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

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

Alphabetical for now! Sample collectors, goldson's dissection crew?

[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].

- WGA of single indiv
- ONT stuff
- Assembly tricks

Genome assemblies were assessed using assembly size and contiguity statistics and BUSCO analysis [6]. We used RepeatModeler [7] and RepeatMasker [8] to estimate the repeat content of the long read genomes.

The code we used to assemble and assess the ASW genome is hosted at github.com/TomHarrop/asw-flye-withpool.

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` [9] to arrange analysis steps into workflows and monitor dependencies, and `Singularity` [10] 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 amodified ddRADseq protocol (**Is there an additional ref for the protocol used by AgResearch?**; [11]). We found lots of sweet variation.

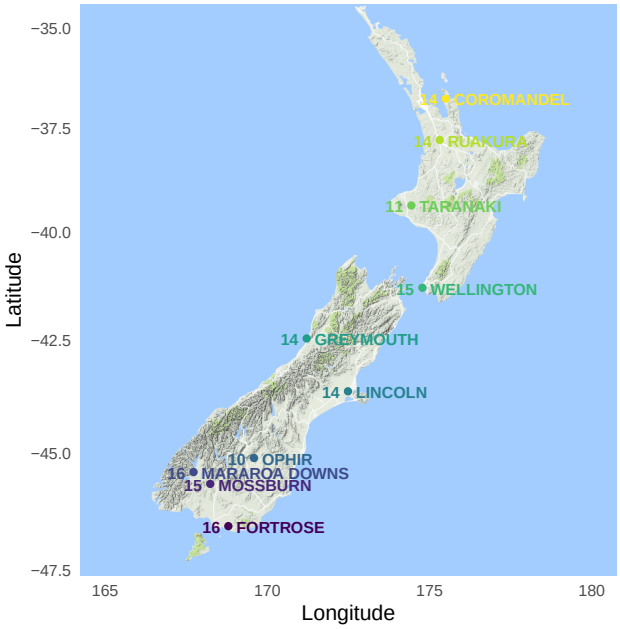


Figure 1A. Weevil sampling locations. We collected Argentine stem weevils from 4 locations in the North Island and 7 locations in the South Island of New Zealand. The number of weevils genotyped from each location is show on the map.

The Argentine stem weevil genome

To determine if between-population variation was related to selection at defined loci, 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). Because of the high heterozygosity in the single-individual short-read library (**Supporting Information**), we 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 is caused by branching of the genomic DNA during WGA by Φ29 DNA polymerase [**ref?**]. Assembling the single individual, long read genome resulted in improved contiguity and BUSCO scores (Table 1). We detected an extreme level of repeats in the single individual, long read genome (Table 1). To improve assembly of long repeat regions, 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 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 read sequencing data from a single individual with the purge_haplotigs pipeline to remove redundant contigs from the combined long read assembly [12]. **Something about the**

