MOLECULAR ECOLOGY
RESOURCES

The design and application of a 50K SNP chip for a threatened Aotearoa New Zealand passerine, the hihi

Journal:	*Molecular Ecology Resources*
Manuscript ID	MER-21-0176.R1
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Lee, Kate; The University of Auckland, School of Biological Sciences Millar, Craig; The University of Auckland, Brekke, Patricia; Zoological Society of London, Institute of Zoology; Imperial College London, Division of Biology Ecology and Evolution section Whibley, Annabel; The University of Auckland School of Biological Sciences, Ewen, John; Zoological Society of London, Institute of Zoology Hingston, Melanie; The University of Auckland, School of Biological Sciences Zhu, Amy; The University of Auckland, School of Biological Sciences Santure, Anna; University of Auckland, School of Biological Sciences
Keywords:	SNP array, resequencing, linkage disequilibrium, RAD-seq

The design and application of a 50K SNP chip for a threatened Aotearoa New Zealand passerine, the hihi

Running title: Hihi SNP chip

Kate D. Lee[1], Craig D. Millar[1], Patricia Brekke[2], Annabel Whibley[1], John G. Ewen[2], Melanie Hingston[1], Amy Zhu[1] and Anna W. Santure[1]

1. School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

2. Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK

Correspondence

Anna Santure, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Email: a.santure@auckland.ac.nz

Key words: SNP array, resequencing, linkage disequilibrium, RAD-seq

Abstract

Next generation sequencing has transformed the fields of ecological and evolutionary genetics by allowing for cost-effective identification of genome-wide variation. Single nucleotide polymorphism (SNP) arrays, or 'SNP chips', enable very large numbers of individuals to be consistently genotyped at a selected set of these identified markers, and also offer the advantage of being able to analyse samples of variable DNA quality. We use reduced representation restriction-aided digest sequencing (RAD-seq) of 31 birds of the threatened hihi (*Notiomystis cincta*; stitchbird) and low-coverage whole genome sequencing (WGS) of ten of these birds to develop an Affymetrix 50K SNP chip. We overcome the limitations of having no hihi reference genome and a low quantity of sequence data by separate and pooled *de novo* assembly of each of the ten WGS birds. Reads from all individuals were mapped back to these *de novo* assemblies to identify SNPs. A subset of RAD-seq and WGS SNPs were selected for inclusion on the chip, prioritising SNPs with the highest quality scores whose flanking sequence uniquely aligned to the zebra finch (*Taeniopygia guttata*) genome. Of the 58,466 SNPs manufactured on the chip, 72% passed filtering metrics and were polymorphic. By genotyping 1,536 hihi on the array, we found that SNPs detected in multiple assemblies were more likely to successfully genotype, representing a cost-effective approach to identify SNPs for genotyping. We demonstrate the utility of the SNP chip by describing the high rates of linkage disequilibrium in the hihi genome, reflecting the history of population bottlenecks in the species.

Introduction

By enabling the discovery and genotyping of hundreds to millions of variants across the genome, the increasing affordability of short and long read sequencing has led to a transformation in ecological and evolutionary research from genetic (single- or a small number of loci) to genomic (genome-wide) research. These genome-wide markers have enabled more accurate inference of relationships and relatedness, inbreeding, ancestry, population structure and genetic diversity, and an ability to infer the genomic basis of adaptive traits in non-model organisms and to describe features of the genome such as the recombination landscape (Kardos *et al.* 2015; Morin *et al.* 2004; Peñalba & Wolf 2020; Stapley *et al.* 2010; Wellenreuther & Hansson 2016).

In species where a reference genome is not available, the process of identifying large numbers of polymorphisms in a population may require the assembly of genomic sequence data to create a draft genome, and subsequent mapping of sequence reads from a representative subset of the population to detect single nucleotide polymorphisms (SNPs). The generation of sequence data therefore needs to balance the number of individuals sequenced, in order to maximise the chance of detecting alleles at low frequency in the population, and the coverage per individual to enable genome assembly (Fumagalli 2013). Further, it remains a challenge to consistently and cost-effectively genotype large numbers of individuals at polymorphic sites in the genome.

Three main methods are currently employed for large-scale SNP genotyping: 1) 'Genotyping-by-Sequencing' (GBS) methods such as reduced representation restriction-aided digest sequencing (RAD-seq) that often assemble sequences *de novo* and then call SNPs directly

from a subset of the genome, 2) whole genome (re)sequencing (WGS), often at low coverage per individual, which frequently maps reads to an existing genome assembly and also genotypes SNPs directly from read alignment to this assembly, and 3) array-based methods in which flanking probe sequences interrogate pre-identified SNPs (termed 'SNP arrays' or 'SNP chips'). GBS methods offer a cost-effective approach for genotyping large numbers of individuals across a fraction of the genome, but tend to have high rates of missing data per SNP and per individual due to inconsistency in amplifying genomic regions to be sequenced, variation between individuals in terms of DNA quality, and inconsistency in recovering the same loci across separate sequencing batches (Davey *et al.* 2011). Whole genome resequencing, while increasingly affordable, represents a trade-off in terms of coverage and the number of individuals sampled, which can limit the power of individual-based analyses such as genome-wide association, and lead to high rates of missing data (Kim *et al.* 2011). Although also expensive per individual, SNP chips offer a robust and easily replicable way of genotyping samples at a consistent set of SNPs, with very low rates of missing data, with the added advantage that they can be used successfully on degraded DNA (Johnston *et al.* 2013; Mead *et al.* 2008), potentially allowing for museum and other historical samples to be included in the study of wild species (Decker *et al.* 2009).

Medium (~thousands to tens of thousands of loci) and high (~hundreds of thousands of loci) density SNP chips have been routinely developed for commercial species, often with a focus on enabling production gains from genome wide association studies, genomic selection and prediction (https://www.illumina.com/areas-of-interest/agrigenomics/plant-animal-genomics/genotyping.html). In contrast, only a handful of medium-high density SNP chips have been designed for non-commercial wild species such as house sparrow (Hagen *et al.* 2013; Lundregan *et al.* 2018), great tit (van Bers *et al.* 2012; Kim *et al.* 2018), polar bear

(Malenfant *et al.* 2015), flycatcher (Kawakami *et al.* 2014), fur seal (Humble *et al.* 2020), Florida scrub-jay (Chen *et al.* 2016) and bald eagle (Judkins *et al.* 2020). These SNP chips for wild species generate a quantity of genomic data that can be used, for example, to infer relatedness (Humble *et al.* 2020) and population structure (Hagen *et al.* 2013; Judkins *et al.* 2020; Malenfant *et al.* 2015; Viengkone *et al.* 2016), analyse the genomic architecture of traits (Duntsch *et al.* 2020; Husby *et al.* 2015; Kim *et al.* 2018; Laine *et al.* 2019; Lundregan *et al.* 2018; Santure *et al.* 2013; Santure *et al.* 2015; Silva *et al.* 2017), characterise copy number variants in the genome (da Silva *et al.* 2018; Kim *et al.* 2018) and investigate the genomic landscape of linkage and linkage disequilibrium (Hagen *et al.* 2020; Kawakami *et al.* 2014; van Oers *et al.* 2014). SNP chips developed for model organisms or agricultural systems have also been utilised to address evolutionary, ecological, and conservation questions. For example, they have been used to identify signatures of adaptation in cattle (Gautier *et al.* 2009), infer the genomic basis of recombination rate variation in Soay sheep (Johnston *et al.* 2016), and identify genetic structure and genes under selection in North American grey wolves (Schweizer *et al.* 2016).

Hihi or stitchbird *(Notiomystis cincta*) is a threatened endemic Aotearoa (New Zealand) passerine that, since the 1980s, has undergone a program of translocations to predator-free sites across Te Ika-a-Maui (the North Island) of Aotearoa (www.hihiconservation.com; Ewen *et al.* 2013). Hihi are of cultural importance to Indigenous Māori people, with their presence seen as an indicator of a healthy, mature forest system. Hihi were once widespread across Te Ika-a-Maui but were extirpated to a single offshore island, Te Hauturu-o-Toi (36°12’S, 175°05’E) by the 1880s, likely due to habitat loss and introduced mammalian predators (Taylor *et al.* 2005). All six reintroduced populations trace back to this remnant population via first- and second- degree translocations. Although historic population sizes are unknown,

the remnant population size is estimated at approximately 2,000 individuals, while other sites vary from ~40 - 250 birds (Parlato *et al*. 2021). While two reintroduced populations, Tiritiri Mātangi Island (36°36'S, 174°53'E) and Zealandia Wildlife Sanctuary (41°17'S, 174°45'E), are currently monitored, with all individuals in these populations banded and sampled at 21 days (Brekke *et al.* 2011; de Villemereuil *et al.* 2019; Rutschmann *et al.* 2020), sampling of other populations has been sporadic. This has resulted in variable and small sample sizes across time, unbalanced sample sizes across populations, and inconsistencies in the collection and preservation of samples. For these reasons, we sought to design a SNP chip to genotype a selection of hihi individuals sampled over many years and populations, aiming to minimise missing data and to maximise the likelihood of successfully genotyping individuals with poor quality DNA. A large dataset of genomic markers would enable, for example, overall genetic diversity and levels of inbreeding to be contrasted across populations, the impact of different management strategies such as artificial migration or assisted colonization, and the genetic basis of adaptive traits to be elucidated, to better inform management actions in this threatened species. To develop this resource for hihi, we made use of a small amount of resequencing and reduced representation data to identify SNPs for inclusion on the array.

In this study, we describe how we overcame the limitations of low coverage genome sequencing in order to identify polymorphisms. We outline the sequencing, assembly and SNP detection from next generation sequencing reads from two library types; RAD-seq of 31 individuals, and low coverage WGS from ten of these individuals. A subset of detected SNPs was selected for inclusion on a custom SNP chip and this was used to genotype 1,536 samples of varying quality from across different hihi populations. We test the conversion rates, also termed genotyping success rates, i.e. the proportion of SNPs included on the array that are polymorphic and successfully genotyped, of SNPs detected from RAD-seq and WGS data.

We also consider the effects of pooling WGS data for assembly on downstream variant calling, and discuss the effects of DNA quality and sample type on genotyping success rates. We demonstrate the utility of this SNP chip to infer linkage disequilibrium in the genome of this threatened species.

Materials and methods

Restriction-site associated DNA sequencing (RAD-seq), assembly, and SNP detection

Hihi individuals sampled from Te Hauturu-o-Toi (26 individuals) and Tiritiri Mātangi (5 individuals) were selected for RAD-seq, with an aim to identify polymorphism both between islands and within Te Hauturu-o-Toi. DNA was isolated from the 31 hihi blood samples using an ammonium acetate precipitation method at the NERC Biomolecular Analysis Facility, University of Sheffield. Samples were inspected visually on an agarose gel for degradation, quantified with a DNA fluorometer (Hoefer DynaQuant 200), and normalised to approximately 50 ng/μL. Samples were submitted to Floragenex (Inc.), Portland, Oregon, for RAD-seq, with one duplicate sample to assess genotyping reproducibility. Samples were digested with the restriction enzyme *Sbf* I, sample libraries prepared and pooled, and single-end 90 bp fragments were sequenced across two lanes of an Illumina HiSeq™. A total of 257,833,998 reads was generated, with a median of 6,709,382 reads per sample (Supplementary Table S1).

The quality of demultiplexed raw reads received from Floragenex (Inc.) was evaluated using FastQC (Andrews 2014). The software Stacks version 1.32 (Catchen *et al.* 2013; Catchen *et al.* 2011) was used to remove reads with low quality scores, assemble sequences, and call SNPs. Raw reads from the replicated individual were merged into one file. *Process_radtags* was then run on each of the 31 samples to clean the data and remove any read with an uncalled base (-c option), and discard reads (-q) where the average score within a sliding window of 15% (-w 0.15) of the read length dropped below 15 (-s 15). Reads were further filtered using the *kmer_filter* module to remove reads that contained very rare (--rare) or very abundant (--abundant) k-mers.

Filtered reads were then assembled *de novo* per individual using the *ustacks* program, with a minimum depth of coverage of six reads required to create a stack (-m 6; chosen because of the high sequencing coverage per individual), and the default of up to two nucleotide mismatches (-M 2) allowed between stacks. The deleveraging algorithm was enabled (-d) to help resolve over-merged tags. A catalogue of loci across individuals was assembled using *cstacks*, with the default one mismatch allowed when merging loci in the catalogue (-n 1). Individual reads were matched back to this catalogue using *sstacks*. The *populations* program was then used to create an output vcf file of SNPs, with all individuals assigned to the same population and SNPs filtered so that SNPs were present in at least 5 of the 31 individuals (-r 0.16), individuals had at least eight reads mapping to the locus (-m 8), and heterozygosity at the locus did not exceed 75% (--max_obs_het 0.75). A total of 30,835 SNPs were detected.

Whole genome sequencing (WGS) and assembly

Low coverage whole genome sequencing of ten individuals (a subset of the samples used in RAD-seq) were used to identify further polymorphisms. To maximise the variation captured, seven of the samples were from the remnant population on Te Hauturu-o-Toi and three were from the reintroduced population on Tiritiri Mātangi. Samples were multiplexed and two PCR-free DNA libraries were prepared by New Zealand Genomics Limited and used to generate 100 bp paired-end Illumina reads over two lanes of Illumina HiSeq™ sequencing. This resulted in a total of 879,894,554 reads with a median of 44,782,143 per sample (Supplementary Table S2).

