lloyd and Powlesland 1994; Veltman 1996). Populations continued to decline across the mainland and are believed to have gone extinct on the North Island in the 1930's. The last known South Island population was lost in the 1980's (Lloyd and Powlesland 1994). A relict population was discovered on Rakiura (Stewart Island) in 1977 and a translocation of a small handful of kākāpō found in Fiordland National Park on the West Coast of the South Island was attempted (Best and Powlesland 1985; Lloyd and Powlesland 1994). However, only one individual from Fiordland successfully bred with individuals from Rakiura. After intensive conservation management interventions, the kākāpō population has grown from a record low of 51 individuals in 1995 to ~200 individuals as of the 2021/2022 breeding season (Kākāpō Recovery Group 2017; Kākāpō Recovery Group *personal communications*). In fact, of the ~200 birds discovered on Rakiura and in Fiordland National Park, the extant kākāpō population can be traced back to only 35 founding individuals (Kākāpō Recovery Team *personal communications*). In an effort to mitigate the effects of small population size and inbreeding in kākāpō, island translocations are partially informed by pedigree data and more recently, genomic estimates of relatedness as a result of the Kākāpō125+ consortium (Guhlin *et al.* 2022 preprint). Briefly, as described in Guhlin *et al*. (2022), to inform kākāpō conservation efforts, the Kākāpō125+ project was initiated in 2015 to sequence all 125 living kākāpō at the time. Between 2015 and 2018, whole-genome short-read sequence data for these 125 individuals, and an additional 44 deceased adults and chicks, were generated for a total of 169 sequenced individuals. The Kākāpō125+ project has established a near-whole species high-quality variant dataset for a species of conservation concern and presents an exciting opportunity to explore strategies for SV discovery and genotyping in a non-model organism. Here, we combine these data with long-read sequence data for a subset of highly represented individuals, a highly contiguous reference genome (Rhie *et al.* 2021), and extensive life history data for all individuals, including verified pedigree relationships (Bergner *et al.* 2014; Galla *et al.* 2021) to compare four short-read and two long-read SV discovery and genotyping strategies to assess how each impacts inferences about SV frequency and size distributions in kākāpō. This study represents a critical first step towards our understanding the eco-evolutionary dynamics of SVs in small populations (Wold *et al.* 2021).

Materials and Methods

All details regarding read processing, variant discovery, genotyping and analyses may be found in the followingGitHub repository: https://github.com/janawold1/2022_MER_Submission.

Read processing and alignment

A highly contiguous reference genome, assembled by the Vertebrate Genome Project (VGP), is available for a single female kākāpō, 'Jane' (Rhie *et al.* 2021). As part of the Kākāpō125+ project, paired-end sequence libraries for 94 males and 75 females were sequenced to a target depth of 30x coverage on multiple Illumina platforms, including MiSeq2500, TruSeq Nano, and HiSeqX. Read lengths varied from 125 - 150bp. All preprocessing of raw sequence data was conducted by JG to maintain consistency across Kākāpō125+ subprojects. Briefly, reads were trimmed, adaptor content removed, and overlapped reads were collapsed into a single read using the default quality thresholds (minimum quality of 2) for fastp v0.20.0 (Chen *et al.* 2018) and AdapterRemoval v2.2.4 (Schubert *et al.* 2016). These processed reads were aligned to the reference genome and a machine learning program, DeepVariant (Poplin *et al.* 2018), employed to generate high quality SNPs for downstream analyses led by the Kākāpō125+ consortium (Guhlin *et al.* 2022 preprint). For short-read based SV discovery, reads were aligned to the reference genome using Burrows-Wheeler Aligner v0.7.17 (BWA; Li & Durbin, 2009).

In addition to the near-whole species resequence data, ten individuals highly represented in the extant population (5 male, 5 female), were targeted for long-read sequencing on the Oxford Nanopore Technologies platform. All individuals were sequenced on a MinION using R9 flow cells using the PCR-free LSK-110 ligation sequencing kit. Basecalling was performed using Guppy v6.3.7 (Anon n.d.) using the 'super' accuracy model (dna_r9.4.1_450bps_sup). Adapters were trimmed using Porechop v0.2.4 (Wick 2022), lambda DNA removed using NanoLyse v1.2.0 (De Coster *et al.* 2018) and reads were filtered for a minimum Q-score of 10 and read length of 3kb using NanoFilt v2.8.0 (De Coster *et al.* 2018). Both the raw and filtered long-read quality were visualised using NanoPlot v1.39.0 (De Coster *et al.* 2018). For long-read based SV discovery, reads were aligned to the reference genome using Winnowmap v2.03 (Jain *et al.* 2020). Read mapping quality was assessed for both short- and long-read alignments using Mosdepth v0.3.3 (Pedersen and Quinlan 2018) and qualimap v2.2.2 (García-Alcalde *et al.* 2012), with summaries of outputs from these tools visualised using MultiQC v1.13 (Ewels *et al.* 2016). A minimum alignment depth of 4x was required for inclusion in long-read-based SV discovery.

The highly contiguous VGP reference genome assembly (Jane's genome) represents a female kākāpō and thus includes both the Z and W sex chromosomes. This may be problematic for SV discovery as the W sex chromosome contains highly repetitive content homologous with content throughout the genome (Rhie *et al.* 2020). A preliminary analysis of SNPs indicated that this homology resulted in sufficient numbers of reads mapping to the W chromosome that erroneous heterozygous SNP calls were produced in both females and males (data not shown). Given that males are the homogametic sex (ZZ) and females are heterogametic (ZW), heterozygous SNP calls on the W for either sex indicate mis-mapping. To address these challenges, reads were realigned for all individuals excluding single-end reads and excluding the W chromosome from male alignments. Alignment for females also excluded single-end reads, but included the W chromosome scaffold to ensure that reads belonging to the W did not interfere with SV discovery on other chromosomes. For joint analyses of the kākāpō population, the Z and W chromosomes and all unplaced scaffolds were excluded from downstream analyses due to low confidence in variant discovery for these scaffolds.

Structural variant discovery and genotyping

Short-read structural variant discovery was conducted with Delly v0.8.7 (Rausch *et al.* 2012), Manta v1.6.0 (Chen *et al.* 2016) and the wrapper programme Smoove v0.2.8 (Pedersen *et al.* 2020a), which implements Lumpy-sv v0.2.13 for SV discovery (Layer *et al.* 2014), annotates variants with Duphold v0.2.1 (Pedersen and Quinlan 2019) and genotypes SVs with SVTyper v0.7.0 (Chiang *et al.* 2015). Long-read SV discovery was conducted using CuteSV v1.0.11 (Jiang *et al.* 2020) and Sniffles v2.0.7 (Sedlazeck *et al.* 2018), and raw individual calls were refined for population genotyping using Jasmine v1.1.5 (Kirsche *et al.* 2021).

Each SV discovery tool differs in approach. For the short-read based discovery approaches, both Delly and Smoove (i.e., Lumpy-sv) implement two algorithms (paired-end and split-read), while Manta implements three (paired-end, split-read and assembly-based). The short-read tools also differ in the suggested strategy for population-level SV discovery. Both Delly and Smoove iterate through individual samples and subsequently merge SV calls for individual genotyping, whereas Manta recommends conducting SV discovery with all available samples at once to increase power (Chen *et al.* 2016). However, due to the assembly-based algorithm, Manta is computationally resource-heavy, and running >10 individuals at ~30x sequence coverage set can often exceed 125 Gb RAM (as observed in the Kakāpō125+ data). In instances where computational resources are limited, samples may be run in batches or individually, although this is not recommended due to the loss of power to resolve SVs and the challenges associated with merging variants called in different sample batches (Anon 2016b; Anon 2016a).

To assess the impacts of using a batched vs. joint calling strategy for SV discovery, Manta was run in two ways: 1) a batched approach where samples were divided into 14 batches (7 male batches and 7 female batches) with an average of 11 individuals per batch (Manta-Batch); and 2) a joint approach where all males were run together and all females were run together. For both datasets, male and female SV discovery was conducted separately due to the inclusion of the W chromosome in female alignments (Manta-Joint). In both cases, variants were merged into 'batched' and 'joint' datasets with BCFtools v1.12 (Danecek *et al.* 2021) with the merge -m all flag.

Long-read SV discovery approaches must incorporate methods to account for the low accuracy associated with long-read sequence data (Sedlazeck *et al.* 2018; Jiang *et al.* 2020). The two tools included here (CuteSV and Sniffles) also attempt to address two challenges associated with alignment-based SV discovery. For example, CuteSV uses multiple signature extraction methods to distinguish SVs from the background noise of long-read data, then implements clustering and refinement approaches to increase sensitivity and identify the signature of heterozygous SVs (Jiang *et al.* 2020). Sniffles similarly identifies the signature of different SV classes, but implements additional methods to resolve nested SVs (Sedlazeck *et al.* 2018). SV discovery for both tools is performed on an individual-basis. Jasmine, which implements a modified minimum spanning forest algorithm, was used to merge SVs detected in individual kākāpō in each call set in preparation for population-scale genotyping with the available short-read data.

Regardless of discovery strategy, nominal genotype outputs from SV discovery tools are generally regarded as unreliable (Chander *et al.* 2019). To address this, both Delly and Smoove include genotyping programs (delly genotype, and SVTyper respectively), yet Manta, CuteSV and Sniffles do not. To genotype these call sets at the population-scale, SVs were filtered (as described below) and genotyped using the aligned kākāpō125+ short-reads with the genotyping tool BayesTyper v1.5 (Sibbesen *et al.* 2018). BayesTyper uses alignments of k-mers to a variant graph and reference genome, then implements a probabilistic model of k-mer counts to genotype individuals. BayesTyper has the benefit of being able to genotype a wide range of genomic variants (e.g., SNPs, small INDELs and SVs), in fact the inclusion of SNP data is recommended as it aids in matching relevant k-mers to sequence reads (Anon 2019). Each VCF output from Manta was processed with the program BayesTyperTools convertAllele to convert symbolic allele notations to REF and ALT sequences. This step was not necessary for the long-read based call sets as they already provided REF and ALT sequences. For both Manta call sets (batch and joint), CuteSV and Sniffles, a SNP call set generated with DeepVariant (Guhlin *et al.* 2022 preprint) was used to aid SV genotyping. All VCFs were normalised, variants left-aligned and any multiallelic sites split with BCFtools norm prior to merging variants with BayesTyperTools combine. Finally, BayesTyper requires the generation of large intermediate files (>2Tb for this dataset) with the tool KMC (Kokot, Długosz, & Deorowicz, 2017). As recommended, KMC v3.1.1 was run with k=55 and singleton k-mers included (-ci1) and a k-mer bloom filter for each individual was generated with BayesTyperTools makeBloom. Since BayesTyper cannot genotype more than 30 individuals at once, samples were batched into 5 groups of 30 and 1 group of 19 individuals prior to identifying variant clusters with BayesTyper cluster and genotyping with BayesTyper genotype under default settings.

Filtering Parameters

Once SV discovery and genotyping were complete, filtering for each SV dataset was conducted in two stages for: 1) SV call quality; and 2) individual genotype quality. The outputs from SV call quality filters were used for comparisons of SV type frequency, size distributions and location (i.e., frequency per chromosome) between tools (described further in the *Structural variant analyses analyses* section below). For comparisons of genotype consistency and variability among individual kakāpō, the outputs from genotype quality filters were used (see *Structural variant analyses* below).

Upon completion of SV discovery, removal of SVs marked as low quality, and additional recommended filtering parameters specific to each tool, were implemented using BCFtools. A standardised filtering approach was not applied to outputs from all three short-read tools, since each program recommends different statistics to assess the quality of SVs and genotypes (Pedersen *et al.* 2020b; Anon 2022a; Anon 2022b). Structural variant filtering for all short-read tools excluded all breakends, and SVs ≥50kb in length as these likely represent unresolved complex variants, mapping error, and/or reference bias. Additional filtering for Delly excluded duplications and inversions <300bp, and deletions <50bp using the delly merge -m option. All remaining SVs that did not pass all variant call quality filters were removed with BCFtools (i.e., INFO/FILTER = "PASS"'). This excludes all SVs where paired-end support was <3 and a MAPQ score <20 (Anon 2022a). Finally, genotype filtering for Delly SVs excluded all sites where <80% of variable genotypes passed all genotype filters with BCFtools (i.e., FMT/FT="PASS").

For Smoove, the lumpy_filter program identifies and discards interchromosomal read pair mismatches >3, and those with alternative matches, unless the identified split matches the location of the mate pair. This inbuilt filtering programme also removes reads where the depth is greater than 1,000x, as well as any orphaned reads. Variants are then genotyped and ready for annotation with the Smoove annotate programme. Once these steps were complete, all breakends, deletions that did not have a depth fold-change relative to flanking regions (FORMAT/DHFFC) < 0.7, and duplications that did not have a depth fold-change relative to bins in the genome with similar GC-content (INFO/DHBFC) > 1.3 were excluded using BCFtools (Pedersen 2022). For genotype filtering, an overall Mean Smoove Het Quality (INFO/MSHQ) ≥ 3 was implemented with BCFtools (Pedersen *et al.* 2020b). The Smoove Het Quality (INFO/SHQ) metric scores confidence in individual heterozygous genotypes where 1 is a low confidence call and 4 is highest, with MSHQ representing the mean score for all heterozygous genotypes (Pedersen *et al.* 2020b).

Variants for both the Batch and Joint Manta outputs were filtered using BCFtools to exclude all variants <50bp in length, all breakend calls and all variants that did not pass all record-level filters (i.e. INFO/FILTER=PASS). Specifically, this excluded: all sites with a QUAL score <20; deletions and duplications not consistent with diploid expectations; SVs with breakpoint depths >3x the median chromosome depth; SVs <1kb in size where >40% of samples contained a MAPQ score of 0 around either breakend; all SVs that were significantly larger than the paired-read fragment size and did not have paired-read support for the alternate allele in any individual; and finally, SVs where no sample passed all sample-level filters.

