Genetic variation associated with a geographical cline in New Zealand populations of Argentine Stem Weevil

Thomas W.R. Harrop[1], Marissa F. Le Lec[1], Ruy Juaregui[2], Rachael Ashby[3], Shannon Taylor[1], Sarah Inwood[1], Jeanne Jacobs[3], Stephen Goldson[4], Peter K. Dearden[1]

Goldson's dissection ppl?

[1] University of Otago
[2] AgResearch Palmerston North
[3] AgResearch Lincoln?
[4] BPRC

Abstract

The abstract should outline the purpose of the paper and the main results, conclusions and recommendations, using clear, factual, numbered statements

- context and need for the work
- approach and methods used
- main results (2-3 points)

Synthesis and applications

- wider implications and relevance to management or policy

Keywords

Naughty weevils, Invasive species, Molecular evolution, ???

Introduction

Materials and methods

Collections *etc.*

Weevils were collected from ...

Reduced-representation genome sequencing and processing

DNA was extracted ...

The code we used to process the genotyping data is hosted at github.com/TomHarrop/stacks-asw and github.com/MarissaLL/asw-para-matched.

Map was plotted with the ggmap package for ggplot2 [1].

Genome assembly

To produce the short read dataset, an Illumina TruSeq PCR-free 350bp insert library was generated from DNA extracted from a single, male Argentine stem weevil collected from endophyte-free hybrid ryegrass (*Lolium perenne* × *Lolium multiflorum*) at **Lincoln, New Zealand (?)**. Library preparation and sequencing were performed by Macrogen Inc. (Seoul, Republic of Korea). A total of 158 GB of 100 b and 150 b paired-end reads were generated from the TruSeq PCR-free library. After removing common sequencing contaminants and trimming adaptor sequences using BBTools [2], the short-read-only genome was assembled with meraculous [3–5]. Reproducible code for assembling the short-read dataset and assessing the assemblies is hosted at github.com/tomharrop/asw-nopcr.

To produce long reads from a single individual, we produced high molecular weight DNA from a single, male ASW collected from **where?** using a modified QIAGEN Genomic-tip 20/G extraction protocol [6]. We amplified the DNA using Φ29 multiple displacement amplification (QIAGEN REPLI-g Midi Kit) and debranched the amplified DNA using T7 Endonuclease I (New England Biolabs) according to the Oxford Nanopore Technologies Premium whole genome amplification protocol version WGA_kit9_v1. Amplified DNA was sequenced on **X** R9.4.1 flowcells using a **version** MinION sequencer (Oxford Nanopore Technologies). We also produced reads from high molecular weight DNA from a pool of 20 unsexed individuals collected from **where?**. We sequenced this DNA on **X** R9.4.1 flowcells, following the Genomic DNA by Ligation protocol (SQK-LSK109; Oxford Nanopore Technologies). We removed adaptor sequences from the long reads with Porechop 0.2.4 [7] and assembled with Flye 2.6 [8]. Reproducible code for assembling and assessing the long-read ASW genomes is hosted at github.com/TomHarrop/asw-flye-withpool.

All genome assemblies were assessed using assembly size and contiguity statistics and BUSCO analysis [9]. Assemblies that had a high rate of duplicated BUSCO genes were curated with Purge Haplotigs 0b9afdf [10] using a low, mid and high cutoff of 60, 120 and 190, respectively. We used RepeatModeler [11] and RepeatMasker [12] to estimate the repeat content of the long read genomes.

Genome-based analyses, F_{ST}, etc. etc.

- Catalog mapping *e.g.* `bwa mem`

Reproducibility and data availability

Raw sequence data for the ASW genome are hosted at the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) under accession **TBA**. We used `snakemake` [13] to arrange analysis steps into workflows and monitor dependencies, and `Singularity` [14] to capture the computing environment. Using the code repositories listed in each methods section, the final results can be reproduced from the raw

data with a single command using `snakemake` and `Singularity`. The source for this manuscript is hosted at github.com/TomHarrop/asw-gbs-genome-paper.

Results

Variation in NZ populations of Argentine stem weevil

To measure the variation in NZ populations of ASW, we collected individuals from 7 sites in the North Island and 5 sites in the South Island of New Zealand (Figure 1A). We genotyped each individual separately using a modified genotyping-by-sequencing (GBS) protocol (**Are we calling it gbs? Ref for the protocol used by AgResearch?**; [15]). We found lots of variation.