Sequence quality was assessed using FastQC. Adapters and poor quality reads were removed with Trimmomatic-0.33 (Bolger *et al.* 2014) under strict conditions (ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:20, MINLEN:70, CROP:110), over-represented reads identified in FastQC were also removed by appending them to the TruSeq3-PE-2.fa file. The sample with the largest number of filtered reads (sample 6) was used to run SOAPdenovo2 version 1.5.14 (Luo *et al.* 2012) at k-mer sizes ranging from 25 - 95. The optimum k-mer length was chosen as 36 bases, based on N50 value and the recovered/estimated length of the assembly. Each of the samples was assembled using SOAPdenovo2 with k-mer 36 and insert size 210. The samples with the three largest sets of reads – samples 6, 9 and 10 – were also assembled together ('3 in 1' assembly), as well as an assembly of all ten samples together ('10 in 1' assembly). Each assembly was mapped back to the Ensemble 86 zebra finch (*Taeniopygia guttata*) genome with bwa-mem version 0.7.12 (Li & Durbin 2009) and the zebra finch coverage was calculated using bedtools genomecov (Quinlan & Hall 2010). Genome completeness was assessed using CEGMA v3 (Parra *et al.* 2007). To determine if the k-mer properties of the trimmed reads represented the k-mer properties of the assemblies, k-mer profiles were computed using KAT version 2.4.2 (Mapleson *et al.* 2017). The "comp" function was used to compare the k-mer properties of filtered reads to assemblies, with their similarities summarised using a Jaccard coefficient.

Mapping and variant calling

For each of the twelve WGS assemblies, filtered reads from all individuals were combined and mapped back to the assemblies using bwa-mem version 0.7.12 adding read group headers (-R) and marking shorter reads as secondary (-M).

Local realignment was carried out with GATK version 1.3 (DePristo *et al.* 2011; McKenna *et al.* 2010) RealignerTargetCreator and IndelRealigner commands. Quality scores were then adjusted using the BaseRecalibrator and Printreads commands. As hihi is not a model species with known variable sites, GATK targetrealigner requires months to run and cannot be parallelised. To overcome this, assembly contigs over 200 bp from the hihi draft genomes were grouped into 50 lists of approximately similar summed length, and these lists were used to split the bam files using a perl script (this and other perl scripts mentioned are available on github; see link in the Data Accessibility section). The list of split bam files for each assembly was then used to merge these files back to a single re-alignment file with the SAMtools version 1.3 (Li 2011; Li *et al.* 2009) using merge, sort and index commands. Variants were called with SAMtools mpileup version 1.3.

SNP filtering from WGS

SAMtools output was parsed and annotated with a perl script, and the genotypes and alleles represented in each location (Tiritiri Mātangi and Te Hauturu-o-Toi) were recorded. BLAST 2.3.0 (Altschul *et al.* 1990; Camacho *et al.* 2009) was used to map assembly contigs back to the Ensemble 86 zebra finch genome with an e-value cut off of five. The BLAST output was parsed with a perl script which calculated the proportion of the query that aligned, retrieved the hit with the lowest e-value for each contig, checked the number of hits, skipped matches of less than 80% of the total query length, and checked if the SNP position on the hihi contig aligned to the zebra finch genome was in a gene using Ensemble biomart 86 gene positions. All this information was then added to the SAMtools vcf output file and output subsequently processed using perl and bash scripts. An initial filtering of SNPs with read depth <10 was carried out. Further filtering resulted in SNPs being discarded if their hihi assembly contig

aligned to more than one zebra finch chromosome; if the number of BLAST hits was greater than ten (in order to exclude repetitive regions); if the variants were indels, monoallelic or multiallelic; if the distance to the nearest identified SNP was less than 40 bp; if the SNP was not polymorphic in the Tiritiri Mātangi population (expected to have a reduced genetic diversity as a result of the reintroduction bottleneck); if the read depth < 19, minimum flanking contig sequence < 6 or maximum flanking contig sequence < 35; and if SNP types weren't A/G, C/T, G/T, or A/C (as these types only require one probe on the SNP chip). SNPs from each assembly were then merged into a single file. Where SNPs mapped to identical positions on the zebra finch genome, only the best version of the SNP was included in the file; defined as the version with the highest quality score and the required minimum flanking sequence. Distances to the next SNP or end of chromosome on the zebra finch genome were then calculated (including the potential differences due to gap size present in the BLAST output). A total of 9,403,082 SNPs remained after the above quality filters and merging (Table 1).

Table 1: SNPs per assembly before and after filtering. For each assembly, the size of the assembled draft genome (excluding N), the N50 value, and the coverage of the zebra finch genome is shown along with gene completeness as assessed with CEGMA. The number of SNPs identified in SAMtools / Stacks after filtering and the number of SNPs that are only found in that assembly (Unique) are also detailed.

Assembly	Size without N	N50	% zebra finch	CEGMA % completeness	Filtered SNPs	Unique SNPs
1	979,320,412	1,175	63.5	7.7	1,050,028	734,870
2	929,250,757	676	59.0	1.2	963,210	690,776
3	836,706,703	379	51.3	0.0	720,939	527,184
4	967,240,548	989	62.3	6.1	1,041,331	735,443
5	971,406,841	1,086	62.8	5.7	1,058,210	745,124
6	991,036,624	1,559	65.0	10.9	1,046,922	719,408
7	988,256,031	1,371	65.2	7.7	1,093,037	767,461
8	951,613,123	834	60.9	2.8	1,014,469	719,369
9	991,536,846	1,482	65.6	6.1	1,088,067	755,084
10	1,002,019,675	1,928	66.0	16.1	1,019,426	685,923
3 in 1	1,048,884,582	3,137	68.3	23.8	806,025	420,806
10 in 1	1,046,305,858	1,960	67.6	15.3	764,792	416,123

RAD	11,827,080	90		9,484	2,388

Selection of SNPs for the hihi SNP chip

A total of ~200,000 SNPs were selected for consideration for the SNP chip, from a combination of the RAD-seq and WGS SNPs. From the 30,835 RAD-seq SNPs, 9,484 SNPs were selected for consideration on the custom SNP chip by setting the minimum number of individuals genotyped for a SNP to 10. RAD-seq contigs containing SNPs were aligned to the zebra finch genome as described above for the WGS SNPs. From the WGS, three steps were used to select SNPs for consideration. First, a target number of SNPs that aligned to each zebra finch chromosome were determined, proportional to the length of the chromosome. Chromosomes were categorised into three SNP-density classes: high (chromosomes 10-28, LG1, LG5, LGE22, Z, and the mitochondria), medium (chromosomes 1-9, 1A, 1B, and 4A) and low (chromosome Un, which indicates zebra finch sequence of unidentified chromosome location), to reflect higher gene densities and recombination rates on avian micro- compared to macro- chromosomes (Axelsson *et al.* 2005; van Oers *et al.* 2014). Densities were adjusted so that high density chromosomes had approximately nine times more selected SNPs per megabase than chromosome Un, and medium density chromosomes had approximately 5.5 times more SNPs per megabase than chromosome Un. SNPs were further filtered to be at least 80 base pairs from the next identified SNP. SNPs were ranked based on SAMtools quality score and the appropriate number of SNPs taken based on this ranking for each chromosome, with a total of 185,647 selected in this step. Second, a total of 4,000 SNPs which mapped to so-called 'random' zebra finch chromosomes (e.g. 1_random) were also included based on their quality score ranking, proportional to the total number of SNPs

detected on each chromosome. Random chromosomes are made up of collections of contigs and scaffolds where there is evidence that they belong to that chromosome, but not enough evidence to place them in order on the chromosome (https://www.ncbi.nlm.nih.gov/grc/help/definitions/). Third, from a list of 1,185 high quality SNPs which did not map to the zebra finch genome, 560 were selected to represent each of the contigs with only one high quality SNP. Flanking sequence each side of the SNP was extracted for all SNPs and formatted for Affymetrix according to their specifications using a perl script, with 35 bases both upstream and downstream of the SNP extracted where possible; if not possible, 35 bases on one side and a minimum of one base on the other side were extracted.

A total of 199,691 SNPs were submitted to Affymetrix, Thermo Fisher Scientific, Santa Clara, California for assessment of the suitability of the SNPs for inclusion on a custom AXIOM 384HT SNP chip (assessment includes a check for duplicate flanking information suggesting repetitive elements, and an assessment of the complexity of the flanking sequence). Of the submitted SNPs, a total of 79,451 were deemed by Affymetrix to not be 'designable' in either the forward or reverse flanking sequence. A total of 8,563 SNPs (4.29%) were repetitive in either their upstream or downstream flanking sequence, with a further 388 (0.19%) SNPs repetitive in both, which may represent the same SNP identified from two or more assemblies but with different contig mappings. From the remaining 120,240 designable SNPs, 59,928 SNPs were selected for inclusion on the SNP chip. All 654 RAD-seq SNPs designable in both forward and reverse directions were included, along with a further 73 RAD-seq SNPs designable in one direction and neutral in the other, selected by ranking SNPs based on their combined Affymetrix pconvert score. From the WGS SNPS, 48,220 were selected using a similar approach to the previous density selection. Densities

were again adjusted such that high and medium density chromosomes had approximately nine times and 5.5 times more SNPs per megabase than chromosome Un respectively. All 2,112 SNPs that were the only SNP within an annotated gene but failed to be selected among the best SNPs on a chromosome were also added, along with 559 high quality WGS SNPs that failed to map to zebra finch. A further 4,155 SNPs were tiled in both directions as both their forward and reverse flanking sequence was assessed to be neutral. The 59,928 SNPs were submitted to Affymetrix, and 58,466 of these were manufactured on the custom Hihi50K AXIOM 384HT array, which included 727 RAD-seq SNPs, 9,056 SNPs within annotated zebra finch genes (including 528 duplicates tiled in both directions) and an overall total of 4,112 SNPs tiled in both directions. An overview of the process to detect RAD-seq and WGS SNPs and select SNPs for inclusion on the array is provided in Supplementary Figure S1.

Finally, following submission of SNPs to Affymetrix, WGS SNP probe flanking sequences were re-aligned to the zebra finch genome using BLAST. This additional BLAST was done to determine whether the short flanking sequence gave the same predicted genome position as the full contig from which the SNP was originally detected. It is expected that these shorter flanking sequences may give a more accurate genome position than the whole contig, because we only required contigs to have >80% of their sequence aligning to the zebra finch genome in order to allocate them a genome position (i.e., the SNP could have been in a section of the contig that did not align). Flanking probe genome positions were assessed as (i) aligning where expected, (ii) no longer aligning to the expected chromosome but instead aligning to the corresponding random chromosome, (iii) aligning to a different chromosome, (iv) aligning to both the expected and random chromosomes or (v) aligning to both the expected and an alternative chromosome.

Note that software and methods described above were selected based on best practice at the time of the design of the array; many are no longer current or best practice.

Samples for genotyping

This study used blood samples collected from Tiritiri Mātangi between the 1996/97 and 2014/15 austral breeding seasons, blood samples of birds translocated from the remnant population in Te Hauturu-o-Toi Island in the 2003/04, 2006/07, 2008/09 and 2010/11 breeding seasons, blood samples from Kāpiti Island (40°51'S 174°55'E) in the 2003/04 breeding season, blood samples from Sanctuary Mountain Maungatautari (38°03'S 175°34'E) in the 2011/12 breeding season and feather samples from Zealandia Wildlife Sanctuary from the 2013/14 and 2014/15 breeding seasons (Supplementary Figure S2, Supplementary Table S3).

For all except the Zealandia population, blood samples were collected by brachial venipuncture (approximately 70 µL) and stored in 95% ethanol as described previously (Brekke *et al.* 2011). For Zealandia, two or three downy feathers were plucked from the underside of 21 day old nestlings and stored in 95% ethanol. DNA for ~2,500 individuals was extracted from the blood and feather samples using Qiagen DNeasy Blood and Tissue kits as recommended by the manufacturer. DNA was quantified on a NanoDrop 8000. A total of 1,536 samples were chosen for genotyping on the hihi SNP chip based on their DNA quality (260/280 ratio of ~1.8 – 1.9 where possible), concentration (≥30 ng/µL where possible; Affymetrix recommendations are a minimum of 25 µL at a minimum concentration of 23ng/µl, with a recommended concentration of 30 ng/µL) and ensuring representation across cohorts. In total, 1,290 Tiritiri Mātangi, 55 Te Hauturu-o-Toi, 14 Kāpiti, 12 Sanctuary

Mountain Maungatautari and 163 Zealandia samples were genotyped, plus two samples of unknown origin (Supplementary Table S3). Samples were quantified before genotyping by Affymetrix using PicoGreen.

Population statistics and linkage disequilibrium

Following array hybridisation and imaging, genotypes were called using default settings in the Axiom Analysis Suite software and exported from the software in *plink* (Purcell *et al.* 2007) format. We then used *plink* v1.9 (www.cog-genomics.org/plink2) to further filter individuals and SNPs, calculate allele frequencies, infer population structure and calculate linkage disequilibrium, using the mapping of the SNP probe flanking sequences to the zebra finch genome as a proxy for their positions relative to each other on the hihi genome. The 1,475 successfully genotyped samples were assumed to have unknown parents and first filtered to remove duplicates (keeping the duplicate sample with the highest genotyping rate; 1,469 individuals). SNPs were filtered to only those mapping to autosomes in the zebra finch genome with a minor allele frequency of greater than 0.01, leaving 40,616 variants. Allele frequencies were calculated per population for each SNP. We also partitioned SNPs into those detected from WGS and from RAD-seq to assess any difference in the minor allele frequency distribution. Next, a principal component analysis of population structure for all individuals across the five populations was calculated by first pruning SNPs in a sliding window of 100kb with $r^2 > 0.5$ (*plink* option 'indep-pairwise 100 10 0.5') and then inferring all principal components of relatedness between individuals ('make-rel' and 'pca 1469'). The set of individuals was then further filtered using 'rel-cutoff 0.25' to exclude one member of each pair of samples with observed genomic relatedness greater than 0.25, leaving 401 samples across populations. Linkage disequilibrium (measured as the correlation coefficient r^2)

between all pairs of SNPs on the same chromosome was then calculated using the *plink* options 'r2', 'ld-window-r2 0', 'ld-window-kb 200000' and 'ld-window 20000' to calculate r2 between all pairs of variants. The r^2 values were binned into 10 kb units and per-bin averages calculated using *R* for all chromosomes and separately for macro- (chromosomes 1-9 and 1A) and micro- chromosomes (chromosome 10-15, 17-26, 1B, 4A and LGE22). The decay of linkage disequilibrium over physical distance was then plotted in *R* (R Core Team 2019). The dataset excluding close relatives was also subset to the 358 individuals from Tiritiri Mātangi and linkage disequilibrium calculated as above.