Filtering of the CuteSV and Sniffles call sets was relatively simple, with all imprecise sites excluded from both call sets. However, it is notable that while the CuteSV had sufficient read depth to filter for SV specificity (i.e., INFO/IS_SPECIFIC=1), Sniffles did not retain any SVs once this metric was implemented. As a result, the Sniffles call set was not filtered on SV specificity in this study.

The SV call sets for both Manta datasets, CuteSV and Sniffles were genotyped using BayesTyper, which implements four ‘hard’ genotype filtering parameters by default. This includes the exclusion of variants with fixed heterozygous genotypes, alleles with <1 sampled k-mer, genotypes with a posterior probability <0.99, and alleles with k-mer coverage that fall below $1\text{-}e^{-0.275x}$. Here, x represents the mean of the negative binomial distribution for k-mer coverage for a specific sample (Sibbesen, 2018 GitHub). All variants with >20% genotypes missing and variants where <80% of genotypes passed all four BayesTyper quality metrics were excluded. Although BayesTyper ships with a programme for converting allele sequences to symbolic alleles (bayesTyperTools convertSeqToAlleleID), we found It challenging to resolve the class of all genotyped variants with this approach (i.e., insertions are incompatible and additional SV classes were changed or remained unresolved). To relate genotyping results back to the called SV class, BCFtools was used to identify the chromosomal positions of the genotyped variants and compared with the locations of SVs prior to file conversion with bayesTyperTools convertAllele.

Structural variant analyses

Structural variants were counted for each SV discovery tool prior to and after filtering. To explore the level of call consensus between these outputs, the number of overlapping SVs were identified using SURVIVOR v1.0.7 (Jeffares *et al.* 2017) in 1kb, 500bp, 50bp windows and for exact overlaps. To count as a consensus call, SV type and strand were required to match and a minimum variant length of 50bp were required. To assess whether some chromosomes carried more SVs relative to their size, we estimated the number of SVs per chromosome and the proportion of base-pairs of each chromosome within an SV (i.e., the sum of all SV lengths for a given chromosome / chromosome size).

Following SV discovery across the six approaches, all individuals were genotyped using the aligned kākāpō125+ short-read dataset. The genotype filtered SV data for all six variant call sets were used for comparisons of individual variability, assessing shifts in the the number of SVs per generation, and to assess population structure of SVs. When reporting the number of SVs per individual and number of SVs among kākāpō cohorts, we use presence or absence of SVs per individual. That is, we consider genotypes as evidence of whether or not the individual carries the SV (0/1 & 1/1 = carrier; 0/0 = non-carrier). Both Fiordland- and Rakiura-derived birds (herein, founders) were used for comparisons across three cohorts (n = 1, 3, 4 for Fiordland founders, F1 and F2 and n = 40, 60, 10 for Rakiura founders, F1 and F2 respectively). Due to the lek mating system and a relatively long lifespan, the kākāpō population has had significant backcrossing through the generations. Therefore, the F1 and F2 generations represented here excluded all individuals with backcrossed lineages, as this may bias true generational patterns in SVs carried by individuals. Finally, to compare variability in the SVs carried by individual kākāpō, genotypes from the genotype filtered SV data for all four strategies was used to conduct a discriminant analysis of principal components (DAPC) with the adegenet R package (Jombart 2008). Only individuals used for generational comparisons (n = 118) were used to assess individual variability and SV population structure.

In the absence of a previously validated catalogue of SVs, neither a 'true' nor 'false' positive rate of detection could be assessed. Despite not being able to estimate the precision and accuracy of SVs called in our data, we aimed to test the consistency of genotyping results using Mendelian Inheritance tests with parent-offspring trios. Although this does not eliminate the possibility of systematic error, nor does it provide an indication of the precision or accuracy of SV detection, departures from Mendelian Inheritance may indicate inconsistency of genotyping within a given SV call. Genotyping consistency is an important consideration for population studies as patterns of population structure or inferences about local adaptation may be impacted by inconsistencies.

To identify SVs that violate Mendelian Inheritance patterns, the BCFtools +mendelian plugin was used. Pedigree data provided by the New Zealand Department of Conservation identified 120 parent-offspring trios consisting of 158 unique individuals in the Kākāpō125+ sequence data. We tested SV genotypes by calculating the proportion of Mendelian Inheritance errors relative to the number of non-missing genotypes (i.e., GT != "mis"). Four thresholds were tested where adherence to Mendelian Inheritance expectations were either 100%, ≥95%, ≥90% and ≥80% of genotypes passed. It is important to note that not all 169 sequenced individuals were represented in pedigree trios, as they may not have descendants or antecedents represented in the short-read data analysed here. In addition, some individuals are represented multiple times in different family groups. This bias towards highly represented individuals in the kākāpō breeding population may not adequately capture all SVs called within the population. As such, we did not filter SVs using Mendelian Inheritance errors for downstream analysis.

Rather, these tests may provide some insights into the relative performance of genotyping approaches among the pipelines used here.

Results

The mean individual mapping depth of short-reads for autosomal chromosomes was ~18x, and ranged from ~9x to ~38x. Of the 10 individuals sequenced using long-reads, 7 met the minimum depth threshold of 4x coverage for long-read SV discovery. The mean individual mapping depth of long-reads for autosomal chromosomes was ~10x, and ranged from ~4x to ~16x. There was considerable variability in the number of SVs initially detected by each of the six approaches (herein datasets), with the most being the Manta-Batch and fewest being the CuteSV dataset (Table 1). In addition, Inversions were the most common SV type detected in short-read discovery tools, while Deletions were more common in long-read SV discovery tools. This pattern was consistent across call quality and genotype filtering thresholds (Table 1). The proportion of SVs passing call quality thresholds also varied, with Delly retaining the lowest proportion of SVs (~4%). Both the Manta-Batch and -Joint call quality filters retained roughly 26% of variants, whereas 27% of CuteSV and 32% of Sniffles variants were retained. The Smoove call set retained the highest proportion of SVs with ~68% passing call quality thresholds (Table 1). Although the size distribution for each filtered SV type varied somewhat between each of the SV discovery tools. It is notable that although a minimum size threshold of 50bp was implemented in Delly, all reported insertions were under this threshold (Table 2).

Table 1. Counts of structural variants (SVs) by type for unfiltered variants, those retained after SV quality filters and after genotype quality filters specific to each of the structural variant discovery tools Delly, Manta and Smoove.

		Unfiltered	**Call Quality Filters**	**Genotype Filters**
Delly	Breakends	9,672	0	0
	Deletions	5,167	696	57
	Duplications	2,099	73	12
	Insertions	473	441	228
	Inversions	35,397	753	437
	Total	**52,808**	**1,963**	**734**
Manta - Batch[1]	Breakends	71,872	0	0
	Deletions	4,236	1,614	515
	Duplications	1,907	510	70
	Insertions	1,803	749	177
	Inversions	60,434	32,959	342
	Total	**140,252**	**35,832**	**1,104**
Manta - Joint[2]	Breakends	63,740	0	0
	Deletions	2,915	1,194	495
	Duplications	1,246	294	73
	Insertions	1,538	221	74
	Inversions	58,393	30,363	301
	Total	**127,832**	**32,072**	**943**
Smoove	Breakends	4,635	0	0
	Deletions	1,899	1,505	1,023
	Duplications	973	435	183
	Insertions	N/A	N/A	N/A
	Inversions	10,068	10,037	2,825
	Total	**17,575**	**11,977**	**4,031**

CuteSV	Breakends	1,048	0	0
	Deletions	3,864	1,209	72
	Duplications	254	138	0
	Insertions	2,972	879	6
	Inversions	18	12	0
	Total	**8,156**	**2,238**	**78**
Sniffles	Breakends	5,068	0	0
	Deletions	2,624	2,734	87
	Duplications	99	61	0
	Insertions	3,893	2,339	39
	Inversions	253	95	0
	Total	**11,937**	**5,229**	**126**

[1] Samples divided into 14 batches (7 male batches and 7 female batches) for SV discovery

[2] Samples divided into a male specific and female specific batch for SV discovery

Table 2. Summary of structural variant size characteristics for Delly, Manta and Smoove data sets filtered for SV call quality.

Data	Structural Variant Type	Count	Size Range (bp)	Median Size (bp)	Mean Size (bp)
Delly	Deletions	696	49 - 26,273	922	2,181
	Duplications	73	355 - 34,273	3,592	6,476
	Insertions	441	22 - 46	29	30
	Inversions	753	300 - 48,626	369	1,088
Manta - Batch	Deletions	1,614	50 - 47,230	623	3,216
	Duplications	510	52 - 40,508	1,976	5,919
	Insertions	749	51 - 1,704	575	461
	Inversions	32,959	51 - 49,035	202	458
Manta - Joint	Deletions	1,194	50 - 47,230	329	1,773
	Duplications	294	52 - 44,414	307	4,588
	Insertions	221	56 - 888	315	341
	Inversions	30,363	51 - 49,035	192	383

Smoove	Deletions	1,505	53 - 47,780	696	3,123
	Duplications	435	148 - 47,433	4,108	8,873
	Insertions	N/A	N/A	N/A	N/A
	Inversions	10,037	76 - 45,629	686	1,039
CuteSV	Deletions	1,209	39 - 47,874	170	847
	Duplications	150	198 - 97,051	9,420	12,380
	Insertions	879	36 - 32,549	151	578
	Inversions	12	258 - 31,628	1,350	6,190
Sniffles	Deletions	2,734	49 - 47,873	135	690
	Duplications	61	211 - 87,106	9,118	14,928
	Insertions	2,339	45 - 24,610	130	526
	Inversions	96	50 - 67,769	208	6,452

Consensus between the six call quality filtered datasets was relatively low, except when considering the two Manta datasets (~76%, n = 29,219 SVs). The next two tools with the highest proportion of agreement were the two long-read based call sets for CuteSV and Sniffles (~17 - 49% agreement, n = 1,099 SVs). The overall agreement between datasets tends to decrease as more tools are included in comparisons, leaving only 94 SVs (90 deletions, 4 duplications) overlapping in all six datasets (Figure 1). These SVs, ranging in size from 314bp to more than 20kb, were challenging to consistently genotype. Few passed genotype thresholds in each dataset, this included twelve deletions and two duplications in both Manta datasets, five deletions in the Smoove dataset and one deletion in the CuteSV dataset. It is challenging to glean a pattern in the overall agreement between datasets given the variability in the number of SVs passing call quality thresholds. For example, Sniffles tended to have a higher degree of overlap with short-read based call sets than CuteSV. However, the filtered Sniffles call set was more than twice the size of the filtered CuteSV call set.

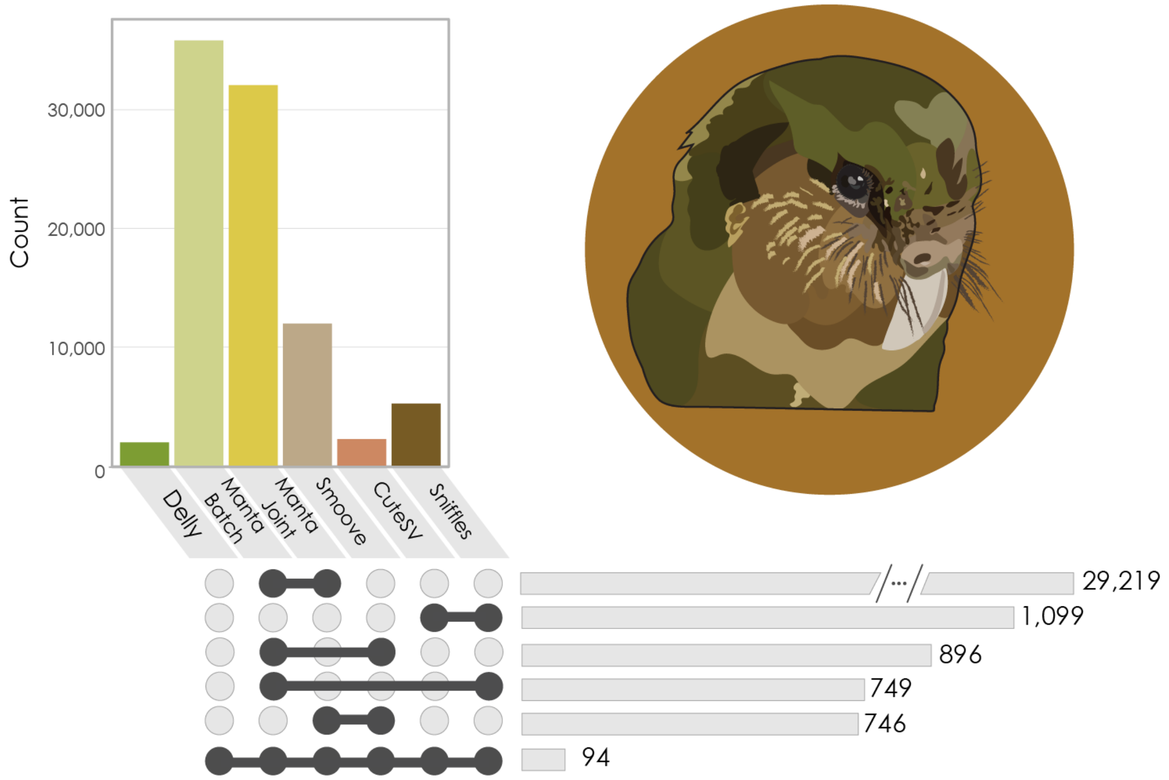


Figure 1. Counts of consensus calls between SV type and strand within a 50 bp window for the top five comparisons and the number of overlapping calls in all of the six datasets (i.e., Delly, Manta - Batch, Manta - Joint, Smoove, CuteSV and Sniffles) for kākāpō. The colored barchart on the upper left represents the number of SVs passing call-quality thresholds in each of the six datasets. Dark green circles with lines between denote which datasets have consensus SV calls. Bars to the right represent the number of SVs overlapping between these datasets. See Supplementary Figure 1 for a full comparison of all consensus calls and Supplementary Table 1 for a summary of the number and type of overlapping SVs.

