Developing a pipeline for Genotyping-by-sequencing (GBS) of the New Zealand Greenshell^{TM} Mussel

Rachael Ashby[1,2], Rudiger Brauning[1], Tracey van Stijn[1], Hayley Baird[1], John McEwan[1], Chris Brown[2], Neil Gemmell[2], and Shannon Clarke[1]

[1] AgResearch, Invermay

[2] University of Otago, Dunedin

agresearch

The New Zealand Greenshell$^{\text{TM}}$ Mussel

- Economically important species for the NZ aquaculture industry
- Exports of Greenshell™ Mussels had a revenue of $211m in 2011
- Current farming methods heavily rely on wild spat collection
- Lifecycle of the mussel has been established
- Cawthron now have a mussel hatchery and is successfully breeding mussels

The New Zealand Greenshell^{TM} Mussel

- Next stage is to selectively breed mussels
- Aim to generate a genomic toolbox for the Greenshell^{TM} Mussel
  - To further understand the genomics of the mussel
  - To aid mussel breeding
- Predicted difficulties:
  - Repeats
  - Heterogeneity e.g., Chilean Mussel 1:25bp SNP rate
- Closest sequenced genomes are the Pacific and Pearl Oysters

The New Zealand Greenshell^{TM} Mussel

- Transcriptome
  - RNA from mantle, foot, gill and adductor muscle
  - Manuscript in preparation
- V1 draft genome

Num. of Scaffolds	Length of Scaffolds (mb)	Min. Scaffold (bp)	Max. Scaffold (bp)	N50 (bp)	Average Length (bp)	Complete Genes (%)	Partial Genes (%)
332,002	1,159	500	165,912	7,018	3,492	39	77

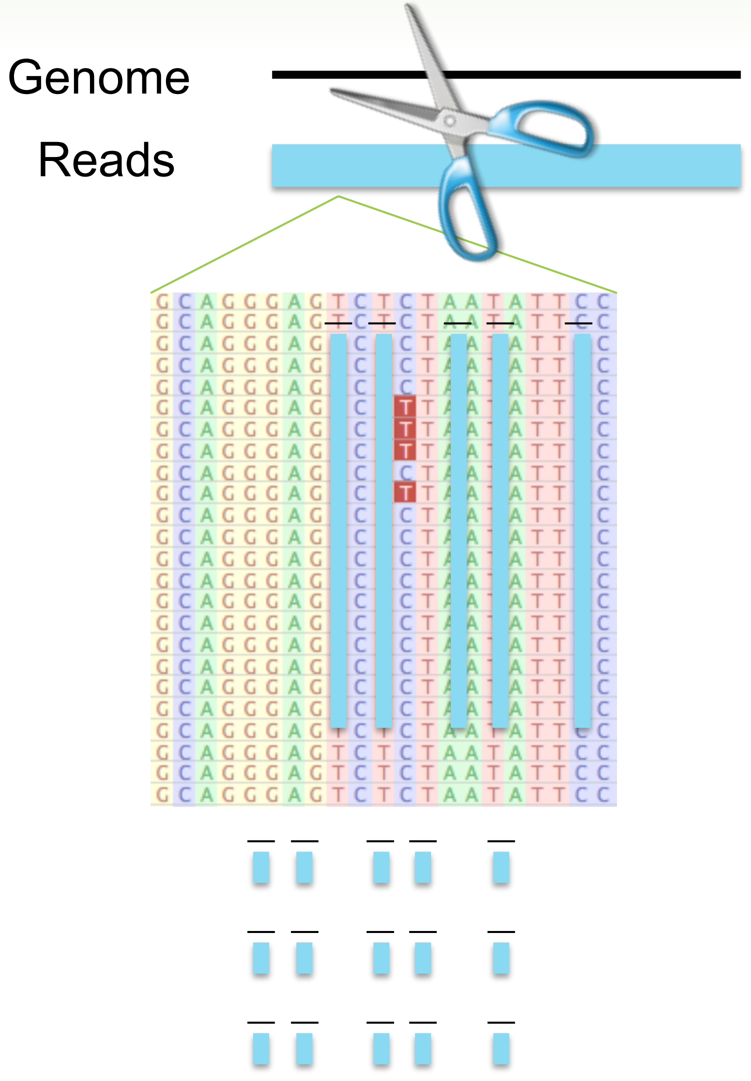
agresearch

Next Stage – Developing a GBS Pipeline

- Our aim is to have a high throughput, reproducible and cost effective GBS method.
- Haemolymph from breeding stocks used for sample collection
- Need high quality DNA

- Analysis
  - Genomic selection
  - Genome wide association studies
  - Parentage
  - Linkage disequilibrium

Genotyping-by-Sequencing (GBS)



