

# Collectively Improving Robustness and Reproducibility in Genotyping-by-Sequencing (GBS) Experiments

Marcus Davy<sup>1</sup>, Cecilia H. Deng<sup>2</sup>, Helge Dzierzon<sup>3</sup>, Robert J Elshire<sup>4</sup>, Stephanie Galla<sup>5</sup>, Elena Hilario<sup>2</sup>, Robyn Johnston<sup>4</sup>, Tammy Steeves<sup>5</sup>, Roy Storey<sup>1</sup>, and Susan Thomson<sup>6</sup>

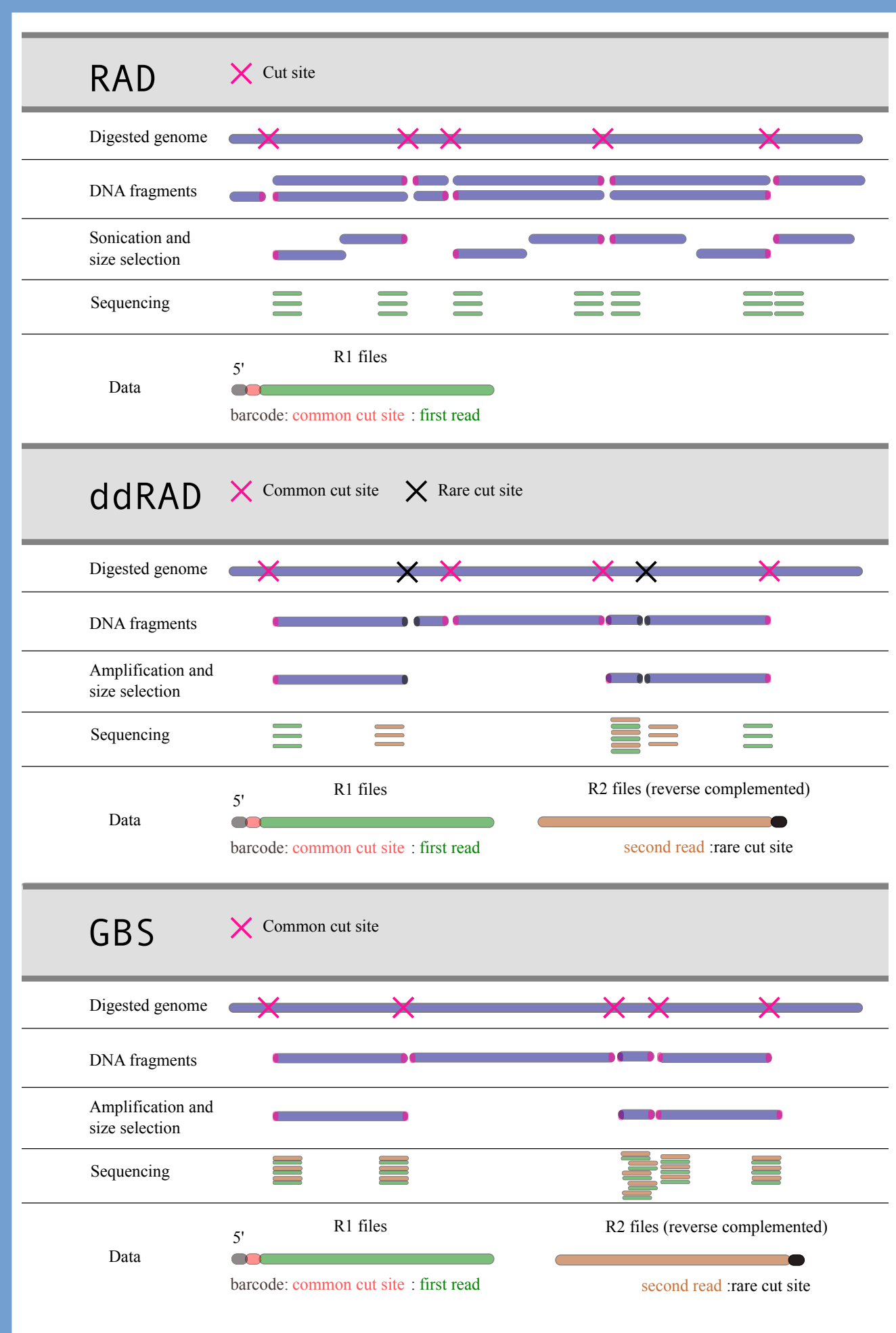
<sup>1</sup>The New Zealand Institute for Plant & Food Research Limited, Te Puke, New Zealand, <sup>2</sup>The New Zealand Institute for Plant & Food Research Limited, Auckland, New Zealand, <sup>3</sup>The New Zealand Institute for Plant & Food Research Limited, Palmerston North, New Zealand, <sup>4</sup>The Elshire Group Limited, Palmerston North, New Zealand, <sup>5</sup>University of Canterbury, Christchurch, New Zealand, <sup>6</sup>The New Zealand Institute for Plant & Food Research Limited, Lincoln, New Zealand