The number of SVs found on each autosomal scaffold correlated with chromosome size in all six datasets (Figure 2a). This pattern was consistent when considering the proportion of chromosome base pairs impacted by SVs. However, there appeared to be variability in the type of SV impacting these chromosomes with inversions tending to impact the largest proportion of base pairs in the short-read datasets. In contrast, duplications tended to affect the largest proportion of the smaller chromosomes in the long-read datasets. Further, there was some variability in which of the smaller chromosomes were most impacted (Figure 2b).

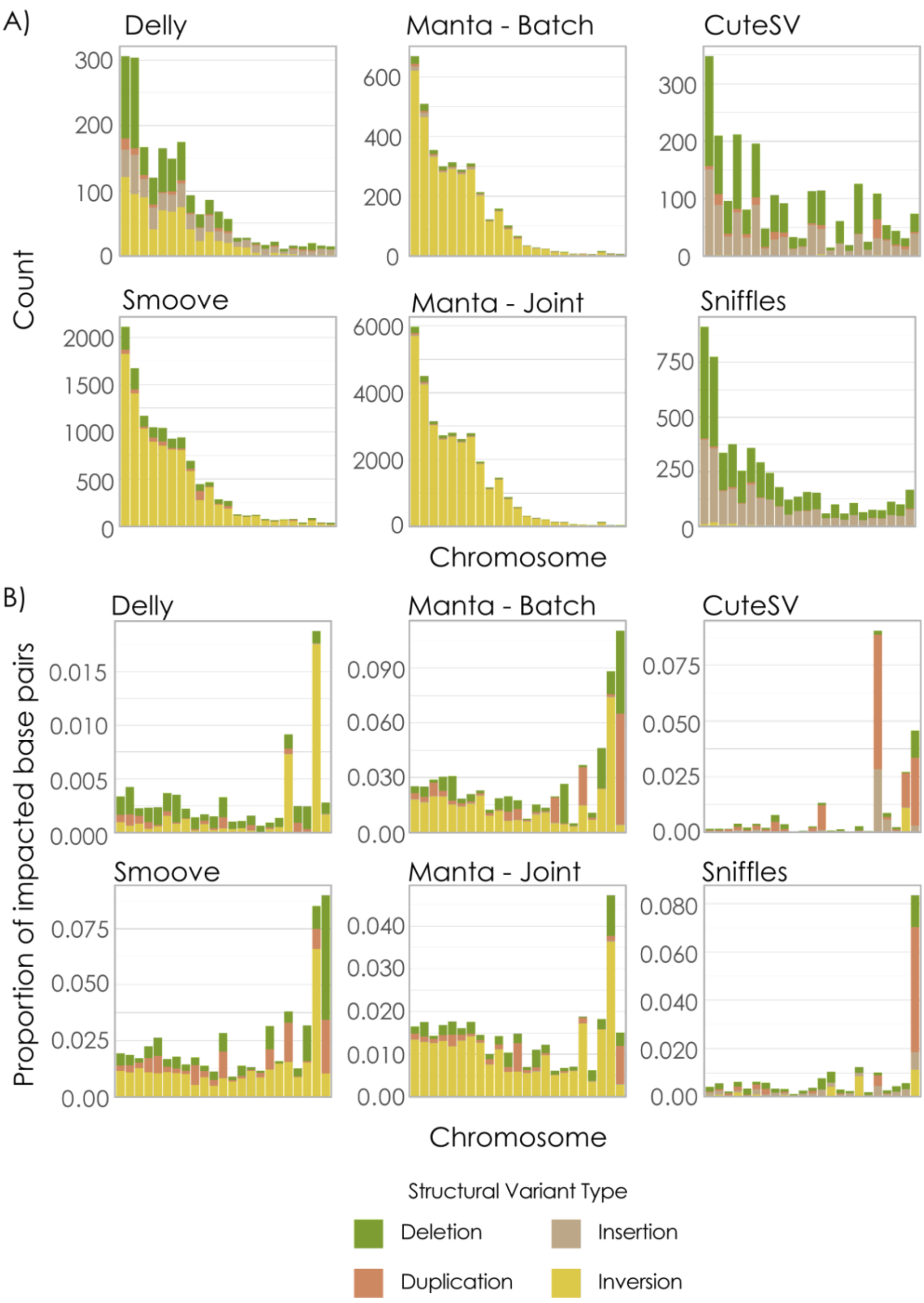
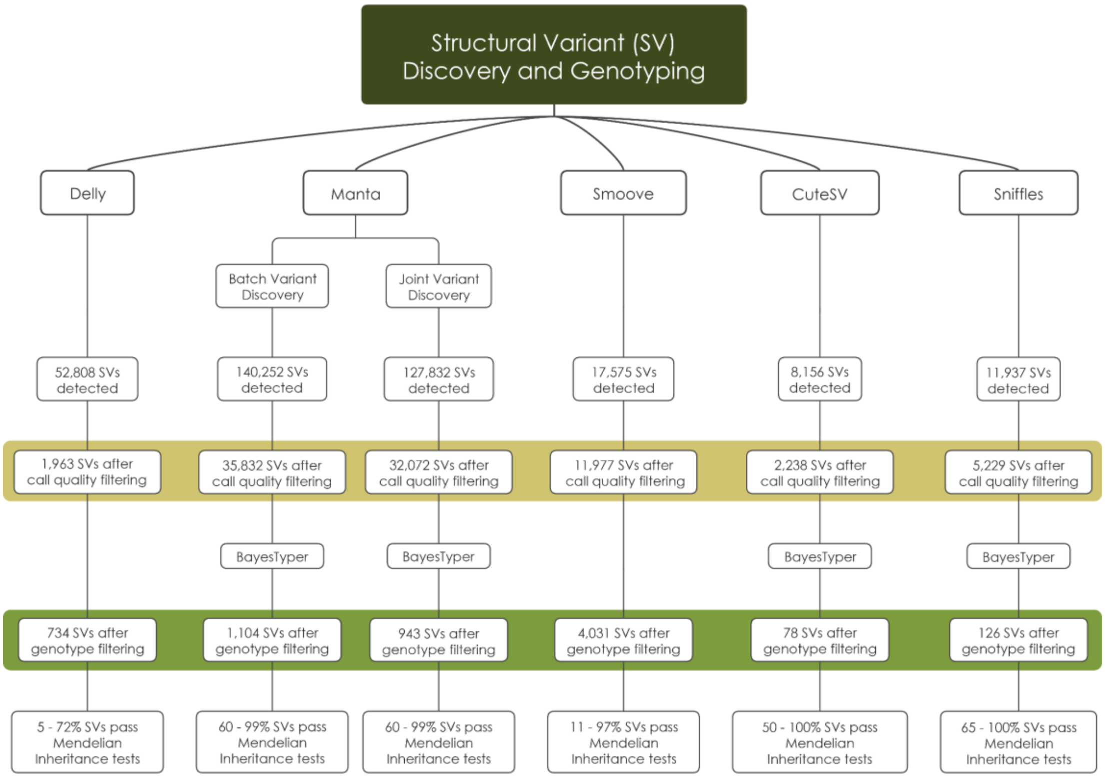


Figure 2. Structural variant (SV) counts per chromosome as called in the short-read SV discovery tools Delly, Manta-Batch, Manta-Joint and Smoove, as well as the long-read SV discovery tools CuteSV and Sniffles. For each, call sets were filtered for SV quality and the number of SVs per chromosome (A), and the proportion of chromosome base-pairs impacted by structural variants (B) were estimated. Chromosomes are ordered left to right by size, excluding the Z and W sex chromosomes. The largest chromosome, chromosome 1, consistently carried the highest number of SVs detected in all six datasets. However, the smallest chromosomes consistently had the highest proportion of base pairs impacted by SVs (i.e., sum of all SV lengths / chromosome size) in all six datasets.

The results reported thus far have focused on the SVs retained after overall 'call quality filtering', or those SVs that passed quality thresholds irrespective of individual genotype quality. Figure 3 summarises the results of SVs that passed both call quality thresholds and genotype quality thresholds.



Structural Variant (SV) Discovery and Genotyping
Delly
Manta
Smoove
CuteSV
Sniffles
Batch Variant Discovery
Joint Variant Discovery
52,808 SVs detected
140,252 SVs detected
127,832 SVs detected
17,575 SVs detected
8,156 SVs detected
11,937 SVs detected
1,963 SVs after call quality filtering
35,832 SVs after call quality filtering
32,072 SVs after call quality filtering
11,977 SVs after call quality filtering
2,238 SVs after call quality filtering
5,229 SVs after call quality filtering
BayesTyper
BayesTyper
BayesTyper
BayesTyper
734 SVs after genotype filtering
1,104 SVs after genotype filtering
943 SVs after genotype filtering
4,031 SVs after genotype filtering
78 SVs after genotype filtering
126 SVs after genotype filtering
5 - 72% SVs pass Mendelian Inheritance tests
60 - 99% SVs pass Mendelian Inheritance tests
60 - 99% SVs pass Mendelian Inheritance tests
11 - 97% SVs pass Mendelian Inheritance tests
50 - 100% SVs pass Mendelian Inheritance tests
65 - 100% SVs pass Mendelian Inheritance tests

Figure 3. Overview of structural variant (SV) discovery and genotyping strategies in the Delly, Manta-Batch, Manta-Joint and Smoove call sets for kākāpō. Delly and Smoove each have their own in-built genotyping programs, while Manta, CuteSV and Sniffles do not. Variants called by Manta, CuteSV and Sniffles were genotyped using the BayesTyper genotyping software package. Data were analysed in two steps: 1) An initial filtering threshold(s) for call quality used for comparisons of SV type, size distributions and overlaps (in gold); and 2) genotype quality threshold(s) used to explore variability in number of SVs carried by individuals and genotype consistency among tools (in green). The proportion of SVs passing Mendelian Inheritance were estimated across a range of thresholds (Table 3).

Overall, the relative proportion of SV classes that pass genotype filtering thresholds followed a similar pattern to those that passed SV call quality thresholds with the most SVs being retained on the largest chromosomes. While the size distribution of SVs was somewhat similar to those filtered for call quality (Supplementary Figure 2 and Supplementary Table 2), the proportion of individual chromosomes impacted was no longer consistent among tools and did not follow a clear pattern (Supplementary Figure

3). Of the genotyped filtered datasets tested for Mendelian Inheritance, Sniffles had the highest proportion of genotypes pass all parent-offspring trios tested for Mendelian Inheritance while the Delly dataset had the lowest proportion of SV pass at this threshold (Table 3). As the stringency of the Mendelian Inheritance tests were relaxed, the proportion of passing SVs increased for all datasets (Table 3).

Table 3. Number of SVs by type adhering to Mendelian Inheritance expectations in 100%, 95%, 90% and 80% of trios tested. Conversion of BayesTyper genotypes from long sequence format to symbolic alleles could only resolve a subset of all genotypes reported. As such, the exact proportion of these SVs exhibiting Mendelian Inheritance patterns could not be reported. Smoove does not call or genotype insertions.

	Deletions	**Duplications**	**Insertions**	**Inversions**	**Total**
Delly Genotype Filtered Counts	**57**	**12**	**228**	**437**	**734**
100% trios pass	16	2	15	1	34
≥95% trios pass	48	6	25	420	499
≥90% trios pass	54	8	27	436	525
≥80% trios pass	56	10	28	437	531
Manta / BayesTyper - Batch	**515**	**70**	**177**	**342**	**1104**
100% trios pass	320	30	122	190	662
≥95% trios pass	513	50	177	335	1075
≥90% trios pass	515	56	177	341	1089
≥80% trios pass	515	62	177	342	1096
Manta / BayesTyper - Joint	**495**	**73**	**74**	**301**	**943**
100% trios pass	311	33	64	159	567
≥95% trios pass	490	54	74	289	907
≥90% trios pass	494	57	74	300	925
≥80% trios pass	495	63	74	301	933
Smoove Genotype Filtered	**1023**	**183**	**N/A**	**2825**	**4031**
100% trios pass	347	44	N/A	56	447
≥95% trios pass	772	90	N/A	2556	3418
≥90% trios pass	894	115	N/A	2700	3709
≥80% trios pass	965	148	N/A	2800	3913
CuteSV / BayesTyper - Genotype	**72**	**0**	**6**	**0**	**78**
100% trios pass	36	0	3	0	39

≥95% trios pass	71	0	6	0	77
≥90% trios pass	72	0	6	0	78
≥80% trios pass	72	0	6	0	78
Sniffles / BayesTyper - Genotype	**87**	**0**	**39**	**0**	**126**
100% trios pass	57	0	25	0	82
≥95% trios pass	87	0	39	0	126
≥90% trios pass	87	0	39	0	126
≥80% trios pass	87	0	39	0	126

In general, the individual kākāpō that carried the highest number of SVs in one dataset also appeared to carry a relatively high number of SVs in other datasets (Figure 4). Depending on the dataset, there appeared to be either high variability in the number of SVs per individual (Delly & Smoove), or relatively little variability (both Manta datasets, CuteSV and Sniffles). Another interesting note is variability in SV type underlying these individual differences. For example, inversions are the dominant SV type among individuals carrying the most SVs in the Delly datasets, whereas deletions dominate in both Manta datasets, CuteSV and Sniffles. For the Smoove data, inversions are the most common SV type in individuals carrying the most SVs, despite deletions being more consistently observed across the population.