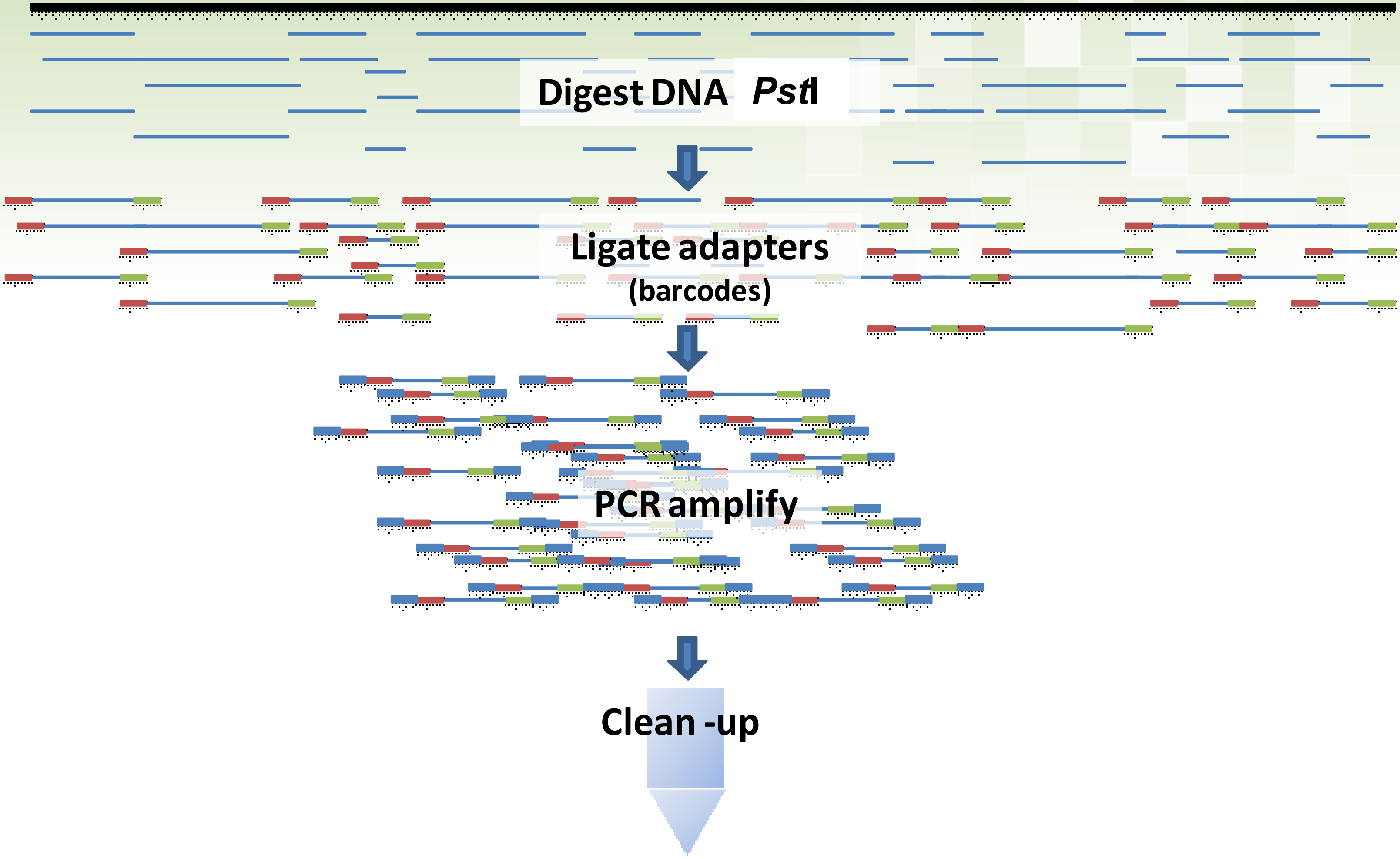
Genome	Individuals
100%	1
1% (*100)	1

1%

1%

1%

100

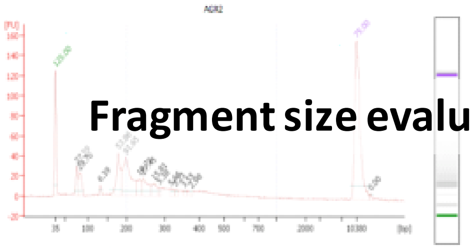


Digest DNA *PstI*

Ligate adapters
(barcodes)