Results

RAD-seq and WGS assemblies

After read filtering and quality control, between 90.8 – 93.7% RAD-seq reads and between 63.5 – 82.0% WGS reads were retained (Supplementary Table S1, Supplementary Table S2). The RAD-seq assembly contained 131,412 contigs, all of length 90 bases (Table 1). Filtering and assembly of WGS draft genomes from reads of single samples and pooled reads from three (3 in 1) or ten samples (10 in 1) resulted in genomes slightly smaller than the median bird genome length, but well within their known range of ~0.96 – 2.2 Gb (Kapusta & Suh 2017). Draft genomes of pooled reads were marginally larger than those from single samples, and single sample assembly sizes were well correlated with the total number of filtered reads that went into each assembly (Table 1; Supplementary Table S2). CEGMA completeness ranged from 0 – 23.8%, reflective of the low contiguity and hence very small N50 of all assemblies (range 379 – 3,137). The assemblies were however a good representation of the k-mer properties of the reads, with Jaccard coefficients ranging from 83.3 – 93.8% for all k-mers and from 86.0 – 98.5% for shared k-mers. Jaccard coefficients calculated from shared k-mers were generally higher when comparing reads to their own assembly (for example, the k-mer profile of reads from sample 1 when compared to the assembly of these sample 1 reads was 97.5%), but k-mer profiles of samples with higher numbers of filtered reads tended to be well-representative of all assemblies (Supplementary Table S4).

Assembled genomes were mapped to zebra finch to ascertain their contig positions and proportion of the genome that was captured (genome coverage). Coverage of the zebra finch genome ranged from 51.3% to 68.3% (Table 1). Fifty four percent of contigs overlapped with another contig when mapped to the zebra finch genome, with a median overlap between these contigs in all single assemblies at -30 bases with a range from -110 to -1, and the median overlap between contigs in all pooled assemblies at -30, with a range from -109 to -1 (Supplementary Figure S3a). Neighbouring contigs for all assemblies had gaps between them 46% of the time. In single assemblies, the median gaps between neighbouring contigs was 17,878 with a range between 0 and 1,083,088 bp, while the median gaps in pooled assemblies were smaller at 6,431 bp, with a range of 0 to 610,970 bp (Supplementary Figure S3b).

SNP identification and characterisation

Of the 9,484 RAD-seq SNPs which passed the quality filter and were considered for inclusion on the array, 2,388 mapped to zebra finch and were detected only in the RAD-seq data set (i.e. were not found among the WGS SNPs; Table 1). Pooled sample draft genome assemblies resulted in a smaller number of detected SNPs before and after SNP chip design filtering (Table 1). Across all assemblies, a total of 9,403,082 WGS SNPs remained after quality filtering and merging of SNPs with homologous zebra finch positions (note that some of these SNPs are represented across multiple assemblies; Table 1).

A total of 58,466 SNP markers were tiled on the hihi SNP chip, with the number of SNPs per chromosome listed in Supplementary Table S5. Chromosomes selected to have a high density of SNPs had 74.5 SNPs per Mb, those with medium density had 43.4 SNPs per Mb, and those with low density had 13.4 SNPs per Mb (Figure 1). The distribution of gap length between adjacent SNPs in each of these groups is illustrated in Supplementary Figure S4.

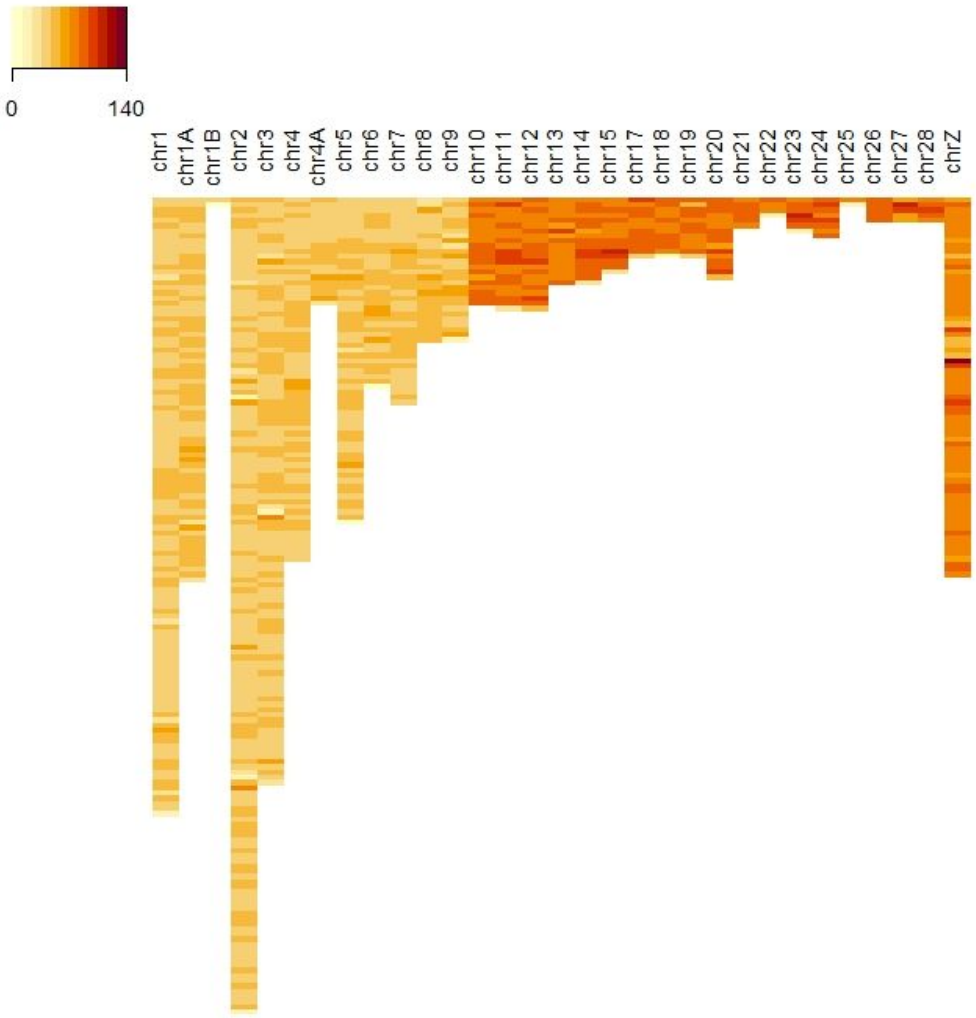


Figure 1. Density of the hihi SNP array SNPs across the zebra finch genome. Density is shown as the total number of SNPs per Mb window for 54,341 SNP markers included on the hihi SNP array, after excluding those that mapped to random chromosomes, chromosome Un, or did not map to the zebra finch genome.

WGS SNP probes re-aligned to the zebra finch genome showed that the majority (93.2%) of SNPs exhibited the same mapping as the hihi contigs from which they were identified. Only a small number (4.9%) aligned to a different chromosome and many of these were on the 'Un' chromosome. A small number (0.2%) mapped to the corresponding 'random' chromosome;

on both the expected chromosome and either the 'random' section of the same chromosome (0.3%) or a second chromosome (1.4%); or on the expected 'random' chromosome as well as an alternative chromosome (<0.01%) (Figure 2).

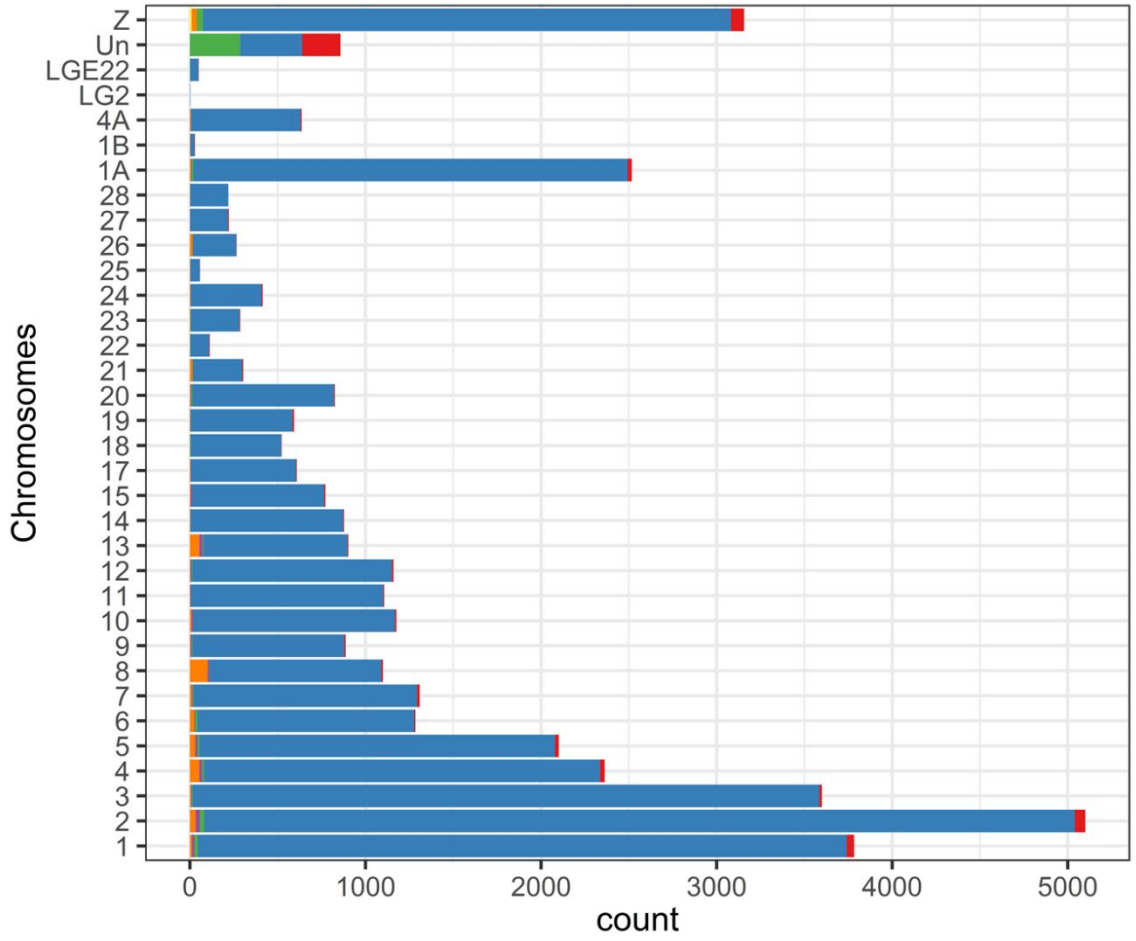


Figure 2: Verifying WGS SNP probe positions on the zebra finch genome. Homologous SNP positions were initially estimated from their position within their hihi contig's best-hit to the zebra finch genome. To ensure they were placed correctly, once the probes were designed, they were re-aligned to the zebra finch using only the SNP and its flanking sequences (37–71 bases). The majority aligned where expected (blue), a small number were found on the 'random' part of the same chromosome (orange), on an alternative chromosome (red), or on both the expected and random chromosomes (purple), or on both the expected and an alternative chromosome (green).

SNP chip

Based on genotyping 1,536 hihi samples, of the 58,466 SNPs on the custom Hihi50K AXIOM 384HT array, 42,212 markers (72.2%) were polymorphic, had individuals with all three

genotypes, and passed Affymetrix filtering metrics in the Axiom Analysis Suite software as determined by the Recommended.ps file from Axiom filtering (hereafter termed 'successfully genotyped' or 'successfully converted'). Of those that 'failed' (i.e. are not informative for this dataset), 7,898 (13.1%) passed filtering metrics but were monomorphic, 1,131 (1.9%) passed filtering metrics but the minor allele homozygote was missing, and the remainder (12.4%) failed due to low call rates or other quality filters. SNPs that were originally assessed by Affymetrix as being 'neutral' to design were significantly more likely to fail than those that were 'recommended' (neutral: 6,563 failed / 6,478 successful, recommended: 9,691 failed / 35,734 successful; Pearson's Chi-squared test with Yates' continuity correction: Chi-squared = 4241.5, df = 1, $p < 2.2e\text{-}16$). As expected, given the selection of WGS SNPs that were polymorphic in Tiritiri Mātangi birds, and of RAD-seq SNPs that were observed in at least five of the 31 individuals, the minor allele frequency distribution was skewed to common alleles (Supplementary Figure S5a, Supplementary Figure S5b).