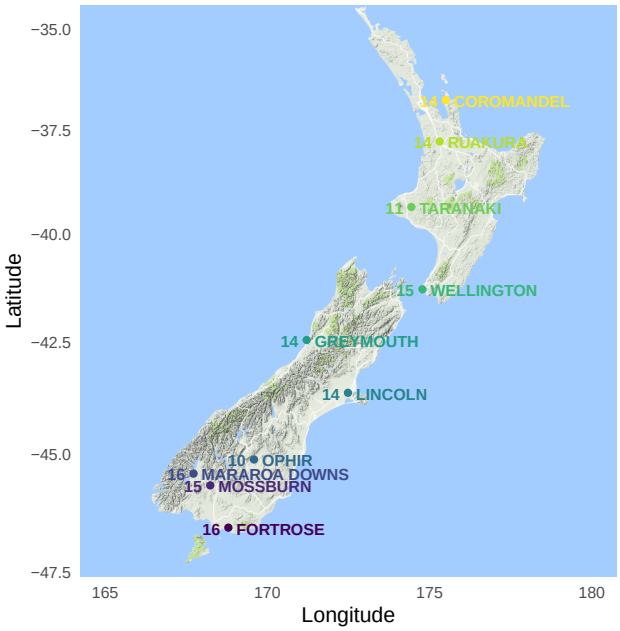


Figure 1A. Weevil sampling locations. We collected Argentine stem weevils from 4 locations in the North Island and 7 locations in the South Island of New Zealand. The number of weevils genotyped from each location is show on the map.

The Argentine stem weevil genome

To find genomic loci associated with between-population variation, we constructed a draft assembly of the ASW genome. We initially attempted assembly from a single individual using PCR-free, short read sequencing. This resulted in a fragmented assembly with low BUSCO scores (Table 1). *k*-mer analysis on the raw reads suggested genomic repeat content of at least 28% and 2.1% heterozygosity (**Supporting Information**). We then attempted to produce a long-read genome assembly using whole-genome amplification (WGA) of high molecular weight (HMW) DNA from a single individual, followed by sequencing on the Oxford Nanopore Technologies (ONT) MinION sequencer. We produced 29.8 GB of quality-filtered reads with an N_{50} length of 9.0 KB. The low read N_{50} length was caused by debranching of the amplified DNA by T7 Endonuclease I, which is necessary following multiple displacement amplification (see methods). Assembling the single individual, long read genome resulted in improved contiguity and BUSCO scores (Table 1). Consistent with the raw short read data, we detected an **extreme level (how much?)** of repeats in the single individual, long read genome (Table 1). To improve assembly across long repeats, we produced a second ONT dataset with longer reads from HMW DNA from a two pools of 20 individuals each. Sequencing these samples on the MinION sequencer produced a total of 12.0 GB of quality-filtered reads with an N_{50} length of 19.5 KB. **For completeness, assemble the pooled genome alone?**. We constructed a combined, long-read genome using the pooled, long-read dataset for contig construction, and the single-individual, long-read dataset for assembly polishing. This resulted in a more contiguous assembly, but a large number of redundant contigs (Table 1), presumably because of the high rate of heterozygosity in the pooled, long-read dataset. We produced a final draft assembly of 1.1 GB (Table 1) by using the PCR-free, short