PCR amplify

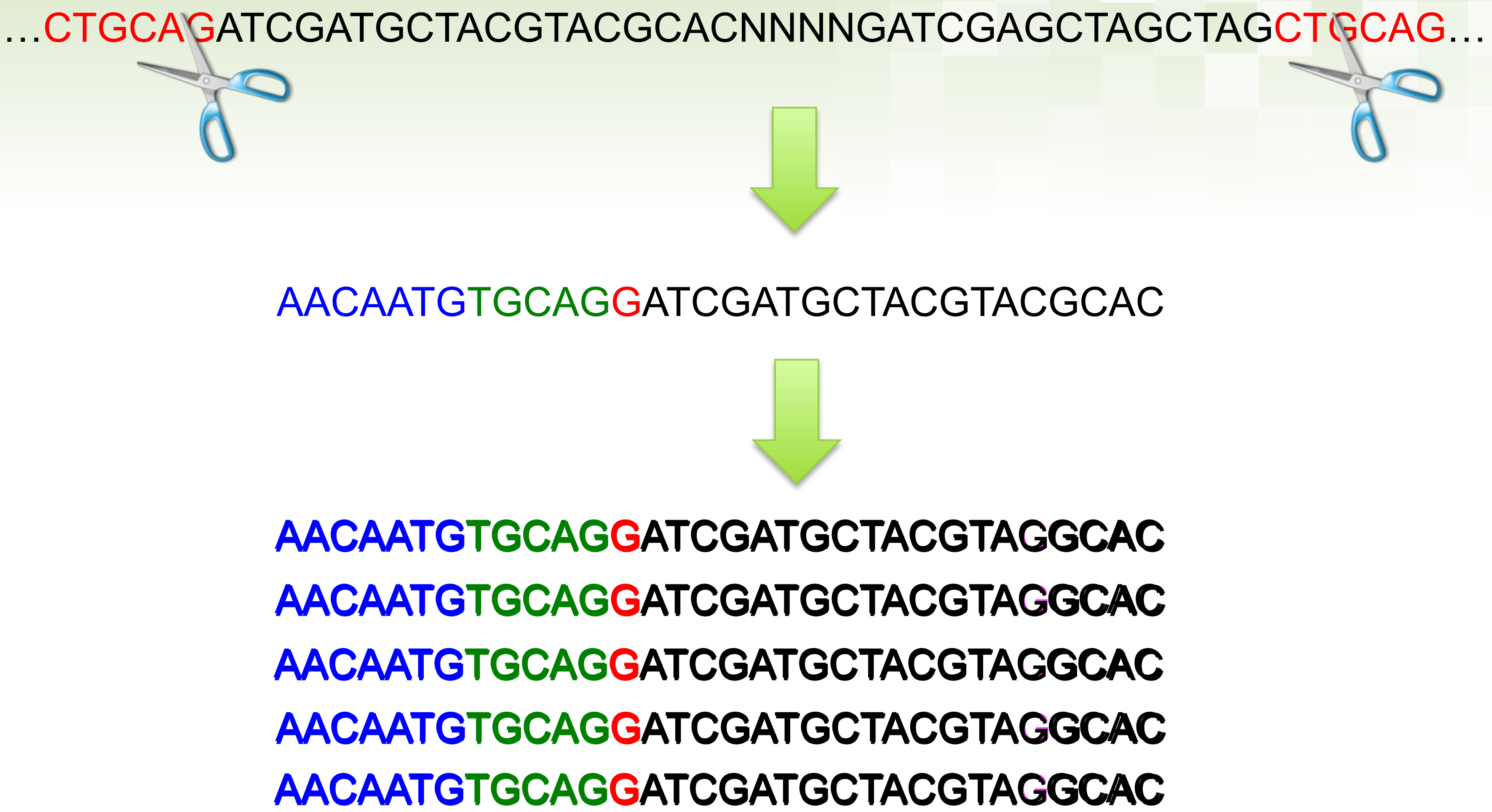
Clean -up



Fragment size evaluate and sequence

Sequencing by Synthesis
~1.5 to 11 days

agresearch



…CTGCAGATCGATGCTACGTACGCACNNNNGATCGAGCTAGCTAGCTGCAG…

AACAATGTGCAGGATCGATGCTACGTACGCAC

AACAATGTGCAGGATCGATGCTACGTAGGCAC

AACAATGTGCAGGATCGATGCTACGTAGGCAC

AACAATGTGCAGGATCGATGCTACGTAGGCAC

AACAATGTGCAGGATCGATGCTACGTAGGCAC

AACAATGTGCAGGATCGATGCTACGTAGGCAC

New species considerations

- Restriction enzyme
  - ApeKI generates larger coverage across genome
  - PstI reduces complexity generating higher depth of individual SNPs
  - PstI-MspI double digest, combines a rare cutsite (PstI) with a more common cutsite (MspI)
- Reference vs *de novo*?
  - Reference based protocols produce better quality SNPs
  - *De novo* faster and cheaper when a reference is unavailable
- 96 Mussels
  - 95 samples + positive control
  - Samples mix of parents, progeny and unrelated

agresearch

Results - UNEAK

	Number of SNPs	HWdgm > 0.05	Proportion of missing genotypes	Mean sample depth	Mean self-relatedness
ApeKI	43,097	42,435	0.53	2.54	1.36
PstI	7,812	7,496	0.53	21.41	1.49
PstI-MspI	30,068	29,629	0.56	7.89	1.43