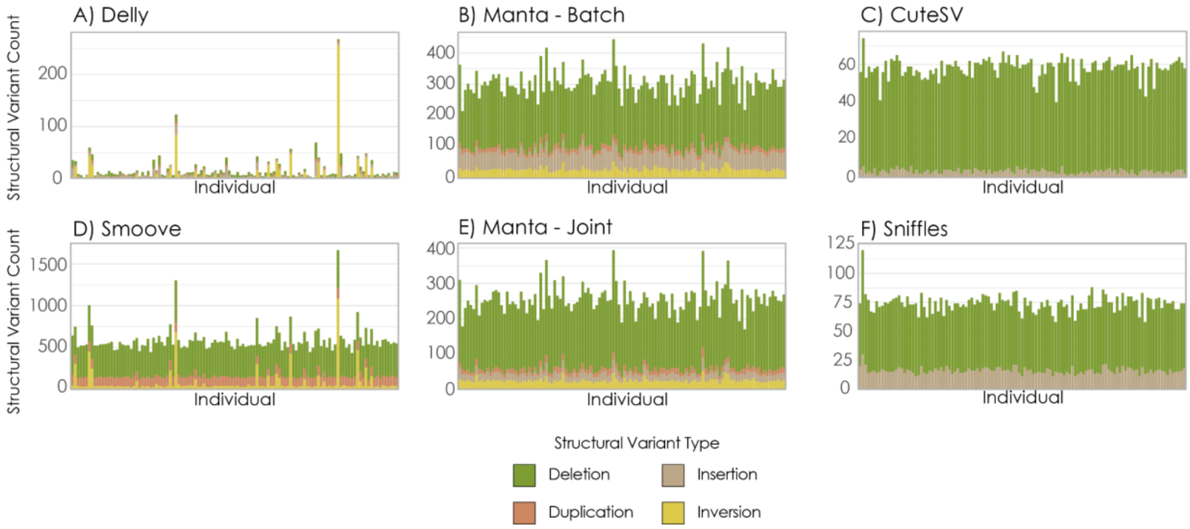


Figure 4. Relative counts of putative SV types carried by individual kākāpō. Individual kākāpō along the x-axis are in the same order in all four plots. There is some agreement among the four data types as to individual kākāpō carrying the highest number of SVs. For example, the individual carrying the highest number of SVs in the Delly dataset (A), is the same individual carrying the second highest number of SVs in the Smoove dataset (D). Upon closer inspection we found that the 3 individuals that consistently carried the most SVs in the Delly and Smoove datasets were not read mapping outliers (22.8x, 23.12x and 26.5x).

When evaluating generational trends in the number of SVs observed, there appears to be some agreement between the six datasets (Figure 5). Kākāpō that are descended from the individual successfully recovered from Fiordland tend to carry more SVs overall than birds with only Rakiura lineages. However, the number of SVs carried by Fiordland lineage kākāpō appears to decrease with each subsequent generation in both Manta datasets and Sniffles, while the number of SVs carried by Rakiura generations remains relatively stable.

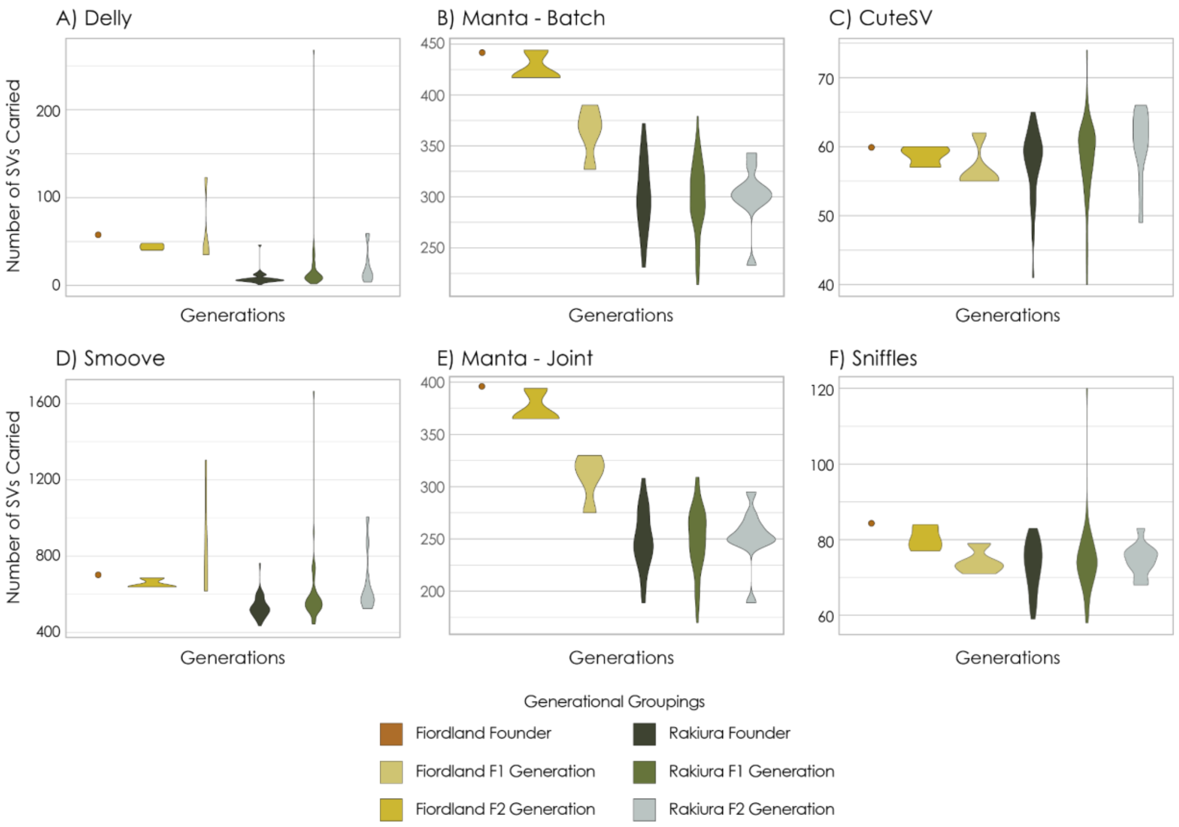


Figure 5. Distribution of SV counts per individual across kākāpō generations. Of the 41 founding individuals, only one originates from the mainland of New Zealand (Fiordland founder; Richard Henry). The sole representative of the Fiordland population had three offspring (Fiordland F1 Generation), one of which had four offspring (Fiordland F2 Generation). In contrast, the 40 founding individuals discovered on Rakiura have had a cumulative 60 offspring (Rakiura F1 Generation), who have in turn had 10 offspring (Rakiura F2 Generation) represented in this figure. First (F1) and second (F2) generation individuals exclude any backcrossed individuals.

Finally, the results of each discriminant analysis of principal components (DAPC) indicated that PC1 was driven by high variability among a few individuals for all six datasets. This variability largely reflected individuals of Fiordland lineage becoming more similar to Rakiura lineages with each successive generation. This pattern was consistent in both the CuteSV and Sniffles datasets, despite many fewer SVs passing genotype filtering thresholds (Table 1; Figure 6).

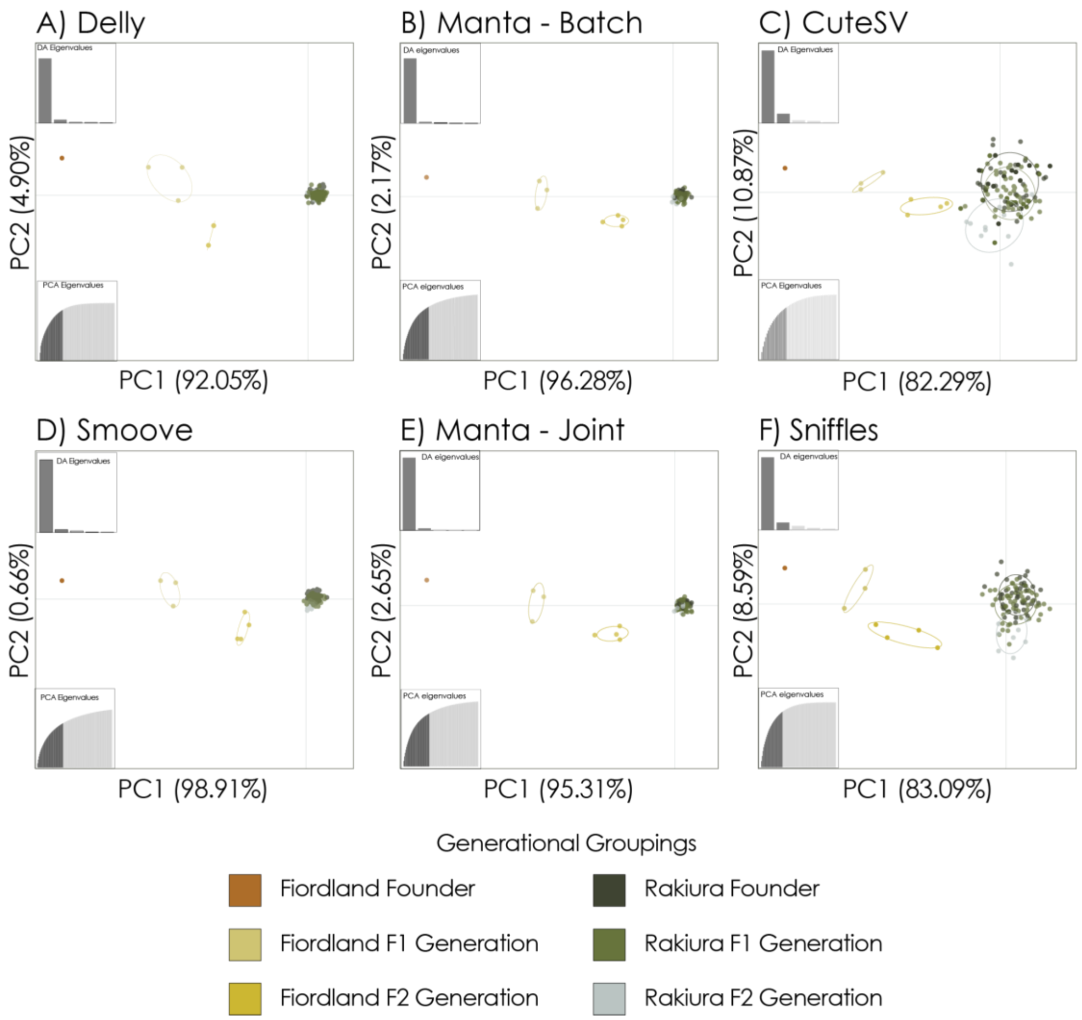


Figure 6. Genotypes from the genotype filtered data for Delly, Manta-Batch, Manta-Joint, Smoove, CuteSV and Sniffles datasets were used to construct a discriminant analysis of principal components (DAPC). Fiordland lineage birds form separate cluster(s) in each DAPC, but become more similar to Rakiura lineage birds with each successive generation.

Discussion

We explored six strategies for SV discovery and genotyping with short- and long-read data in the critically endangered kākāpō. We found that the choice of SV discovery tool heavily impacted the overall count, location, and size distribution of SV types characterised. Further, the proportion of SVs retained after filtering for SV call quality and genotype quality varied across all six datasets. Finally, after leveraging a meticulously curated pedigree, we also found that each genotyping tool had variable success in consistently genotyping high quality SVs. As a result, the number and type of SVs carried by individual kakāpō also differed. Nevertheless, there was some agreement between datasets as to which individuals carried a relatively high number of SVs. The general consensus among datasets was also reflected in the consistency of the number of SVs carried by each generation. Our combined results indicate that whereas inferences about population-scale trends are appropriate for kākāpō, direct comparisons between individuals birds are best avoided.

Implications of SV discovery strategies

The six SV discovery tools used here vary in the overall number of SVs detected, SV type, and their location. This variability may indicate that all six tools are sensitive to different mapping characteristics within the kākāpō short-read data, and suggests some complementarity between tools. Further, the lack of complete overlap in the location of SVs between the Manta datasets is interesting given the overall similarity in the number of SVs per chromosome and the overall counts of each SV type. The strategies used to call SVs with Manta differ only in the way that individuals were grouped during the initial SV discovery (i.e., samples divided into 14 batches, versus all males analysed jointly together and all females analysed jointly). Given that Manta incorporates local assembly of reads when detecting SVs, it is possible that different read sets have therefore led to differences in both the power and precision to accurately locate SVs in these analyses. Randomisation of sample batches would have aided in resolving this, however this was not possible due to computational resource limitation. Given the lack of consensus on the total number, location, or size of SVs called between methods, caution should be exercised when drawing conclusions about the specific characteristics of SVs identified here (e.g., size, relative frequency, proximity to genes/gene regions). Further work is needed to resolve the relative precision of each tool to identify population trends and the potential impacts of merging outputs from multiple tools.

All four short-read call quality filtered datasets had a very high prevalence of inversions. Both the individual-based strategy implemented by Delly and Smoove, as well as the multi-sample approach implemented by Manta, likely over-represented the number of inversions relative to other SV types. This is not surprising given the challenges associated with resolving inversion breakpoints, even after the merging of a consensus call set (Mahmoud *et al.* 2019; Ho *et al.* 2020). In addition, no clear filtering approach for consistently resolving well-supported inversion breakpoints emerged for the tools used here. It is notable that very few inversions are retained after genotype filtering, suggesting that this SV type may be particularly challenging to genotype using short read data. In some cases, this may be due to an inability to differentiate between one large inversion and overlapping inversion haplotypes when using short-read data (e.g., Kim *et al.* 2017; Knief *et al.* 2017; Hallast *et al.* 2021).

Overall, long-read based discovery strategies retained a relatively higher number of insertions than short-read discovery tools. This is not surprising given the known limitations of short-read data when characterising insertions (Delage *et al.* 2020). Another interesting observation from this study was the lack of duplications and inversions that passed genotyping quality thresholds in both long-read based callsets, despite overlaps between short- and long-read based discovery tools. On one hand, the long-read data may better characterise insertions and duplications, while genotyping these variants with short-read data may be somewhat problematic due to the low precision around variant breakpoints as a result of small long-read sample size and/or sequencing depth. Despite the small sample size used for long-read SV discovery, these approaches appear useful for assessing SV diversity of small populations.