RAD and WGS success rates

There was a lower success rate among the WGS SNPs, with 83.1% of 727 SNPs generated from RAD data successfully converting (i.e. successfully passing genotyping and being polymorphic) compared to 72.1% of the 57,739 SNPs from WGS data (Pearson's Chi-squared test with Yates' continuity correction: Chi-squared = 42.88, df = 1, $p = 5.812\text{ e-}11$). WGS SNPs that mapped to assembled zebra finch chromosomes had higher conversion rates (72.6%) than SNPs that mapped to random chromosomes, chromosome Un or did not map at all (63.8%) (Pearson's Chi-squared test with Yates' continuity correction: Chi-squared = 122.71, df = 1, $p < 2.2e\text{-}16$). This trend was driven by very high failure rates of SNPs that failed to map to the zebra finch genome (71.7%, compared to 29.2% for each of random and Un; Supplementary Table S5).

SNP success rate per assembly

Of the twelve assemblies, seven assemblies had conversion rates around 80%, while the other five assemblies had lower conversion rates (59-74%). All five assemblies with low conversion rates were assemblies of single individuals (Table 2, Supplementary Figure S6). As a consequence, on average, the two pooled assemblies showed higher success rates (80.9%) than the ten single assemblies (average of 71.5%) (Pearson's Chi-squared test with Yates' continuity correction: Chi-squared = 156.54, df = 1, $p < 2.2e\text{-}16$; Figure 3).

Table 2: SNP probe performance on Affymetrix SNP array. Absolute numbers for the performance of the SNPs from each assembly on the Affymetrix SNP array. Each assembly had some overlapped SNPs shared with single assemblies and also some that were shared with pooled assemblies. A large number of the total SNPs were only found in one assembly (i.e. Unique), either single or pooled.

	Total			Unique			Shared with an assembly from a single individual's data			Shared with an assembly from pooled data		
Assembly	**Successful**	**Fail**	**Conversion rate (%)**	**Successful**	**Fail**	**Conversion rate (%)**	**Successful**	**Fail**	**Conversion rate (%)**	**Successful**	**Fail**	**Conversion rate (%)**
1	4,753	2,492	66	3,187	1,960	62	1,399	486	74	714	241	75
2	3,562	903	80	2,516	556	82	982	325	75	442	114	79
3	2,276	811	74	1,773	564	76	476	240	66	191	62	75
4	4,447	3,092	59	3,096	2,576	55	1,238	480	72	605	213	74
5	4,629	3,036	60	3,187	2,502	56	1,325	495	73	671	230	74
6	5,125	2,507	67	3,323	1,940	63	1,514	503	75	936	285	77
7	4,894	1,198	80	3,313	679	83	1,380	478	74	735	226	76
8	3,943	979	80	2,755	536	84	1,123	421	73	527	175	75
9	4,910	1,209	80	3,285	656	83	1,419	494	74	811	281	74
10	5,049	1,174	81	3,149	596	84	1,540	499	76	1,042	334	76
3in1	5,447	1,274	81	2,493	487	84	2,142	625	77	5,447	1,274	81
10in1	5,145	1,222	81	2,433	488	83	1,900	572	77	5,145	1,222	81

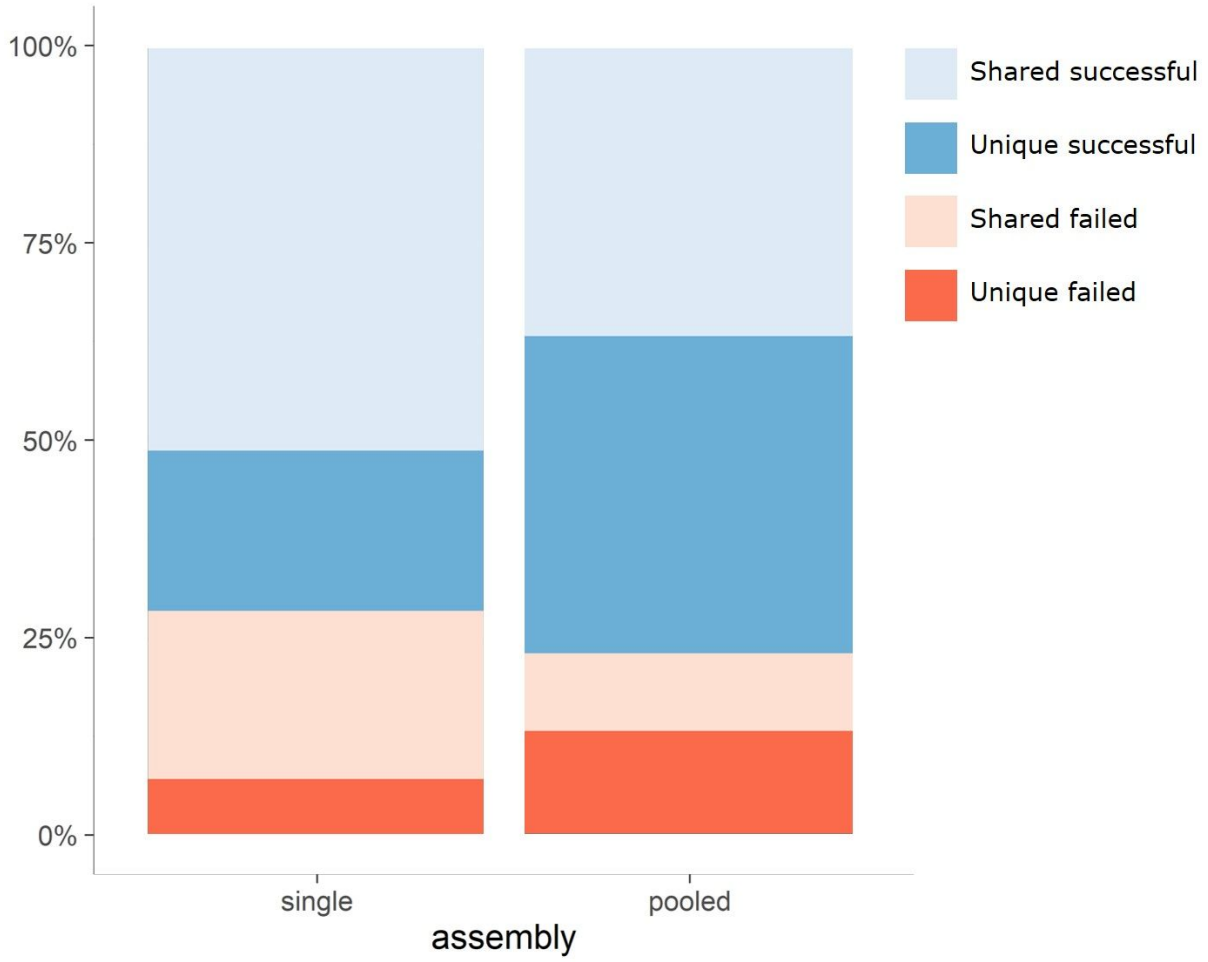


Figure 3: Averaged SNP performance for probes from single and pooled assemblies. The single column represents all array SNPs on all draft assemblies for each of the ten samples assembled separately. The pooled column represents all array SNPs from the assembly of pooled reads from samples 6, 9 and 10 (3 in 1) and pooled reads from all ten samples (10 in 1). The graph shows the proportion of SNPs that were found only in one assembly and successfully genotyped (light blue), SNPs that were also found in other assemblies that successfully genotyped (dark blue), SNPs that were found only in one assembly and failed genotyping (light orange), and SNPs that were also found in other assemblies that failed genotyping (dark orange).

A large proportion of SNPs were found only in one assembly (as shown by the 'Unique successful' and 'Unique failed' segments of Supplementary Figures S6, Figure 3 and also in the 'Unique SNPs' in Table 2). In general, there were marginally more SNPs shared with pooled assemblies than with single assemblies (Table 2). SNPs in single assemblies had a significantly greater chance of success if they were also found in other single assemblies or in pooled assemblies (see Supplementary Table S6 for test statistics). Pooled assembly SNPs

519 had a significantly greater chance of success if they were also found in one or more single
520 assemblies, but if they were also found in the second pooled assembly this had no significant
521 impact on their success rate (Supplementary Table S6).
522

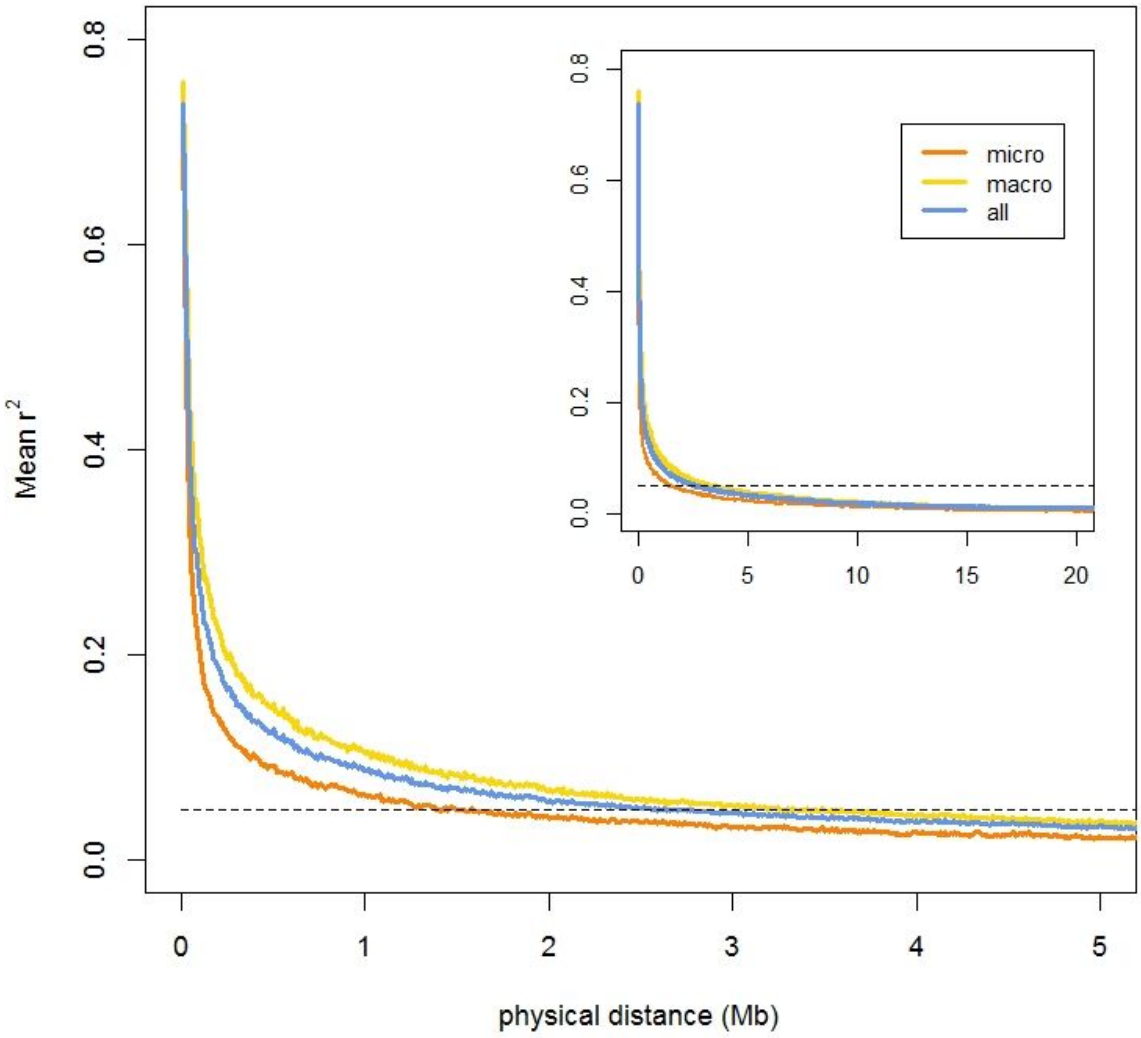
523 **Sample type, quantity and quality**

524 Of the total 1,536 samples, 96.03% were successfully genotyped according to Axiom
525 Analysis Suite filtering metrics. Although no duplicate samples were intentionally included
526 on the chip, Axiom Analysis Suite identified six replicated samples, likely due to plating
527 errors during sample extractions. From these samples, genotyping reproducibility could be
528 calculated and was very high at 99.98%. There was no significant difference in success rate
529 between blood and feather samples processed on the array (Pearson's Chi-squared test: Chi-
530 squared = 0.040316, df = 1, *p-value* = 0.8409; there were 1,318 blood samples that passed and
531 55 that failed, feathers had 157 pass and 6 fail).
532

533 DNA sample concentration as measured by Affymetrix using PicoGreen had an impact on
534 SNP genotyping success rates (Supplementary Figure S7a). A Mann-Whitney-Wilcoxon Test
535 demonstrates that the mean DNA concentration of the fail group (30.5 ng/µL) is significantly
536 smaller than that of the pass group (60.4 ng/µL), (W = 22557, p = 3.915e-11). In contrast,
537 NanoDrop measurements of DNA quality showed no significant impact of being outside the
538 recommended 260/280 ratio (1.8 – 1.9) on the pass rate of the samples (Pearson's Chi-squared
539 test: Chi-squared = 3.6412, df = 1, p = 0.05637; of the DNA that fell in the recommended
540 260/280 ratio, 558 passed and 16 failed, of the DNA outside the recommended 260/280 ratio,
541 898 passed and 45 failed; Supplementary Figure S7b).
542

Population statistics and linkage disequilibrium

Hihi populations show some population structure and slight differences in allele frequency distribution, reflective of bottleneck histories (Supplementary Figure S5b, Supplementary Figure S8). Linkage disequilibrium, measured as the correlation coefficient r^2 between all pairs of SNPs on the same chromosome, was very high, but decayed more rapidly in the micro-chromosomes compared to the macro-chromosomes, with r^2 reaching 0.05 by approximately 1.52 Mb and 3.34 Mb respectively (Figure 4). The average (median) inter-marker distance was 23,284 (17,341) and the average (median) r^2 between neighbouring markers was 0.558 (0.565). Results from only Tiritiri Mātangi individuals were almost identical (average (median) r^2 = 0.560 (0.573); Supplementary Figure S9).