Results – Tassel 5

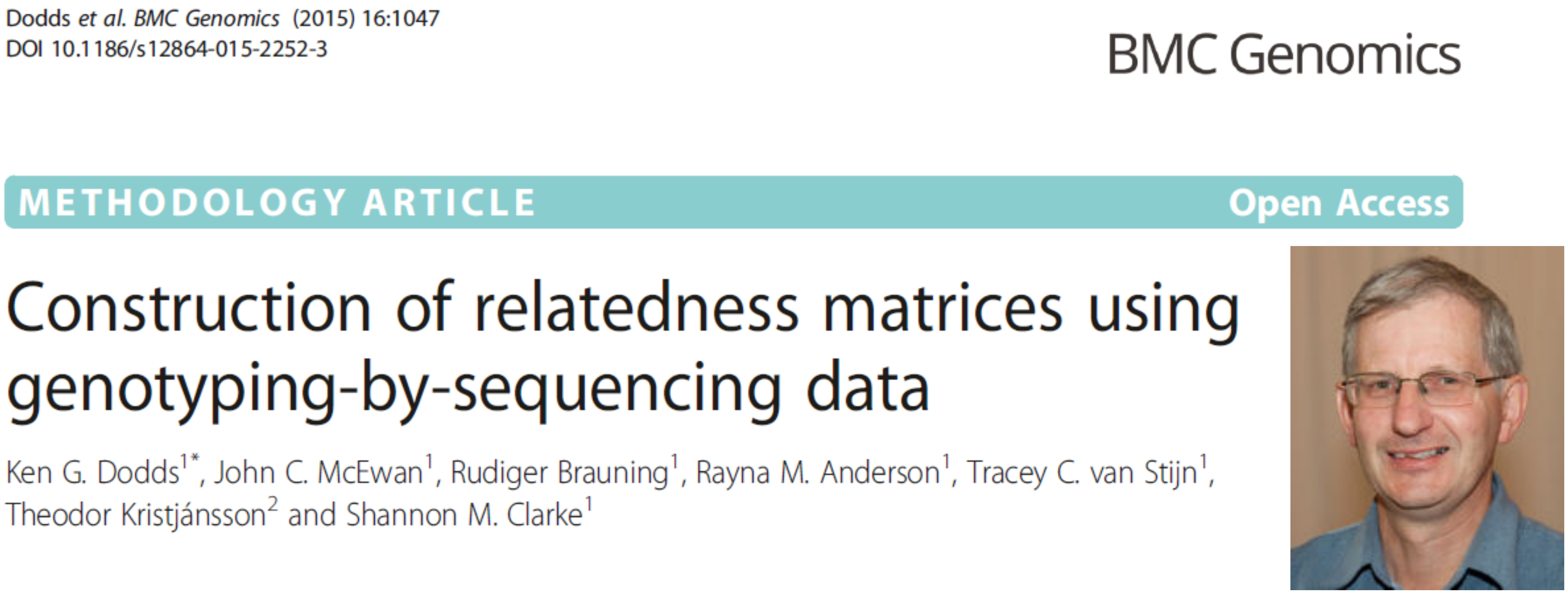
	Number of SNPs	HWdgm > 0.05	Proportion of missing genotypes	Mean sample depth	Mean self-relatednes s
ApeKI	35,953	33,603	0.39	8.05	1.48
PstI	14,085	13,158	0.36	51.5	2.25
PstI-MspI	19,592	18,633	0.38	19.05	1.75

Mussel GBS Issues

- UNEAK
  - Only looks for 1 SNP in a 64bp read
  - Under calls SNPs
  - 50% of Tags missing across individuals
- Tassel 5
  - Low mapping rates to reference genome
  - Low SNP calling rate
  - High percentage of tags missing across all samples
  - Large variation between the self-relatedness

Improving Mussel GBS SNP Calling

- Filter SNPs using KGD
- Improve reference genome
- Try other tools to improve the SNP calling
  - How do we determine which tool is the best?



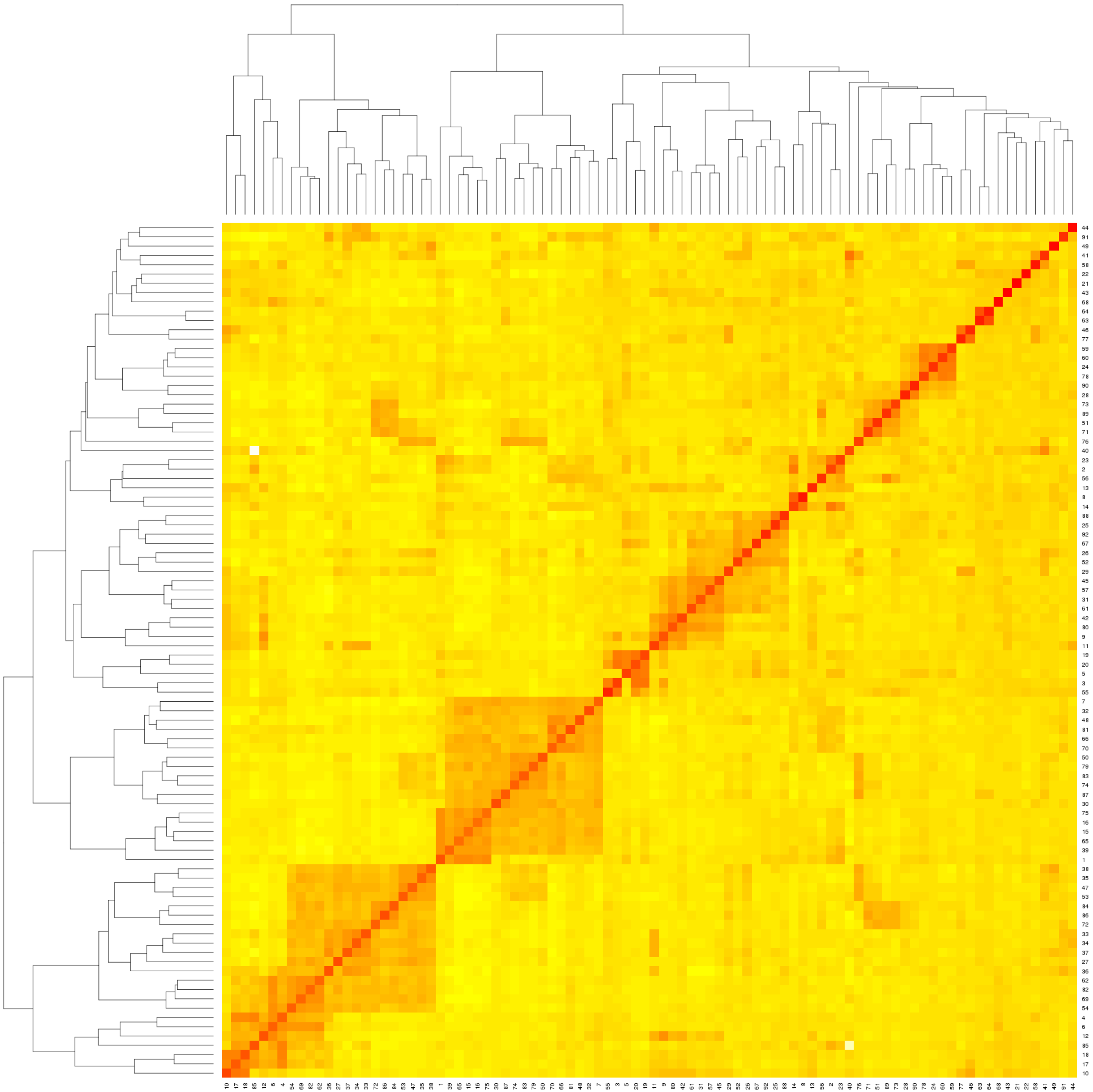
Dodds *et al. BMC Genomics* (2015) 16:1047
DOI 10.1186/s12864-015-2252-3