When considering relative levels of individual SV diversity, there is some concordance between Delly and Smoove when identifying individuals with the highest number of SVs. However, it is notable that the SV type largely driving this pattern are inversions, which occur at a much lower frequency in the long-read datasets overall. This is surprising given that the long-read data should better resolve more complex variants like duplications, insertions and inversions (Alkan *et al.* 2011; Mahmoud *et al.* 2019; Chaisson *et al.* 2019; Mérot *et al.* 2022). Further work is needed to determine whether the small sample size and relatively low sequence depth for the long-read data impeded discovery of inversions, or whether these calls are largely false-positives in the short-read based datasets. However, the three kākāpō (two male, one female) that consistently had the most SVs in the Delly and Smoove datasets did not have obvious read-depth, or insert-length differences and were not outliers in the DAPCs presented here as they each clustered with their respective cohorts (Fiordland F2, Rakiura F1).

Addressing the challenges associated with quantifying and characterising individual SV diversity is important for kākāpō conservation. For example,we are able to infer population structure between the only founding individual successfully recovered from Fiordland, and his descendents, from Rakiura lineage birds. This is notable as it is consistent with SNP-based analyses (Guhlin *et al*. 2022 preprint). While this individual carries a higher number of SVs on average than birds solely from Rakiura lineages, our ability to detect and genotype SVs for this lineage may partly be accounted for by the fact that the kākāpō reference genome was assembled using a bird with pure Rakiura lineage. The high number of SVs detected in the Fiordland founder may be attributable to the comparison of groups of more- and less- related birds against a single reference. Given that the Fiordland founder is the only individual without direct relation to the Rakiura lineage, it is likely that he carries more genetic differences in comparison to the reference genome, and these differences are likely to be inherited by his descendents. A key question for ongoing conservation efforts is whether there are a number of SVs unique to the Fiordland lineage that have been lost in subsequent generations.

Conservation implications

One significant challenge for studying SVs in many species of conservation concern is the lack of resources available to generate independent data for SV validation (e.g., PCR amplification and Sanger sequencing, Optical Mapping). Without the ability to estimate a false-discovery rate, or verify the accuracy of specific tools, it is challenging to interpret these results or draw conclusions about the frequency and/or size of SVs in non-model species. However, we have been able to leverage the extensive pedigree data for kākāpō to assess the proportion of SVs adhering to Mendelian inheritance. Although concordance across all (100%) trios was low for some tools, it is promising to note that call and genotype-filtered SVs had between 72-100% concordance in at least 80% of trios. Where pedigree data is available, as will be the case for many intensively-managed threatened species, this additional filtering step is likely to enrich a SV set for true positives.

It remains difficult to draw reliable conclusions about the SVs characterised in any of the six datasets described here, but there is preliminary evidence that the overall number of SVs may be relatively stable from one generation to the next within the Rakiura individuals. This is exciting as the generations captured in this study cover the duration that the extant kākāpō population has been under active management. To date, conservation practitioners actively use pedigree and genetic/genomic data to inform translocations to off-shore islands, increase offspring contributions from relatively underrepresented lineages, and prioritise nests that are from relatively unrelated pairings (Cresswell 1996). As a result, the maintenance of genetic diversity in terms of overall SV counts per individual may be reflecting these efforts (Guhlin *et al.* 2022 preprint). Promisingly, these preliminary results suggest that SVs may provide a sensitive metric for monitoring the impacts of conservation actions on genome-wide diversity in species of conservation concern.

Future Directions

The factors driving reduced costs associated with generating short-read WGS data are also increasing the accessibility of long-read sequence data. Further, with advancements in bioinformatic approaches, such as pangenomes and genome graphs, many of the challenges associated with SV discovery with short-read data may be alleviated

(Hurgobin and Edwards 2017; Bayer *et al.* 2020; Ebler *et al.* 2020; Eizenga *et al.* 2020). For SV studies in species of conservation concern, it may be more economical to target a subset of highly represented individuals for long-read sequencing and the construction of genome graphs for SV discovery. Similar approaches are underway to better inform breeding and selection in agriculturally significant species such as cattle, soybean and tomato (Alonge *et al.* 2020; Cappetta *et al.* 2020; Liu *et al.* 2020; Talenti *et al.* 2022). Population-scale and individual-scale genotyping may then be possible with short-read data and assessments of population diversity may include both SNPs and SVs to better inform conservation management. In parallel with the increased application of these sequencing and bioinformatic approaches, we anticipate the inclusion of metrics tailored to SVs and their characteristics (e.g., size, type, location, genotype) into estimates of genome diversity across threatened individuals and populations, and any associated fitness consequences will be an area of active research with broad applicability to the conservation genomics space.

Data accessibility and benefit sharing

This research was undertaken as part of the Kākāpō125+ Project that includes research partnerships between the University of Canterbury's Conservation, Systematics and Evolutionary Research Team (ConSERT, including JRW,TES), Genomics Aotearoa (including AWS, JGG, PKD, TES), New Zealand Department of Conservation (DOC) and Te Rūnanga o Ngāi Tahu (TRONT). The goal of the Kākāpō125+ Project is to facilitate the development and implementation of conservation management strategies to enhance the recovery of this critically endangered taonga, or treasured, species. Approval to access the Kākāpō125+ short-read data used in this study was granted to TES and her research team by DOC and TRONT. The Kākāpō125+ Project short-read data is stored in the Aotearoa Genomic Data Repository (AGDR): https://data.agdr.org.nz/ and is is subject to the Kākāpō125+ Genomics Data Sharing Terms and Conditions described here: https://www.doc.govt.nz/our-work/kakapo-recovery/what-we-do/research-for-the-future/kakapo125-gene-sequencing/request-kakapo125-data/ . The generation of the long-read data was conducted under DOC authorisation (authorisation number: 97814-FAU) and enabled by High Quality Genomes and Population Genomics at Genomics Aotearoa. In accordance with FAIR and CARE data principles (Carroll *et al.* 2020; Carroll *et al.* 2021; Mc Cartney *et al.* 2022), the long-read data is also stored in the AGDR and data sharing subject to approval by DOC and TRONT.

Acknowledgements

We are grateful to the Kākāpō125+ Project led by Genomics Aotearoa in partnership with the New Zealand Department of Conservation (DOC) and Te Rūnanga o Ngāi Tahu (TRONT). The generation and availability of the short-read data used in this study is owed to the following additional parties: Genetic Rescue Foundation, University of Otago, New Zealand Genomics Limited, Rockefeller Institute, Duke University, Science Exchange and Experiment.com. We also thank DOC and TRONT for providing the opportunity to explore structural variants in kākāpō using both short- and long-read data. JRW and TES are grateful to theShorebirds Trust, DOC, Genomics Aotearoa, and the University of Canterbury (UC) for their financial support and also extend their thanks to the UC Conservation, Systematics and Evolution Research Team (ConSERT), the UC Research Computational Cluster, especially Francois Bissey, and the New Zealand eScience Infrastructure for support in accessing the computational resources necessary for generating this manuscript.

References

Alkan C, Coe BP, Eichler EE (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363–376. doi:10.1038/nrg2958

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR, Lemmon ZH, Kim J, Robitaille G, Kramer M, Goodwin S, McCombie WR, Hutton S, Van Eck J, Gillis J, Eshed Y, Sedlazeck FJ, van der Knaap E, Schatz MC, Lippman ZB (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145-161.e23. doi:10.1016/j.cell.2020.05.021

Anon (2019). BayesTyper Issue #19. *GitHub*. Available at: https://github.com/bioinformatics-centre/BayesTyper/issues/19 [accessed 17 January 2022]

Anon (2022a). Delly. Available at: https://github.com/dellytools/delly [accessed 22 February 2022]

Anon (2016a). Issue #30 · Illumina/manta. *GitHub*. Available at: https://github.com/Illumina/manta/issues/30 [accessed 22 February 2022]

Anon (2016b). Issue #53 · Illumina/manta. *GitHub*. Available at: https://github.com/Illumina/manta/issues/53 [accessed 22 February 2022]

Anon (2022b). Manta Structural Variant Caller. Available at: https://github.com/Illumina/manta [accessed 22 February 2022]

Anon Oxford Nanopore Community - Sequencing low quantities of gDNA. Available at: https://community.nanoporetech.com/posts/low-input-run-genomic-dna [accessed 19 May 2021]

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020). Plant pan-genomes are the new reference. *Nature Plants* **6**, 914–920. doi:10.1038/s41477-020-0733-0

Berdan EL, Mérot C, Pavia H, Johannesson K, Wellenreuther M, Butlin RK (2021). A large chromosomal inversion shapes gene expression in seaweed flies (Coelopa frigida). *Evolution Letters* **5**, 607–624. doi:10.1002/evl3.260

Bergner LM, Jamieson IG, Robertson BC (2014). Combining genetic data to identify relatedness among founders in a genetically depauperate parrot, the Kakapo (Strigops habroptilus). *Conservation Genetics* **15**, 1013–1020. doi:10.1007/s10592-014-0595-y

Best HA, Powlesland R (1985). 'Kakapo'. (New Zealand Wildlife Service: Wellington, New Zealand)

Cameron DL, Di Stefano L, Papenfuss AT (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications* **10**, 3240. doi:10.1038/s41467-019-11146-4

Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research*. doi:10.1101/gr.222109.117

Campbell CR, Poelstra JW, Yoder AD (2018). What is Speciation Genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biological Journal of the Linnean Society* **124**, 561–583. doi:10.1093/biolinnean/bly063

Cappetta E, Andolfo G, Di Matteo A, Barone A, Frusciante L, Ercolano MR (2020). Accelerating Tomato Breeding by Exploiting Genomic Selection Approaches. *Plants* **9**, 1236. doi:10.3390/plants9091236

Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, Sara R, Walker JD, Anderson J, Hudson M (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal* **19**, 43. doi:10.5334/dsj-2020-043

Carroll SR, Herczog E, Hudson M, Russell K, Stall S (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data* **8**, 108. doi:10.1038/s41597-021-00892-0

Cayuela H, Dorant Y, Mérot C, Laporte M, Normandeau E, Gagnon-Harvey S, Clément M, Sirois P, Bernatchez L (2021). Thermal adaptation rather than demographic history drives genetic structure inferred by copy number variants in a marine fish. *Molecular Ecology* **30**, 1624–1641. doi:10.1111/mec.15835

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D, Anantharaman T, Audano PA, Brand H, Cantsilieris S, Cao H, Cerveira E, Chen C, Chen X, Chin C-S, Chong Z, Chuang NT, Lambert CC, Church DM, Clarke L, Farrell A, Flores J, Galeev T, Gorkin DU, Gujral M, Guryev V, Heaton WH, Korlach J, Kumar S, Kwon JY, Lam ET, Lee JE, Lee J, Lee W-P, Lee SP, Li S, Marks P, Viaud-Martinez K, Meiers S, Munson KM, Navarro FCP, Nelson BJ, Nodzak C, Noor A, Kyriazopoulou-Panagiotopoulou S, Pang AWC, Qiu Y, Rosanio G, Ryan M, Stütz A, Spierings DCJ, Ward A, Welch AE, Xiao M, Xu W, Zhang C, Zhu Q, Zheng-Bradley X, Lowy E, Yakneen S, McCarroll S, Jun G, Ding L, Koh CL, Ren B, Flicek P, Chen K, Gerstein MB, Kwok P-Y, Lansdorp PM, Marth GT, Sebat J, Shi X, Bashir A, Ye K, Devine SE, Talkowski ME, Mills RE, Marschall T, Korbel JO, Eichler EE, Lee C (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications* **10**, 1784. doi:10.1038/s41467-018-08148-z

Chander V, Gibbs RA, Sedlazeck FJ (2019). Evaluation of computational genotyping of structural variation for clinical diagnoses. *GigaScience* **8**. doi:10.1093/gigascience/giz110

Chen S, Zhou Y, Chen Y, Gu J (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890. doi:10.1093/bioinformatics/bty560

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222. doi:10.1093/bioinformatics/btv710

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods* **12**, 966–968. doi:10.1038/nmeth.3505

Chueca LJ, Schell T, Pfenninger M (2021). Whole-genome re-sequencing data to infer historical demography and speciation processes in land snails: the study of two Candidula sister species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **376**, 20200156. doi:10.1098/rstb.2020.0156

Cresswell M (1996). Kākāpō Recovery Plan 1996-2005. Threatened Species Recovery Plan No. 21. Department of Conservation, Wellington, New Zealand.