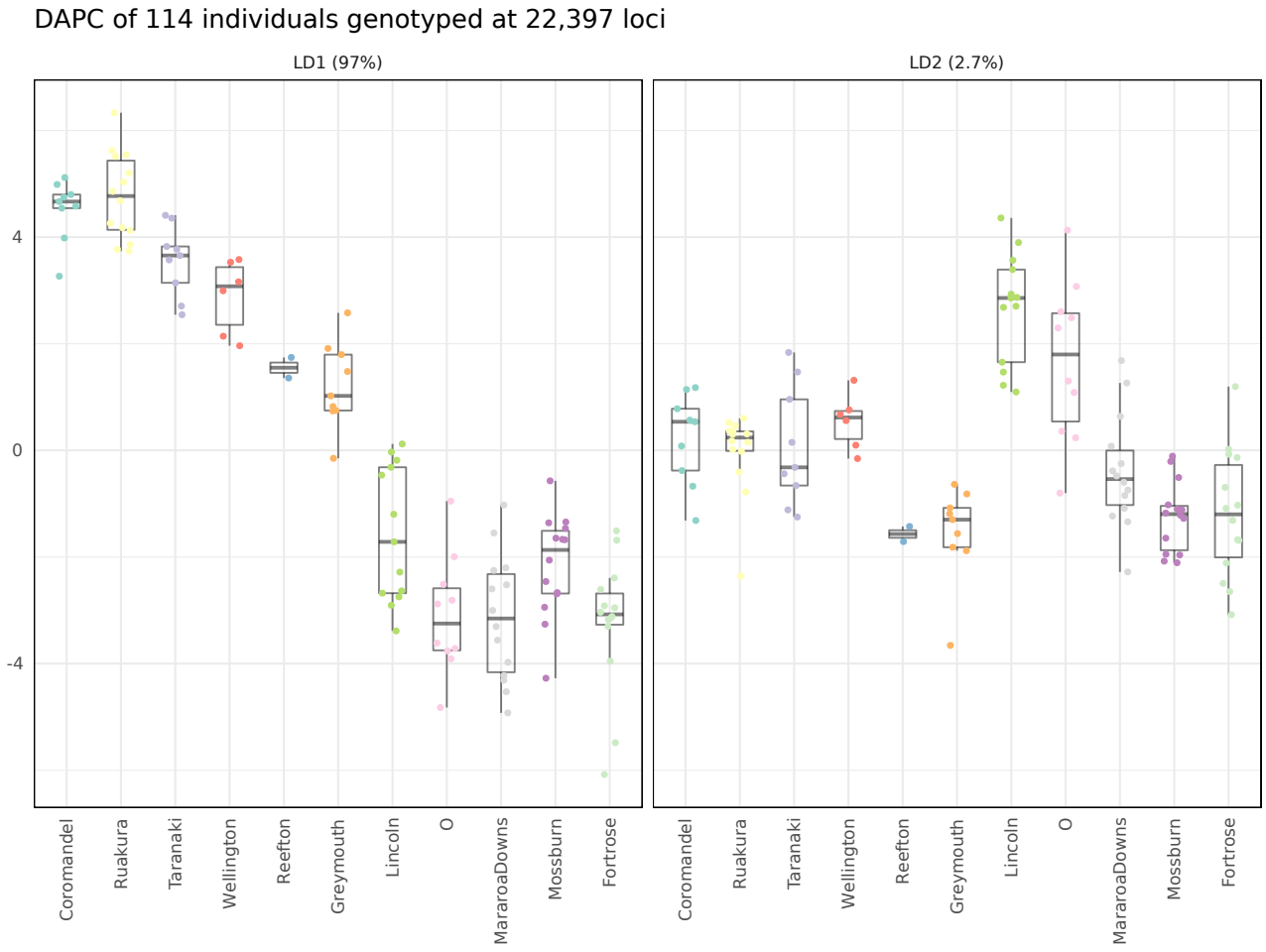


Figure 1B. B. Pricipal components analysis showing first two principal components. C. Some figure showing the high heterozygosity.

74 read sequencing data from a single individual with the purge_haplotigs pipeline to remove redundant contigs
75 from the combined long read assembly [10]. **Something about the repetitiveness**. We used our final
76 draft genome for all subsequent analyses.

Table 1. Assembly statistics for draft and intermediate assemblies.

	Short read	Single individual, long read	Pooled, long read	Combined, long read	Final draft
Assembly length (Gb)	1.3	x	x	x	1.1
N_{50}	53046	x	x	x	2681
N_{50} length (kb)	7.1	x	x	x	122.3
Complete single-copy BUSCOs (%)	32.7	72.2	x	69.2	78.8
Complete multiple-copy BUSCOs (%)	17.2	7.5	x	17.4	5.1
Repeat fraction	n.d.	x	x	x	x

Variation associates with a North-South cline

77 etc. etc.

Discussion

Authors' contributions

Acknowledgements

Data availability

References

1. Kahle, D.; Wickham, H. Ggmap: Spatial Visualization with ggplot2. *The R Journal* **2013**, *5*, 144. doi: 10.32614/RJ-2013-014.

2. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*; Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), 2014;Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

3. Chapman, J.A.; Ho, I.; Sunkara, S.; Luo, S.; Schroth, G.P.; Rokhsar, D.S. Meraculous: De Novo Genome Assembly with Short Paired-End Reads. *PLoS ONE* **2011**, *6*. doi: 10.1371/journal.pone.0023501.

4. Chapman, J.A.; Ho, I.Y.; Goltsman, E.; Rokhsar, D.S. Meraculous2: Fast accurate short-read assembly of large polymorphic genomes. *arXiv:1608.01031 [cs, q-bio]* **2016**. Retrieved from http://arxiv.org/abs/1608.01031.

5. Goltsman, E.; Ho, I.; Rokhsar, D. Meraculous-2D: Haplotype-sensitive Assembly of Highly Heterozygous genomes. *arXiv:1703.09852 [q-bio]* **2017**. Retrieved from http://arxiv.org/abs/1703.09852.