BMC Genomics

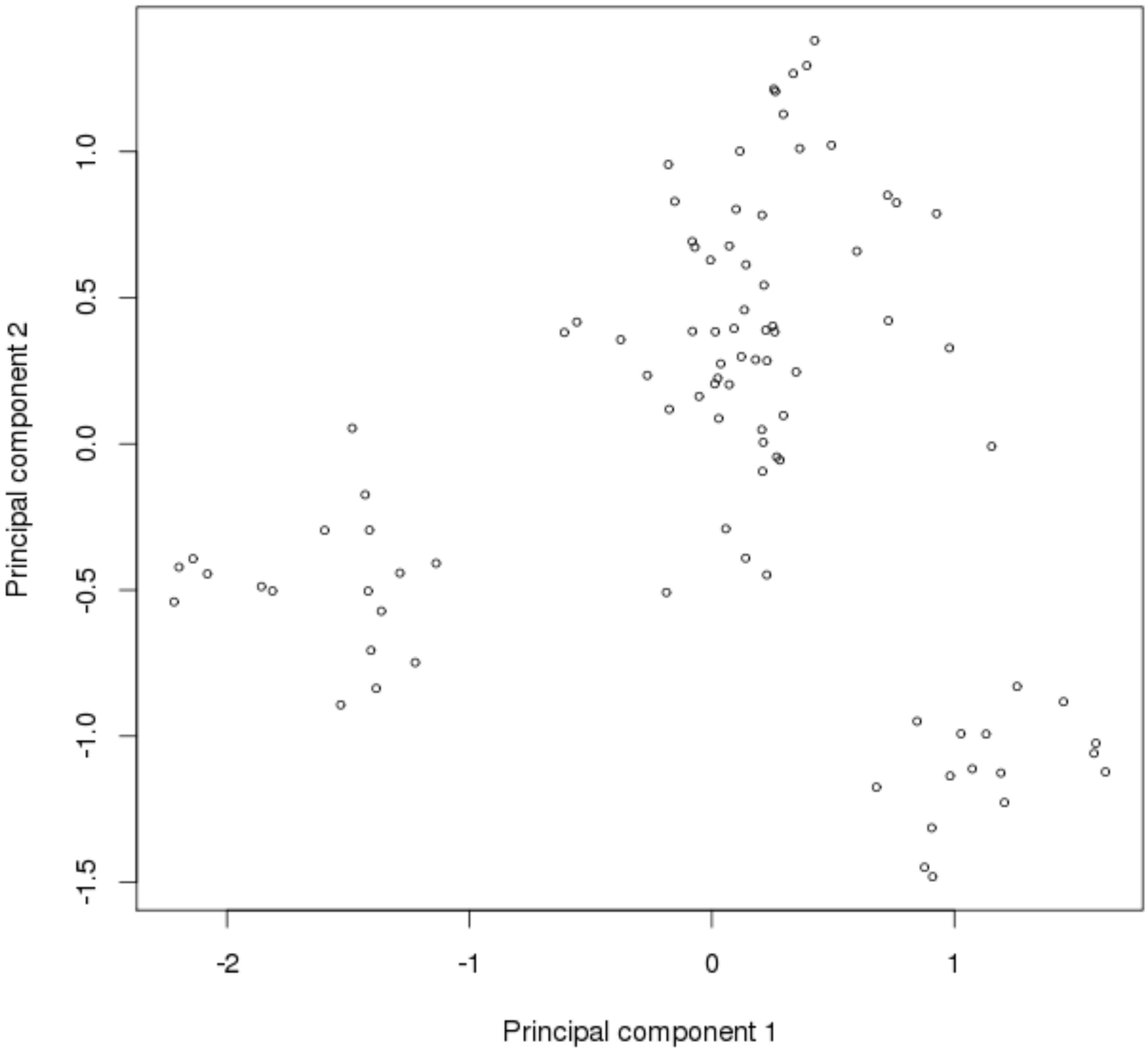
METHODOLOGY ARTICLE Open Access

Construction of relatedness matrices using genotyping-by-sequencing data

Ken G. Dodds[1*], John C. McEwan[1], Rudiger Brauning[1], Rayna M. Anderson[1], Tracey C. van Stijn[1], Theodor Kristjánsson[2] and Shannon M. Clarke[1]