553

554 **Figure 4. Linkage disequilibrium in the hihi genome.** Decline of linkage disequilibrium,
555 measured as the correlation coefficient r^2, between pairs of SNPs for micro-chromosomes
556 (chromosome 10-15, 17-26, 1B, 4A, and LGE22), macro-chromosomes (chromosomes 1-9
557 and 1A) and all chromosomes, with physical distance based on alignment of SNPs to the
558 zebra finch genome. The main graph shows decay from 0 – 5 Mb between marker pairs, while
559 the inset zooms out to 0 – 20 Mb. Dotted horizontal lines correspond to an r^2 of 0.05. Linkage
560 disequilibrium was calculated after excluding highly related individuals; final input was 401
561 individuals genotyped at 13,126 micro-chromosome SNPs and 27,490 macro-chromosome
562 SNPs.

Discussion

We have demonstrated that by combining sequencing reads from individual samples and from pooled samples for assembly and SNP detection, we were able to identify a large number of SNPs in a cost-effective manner from low coverage sequencing, in the absence of a high-quality hihi genome assembly. A subset of the identified WGS and RAD-seq SNPs were tiled on an Affymetrix 50K SNP chip, with 1,475 individuals successfully genotyped at 42,212 polymorphic SNPs across the genome. Genotype data from this array has been used to infer some degree of population structure and high levels of linkage disequilibrium in the species (this study), reflective of the establishment history of hihi populations from a single remnant population. The array has also enabled the genetic basis of adaptive morphological traits in the Tiritiri Mātangi population to be determined (Duntsch *et al.* 2020). Our future work will focus on better understanding the differences in genetic diversity between populations, which may inform conservation translocations in order to maintain variation in these small, isolated populations. The array data also provides a valuable resource to investigate inbreeding and inbreeding depression, particularly in the extensively monitored Tiritiri Mātangi population. We acknowledge that, by design, the minor allele frequency distribution of genotyped SNPs is skewed to common alleles, and variation has been predominantly sampled from Tiritiri Mātangi. While this may limit or place caveats on some analyses, the array will be an invaluable genomic resource for our ongoing work investigating adaptive potential in this threatened species.

The low coverage whole genome short read sequencing resulted in assemblies that were highly fragmented. As a consequence, we also assembled draft genomes by pooling reads from the three individuals with the largest number of reads after quality control and by

pooling all ten individuals. Pooling individuals before sequencing is considered an effective strategy to reduce overall costs (Wang *et al.* 2013). Here, we demonstrate that pooling low-coverage sequencing data may also offset the low per-individual coverage to some degree. The much higher genome coverage of sequence reads used in the pooled assemblies resulted in larger N50 values compared to the single assemblies (Table 1). However, there is a risk that the much higher overall level of polymorphism from pooled samples is likely to have led to regions being duplicated in the assembly, as suggested by the larger estimated genome size of both of the pooled assemblies. This is further supported by slightly lower Jaccard similarity indices when comparing how well these assemblies represented the k-mer profiles of their input reads, compared to similarities of the single-individual assemblies (Supplementary Table S4).

Interestingly, despite the potential duplication of genomic regions in the pooled assemblies, the assemblies from pooled data yielded SNPs that were on average much more likely to successfully genotype on the SNP chip than SNPs identified from single assemblies (Figure 3). It should be noted that the overall difference in success rates between single and pooled assemblies was due to poor conversion rates of SNPs from half of the single assemblies. However, our data suggests that pooled assemblies may be able to attenuate the effects of variation in the quality of low coverage individual assemblies, because the 3 in 1 assembly included data with a high failure rate from sample 6 in addition to data with lower failure rates from samples 9 and 10, and the 10 in 1 assembly had half the samples with higher and half with lower failure rates. Despite larger assembly sizes, the relatively small number of SNPs detected from these pooled assemblies may be a consequence of duplicated regions in the assembly translating into lower downstream mapping scores and lower numbers of variants called, as reads mapped back to the assembly can match more than one location.

612 Lower mapping scores may then result in lower quality scores for these SNPs and fewer SNPs
613 from these regions reaching the quality threshold and being included on the array. This in turn
614 may have contributed to higher rates of SNP conversion overall, as polymorphisms in
615 flanking regions that interfere with the SNP probe binding will be minimised.
616
617 Each of the draft hihi genomes from low coverage data resulted in a large proportion of SNP
618 discoveries that were found only in that assembly, regardless of whether they were from
619 single or pooled samples. The ten WGS assemblies each covered between 51-68% of the
620 zebra finch genome, and CEGMA estimates that very few gene sequences were present in-full
621 in any of the assemblies, and so each could be representing a large proportion of the genome
622 not assembled in the others. As bird genomes are highly conserved in gene synteny and
623 chromosomal structure (Zhang *et al.* 2014), it is expected that the zebra finch genome
624 coverage will represent a good estimate of how much of the whole hihi genome each of these
625 draft assemblies cover. The pooled assemblies shared only a small proportion of the SNPs
626 discovered in the single-individual assemblies. Furthermore, pooling reads from ten birds
627 identified different SNPs than pooling reads from three birds. Importantly, these results
628 indicate that assembling and remapping data in different ways can enhance the utility of the
629 dataset and lead to the discovery of high-quality SNPs that would not otherwise be detected
630 from a one-off assembly using a single or pooled sample.
631
632 Although each assembly yielded a relatively high number of SNPs after filtering, combining
633 SNPs across all assemblies enabled us to further filter the dataset to choose SNPs for
634 inclusion on the SNP chip that were of high quality and were at least 80 base pairs from the
635 next identified SNP, and further, were designable by Affymetrix. What was particularly
636 valuable was the ability to identify SNPs that had been detected from mapping reads to more

than one assembly, which had much higher success rates than those that were only identified when mapping to one assembly. van Bers *et al*. (2012) similarly found that SNPs identified in both the United Kingdom and Netherlands great tit populations had higher conversion rates. In both the hihi and great tit studies, the identification of shared SNPs across assemblies was enabled by mapping SNPs to the zebra finch genome, but we note that in the absence of a high quality genome from a related species it would be possible to, for example, extract ~50 flanking bases upstream and downstream of the identified SNP and check for duplicates across datasets.

Despite the relative success of SNPs from pooled assemblies and shared SNPs, the overall success rate of SNPs tiled on the array to high quality genotypes is not as high as we had expected, with 72.2% of the total SNPs polymorphic and passing Affymetrix filtering metrics. If monomorphic sites and SNPs for which no minor allele homozygote was observed are included, 87.6% passed the filtering, which compares well to conversion rates in other non-model avian species. For example, the flycatcher 50K SNP chip reported a 90% conversion rate (Kawakami *et al.* 2014); the 200K house sparrow chip a 92.8% conversion rate (Lundregan *et al.* 2018); the 10K great tit chip an 86% conversion rate (van Bers *et al.* 2012); and the great tit 500K SNP chip reported an 87% conversion rate for SNPs previously typed on the 10K SNP chip and an 82% conversion rate for unvalidated SNPs (Kim *et al.* 2018). Notably, SNP discovery for all but the great tit 10K chip were based on high-coverage and generally contiguous reference genome sequences. The few RAD-seq SNPs included on the chip were more likely to successfully genotype and be polymorphic than the WGS SNPs. It was expected that RAD-seq data would be more robust as it was generated from samples of 31 birds at much higher coverage per site.

Overall individual genotyping success rates were very high (96.03%), with no effect of DNA quality or sample type on genotyping success, although DNA concentration did impact sample success. The Affymetrix recommended concentration of 30ng/µl was relaxed in order to accommodate representation of cohorts and populations with fewer available DNA samples, such that 368 samples fell below the recommended concentration. Given that 333 of these samples genotyped successfully (albeit with a lower success rate than those above the recommended minimum), for important samples it may be worth attempting to genotype them even if DNA concentration is low. Failure rate in our study is much lower than has been reported elsewhere for samples of low DNA quantity (Kim *et al.* 2018).

To maximise cohort and population representation, of the 1,517 samples with 260/280 NanoDrop measurements, 798 were included on the SNP chip even though they had DNA quality measures outside the recommended 260/280 ratio of 1.8-1.9 for DNA (Supplementary Figure S7b). We found that DNA quality had no significant effect on the overall genotyping success of the sample on the SNP array, as has been shown elsewhere with human saliva samples genotyped on an Illumina OmniExpress array (Gudiseva *et al.* 2016) and fish scale samples genotyped on an Illumina iSelect array (Johnston *et al.* 2013).

No significant difference in the genotyping success rate of samples extracted from feather or blood was found. Although taking blood samples is in most cases preferable, taking feathers is useful when handling is difficult, drawing blood might present a danger to the bird, or field workers are not trained to take blood samples (McDonald & Griffith 2011). Further, feathers are relatively easy to store and transport. One limitation is that the DNA extraction uses the whole sample (the plucked shaft from two-three plucked feathers), so there is no opportunity for reanalysis of the sample. Nucleated erythrocytes make bird blood an effective source of

687 DNA, but here, in agreement with previous studies (Harvey *et al.* 2006; Maurer *et al.* 2010),
688 we demonstrate that feathers are sufficient to successfully genotype an individual in cases
689 where obtaining blood is not possible.
690
691 The 42,212 polymorphic SNPs successfully genotyped on 1,475 individuals across five
692 populations represent a valuable tool for ongoing genomic studies of the genetic effects of
693 management practices on the populations, assessment of inbreeding and inbreeding
694 depression, the genomic architecture of traits (Duntsch *et al.* 2020), population structure and
695 linkage disequilibrium (this study), genetic diversity and recombination landscape of the
696 genome, and overall estimation of evolutionary potential. Crucially, we have demonstrated
697 here a very high level of linkage disequilibrium in the hihi genome, including substantial
698 linkage disequilibrium between neighbouring markers, suggesting the density of genotyped
699 SNPs is well-powered to accurately tag most areas of the hihi genome. Genotype data from
700 this array will be an invaluable genomic resource for our ongoing work investigating adaptive
701 potential in this threatened species.
702
703

Acknowledgements

Many thanks to Rachel Tucker and Jon Slate at the NERC Biomolecular Analysis Facility, University of Sheffield, for completing the DNA extractions for the RAD-seq individuals, and to Selina Patel, School of Biological Sciences, University of Auckland, for all her help optimising the DNA extractions for the SNP chip genotyping. Thank you to Klaus Lehnert, School of Biological Sciences, University of Auckland, for bioinformatics advice on the SNP chip design. We thank Isabel Cantera for her initial RAD-seq assembly during her internship. Thank you to Jason Boone and the Floragenex team for the RAD-seq, and to the New Zealand Genomics Limited team for WGS sequencing. We extend many thanks to Alayna Burrett, Millennium Science (NZ) Pty Ltd, for brokering the contract for SNP chip development and genotyping with Affymetrix. We also thank everyone in the Affymetrix Scientific Services and Bioinformatics Services teams who helped with the design, manufacture, genotyping and data delivery of the SNP chip, in particular Christofer Bertani who designed the array. Our thanks to René Malenfant for feedback on an early version of this manuscript. We acknowledge the use of New Zealand eScience Infrastructure (NeSI) high-performance computing facilities. We are thankful to the Hihi Recovery Group, Department of Conservation and to all volunteers, past students and staff that contributed to monitoring and sampling the hihi populations at Tiritiri Mātangi Island, Zealandia Te Māra a Tāne, Kāpiti Island, Sanctuary Mountain Maungatautari and Te Hauturu-o-Toi. We acknowledge Ngati Manuhiri as Mana Whenua and Kaitiaki of Te Hauturu-o-Toi and its taonga, including hihi. AWS, KDL, PB and JGE were supported by a Marsden Grant (UOA1408) awarded to AWS from the New Zealand Royal Society Te Aparangi. AWS was also supported by a New Zealand National Science Challenge Biological Heritage Project Grant, Project 1.4, and PB was also supported by an AXA Fellowship and Research England. Permissions to conduct the

728 research and collect hihi blood samples were granted by the New Zealand Department of
729 Conservation, permit numbers 15073-RES, 13939-RES, 246-RES, 36186-FAU, 24128-FAU,
730 32213-FAU and 44300-FAU.
731
732
733

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Andrews S (2014) *FastQC A Quality Control tool for High Throughput Sequence Data.*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Axelsson E, Webster MT, Smith NGC, Burt DW, Ellegren H (2005) Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Research* **15**, 120-125.

van Bers NE, Santure AW, van Oers K*, et al.* (2012) The design and cross-population application of a genome-wide SNP chip for the great tit *Parus major*. *Molecular Ecology Resources* **12**, 753-770.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Brekke P, Bennett PM, Santure AW, Ewen JG (2011) High genetic diversity in the remnant island population of hihi and the genetic consequences of re-introduction. *Molecular Ecology* **20**, 29-45.

Camacho C, Coulouris G, Avagyan V*,* Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) *Stacks*: Building and Genotyping Loci *De Novo* From Short-Read Sequences. *G3: Genes|Genomes|Genetics* **1**, 171.

Chen N, Cosgrove EJ, Bowman R, Fitzpatrick JW, Clark AG (2016) Genomic Consequences of Population Decline in the Endangered Florida Scrub-Jay. *Current Biology* **26**, 2974-2979.

da Silva VH, Laine VN, Bosse M*, et al.* (2018) CNVs are associated with genomic architecture in a songbird. *BMC Genomics* **19**, 195.