Cruickshank TE, Hahn MW (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* **23**, 3133–3157. doi:10.1111/mec.12796

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021). Twelve years of SAMtools and BCFtools. *GigaScience* **10**. doi:10.1093/gigascience/giab008

Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, Simpson F, Merrill RM, Joron M, Mallet J, Dasmahapatra KK, Jiggins CD (2016). Major Improvements to the Heliconius melpomene Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. *G3 Genes|Genomes|Genetics* **6**, 695–708. doi:10.1534/g3.115.023655

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669. doi:10.1093/bioinformatics/bty149

Delage WJ, Thevenon J, Lemaitre C (2020). Towards a better understanding of the low recall of insertion variants with short-read based variant callers. *BMC Genomics* **21**, 762. doi:10.1186/s12864-020-07125-5

Dorant Y, Cayuela H, Wellband K, Laporte M, Rougemont Q, Mérot C, Normandeau E, Rochette R, Bernatchez L (2020). Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. *Molecular Ecology* **29**, 4765–4782. doi:10.1111/mec.15565

Ebler J, Clarke WE, Rausch T, Audano PA, Houwaart T, Korbel J, Zody MC, Dilthey AT, Marschall T (2020). Pangenome-based genome inference. *bioRxiv*, 37.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E (2020). Pangenome Graphs. *Annual Review of Genomics and Human Genetics* **21**, 139–162. doi:10.1146/annurev-genom-120219-080406

English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S, Carroll A, Veeraraghavan N, Bruestle J, Drees B, Hastie A, Lam ET, White S, Mishra P, Wang M, Han Y, Zhang F, Stankiewicz P, Wheeler DA, Reid JG, Muzny DM, Rogers J, Sabo A, Worley KC, Lupski JR, Boerwinkle E, Gibbs RA (2015). Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genomics* **16**, 286. doi:10.1186/s12864-015-1479-3

Ewels P, Magnusson M, Lundin S, Käller M (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048. doi:10.1093/bioinformatics/btw354

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, Malukiewicz J, Mouton A, Oomen RA, Paez S, Palsbøll PJ, Pampoulie C, Ruiz-López MJ, Svardal H, Theofanopoulou C, de Vries J, Waldvogel A-M, Zhang G, Mazzoni CJ, Jarvis ED, Bálint M, Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Čiampor F, Ciofi C, Crottini A, Godoy JA, Hoglund J, Malukiewicz J, Mouton A, Oomen RA, Paez S, Palsbøll P, Pampoulie C, Ruiz-López MJ, Svardal H, Theofanopoulou C, de Vries J, Waldvogel A-M, Zhang G, Mazzoni CJ, Jarvis E, Bálint M, Aghayan SA, Alioto TS, Almudi I, Alvarez N, Alves PC, Amorim IR, Antunes A, Arribas P, Baldrian P, Berg PR, Bertorelle G, Böhne A, Bonisoli-Alquati A, Boštjančić LL, Boussau B, Breton CM, Buzan E, Campos PF, Carreras C, Castro LFi, Chueca LJ, Conti E, Cook-Deegan R, Croll D, Cunha MV, Delsuc F, Dennis AB, Dimitrov D, Faria R, Favre A, Fedrigo OD, Fernández R, Ficetola GF, Flot J-F, Gabaldón T, Galea Agius DR, Gallo GR, Giani AM, Gilbert MTP, Grebenc T, Guschanski K, Guyot R, Hausdorf B, Hawlitschek O, Heintzman PD, Heinze B, Hiller M, Husemann M, Iannucci A, Irisarri I, Jakobsen KS, Jentoft S, Klinga P, Kloch A, Kratochwil CF, Kusche H, Layton KKS, Leonard JA, Lerat E, Liti G, Manousaki T, Marques-Bonet T, Matos-Maraví P, Matschiner M, Maumus F, Mc Cartney AM, Meiri S, Melo-Ferreira J, Mengual X, Monaghan MT, Montagna M, Mysłajek RW, Neiber MT, Nicolas V, Novo M, Ozretić P, Palero F, Pârvulescu L, Pascual M, Paulo OS, Pavlek M, Pegueroles C, Pellissier L, Pesole G, Primmer CR, Riesgo A, Rüber L, Rubolini D, Salvi D, Seehausen O, Seidel M, Secomandi S, Studer B, Theodoridis S, Thines M, Urban L, Vasemägi A, Vella A, Vella N, Vernes SC, Vernesi C, Vieites DR, Waterhouse RM, Wheat CW, Wörheide G, Wurm Y, Zammit G (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution* **37**, 197–202. doi:10.1016/j.tree.2021.11.008

Funk ER, Mason NA, Pálsson S, Albrecht T, Johnson JA, Taylor SA (2021). A supergene underlies linked variation in color and morphology in a Holarctic songbird. *Nature Communications* **12**, 6833. doi:10.1038/s41467-021-27173-z

Galla SJ, Brown L, Couch-Lewis Y, Cubrinovska I, Eason D, Gooley R, Hamilton JA, Heath JA, Hauser SS, Latch EK, Matocq MD, Richardson A, Wold JR, Hogg CJ, Santure AW, Steeves TE (2021). The relevance of pedigrees in the conservation genomics era. *Molecular Ecology* **31**, 41–54. doi:10.1111/mec.16192

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679. doi:10.1093/bioinformatics/bts503

Guhlin J, Lec MFL, Wold J, Koot E, Winter D, Biggs P, Galla SJ, Urban L, Foster Y, Cox MP, Digby A, Uddstrom L, Eason D, Vercoe D, Davis T, Kākāpō Recovery Team, Howard JT, Jarvis E, Robertson FE, Robertson BC, Gemmell N, Steeves TE, Santure AW, Dearden PK (2022). Species-wide genomics of kākāpō provides transformational tools to accelerate recovery. preprint. Genomics. doi:10.1101/2022.10.22.513130

Hallast P, Kibena L, Punab M, Arciero E, Rootsi S, Grigorova M, Flores R, Jobling MA, Poolamets O, Pomm K, Korrovits P, Rull K, Xue Y, Tyler-Smith C, Laan M (2021). A common 1.6 mb Y-chromosomal inversion predisposes to subsequent deletions and severe spermatogenic failure in humans Ed GH Perry. *eLife* **10**, e65420. doi:10.7554/eLife.65420

Ho SS, Urban AE, Mills RE (2020). Structural Variation in the Sequencing Era: Comprehensive Discovery and Integration. *Nature reviews. Genetics* **21**, 171–189. doi:10.1038/s41576-019-0180-9

Huang K, Andrew RL, Owens GL, Ostevik KL, Rieseberg LH (2020). Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Molecular Ecology* **29**, 2535–2549. doi:10.1111/mec.15428

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-

Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin C-S, Korlach J, Wilson RK, Eichler EE (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research* **27**, 677–685. doi:10.1101/gr.214007.116

Hurgobin B, Edwards D (2017). SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? *Biology* **6**, 21. doi:10.3390/biology6010021

Huynh LY, Maney DL, Thomas JW (2011). Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (Zonotrichia albicollis). *Heredity* **106**, 537–546. doi:10.1038/hdy.2010.85

Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM (2020). Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118. doi:10.1093/bioinformatics/btaa435

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications* **8**, 14061. doi:10.1038/ncomms14061

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* **21**, 189. doi:10.1186/s13059-020-02107-y

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. doi:10.1093/bioinformatics/btn129

Jun G, Sedlazeck F, Zhu Q, English A, Metcalf G, Kang HM, Human Genome Structural Variation Consortium (HGSVC), Lee C, Gibbs R, Boerwinkle E (2021). muCNV: genotyping structural variants for population-level sequencing. *Bioinformatics* **37**, 2055–2057. doi:10.1093/bioinformatics/btab199

Kākāpō Recovery Group (2017). A history of kākāpō. Available at: https://www.doc.govt.nz/our-work/kakapo-recovery/what-we-do/history/

Kess T, Brachmann M, Boulding EG (2021). Putative chromosomal rearrangements are associated primarily with ecotype divergence rather than geographic separation in an intertidal, poorly dispersing snail. *Journal of Evolutionary Biology* **34**, 193–207. doi:10.1111/jeb.13724

Kim K-W, Bennison C, Hemmings N, Brookes L, Hurley LL, Griffith SC, Burke T, Birkhead TR, Slate J (2017). A sex-linked supergene controls sperm morphology and swimming speed in a songbird. *Nature Ecology & Evolution* **1**, 1168–1176. doi:10.1038/s41559-017-0235-2

Kirsche M, Prabhu G, Sherman RM, Ni B, Aganezov S, Schatz MC (2021). Jasmine: Population-scale structural variant comparison and analysis. *bioRxiv*. doi:10.1101/2021.05.27.445886

Knief U, Forstmeier W, Pei Y, Ihle M, Wang D, Martin K, Opatová P, Albrechtová J, Wittig M, Franke A, Albrecht T, Kempenaers B (2017). A sex-chromosome inversion causes strong overdominance for sperm traits that affect siring success. *Nature Ecology & Evolution* **1**, 1177–1184. doi:10.1038/s41559-017-0236-1

Kokot M, Długosz M, Deorowicz S (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761. doi:10.1093/bioinformatics/btx304

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117. doi:10.1186/s13059-019-1720-5

Küpper C, Stocks M, Risse JE, dos Remedios N, Farrell LL, McRae SB, Morgan TC, Karlionova N, Pinchuk P, Verkuil YI, Kitaysky AS, Wingfield JC, Piersma T, Zeng K, Slate J, Blaxter M, Lank DB, Burke T (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics* **48**, 79–83. doi:10.1038/ng.3443

Lado S, Elbers JP, Doskocil A, Scaglione D, Trucchi E, Banabazi MH, Almathen F, Saitou N, Ciani E, Burger PA (2020). Genome-wide diversity and global migration patterns in dromedaries follow ancient caravan routes. *Communications Biology* **3**, 1–8. doi:10.1038/s42003-020-1098-7

Layer RM, Chiang C, Quinlan AR, Hall IM (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology* **15**, R84. doi:10.1186/gb-2014-15-6-r84

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760. doi:10.1093/bioinformatics/btp324

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162-176.e13. doi:10.1016/j.cell.2020.05.023

Lloyd BD, Powlesland RG (1994). The decline of kakapo Strigops habroptilus and attempts at conservation by translocation. *Biological Conservation* **69**, 75–85. doi:10.1016/0006-3207(94)90330-1

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019). Structural variant calling: the long and the short of it. *Genome Biology* **20**, 246. doi:10.1186/s13059-019-1828-7

Mathur S, DeWoody JA (2021). Genetic load has potential in large populations but is realized in small inbred populations. *Evolutionary Applications* **14**, 1540–1557. doi:10.1111/eva.13216

Mc Cartney AM, Anderson J, Liggins L, Hudson ML, Anderson MZ, TeAika B, Geary J, Cook-Deegan R, Patel HR, Phillippy AM (2022). Balancing openness with Indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all of life. *Proceedings of the National Academy of Sciences* **119**, e2115860119. doi:10.1073/pnas.2115860119

Mérot C, Oomen RA, Tigano A, Wellenreuther M (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology & Evolution* **35**, 561–572. doi:10.1016/j.tree.2020.03.002

Mérot C, Stenløkk KSR, Venney C, Laporte M, Moser M, Normandeau E, Árnyasi M, Kent M, Rougeux C, Flynn JM, Lien S, Bernatchez L (2022). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs

(Coregonus sp.) with long and short reads. *Molecular Ecology* **00**. doi:10.1111/mec.16468

Pedersen B (2022). duphold: uphold your DUP and DEL calls. Available at: https://github.com/brentp/duphold [accessed 22 February 2022]

Pedersen BS, Layer R, Quinlan AR (2020a). Smoove: Structural-variant calling and genotyping with existing tools. Available at: https://github.com/brentp/smoove

Pedersen BS, Layer R, Quinlan AR (2020b). ‘smoove: structural-variant calling and genotyping with existing tools’ Available at: https://github.com/brentp/smoove [accessed 22 February 2022]

Pedersen BS, Quinlan AR (2019). Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience* **8**, giz040. doi:10.1093/gigascience/giz040

Pedersen BS, Quinlan AR (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868. doi:10.1093/bioinformatics/btx699

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987. doi:10.1038/nbt.4235

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339. doi:10.1093/bioinformatics/bts378

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Gedman GL, Cantin LJ, Thibaud-Nissen F, Haggerty L, Lee C, Ko BJ, Kim J, Bista I, Smith M, Haase B, Mountcastle J, Winkler S, Paez S, Howard J, Vernes SC, Lama TM, Grutzner F, Warren WC, Balakrishnan C, Burt D, George JM, Biegler M, Iorns D, Digby A, Eason D, Edwards T, Wilkinson M, Turner G, Meyer A, Kautt AF, Franchini P, Detrich HW, Svardal H, Wagner M, Naylor GJP, Pippel M, Malinsky M, Mooney M, Simbirsky M, Hannigan BT, Pesout T, Houck M, Misuraca A, Kingan SB, Hall R, Kronenberg Z, Korlach J, Sović I, Dunn C, Ning Z, Hastie A, Lee J, Selvaraj S, Green RE, Putnam NH, Ghurye J, Garrison E, Sims Y, Collins J, Pelan S, Torrance J, Tracey A, Wood J, Guan D, London SE, Clayton DF, Mello CV, Friedrich SR, Lovell PV, Osipova E, Al-Ajli FO, Secomandi S, Kim H, Theofanopoulou C, Zhou Y, Harris RS, Makova KD, Medvedev P, Hoffman J, Masterson P, Clark K, Martin F, Howe K, Flicek P, Walenz BP, Kwak W, Clawson H, Diekhans M, Nassar L, Paten B, Kraus RHS, Lewin H, Crawford AJ, Gilbert MTP, Zhang G, Venkatesh B, Murphy RW, Koepfli K-P, Shapiro B, Johnson WE, Palma FD, Margues-Bonet T, Teeling EC, Warnow T, Graves JM, Ryder OA, Hausler D, O’Brien SJ, Howe K, Myers EW, Durbin R, Phillippy AM, Jarvis ED (2020). Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv*, 2020.05.22.110833. doi:10.1101/2020.05.22.110833

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, Lee C, Ko BJ, Chaisson M, Gedman GL, Cantin LJ, Thibaud-Nissen F, Haggerty L, Bista I, Smith M, Haase B, Mountcastle J, Winkler S, Paez S, Howard J, Vernes SC, Lama TM, Grutzner F, Warren WC, Balakrishnan CN, Burt D, George JM, Biegler MT, Iorns D, Digby A, Eason D, Robertson B, Edwards T, Wilkinson M, Turner G, Meyer A, Kautt AF, Franchini P, Detrich HW, Svardal H, Wagner M, Naylor GJP, Pippel M, Malinsky M, Mooney M, Simbirsky M, Hannigan BT, Pesout T, Houck M, Misuraca A, Kingan SB, Hall R, Kronenberg Z, Sović I, Dunn C, Ning Z, Hastie A, Lee J, Selvaraj S, Green RE, Putnam NH, Gut I, Ghurye J, Garrison E, Sims Y, Collins J, Pelan S, Torrance J, Tracey A, Wood J, Dagnew RE, Guan D, London SE, Clayton DF, Mello CV, Friedrich SR, Lovell PV, Osipova E, Al-Ajli FO, Secomandi S, Kim H, Theofanopoulou C, Hiller M, Zhou Y, Harris RS, Makova KD, Medvedev P, Hoffman J, Masterson P, Clark K, Martin F, Howe K, Flicek P, Walenz BP, Kwak W, Clawson H, Diekhans M, Nassar L, Paten B, Kraus RHS, Crawford AJ, Gilbert MTP, Zhang G, Venkatesh B, Murphy RW, Koepfli K-P, Shapiro B, Johnson WE, Di Palma F, Marques-Bonet T, Teeling EC, Warnow T, Graves JM, Ryder OA, Haussler D, O'Brien SJ, Korlach J, Lewin HA, Howe K, Myers EW, Durbin R, Phillippy AM, Jarvis ED (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746. doi:10.1038/s41586-021-03451-0