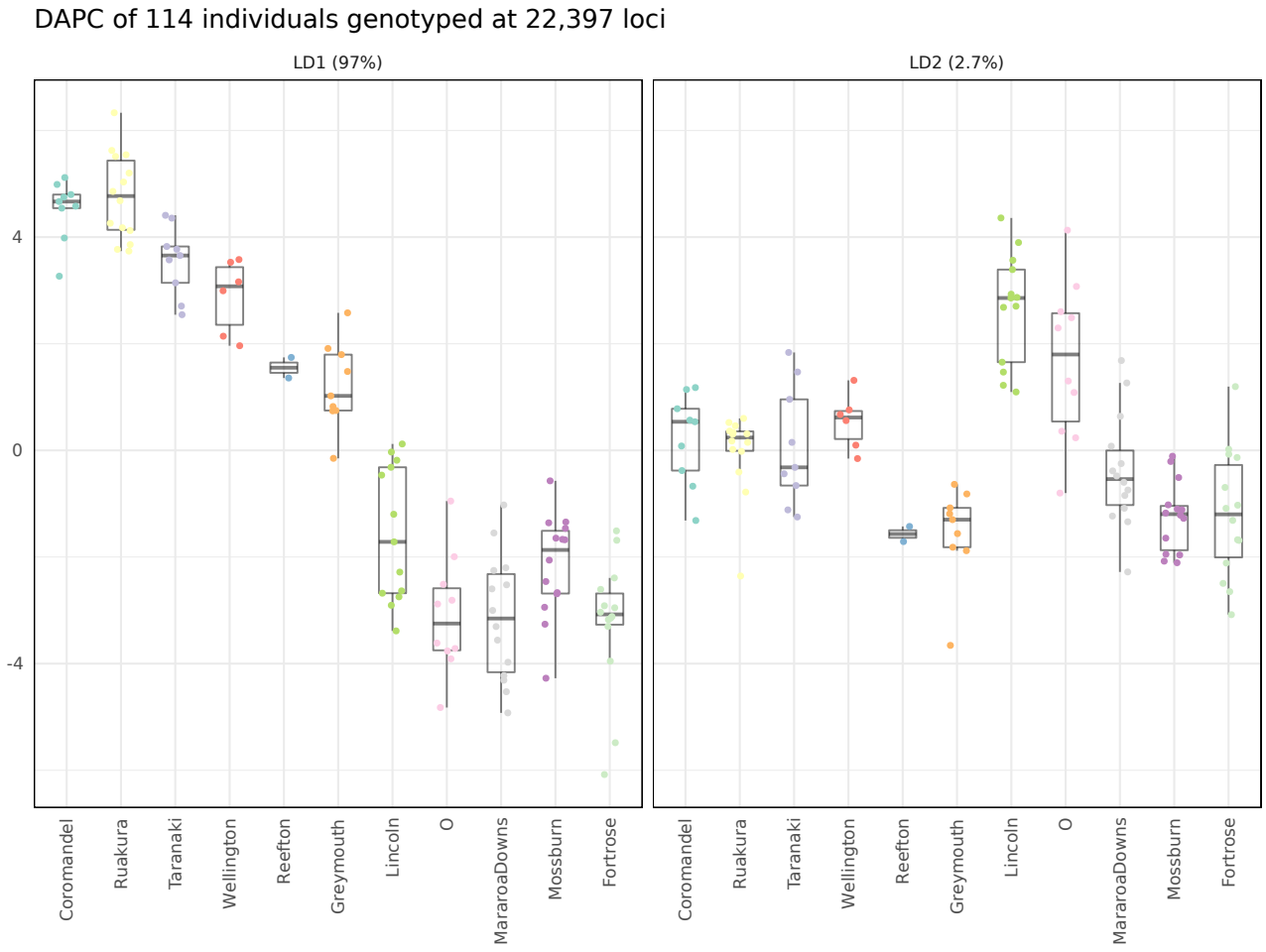


Figure 1B. A. Argentine stem weevil sampling locations. B. Pricipal components analysis showing first two principal components. C. Some figure showing the high heterozygosity.

71 **repetitiveness**. We used our final 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

72 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.

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

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*.

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**.

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

6. 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.

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

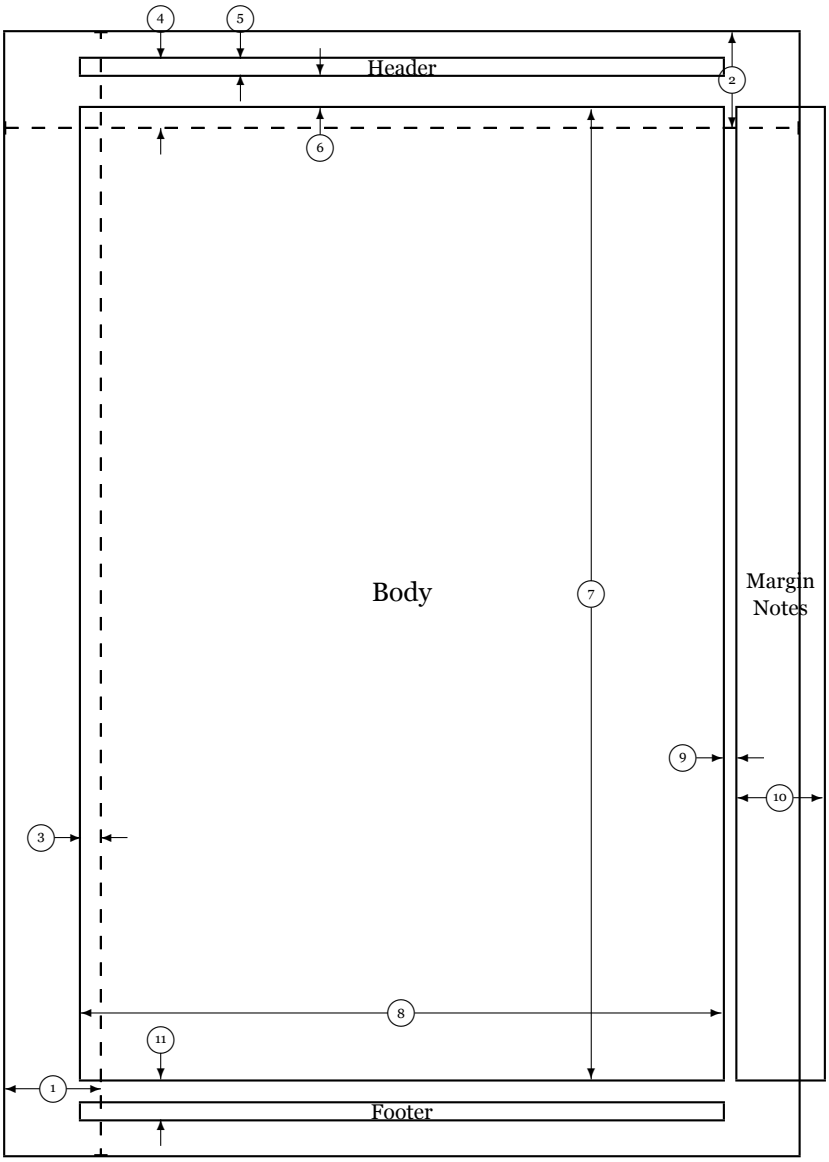
8. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker Open-4.0. 2015.

9. Köster, J.; Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **2012**, *28*, 2520–2522.

10. Kurtzer, G.M.; Sochat, V.; Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLOS ONE* **2017**, *12*, e0177459.

11. 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.

12. 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.



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