Davey J, Hohenlohe P, Etter P, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

de Villemereuil P, Rutschmann A, Lee KD*,* Ewen JG, Brekke P, Santure AW (2019) Little adaptive potential in a threatened passerine bird. *Current Biology* **29**, 889-894.e883.

Decker JE, Pires JC, Conant GC*, et al.* (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences* **106**, 18644.

DePristo MA, Banks E, Poplin R*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498.

Duntsch L, Tomotani BM, de Villemereuil P, Brekke P, Lee KD, Ewen JD, Santure AW. (2020) Polygenic basis for adaptive morphological variation in a threatened Aotearoa | New Zealand bird, the hihi (*Notiomystis cincta*). *Proceedings of the Royal Society B: Biological Sciences* **287**, 20200948.

Ewen J, Renwick R, Adams L, Armstrong D, Parker K (2013) 1980-2011: 31 years of reintroduction efforts of the hihi (stitchbird) *Notiomystis cincta* in New Zealand. In: *Global Reintroduction Perspectives: additional case studies from around the globe* (ed. Soorae P). IUCN/SSC Re-introduction Specialist Group, Abu Dhabi, UAE.

781 Fumagalli M (2013) Assessing the Effect of Sequencing Depth and Sample Size in Population
782 Genetics Inferences. *PLoS ONE* **8**, e79667.
783 Gautier M, Flori L, Riebler A*, et al.* (2009) A whole genome Bayesian scan for adaptive
784 genetic divergence in West African cattle. *BMC Genomics* **10**, 550.
785 Gudiseva HV, Hansen M, Gutierrez L*, et al.* (2016) Saliva DNA quality and genotyping
786 efficiency in a predominantly elderly population. *BMC Medical Genomics* **9**, 17.
787 Hagen IJ, Billing AM, Rønning B, Pedersen SA, Pärn H, Slate J, Jensen H (2013) The easy
788 road to genome-wide medium density SNP screening in a non-model species:
789 development and application of a 10 K SNP-chip for the house sparrow (*Passer*
790 *domesticus*). *Molecular Ecology Resources* **13**, 429-439.
791 Hagen IJ, Lien S, Billing AM*, et al.* (2020) A genome-wide linkage map for the house
792 sparrow (*Passer domesticus*) provides insights into the evolutionary history of the
793 avian genome. *Molecular Ecology Resources* **20**, 544-559.
794 Harvey MG, Bonter DN, Stenzler LM, Lovette IJ (2006) A comparison of plucked feathers
795 versus blood samples as DNA sources for molecular sexing. *Journal of Field*
796 *Ornithology* **77**, 136-140.
797 Humble E, Paijmans AJ, Forcada J, Hoffman JI (2020) An 85K SNP Array Uncovers
798 Inbreeding and Cryptic Relatedness in an Antarctic Fur Seal Breeding Colony. *G3:*
799 *Genes|Genomes|Genetics* **10**, 2787.
800 Husby A, Kawakami T, Rönnegård L, Smeds L, Ellegren H, Qvarnström A (2015) Genome-
801 wide association mapping in a wild avian population identifies a link between genetic
802 and phenotypic variation in a life-history trait. *Proceedings of the Royal Society B-*
803 *Biological Sciences* **282**.
804 Johnston SE, Bérénos C, Slate J, Pemberton JM (2016) Conserved Genetic Architecture
805 Underlying Individual Recombination Rate Variation in a Wild Population of Soay
806 Sheep (*Ovis aries*). *Genetics* **203**, 583.
807 Johnston SE, Lindqvist M, Niemelä E*, et al.* (2013) Fish scales and SNP chips: SNP
808 genotyping and allele frequency estimation in individual and pooled DNA from
809 historical samples of Atlantic salmon (*Salmo salar*). *BMC Genomics* **14**, 439.
810 Judkins ME, Couger BM, Warren WC, Van Den Bussche RA (2020) A 50K SNP array
811 reveals genetic structure for bald eagles (*Haliaeetus leucocephalus*). *Conservation*
812 *Genetics* **21**, 65-76.
813 Kapusta A, Suh A (2017) Evolution of bird genomes—a transposon's-eye view. *Annals of the*
814 *New York Academy of Sciences* **1389**, 164-185.
815 Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of
816 genomics: marker-based measures are better than pedigrees. *Heredity* **115**, 63-72.
817 Kawakami T, Backström N, Burri R*, et al.* (2014) Estimation of linkage disequilibrium and
818 interspecific gene flow in Ficedula flycatchers by a newly developed 50k single-
819 nucleotide polymorphism array. *Molecular Ecology Resources* **14**, 1248-1260.
820 Kim JM, Santure AW, Barton HJ*, et al.* (2018) A high-density SNP chip for genotyping great
821 tit (*Parus major*) populations and its application to studying the genetic architecture of
822 exploration behaviour. *Molecular Ecology Resources* **18**, 877-891.
823 Kim SY, Lohmueller KE, Albrechtsen A*, et al.* (2011) Estimation of allele frequency and
824 association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**,
825 231.
826 Laine VN, Atema E, Vlaming P*, et al.* (2019) The genomics of circadian timing in a wild
827 bird, the great tit (*Parus major*). *Frontiers in Ecology and Evolution* 7, 152.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li H, Handsaker B, Wysoker A*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Lundregan SL, Hagen IJ, Gohli J*, et al.* (2018) Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis. *Molecular Ecology* **27**, 3498-3514.

Luo R, Liu B, Xie Y*, et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18-18.

Malenfant RM, Coltman DW, Davis CS (2015) Design of a 9K illumina BeadChip for polar bears (*Ursus maritimus*) from RAD and transcriptome sequencing. *Molecular Ecology Resources* **15**, 587-600.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ (2017) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574-576.

Maurer G, Beck N, Double MC (2010) A 'feather-trap' for collecting DNA samples from birds. *Molecular Ecology Resources* **10**, 129-134.

McDonald, PG, Griffith, SC (2011) To pluck or not to pluck: the hidden ethical and scientific costs of relying on feathers as a primary source of DNA. *Journal of Avian Biology* **42**, 197-203.

McKenna A, Hanna M, Banks E*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303.

Mead S, Poulter M, Beck J*, et al.* (2008) Successful amplification of degraded DNA for use with high-throughput SNP genotyping platforms. *Human Mutation* **29**, 1452-1458.

Morin PA, Luikart G, Wayne RK, the SNP Workshop Group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208-216.

van Oers K, Santure AW, De Cauwer I*, et al.* (2014) Replicated high-density genetic maps of two great tit populations reveal fine-scale genomic departures from sex-equal recombination rates. *Heredity* **112**, 307-316.

Parlato EH, Ewen JG, McCready M, Parker KA, Armstrong DP (2021). A modelling framework for integrating reproduction, survival and count data when projecting the fates of threatened populations. *Oecologia* **195**, 627–640.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067.

Peñalba JV, Wolf JBW (2020) From molecules to populations: appreciating and estimating recombination rate variation. *Nature Reviews Genetics* **21**, 476-492.

Purcell S, Neale B, Todd-Brown K*, et al.* (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559-575.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.

R Core Team (2019) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rutschmann A, de Villemereuil P, Brekke P, Ewen JG, Anderson N, Santure AW (2020) Consequences of space sharing on individual phenotypes in the New Zealand hihi. *Evolutionary Ecology* **34**, 821-839.

Santure AW, De Cauwer I, Robinson MR, Poissant J, Sheldon BC, Slate J (2013) Genomic dissection of variation in clutch size and egg mass in a wild great tit (*Parus major*) population. *Molecular Ecology* **22**, 3949-3962.

Santure AW, Poissant J, De Cauwer I, *et al.* (2015) Replicated analysis of the genetic architecture of quantitative traits in two wild great tit populations. *Molecular Ecology* **24**, 6148-6162.

Schweizer RM, vonHoldt BM, Harrigan R, *et al.* (2016) Genetic subdivision and candidate genes under selection in North American grey wolves. *Molecular Ecology* **25**, 380-402.

Silva CNS, McFarlane SE, Hagen IJ, *et al.* (2017) Insights into the genetic architecture of morphological traits in two passerine bird species. *Heredity* **119**, 197-205.

Stapley J, Reger J, Feulner PGD, *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology and Evolution* **25**, 705-712.

Taylor S, Castro I, Griffiths R (2005) *Hihi / stitchbird (Notiomystis cincta) Recovery Plan 2004–09* Department of Conservation, Wellington, New Zealand.

Viengkone M, Derocher AE, Richardson ES, *et al.* (2016) Assessing polar bear (*Ursus maritimus*) population structure in the Hudson Bay region using SNPs. *Ecology and Evolution* **6**, 8474-8484.

Wang W, Yin X, Soo Pyon Y, Hayes M, Li J (2013) Rare variant discovery and calling by sequencing pooled samples with overlaps. *Bioinformatics* **29**, 29-38.

Wellenreuther M, Hansson B (2016) Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. *Trends in Genetics* **32**, 155-164.

Zhang G, Li C, Li Q, *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311-1320.

Data Accessibility and Benefit-Sharing Statement

Hihi are of cultural significance to the Indigenous people of Aotearoa New Zealand, the Māori, and are considered a taonga (treasured) species whose whakapapa (genealogy) is intricately tied to that of Māori. For this reason, the raw reads, assemblies and genotypes for hihi will be made available by request on the recommendation of Ngati Manuhiri, the iwi (tribe) that affiliates as kaitiaki (guardians) for hihi. To obtain contact details for the iwi, please contact Dr Anna Santure: a.santure@auckland.ac.nz

Perl scripts used in the design of the hihi SNP array are available at https://github.com/klee8/key-scripts-for-hihi-snpchip-design

Author contributions

AWS, PB and JGE conceived of the study. MH, AZ and KDL contributed to the sample preparation and DNA extractions. KDL performed the WGS filtering, assemblies, SNP detection, SNP filtering and SNP selection for the array and conducted the analysis of the data. AWS performed the RAD-seq and linkage disequilibirium analyses. AW conducted the kmer analysis. CDM coordinated and helped to optimise the DNA extractions for the array. JGE and PB coordinated the data collection. KDL and AWS wrote the paper, with advice from AW and input from all other authors.

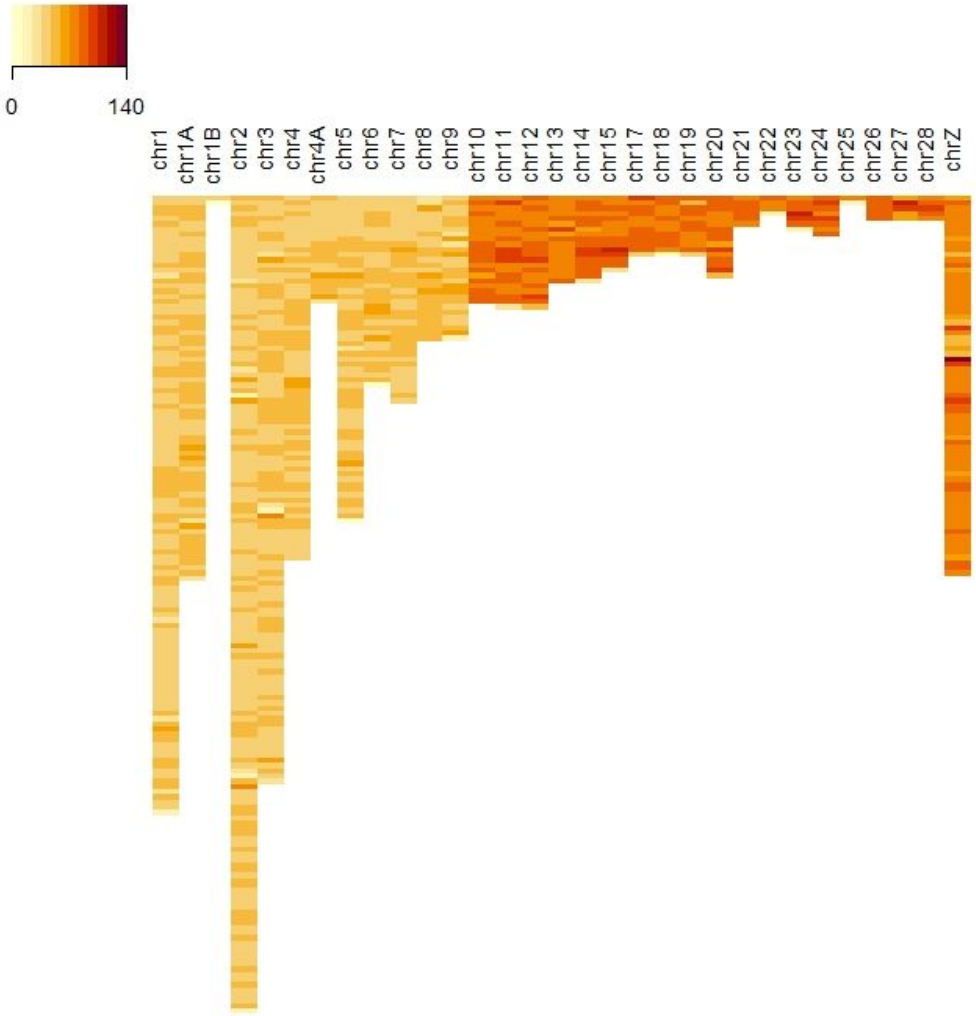
Tables and Figures

Table 1: SNPs per assembly before and after filtering. For each assembly, the size of the assembled draft genome (excluding N), the N50 value, and the coverage of the zebra finch genome is shown along with gene completeness as assessed with CEGMA. The number of SNPs identified in SAMtools / Stacks after filtering and the number of SNPs that are only found in that assembly (Unique) are also detailed.