Schubert M, Lindgreen S, Orlando L (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes* **9**, 88. doi:10.1186/s13104-016-1900-2

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC (2018). Accurate detection of complex structural variations using single molecule sequencing. *Nature methods* **15**, 461–468. doi:10.1038/s41592-018-0001-7

Sibbesen JA (2018). Filtering · bioinformatics-centre/BayesTyper Wiki. *GitHub*. Available at: https://github.com/bioinformatics-centre/BayesTyper [accessed 21 January 2022]

Sibbesen JA, Maretty L, Krogh A (2018). Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics* **50**, 1054–1059. doi:10.1038/s41588-018-0145-5

Talenti A, Powell J, Hemmink JD, Cook E a. J, Wragg D, Jayaraman S, Paxton E, Ezeasor C, Obishakin ET, Agusi ER, Tijjani A, Marshall K, Fisch A, Ferreira BR, Qasim A, Chaudhry U, Wiener P, Toye P, Morrison LJ, Connelley T, Prendergast JGD (2022). A cattle graph genome incorporating global breed diversity. *Nature Communications* **13**, 910. doi:10.1038/s41467-022-28605-0

Tigano A, Jacobs A, Wilder AP, Nand A, Zhan Y, Dekker J, Therkildsen NO (2021). Chromosome-Level Assembly of the Atlantic Silverside Genome Reveals Extreme Levels of Sequence Diversity and Structural Genetic Variation. *Genome Biology and Evolution* **13**. doi:10.1093/gbe/evab098

Todesco M, Owens GL, Bercovich N, Légaré J-S, Soudi S, Burge DO, Huang K, Ostevik KL, Drummond EBM, Imerovski I, Lande K, Pascual-Robles MA, Nanavati M, Jahani M, Cheung W, Staton SE, Muños S, Nielsen R, Donovan LA, Burke JM, Yeaman S, Rieseberg LH (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607. doi:10.1038/s41586-020-2467-6

Veltman CJ (1996). Investigating causes of population decline in New Zealand plants and animals: Introduction to Symposium. *New Zealand Journal of Ecology* **20**, 1–5. Available at: https://www.jstor.org/stable/24053728 [accessed 22 February 2022]

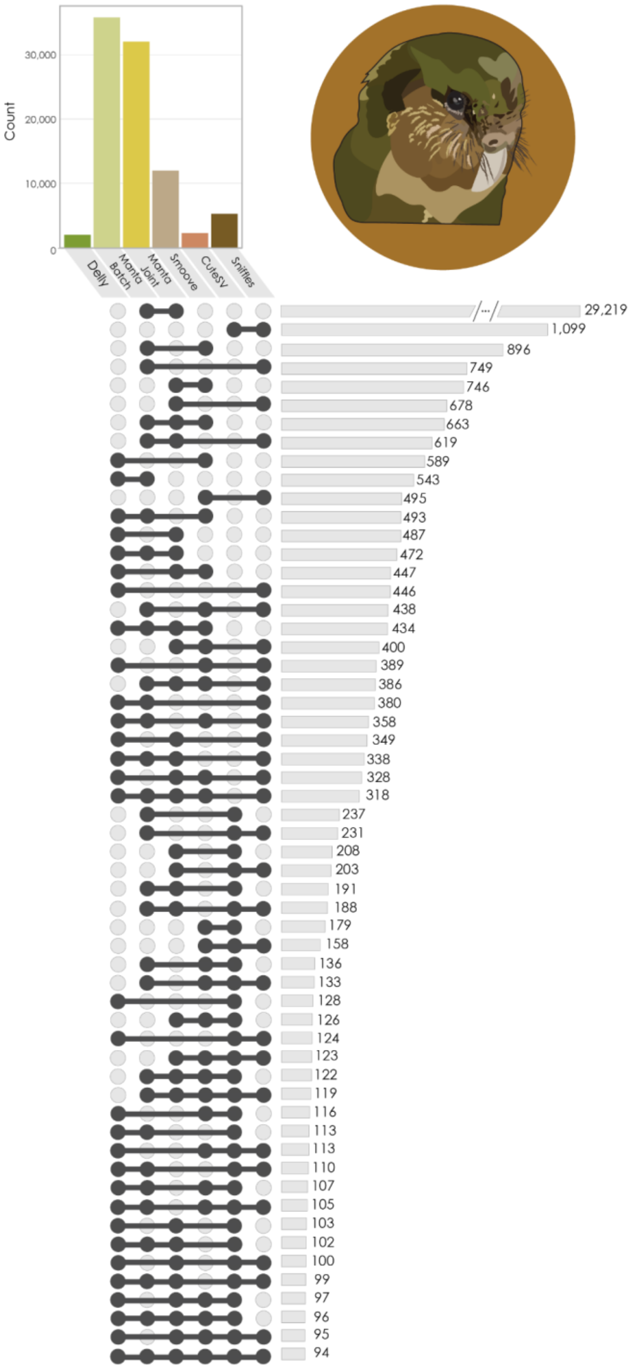
vonHoldt BM, Shuldiner E, Koch IJ, Kartzinel RY, Hogan A, Brubaker L, Wanser S, Stahler D, Wynne CDL, Ostrander EA, Sinsheimer JS, Udell MAR (2017). Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Science Advances* **3**, e1700398. doi:10.1126/sciadv.1700398

Wellenreuther M, Bernatchez L (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology & Evolution* **33**, 427–440. doi:10.1016/j.tree.2018.04.002

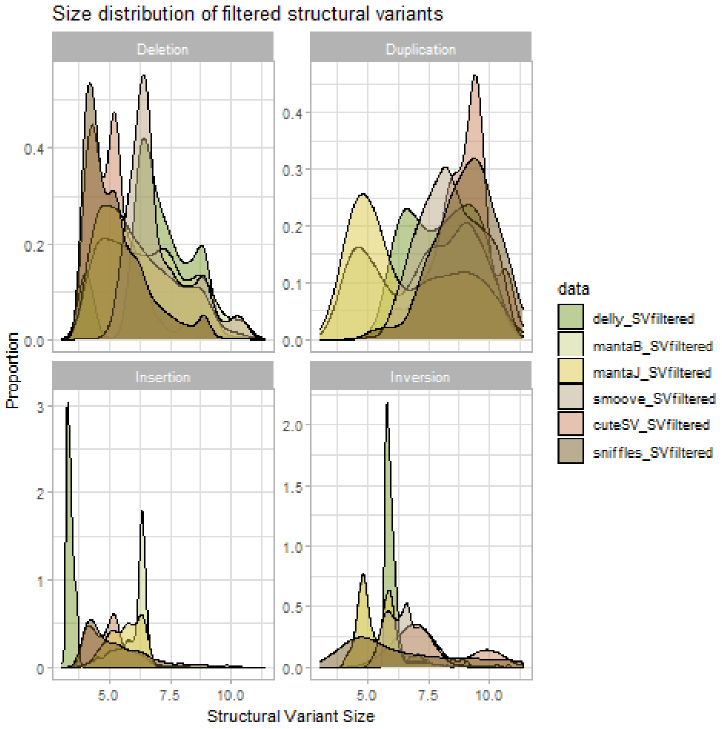
Wick R (2022). Porechop. Available at: https://github.com/rrwick/Porechop [accessed 10 March 2022]

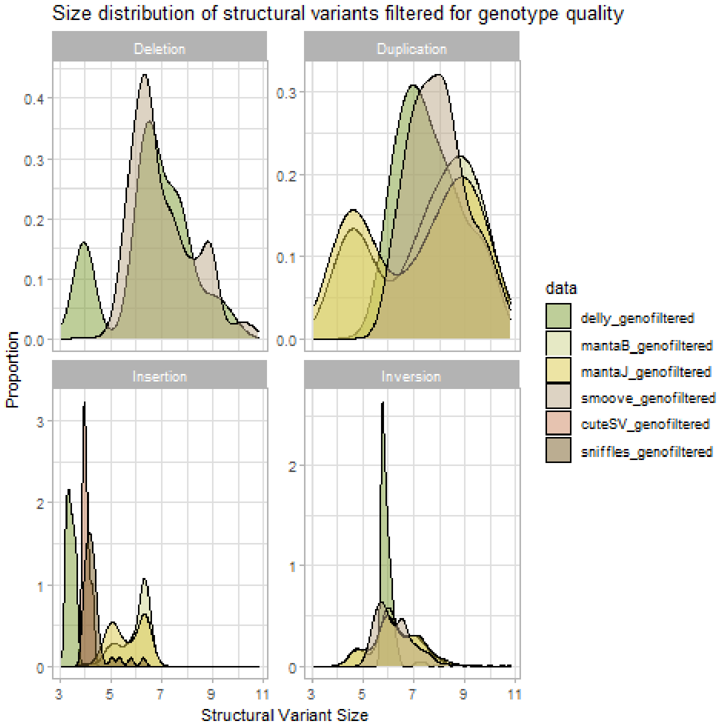
Williams GR (1956). The Kakapo (Strigops habrotilus, Gray): a review and appraisal of a near-extinct species.

Wold J, Koepfli K-P, Galla SJ, Eccles D, Hogg CJ, Le Lec MF, Guhlin J, Santure AW, Steeves TE (2021). Expanding the conservation genomics toolbox: Incorporating structural variants to enhance genomic studies for species of conservation concern. *Molecular Ecology* **30**, 5949–5965. doi:10.1111/mec.16141

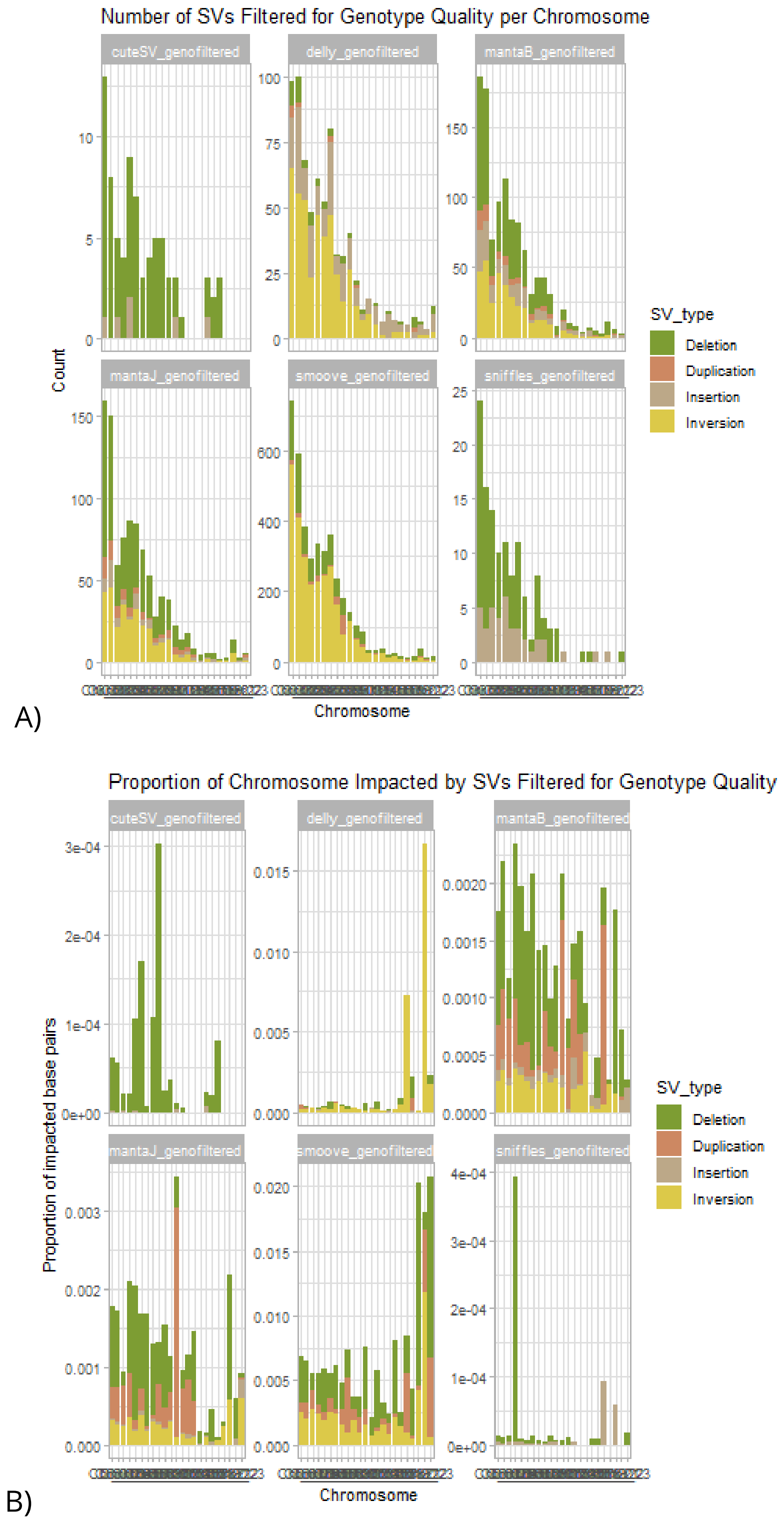


Supplementary Figure 1. Counts of consensus calls between SV type and strand within a 50 bp window for the all comparisons between Delly, Manta - Batch, Manta - Joint, Smoove, CuteSV and Sniffles in kākāpō. Here, the colored barchart on the upper left represents the number of SVs passing call-quality thresholds in each of the six datasets. Dark green circles with lines between denote which datasets have consensus SV calls. Bars to the right represent the number of SVs overlapping between these datasets.





Supplementary Figure 2. Size distribution for SVs that passed call quality thresholds. Due to the high level of variance in SV size, a log transformation using the natural log was used to visualise the size distribution.