44
91
49
41
58
22
21
43
68
64
63
46
77
59
60
24
78
90
28
73
89
51
71
76
40
23
2
56
13
8
14
88
25
92
67
26
52
29
45
57
31
61
42
80
9
11
19
20
5
3
55
7
32
48
81
66
70
50
79
83
74
87
30
75
16
15
65
39
1
38
35
47
53
84
86
72
33
34
37
27
36
62
82
69
54
4
6
12
85
18
17
10
10 17 18 85 12 6 4 54 69 82 62 36 27 37 34 33 72 86 84 53 47 35 38 1 39 65 15 16 75 30 87 74 83 79 50 70 66 81 48 32 7 55 3 5 20 19 11 9 80 42 61 31 57 45 29 52 26 67 92 25 88 14 8 13 56 2 23 40 76 71 51 89 73 28 90 78 24 60 59 77 46 63 64 68 43 21 22 58 41 49 91 44

esearch


Principal component 2
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2
-1
0
1
Principal component 1

search

Simulation Overview

- Simulate GBS data from a well assembled, known reference
- Add SNPs to reference
- Keep reference of the positions of SNPs
- Know the answer before asking the question
- Compare multiple GBS software pipelines to identify most accurate

Simulation Data Generation

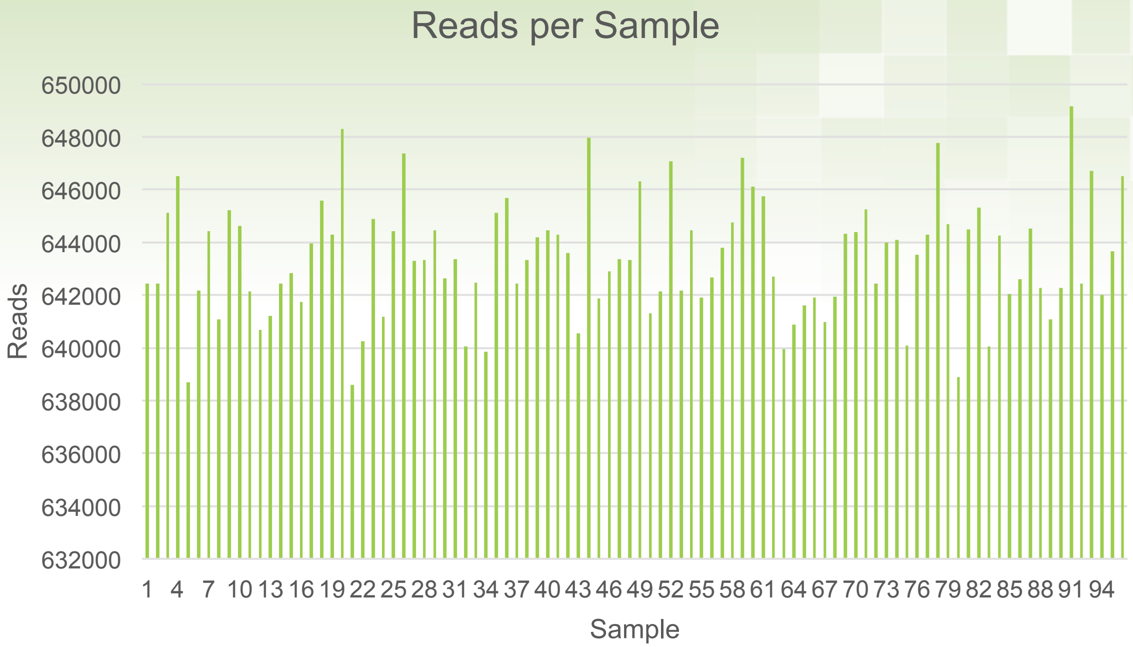
1. Read in whole genome
2. *In silico* restriction enzyme digest
3. *In silico* size selection of fragments
4. Generate SNP positions using an exponential distribution across reference
   - Currently at a rate of 1/300bp
   - Aim to increase rate to 1/100bp and 1/25bp