Assembly	Size without N	N50	% zebra finch	CEGMA % completeness	Filtered SNPs	Unique SNPs
1	979,320,412	1,175	63.5	7.7	1,050,028	734,870
2	929,250,757	676	59.0	1.2	963,210	690,776
3	836,706,703	379	51.3	0.0	720,939	527,184
4	967,240,548	989	62.3	6.1	1,041,331	735,443
5	971,406,841	1,086	62.8	5.7	1,058,210	745,124
6	991,036,624	1,559	65.0	10.9	1,046,922	719,408
7	988,256,031	1,371	65.2	7.7	1,093,037	767,461
8	951,613,123	834	60.9	2.8	1,014,469	719,369
9	991,536,846	1,482	65.6	6.1	1,088,067	755,084
10	1,002,019,675	1,928	66.0	16.1	1,019,426	685,923
3 in 1	1,048,884,582	3,137	68.3	23.8	806,025	420,806
10 in 1	1,046,305,858	1,960	67.6	15.3	764,792	416,123
RAD	11,827,080	90			9,484	2,388

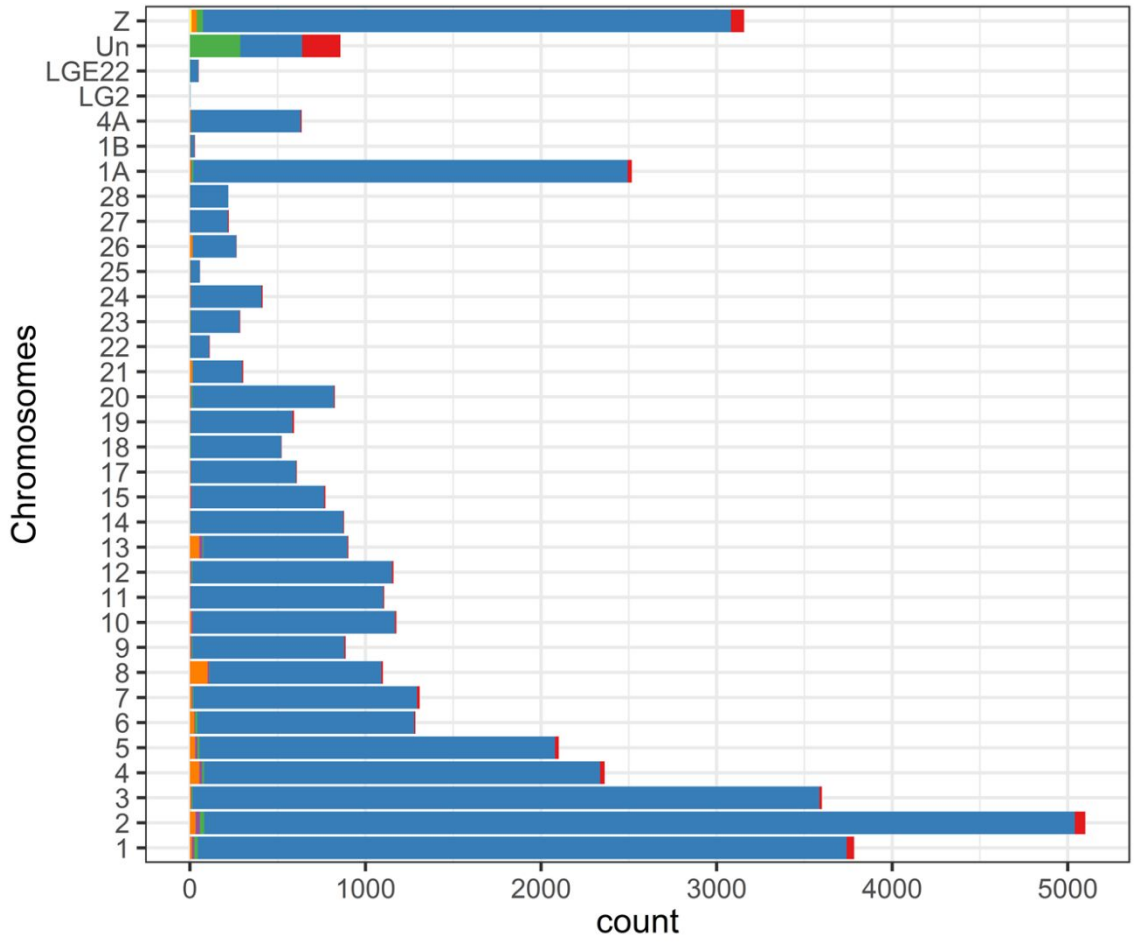
Table 2: SNP probe performance on Affymetrix SNP array. Absolute numbers for the performance of the SNPs from each assembly on the Affymetrix SNP array. Each assembly had some overlapped SNPs shared with single assemblies and also some that were shared with pooled assemblies. A large number of the total SNPs were only found in one assembly (i.e. Unique), either single or pooled.

	Total			Unique			Shared with an assembly from a single individual's data			Shared with an assembly from pooled data		
Assembly	**Successful**	**Fail**	**Conversion rate (%)**	**Successful**	**Fail**	**Conversion rate (%)**	**Successful**	**Fail**	**Conversion rate (%)**	**Successful**	**Fail**	**Conversion rate (%)**
1	4,753	2,492	66	3,187	1,960	62	1,399	486	74	714	241	75
2	3,562	903	80	2,516	556	82	982	325	75	442	114	79
3	2,276	811	74	1,773	564	76	476	240	66	191	62	75
4	4,447	3,092	59	3,096	2,576	55	1,238	480	72	605	213	74
5	4,629	3,036	60	3,187	2,502	56	1,325	495	73	671	230	74
6	5,125	2,507	67	3,323	1,940	63	1,514	503	75	936	285	77
7	4,894	1,198	80	3,313	679	83	1,380	478	74	735	226	76
8	3,943	979	80	2,755	536	84	1,123	421	73	527	175	75
9	4,910	1,209	80	3,285	656	83	1,419	494	74	811	281	74
10	5,049	1,174	81	3,149	596	84	1,540	499	76	1,042	334	76
3in1	5,447	1,274	81	2,493	487	84	2,142	625	77	5,447	1,274	81
10in1	5,145	1,222	81	2,433	488	83	1,900	572	77	5,145	1,222	81



942

943 **Figure 1. Density of the hihi SNP array SNPs across the zebra finch genome.** Density is
944 shown as the total number of SNPs per Mb window for 54,341 SNP markers included on the
945 hihi SNP array, after excluding those that mapped to random chromosomes, chromosome Un,
946 or did not map to the zebra finch genome.



947

948 **Figure 2: Verifying WGS SNP probe positions on the zebra finch genome.** Homologous
949 SNP positions were initially estimated from their position within their hihi contig's best-hit to
950 the zebra finch genome. To ensure they were placed correctly, once the probes were designed,
951 they were re-aligned to the zebra finch using only the SNP and its flanking sequences (37–71
952 bases). The majority aligned where expected (blue), a small number were found on the
953 'random' part of the same chromosome (orange), on an alternative chromosome (red), or on
954 both the expected and random chromosomes (purple), or on both the expected and an
955 alternative chromosome (green).
956

957

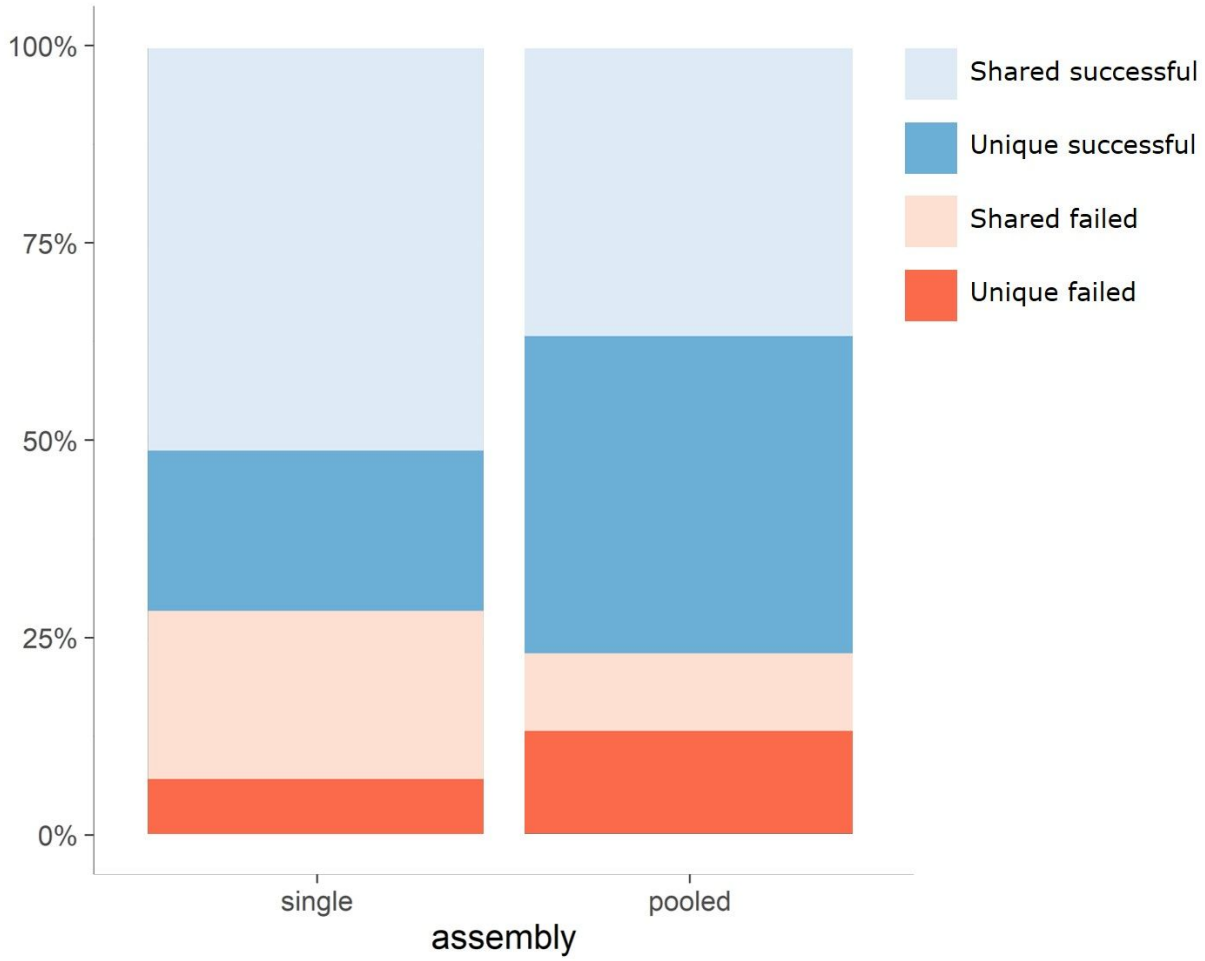
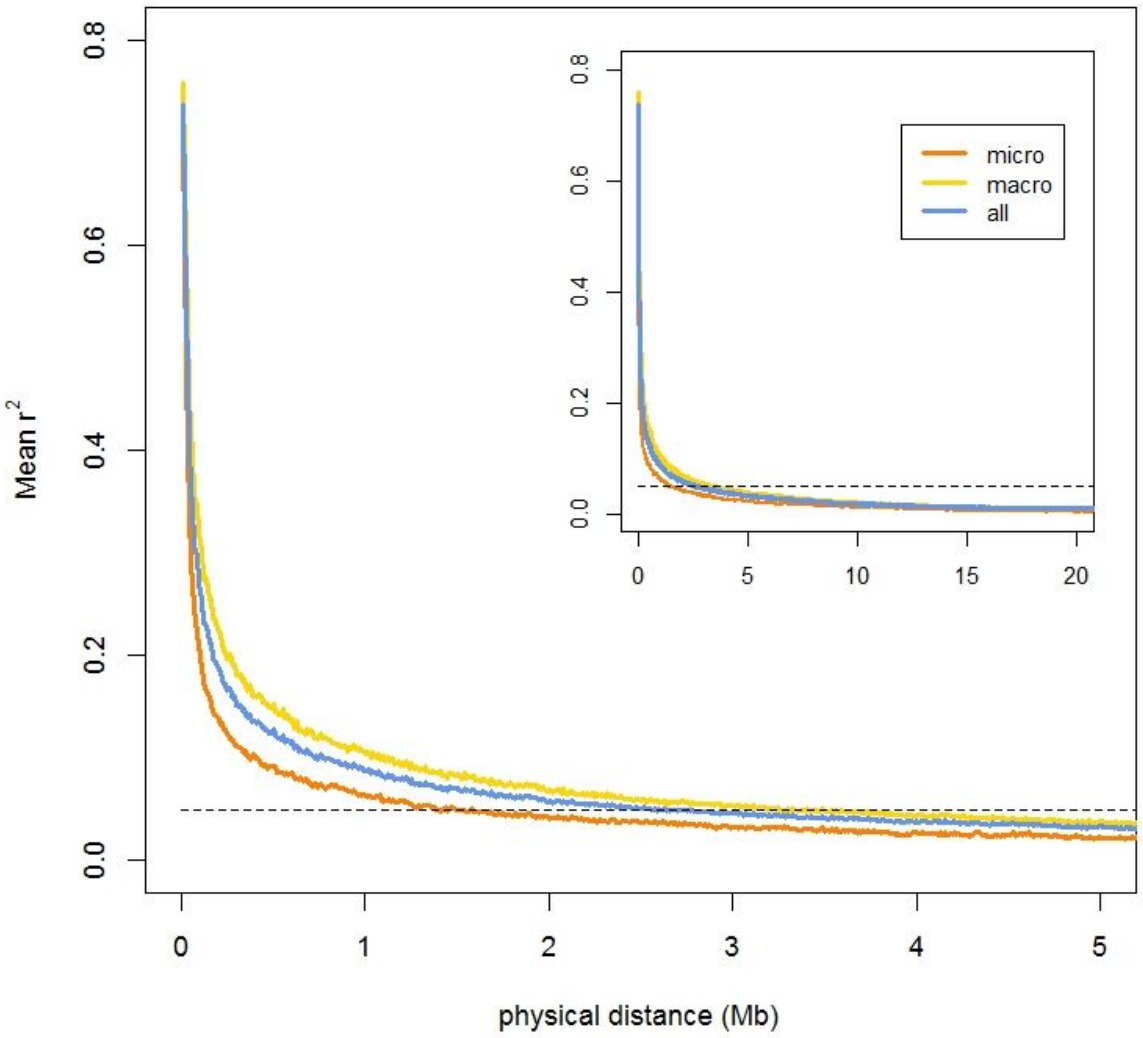


Figure 3: Averaged SNP performance for probes from single and pooled assemblies. The single column represents all array SNPs on all draft assemblies for each of the ten samples assembled separately. The pooled column represents all array SNPs from the assembly of pooled reads from samples 6, 9 and 10 (3 in 1) and pooled reads from all ten samples (10 in 1). The graph shows the proportion of SNPs that were found only in one assembly and successfully genotyped (light blue), SNPs that were also found in other assemblies that successfully genotyped (dark blue), SNPs that were found only in one assembly and failed genotyping (light orange), and SNPs that were also found in other assemblies that failed genotyping (dark orange).