6. Harrop, T. HMW DNA extraction for insects. **2018**. doi: 10.17504/protocols.io.pnwdmfe.

7. Wick, R. Rrwick/Porechop 2020.

8. Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P.A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **2019**, 1. doi: 10.1038/s41587-019-0072-8.

9. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. doi: 10.1093/bioinformatics/btv351.

10. Roach, M.J.; Schmidt, S.A.; Borneman, A.R. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **2018**, *19*, 460. doi: 10.1186/s12859-018-2485-7.

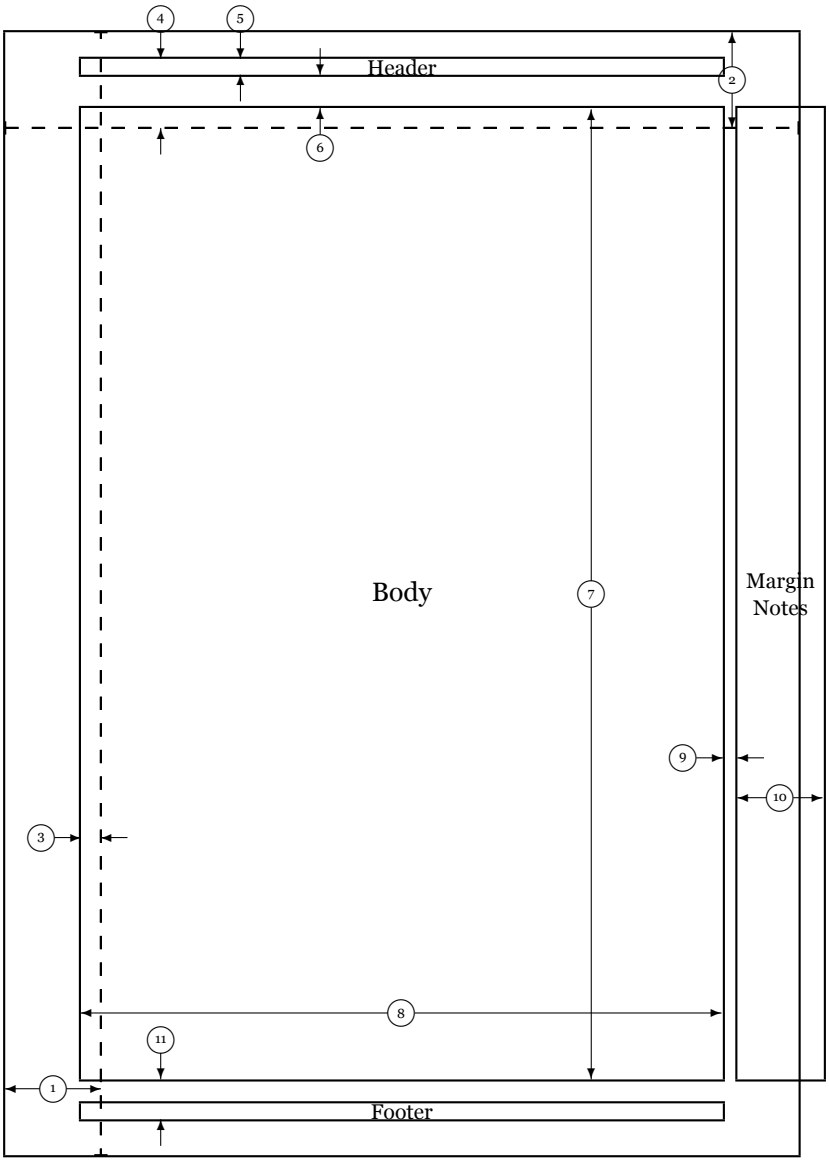
11. Smit, A.F.A.; Hubley, R. RepeatModeler Open-1.0 2015.

12. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker Open-4.0. 2015.

13. Köster, J.; Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **2012**, *28*, 2520–2522. doi: 10.1093/bioinformatics/bts480.

14. Kurtzer, G.M.; Sochat, V.; Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLOS ONE* **2017**, *12*, e0177459. doi: 10.1371/journal.pone.0177459.

15. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE* **2011**, *6*, e19379. doi: 10.1371/journal.pone.0019379.



107

1	one inch + \hoffset	2	one inch + \voffset
3	\oddsidemargin = -15pt	4	\topmargin = -52pt
5	\headheight = 12pt	6	\headsep = 25pt
7	\textheight = 731pt	8	\textwidth = 483pt
9	\marginparsep = 11pt	10	\marginparwidth = 65pt
11	\footskip = 30pt		\marginparpush = 5pt (not shown)
	\hoffset = 0pt		\voffset = 0pt
	\paperwidth = 597pt		\paperheight = 845pt