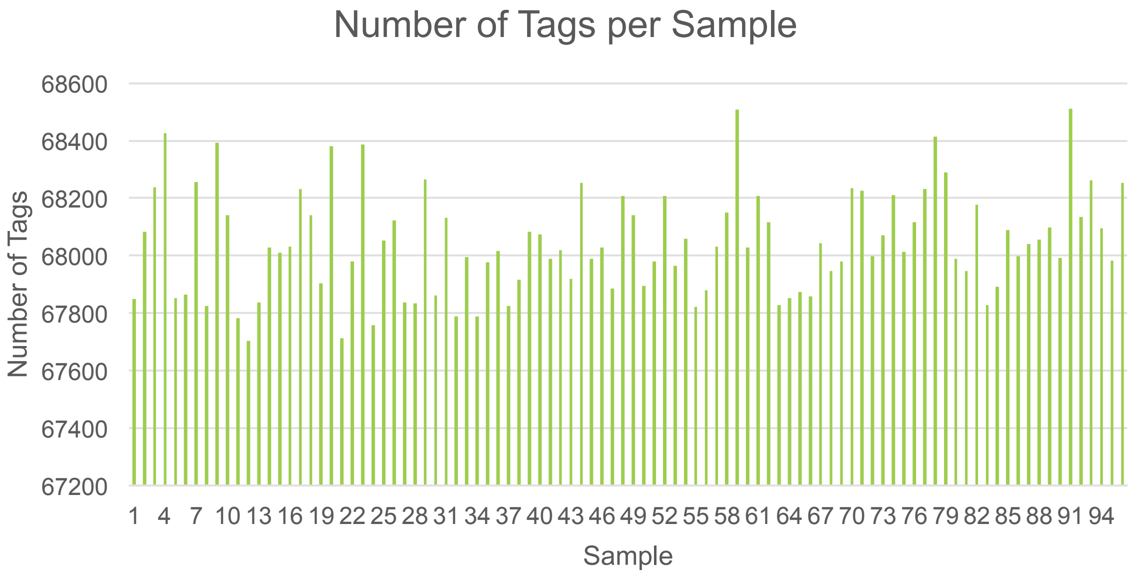
Simulation Data Generation

5. Generate Fastq Reads
   - Generate reads for 96 barcodes
   - % barcodes without reads at that site
   - Homozygous, heterozygous reference, homozygous alternate
   - Depth
6. Compile Annotation File
   - Chromosome
   - Cut site start and finish position
   - SNP position
   - Reference and Alternate Alleles
   - Allele and read depth for each barcode

Simulation Results

- 61,958,118 fastq reads across all chromosomes and 96 barcodes
- 8,639,933 SNPs in total across a 3Gb genome
- 26,458 seen in reads
- UNEAK identified 61,763,960 to be ‘good barcoded reads’


Reads per Sample
650000
648000
646000
644000
642000
640000
638000
636000
634000
632000
Reads
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94
Sample

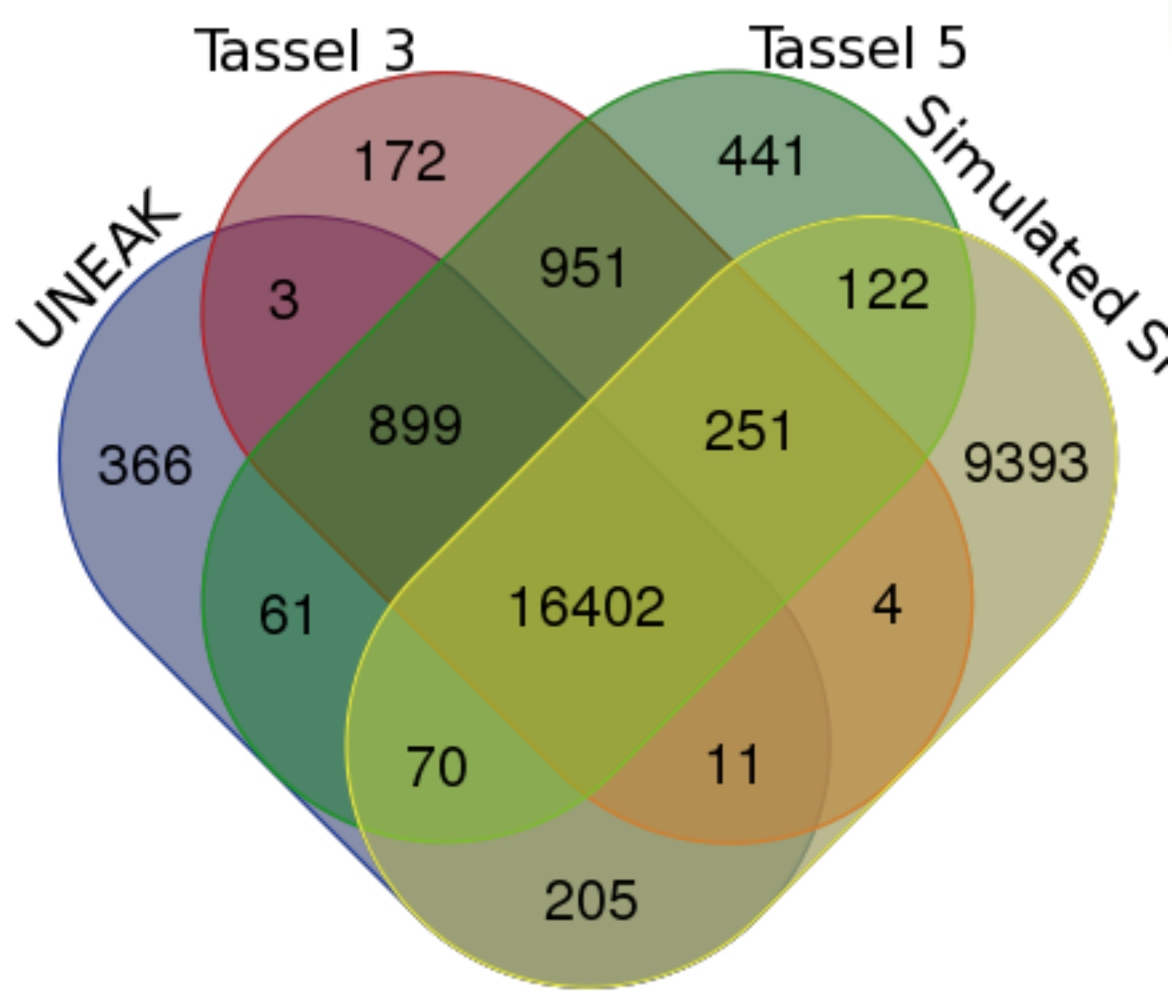

Number of Tags per Sample
68600
68400
68200
68000
67800
67600
67400
67200
Number of Tags
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94
Sample

agresearch

Simulation Results

	Number of SNPs	HWdgm > 0.05	Proportion of missing genotypes	Mean sample depth	Mean self-relatedness
UNEAK	19,510	19,425	0.28	7.5	1.1
Tassel 3	18,813	18,594	0.28	7.4	1.1
Tassel 5	19,197	19,122	0.28	7.5	1.1