969



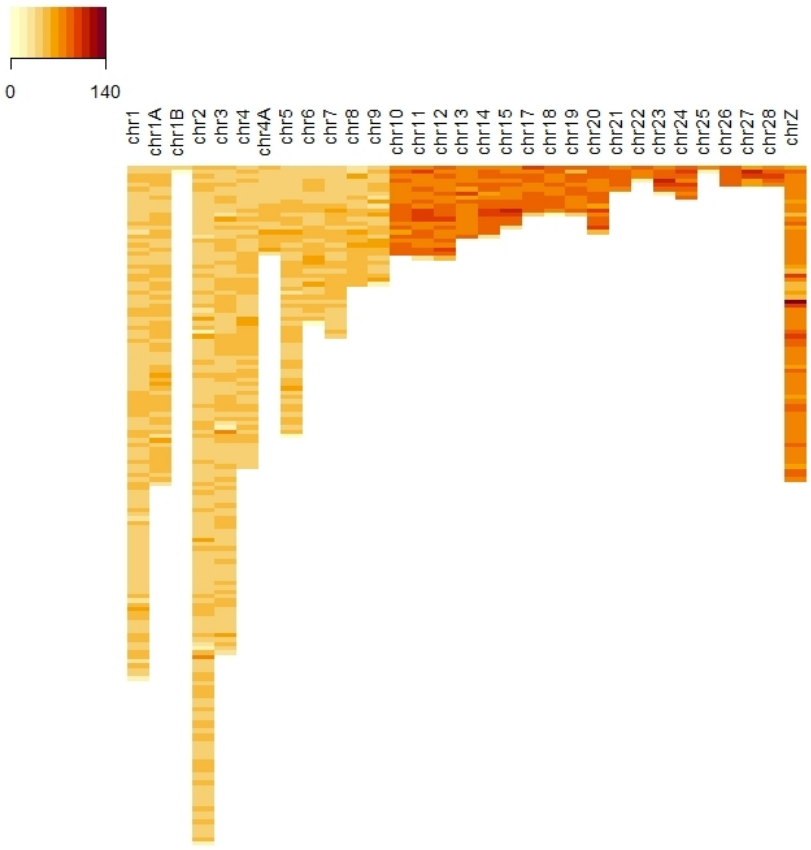
970

971

972 **Figure 4. Linkage disequilibrium in the hihi genome.** Decline of linkage disequilibrium,
973 measured as the correlation coefficient r^2, between pairs of SNPs for micro-chromosomes
974 (chromosome 10-15, 17-26, 1B, 4A, and LGE22), macro-chromosomes (chromosomes 1-9
975 and 1A) and all chromosomes, with physical distance based on alignment of SNPs to the
976 zebra finch genome. The main graph shows decay from 0 – 5 Mb between marker pairs, while
977 the inset zooms out to 0 – 20 Mb. Dotted horizontal lines correspond to an r^2 of 0.05. Linkage
978 disequilibrium was calculated after excluding highly related individuals; final input was 401
979 individuals genotyped at 13,126 micro-chromosome SNPs and 27,490 macro-chromosome
980 SNPs.

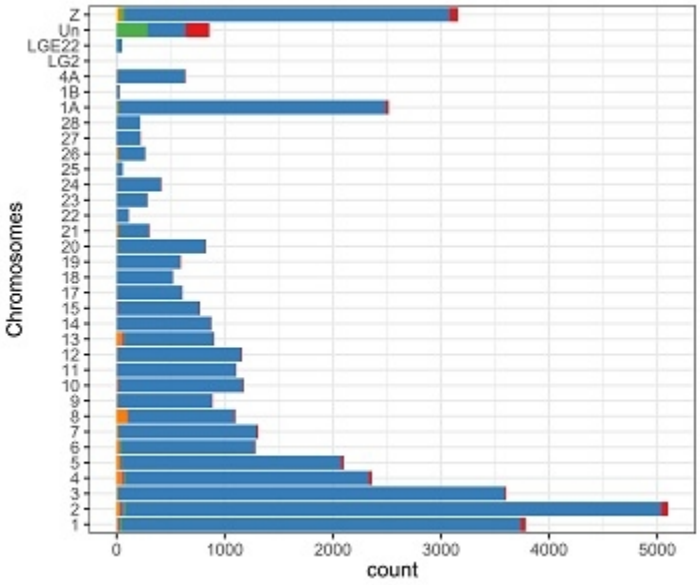
981

982



Density of the hihi SNP array SNPs across the zebra finch genome. Density is shown as the total number of SNPs per Mb window for 54,341 SNP markers included on the hihi SNP array, after excluding those that mapped to random chromosomes, chromosome Un, or did not map to the zebra finch genome.

271x271mm (72 x 72 DPI)



Verifying WGS SNP probe positions on the zebra finch genome. Homologous SNP positions were initially estimated from their position within their hihi contig's best-hit to the zebra finch genome. To ensure they were placed correctly, once the probes were designed, they were re-aligned to the zebra finch using only the SNP and its flanking sequences (37–71 bases). The majority aligned where expected (blue), a small number were found on the 'random' part of the same chromosome (orange), on an alternative chromosome (red), or on both the expected and random chromosomes (purple), or on both the expected and an alternative chromosome (green).

30x25mm (300 x 300 DPI)

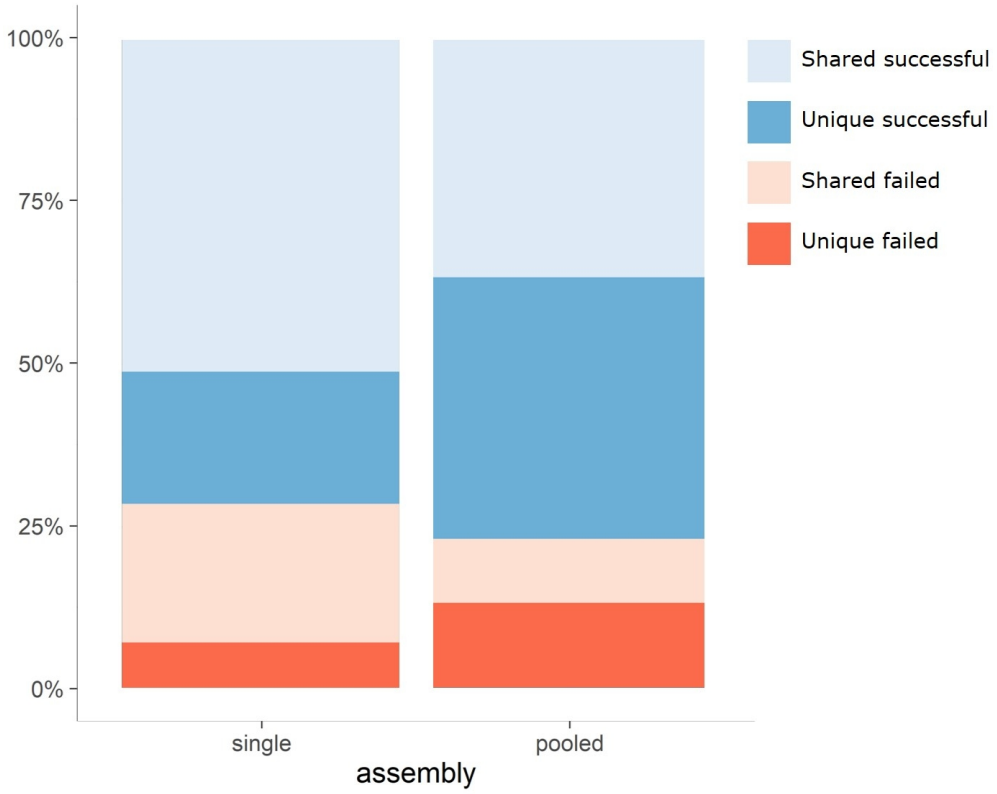


Figure 3: Averaged SNP performance for probes from single and pooled assemblies. The single column represents all array SNPs on all draft assemblies for each of the ten samples assembled separately. The pooled column represents all array SNPs from the assembly of pooled reads from samples 6, 9 and 10 (3 in 1) and pooled reads from all ten samples (10 in 1). The graph shows the proportion of SNPs that were found only in one assembly and successfully genotyped (light blue), SNPs that were also found in other assemblies that successfully genotyped (dark blue), SNPs that were found only in one assembly and failed genotyping (light orange), and SNPs that were also found in other assemblies that failed genotyping (dark orange).

235x190mm (150 x 150 DPI)

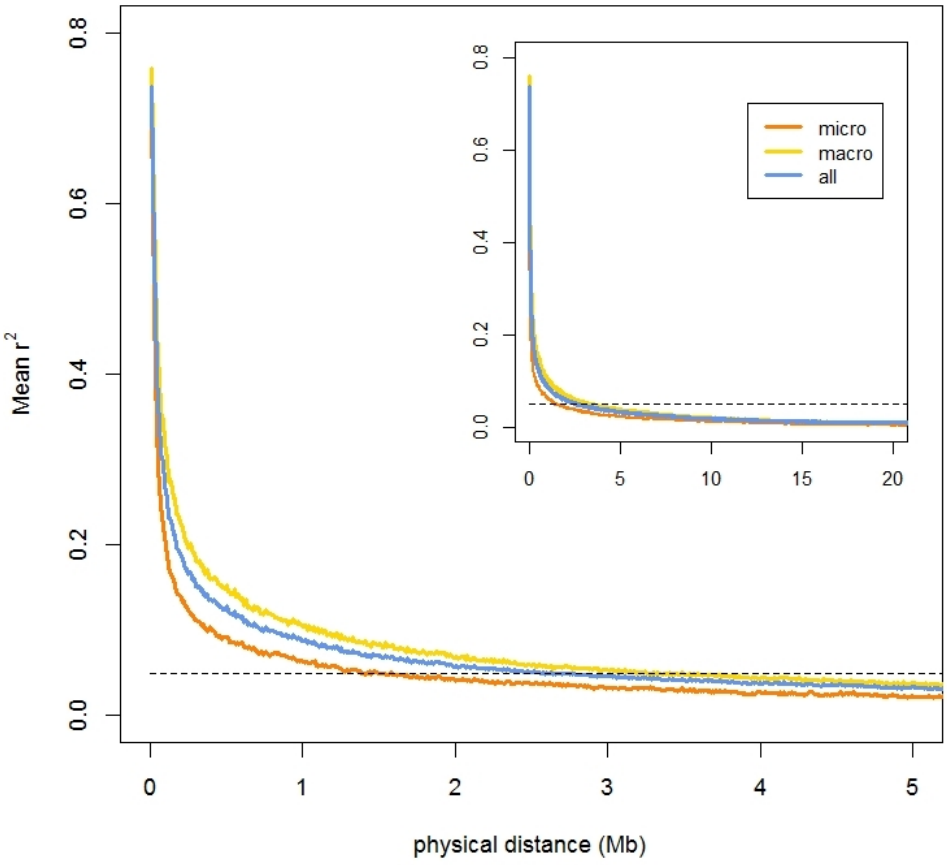


Figure 4. Linkage disequilibrium in the hihi genome. Decline of linkage disequilibrium, measured as the correlation coefficient r2, between pairs of SNPs for micro-chromosomes (chromosome 10-15, 17-26, 1B, 4A, and LGE22), macro-chromosomes (chromosomes 1-9 and 1A) and all chromosomes, with physical distance based on alignment of SNPs to the zebra finch genome. The main graph shows decay from 0 – 5 Mb between marker pairs, while the inset zooms out to 0 – 20 Mb. Dotted horizontal lines correspond to an r2 of 0.05. Linkage disequilibrium was calculated after excluding highly related individuals; final input was 401 individuals genotyped at 13,126 micro-chromosome SNPs and 27,490 macro-chromosome SNPs.

271x271mm (72 x 72 DPI)