Supplementary Figure 3. Number of SVs per chromosome passing genotype quality thresholds (A), and the proportion of each chromosome impacted by SV type (B). As with Figure 2, all chromosomes are ordered by size from largest to smallest (left to right). The Z and W sex chromosomes are excluded.

Supplementary Table 1. Overlaps of SVs passing genotype thresholds. Comparisons were made for exact matches (0bp), 50bp, 500bp and 1kb. Here, D denotes the Delly dataset, B is Manta - Batch, J is the Manta - Joint, S is Smoove, C is CuteSV and Sn is Sniffles.

Comparison	Total	Deletions	Duplications	Insertions	Inversions
allvall_0bp	0	0	0	0	0
allvall_1000bp	0	0	0	0	0
allvall_500bp	0	0	0	0	0
allvall_50bp	0	0	0	0	0
BvC_0bp	0	0	0	0	0
BvC_1000bp	47	47	0	0	0
BvC_500bp	47	47	0	0	0
BvC_50bp	47	47	0	0	0
BvCvSn_0bp	0	0	0	0	0
BvCvSn_1000bp	8	8	0	0	0
BvCvSn_500bp	8	8	0	0	0
BvCvSn_50bp	8	8	0	0	0
BvJ_0bp	0	0	0	0	0
BvJ_1000bp	709	451	56	53	149
BvJ_500bp	711	453	56	53	149
BvJ_50bp	712	453	56	53	150
BvJvC_0bp	0	0	0	0	0
BvJvC_1000bp	46	46	0	0	0
BvJvC_500bp	46	46	0	0	0
BvJvC_50bp	46	46	0	0	0
BvJvCvSn_0bp	0	0	0	0	0
BvJvCvSn_1000bp	7	7	0	0	0
BvJvCvSn_500bp	7	7	0	0	0
BvJvCvSn_50bp	7	7	0	0	0
BvJvS_0bp	0	0	0	0	0
BvJvS_1000bp	381	338	28	0	15
BvJvS_500bp	380	338	28	0	14
BvJvS_50bp	356	325	26	0	5

BvJvSn_0bp	0	0	0	0	0
BvJvSn_1000bp	17	17	0	0	0
BvJvSn_500bp	17	17	0	0	0
BvJvSn_50bp	17	17	0	0	0
BvJvSvC_0bp	0	0	0	0	0
BvJvSvC_1000bp	33	33	0	0	0
BvJvSvC_500bp	33	33	0	0	0
BvJvSvC_50bp	32	32	0	0	0
BvJvSvCvSn_0bp	0	0	0	0	0
BvJvSvCvSn_1000bp	1	1	0	0	0
BvJvSvCvSn_500bp	1	1	0	0	0
BvJvSvCvSn_50bp	1	1	0	0	0
BvJvSvSn_0bp	0	0	0	0	0
BvJvSvSn_1000bp	3	3	0	0	0
BvJvSvSn_500bp	3	3	0	0	0
BvJvSvSn_50bp	3	3	0	0	0
BvS_0bp	0	0	0	0	0
BvS_1000bp	436	375	33	0	28
BvS_500bp	435	375	33	0	27
BvS_50bp	396	359	29	0	8
BvSn_0bp	0	0	0	0	0
BvSn_1000bp	21	21	0	0	0
BvSn_500bp	21	21	0	0	0
BvSn_50bp	21	21	0	0	0
BvSvC_0bp	0	0	0	0	0
BvSvC_1000bp	33	33	0	0	0
BvSvC_500bp	33	33	0	0	0
BvSvC_50bp	32	32	0	0	0
BvSvCvSn_0bp	0	0	0	0	0
BvSvCvSn_1000bp	1	1	0	0	0
BvSvCvSn_500bp	1	1	0	0	0

BvSvCvSn_50bp	1	1	0	0	0
BvSvSn_0bp	0	0	0	0	0
BvSvSn_1000bp	3	3	0	0	0
BvSvSn_500bp	3	3	0	0	0
BvSvSn_50bp	3	3	0	0	0
CvSn_0bp	0	0	0	0	0
CvSn_1000bp	21	20	0	1	0
CvSn_500bp	21	20	0	1	0
CvSn_50bp	21	20	0	1	0
DvB_0bp	0	0	0	0	0
DvB_1000bp	12	9	3	0	0
DvB_500bp	12	9	3	0	0
DvB_50bp	11	8	3	0	0
DvBvC_0bp	0	0	0	0	0
DvBvC_1000bp	0	0	0	0	0
DvBvC_500bp	0	0	0	0	0
DvBvC_50bp	0	0	0	0	0
DvBvCvSn_0bp	0	0	0	0	0
DvBvCvSn_1000bp	0	0	0	0	0
DvBvCvSn_500bp	0	0	0	0	0
DvBvCvSn_50bp	0	0	0	0	0
DvBvJ_0bp	0	0	0	0	0
DvBvJ_1000bp	12	9	3	0	0
DvBvJ_500bp	12	9	3	0	0
DvBvJ_50bp	11	8	3	0	0
DvBvJvC_0bp	0	0	0	0	0
DvBvJvC_1000bp	0	0	0	0	0
DvBvJvC_500bp	0	0	0	0	0
DvBvJvC_50bp	0	0	0	0	0
DvBvJvCvSn_0bp	0	0	0	0	0
DvBvJvCvSn_1000bp	0	0	0	0	0

DvBvJvCvSn_500bp	0	0	0	0	0
DvBvJvCvSn_50bp	0	0	0	0	0
DvBvJvS_0bp	0	0	0	0	0
DvBvJvS_1000bp	11	9	2	0	0
DvBvJvS_500bp	11	9	2	0	0
DvBvJvS_50bp	10	8	2	0	0
DvBvJvSn_0bp	0	0	0	0	0
DvBvJvSn_1000bp	0	0	0	0	0
DvBvJvSn_500bp	0	0	0	0	0
DvBvJvSn_50bp	0	0	0	0	0
DvBvJvSvC_0bp	0	0	0	0	0
DvBvJvSvC_1000bp	0	0	0	0	0
DvBvJvSvC_500bp	0	0	0	0	0
DvBvJvSvC_50bp	0	0	0	0	0
DvBvJvSvSn_0bp	0	0	0	0	0
DvBvJvSvSn_1000bp	0	0	0	0	0
DvBvJvSvSn_500bp	0	0	0	0	0
DvBvJvSvSn_50bp	0	0	0	0	0
DvBvS_0bp	0	0	0	0	0
DvBvS_1000bp	11	9	2	0	0
DvBvS_500bp	11	9	2	0	0
DvBvS_50bp	10	8	2	0	0
DvBvSn_0bp	0	0	0	0	0
DvBvSn_1000bp	0	0	0	0	0
DvBvSn_500bp	0	0	0	0	0
DvBvSn_50bp	0	0	0	0	0
DvBvSvC_0bp	0	0	0	0	0
DvBvSvC_1000bp	0	0	0	0	0
DvBvSvC_500bp	0	0	0	0	0
DvBvSvC_50bp	0	0	0	0	0
DvBvSvCvSn_0bp	0	0	0	0	0

DvBvSvCvSn_1000bp	0	0	0	0	0
DvBvSvCvSn_500bp	0	0	0	0	0
DvBvSvCvSn_50bp	0	0	0	0	0
DvBvSvSn_0bp	0	0	0	0	0
DvBvSvSn_1000bp	0	0	0	0	0
DvBvSvSn_500bp	0	0	0	0	0
DvBvSvSn_50bp	0	0	0	0	0
DvC_0bp	0	0	0	0	0
DvC_1000bp	0	0	0	0	0
DvC_500bp	0	0	0	0	0
DvC_50bp	0	0	0	0	0
DvCvSn_0bp	0	0	0	0	0
DvCvSn_1000bp	0	0	0	0	0
DvCvSn_500bp	0	0	0	0	0
DvCvSn_50bp	0	0	0	0	0
DvJ_0bp	0	0	0	0	0
DvJ_1000bp	12	9	3	0	0
DvJ_500bp	12	9	3	0	0
DvJ_50bp	11	8	3	0	0
DvJvC_0bp	0	0	0	0	0
DvJvC_1000bp	0	0	0	0	0
DvJvC_500bp	0	0	0	0	0
DvJvC_50bp	0	0	0	0	0
DvJvCvSn_0bp	0	0	0	0	0
DvJvCvSn_1000bp	0	0	0	0	0
DvJvCvSn_500bp	0	0	0	0	0
DvJvCvSn_50bp	0	0	0	0	0
DvJvS_0bp	0	0	0	0	0
DvJvS_1000bp	11	9	2	0	0
DvJvS_500bp	11	9	2	0	0
DvJvS_50bp	10	8	2	0	0

DvJvSn_0bp	0	0	0	0	0
DvJvSn_1000bp	0	0	0	0	0
DvJvSn_500bp	0	0	0	0	0
DvJvSn_50bp	0	0	0	0	0
DvJvSvC_0bp	0	0	0	0	0
DvJvSvC_1000bp	0	0	0	0	0
DvJvSvC_500bp	0	0	0	0	0
DvJvSvC_50bp	0	0	0	0	0
DvJvSvCvSn_0bp	0	0	0	0	0
DvJvSvCvSn_1000bp	0	0	0	0	0
DvJvSvCvSn_500bp	0	0	0	0	0
DvJvSvCvSn_50bp	0	0	0	0	0
DvJvSvSn_0bp	0	0	0	0	0
DvJvSvSn_1000bp	0	0	0	0	0
DvJvSvSn_500bp	0	0	0	0	0
DvJvSvSn_50bp	0	0	0	0	0
DvS_0bp	0	0	0	0	0
DvS_1000bp	47	39	4	0	4
DvS_500bp	46	38	4	0	4
DvS_50bp	35	29	4	0	2
DvSn_0bp	0	0	0	0	0
DvSn_1000bp	0	0	0	0	0
DvSn_500bp	0	0	0	0	0
DvSn_50bp	0	0	0	0	0
DvSvC_0bp	0	0	0	0	0
DvSvC_1000bp	0	0	0	0	0
DvSvC_500bp	0	0	0	0	0
DvSvC_50bp	0	0	0	0	0
DvSvCvSn_0bp	0	0	0	0	0
DvSvCvSn_1000bp	0	0	0	0	0
DvSvCvSn_500bp	0	0	0	0	0

DvSvCvSn_50bp	0	0	0	0	0
DvSvSn_0bp	0	0	0	0	0
DvSvSn_1000bp	0	0	0	0	0
DvSvSn_500bp	0	0	0	0	0
DvSvSn_50bp	0	0	0	0	0
JvC_0bp	0	0	0	0	0
JvC_1000bp	46	46	0	0	0
JvC_500bp	46	46	0	0	0
JvC_50bp	46	46	0	0	0
JvCvSn_0bp	0	0	0	0	0
JvCvSn_1000bp	7	7	0	0	0
JvCvSn_500bp	7	7	0	0	0
JvCvSn_50bp	7	7	0	0	0
JvS_0bp	0	0	0	0	0
JvS_1000bp	420	361	31	0	28
JvS_500bp	417	361	30	0	26
JvS_50bp	381	346	27	0	8
JvSn_0bp	0	0	0	0	0
JvSn_1000bp	17	17	0	0	0
JvSn_500bp	17	17	0	0	0
JvSn_50bp	17	17	0	0	0
JvSvC_0bp	0	0	0	0	0
JvSvC_1000bp	33	33	0	0	0
JvSvC_500bp	33	33	0	0	0
JvSvC_50bp	32	32	0	0	0
JvSvCvSn_0bp	0	0	0	0	0
JvSvCvSn_1000bp	1	1	0	0	0
JvSvCvSn_500bp	1	1	0	0	0
JvSvCvSn_50bp	1	1	0	0	0
JvSvSn_0bp	0	0	0	0	0
JvSvSn_1000bp	3	3	0	0	0

JvSvSn_500bp	3	3	0	0	0
JvSvSn_50bp	3	3	0	0	0
SvC_0bp	0	0	0	0	0
SvC_1000bp	36	36	0	0	0
SvC_500bp	36	36	0	0	0
SvC_50bp	34	34	0	0	0
SvCvSn_0bp	0	0	0	0	0
SvCvSn_1000bp	1	1	0	0	0
SvCvSn_500bp	1	1	0	0	0
SvCvSn_50bp	1	1	0	0	0
SvSn_0bp	0	0	0	0	0
SvSn_1000bp	4	4	0	0	0
SvSn_500bp	4	4	0	0	0
SvSn_50bp	4	4	0	0	0

Supplementary Table 2. Summary of structural variant size characteristics for Delly, Manta and Smoove datasets filtered for genotype quality.

Data	Structural Variant Type	Count	Size Range (bp)	Median Size (bp)	Mean Size (bp)
Delly	Deletions	57	49 - 18,651	756	1977
	Duplications	12	456 - 19,889	1459	4366
	Insertions	228	22 - 45	31	32
	Inversions	437	300 - 48,437	359	705
Manta-Batch	Deletions	515	50 - 41,963	578	1820
	Duplications	70	66 - 26,442	3246	5527
	Insertions	177	51 - 1,042	505	441
	Inversions	342	59 - 10,746	462	799
Manta-Joint	Deletions	495	54 - 41,963	577	1842
	Duplications	73	52 - 41,193	1978	5478

	Insertions	74	84 -888	317	354
	Inversions	301	59 - 7,093	463	841
Smoove	Deletions	1023	53 - 47,780	781	2696
	Duplications	183	335 - 47,433	2748	5793
	Insertions	N/A	N/A	N/A	N/A
	Inversions	2825	76 - 30,347	445	729
CuteSV	Deletions	72	49 - 7,497	199	910
	Duplications	0	0	0	0
	Insertions	6	51 - 73	55	58
	Inversions	0	0	0	0
Sniffles	Deletions	87	49 - 30,711	62	456
	Duplications	0	0	0	0
	Insertions	39	50 - 539	68	93
	Inversions	0	0	0	0