How many SNPs are the same?



agresearch

Summary

- Compared different restriction enzymes for Mussel GBS
  - ApeKI
  - PstI
  - PstI-MspI
- Compared *de novo* vs reference tools for SNP calling
- Identified heterozygous undercalling is present in Mussel GBS

Summary

- Developed a pipeline to generate simulated GBS data
- Compared UNEAK, Tassel 3 and Tassel 5 using simulated data
- Tassel 3 and Tassel 5 under calling and miscalling SNPs

Next steps

- Run simulated data through other GBS pipelines
  - FreeBayes
  - STACKS
  - Homebrew Pipeline – ‘Gold Standard’ but very slow
- Increase SNP rate to 1/100bp and 1/25bp
- Identify optimal pipeline for GBS SNP calling
- Apply pipeline to real data and compare results

Acknowledgements

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND
SAPERE AUDE

- Chris Brown
- Neil Gemmell
- Kim Rutherford

- Otago University Strategic Research Grant -*Neil Gemmell and Shannon Clarke*

agresearch

- Shannon Clarke
- Rudi Brauning
- Hayley Baird
- Rayna Anderson
- Tracey Van Stijn
- Ken Dodds
- John McEwan

- MBIE Research Grant -*Genomics for production and Security (AgResearch)*

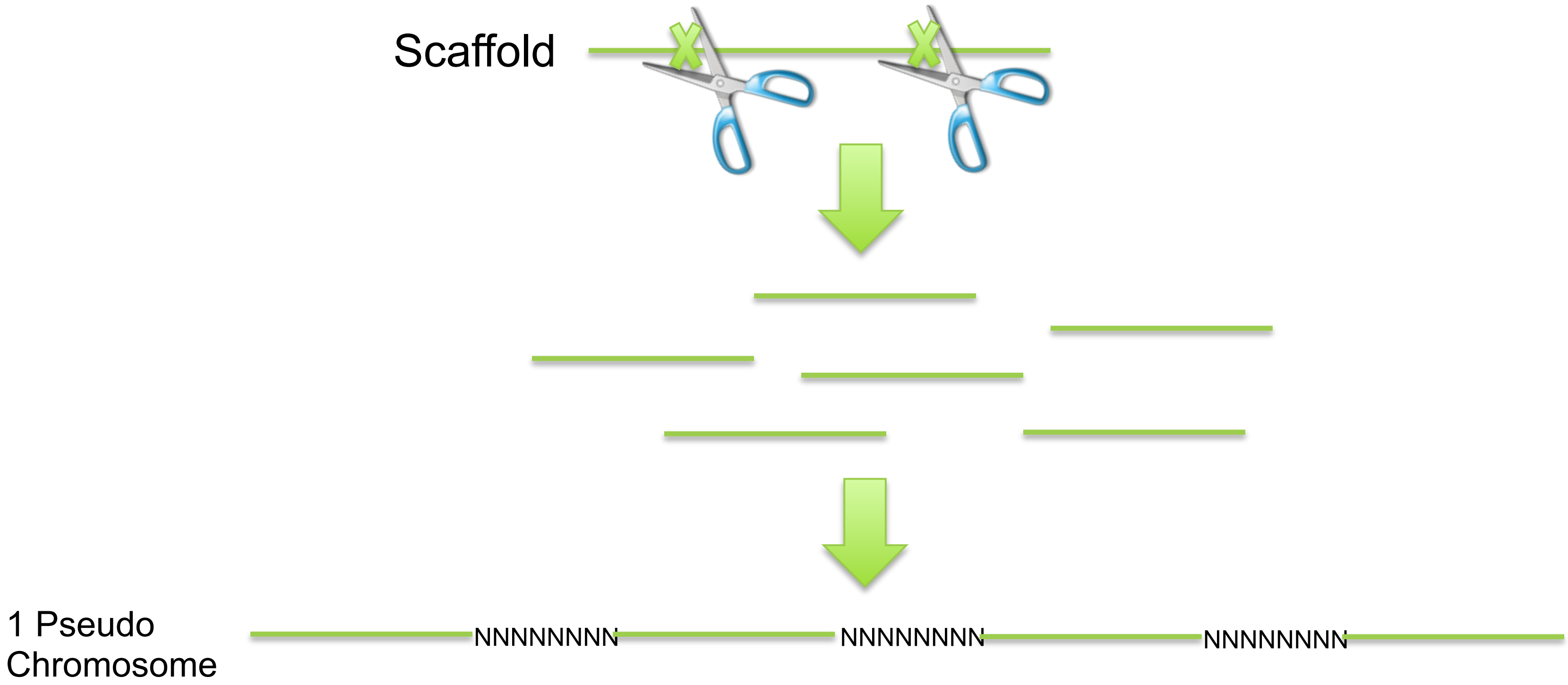
illumina

NZGL

CAWTHRON INSTITUTE

agresearch

In silico Restriction Enzyme Digest – “Chowder Genome”



agresearch

In silico Restriction Enzyme Digest – “Chowder Genome”

- Smaller genome for mapping
- Speeds up mapping
- PstI chowder genome 7,130,077 bases
  - 84.13% overall alignment rate against V1
  - 74.98% overall alignment rate against Chowder
- Losing small number of tags, but decreasing time taken for mapping stage