# Overview

Many researchers have used genotyping-by-sequencing to generate marker data since the method was published in PLoS ONE in 2011. Others have made modifications to the method resulting in different, but related, types of GBS data generated. Additional analysis pipelines have been developed, many of which are licensed under Free / Libre and Open Source licenses -- allowing them to be inspected, tested and improved upon. Taken together these modifications and analysis pipelines demonstrate the power of the scientific method when combined with modern genomics laboratory techniques, open access publishing and open source software to rapidly and democratically advance our ability to conduct research. It has also left researchers wanting to use the technology with several questions which need to be answered before they adopt the approach. First, what laboratory method should one use to make GBS libraries? Second, what sequencing platform should be used and how? Third, what pipeline is best suited for the data generated and the genetics of the system being studied?

A group of like-minded scientists has come together to build software tools and an information repository to help others answer these questions for their own experiments in a project called Biospectra-by-Sequencing. The aim of the project is developing an informational Wiki for geneticists and genomicists as well as a containerized software test rig to evaluate and run many GBS pipelines.

# Wet Lab



Data Types image courtesy of Dean Eaton (CC-BY)

## Representative Methods

Author	DOI	Year	Enzyme Number	Size Selection	Selective Bases	Enzyme Type	DNA source
Elshire et. al.	<a href="https://doi.org/10.1371/journal.pone.0019379">10.1371/journal.pone.0019379</a>	2011	1	no	no	II	native
Poland et. al.	<a href="https://doi.org/10.1371/journal.pone.0032253">10.1371/journal.pone.0032253</a>	2012	2	no	no	II	native
Sonah et. al.	<a href="https://doi.org/10.1371/journal.pone.0054603">10.1371/journal.pone.0054603</a>	2013	1	no	yes	II	native
Peterson et. al	<a href="https://doi.org/10.3390/d6040665">10.3390/d6040665</a>	2014	2	yes	no	II	native
Pan et. al.	<a href="https://doi.org/10.1111/1755-0998.12342">10.1111/1755-0998.12342</a>	2015	4	yes	no	II	native
Hilario et al.	<a href="https://doi.org/10.1371/journal.pone.0143193">10.1371/journal.pone.0143193</a>	2015	1	no	no	II	amplified

## Wiki Resources

[illegible]

### QC and troubleshooting quantification results

## The Standard Curve

The Poisson assay is linear over three orders of magnitude, up to  $10^6$  cells. Data assays (this) are thus able to estimate of sample concentrations up to  $10^6$  cells, with this product. The fit of the standard curve is of importance in accurate quantification. Keeping the sample count level to understand sample concentrations, while an overall standard curve will overestimate most of the samples. Assessing the fit of the standard curve is a critical quality control parameter.

### Assessing the fit

- Check the  $R^2$  value for the standard curve. This should be  $>0.99$
- Visually assess the fit. In particular, check whether the lowest and highest points on the curve appear to fit the linear model.

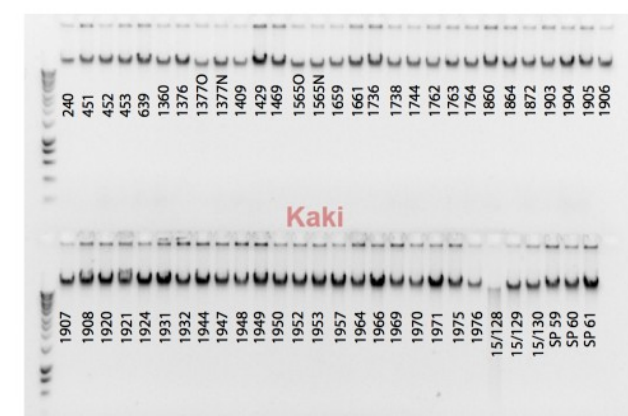
Good (left) and poor (right) standard curves. Note the low  $R^2$  value for the poor standard curve, and how the low counts and the  $10^6$  data point are outliers. The results are an early key factor which leads to overestimation of large sample concentrations and underestimation of small sample concentrations. The poor standard curve is a result of the low counts and the  $10^6$  data point being outliers.

### Guidelines for preparing gel images for DNA QC

Before proceeding with GBS library preparation, it is important to ensure that DNA samples are of high quality. All DNA samples should be run on a gel to ensure that they are not degraded. That degraded samples should also be confirmed to ensure that DNA samples can be digested successfully by restriction enzymes.


**Prepare gel images of all uncut DNAs**

- Prepare a 0.7% agarose gel.
- Be sure to load a DNA ladder.
- Load each individual DNA sample in a separate well.
- Load a total of 120 ng of DNA per well. For example, if the concentration of your DNA sample is 40 ng/μl, then load 3 μl.
- Run the gel using enough voltage to resolve.
- Photograph the gel.
- Label each lane in the photograph with the sample name.
- Understand that DNA samples that have a high molecular weight band (greater than 20 kb) are degraded DNA samples and are likely to appear as a smear. Sample 16/12 in the image below appears degraded. The experiment that it was used in failed.




*Himantopus novaezelandiae* / Kākī / Black Stilt

General Information	
Genome Size	~1.2G (estimated)
Proteome Level	30,000
Genetic Method	Ensemble GBS <sup>1</sup>
DTM	April
Mapping Length	14
Reference Genome	None
Analysis Software	TreeKID, LAMC
Software Version	0.1
Author Number	606
Mating Date	2014-2015
Protein Amount	20-30 mg/ml
Genomic DNA Amount	10 mg
Phenotypic Data	
Interdisciplinary	
Sample Information	Conservation/Breeding Programme



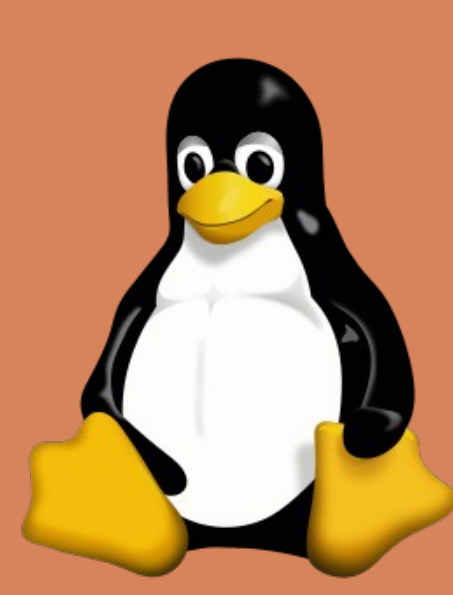
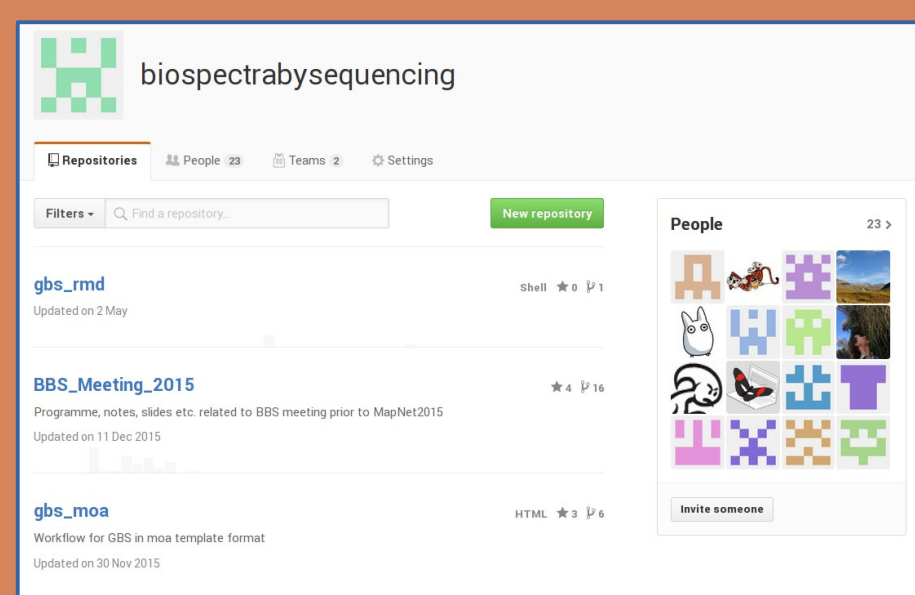
Don 206



Manhattan plot showing the results of a genome-wide association study (GWAS) for the trait 'Male / Female bias / Sex'. The y-axis represents the  $-\log_{10}(p\text{-value})$  and the x-axis represents the chromosome number (1 to 24). A significant peak is observed on chromosome 1, reaching a  $-\log_{10}(p\text{-value})$  of approximately 10. Other smaller peaks are visible on chromosomes 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24.

# Dry Lab

Pipeline Name`	DOI	Year Published	Denovo / Reference	Demultiplexer	Trimmer / Read QC Filter	Aligner	SNP Caller	Hardcoded enzyme	Output file types
TASSEL UNEAK	10.1371/journal.pgen.1003215	2013	Denovo	Built In	None	NA	Built In	Yes	hmp
STACKS	10.1111/mec.12354	2013	Denovo / RefereZnce	Built In	Built In	BWA / Bowtie2 / GSnap	Built In	Yes	Database with export
IGST-GBS	10.1371/journal.pone.0054603	2013	Reference	FastX toolkit	FastX toolkit	BWA	sam tools	Unknown	vcf
TASSLE 3	10.1371/journal.pone.0090346	2014	Reference	Built In	None	BWA / Bowtie 2	Built In	Yes	hmp
pyRAD	10.1093/bioinformatics/btu121	2014	Denovo	Built In	Built in	Muscle	Built In	No	Text (.loci, .phy, .nex, .snps, .vcf and others),
AfrRad	10.1111/1755-0998.12378	2015	Denovo	Built In	Built in	Mafft	Built In	No	Text (translation scripts)
GBS-SNP-CROP	10.1186/s12859-016-0879-y	2016	Denovo / Reference	Built In	Trimmomatic	BWA	sam tools	No	SNP matrix, TASSEL hmp, PLINK tped
GibPSS	10.1111/1755-0998.12510	2016	Denovo	Built In	Built in	Matching Built in	Built In	No	Database with export.
NGSEP	10.1186/s12864-016-2827-7	2016	Reference	Built In	Built in	Bowtie2	Built In	Unknown	vcf
FastGBS	10.1186/s12859-016-1431-9	2017	Reference	Sabre	Cutadapt	BWA	Platypus	Not used	vcf
ipyRAD	NA	NA	Denovo, Reference, Denovo+Reference	Built In	Built in	BWA / SMALT	sam tools / bed tools	No	Vcf + many others
TASSEL 5	NA	NA	Reference	Built In	Built In	BWA / Bowtie2	Built In	Yes	Database with export



# Collaboration

The Biospectra-by-Sequencing project welcomes participation by like minded scientist from around the world regardless of skill level. One way to participate is to use the resources we are developing and provide feedback so they can be improved. If you have conducted GBS experiments you can add information about those experiments to the wiki (<http://biospectrabysequencing.org>). By doing that, you will be helping create a resource for other scientists and yourself. For those with coding skills, we encourage you to assist with the pipeline testing rig. There is plenty to do there with the twelve pipelines in the table above and more likely to come along. Find us on github (<https://github.com/biospectrabysequencing>). For more information or general questions about BBS, email Rob at [Rob@ElshireGroup.co.nz](mailto:Rob@ElshireGroup.co.nz).

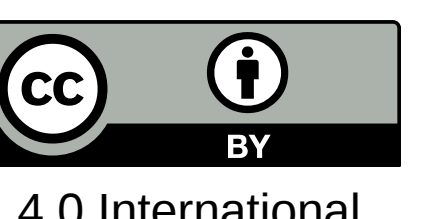


Plant & Food  
**RESEARCH**  
RANGAHALLI AHUMARA KAI



**The Elshire  
Group Ltd.**

**UC**  
**UNIVERSITY OF**  
**CANTERBURY**  
*Te Whare Wānanga o Waitaha*  
CHRISTCHURCH NEW ZEALAND



## 4.0 International