Population genomics of southern right whales: culture and connectivity

Marie Skłodowska-Curie Actions

newton international fellowships

University of St Andrews | FOUNDED 1413

Southern right whales, *then*

60,000 to 100,000 on 11-13 whaling grounds

HISTORICAL COASTAL WINTERING GROUNDS

Photo: W. Rayment

Impact of 200 years of whaling

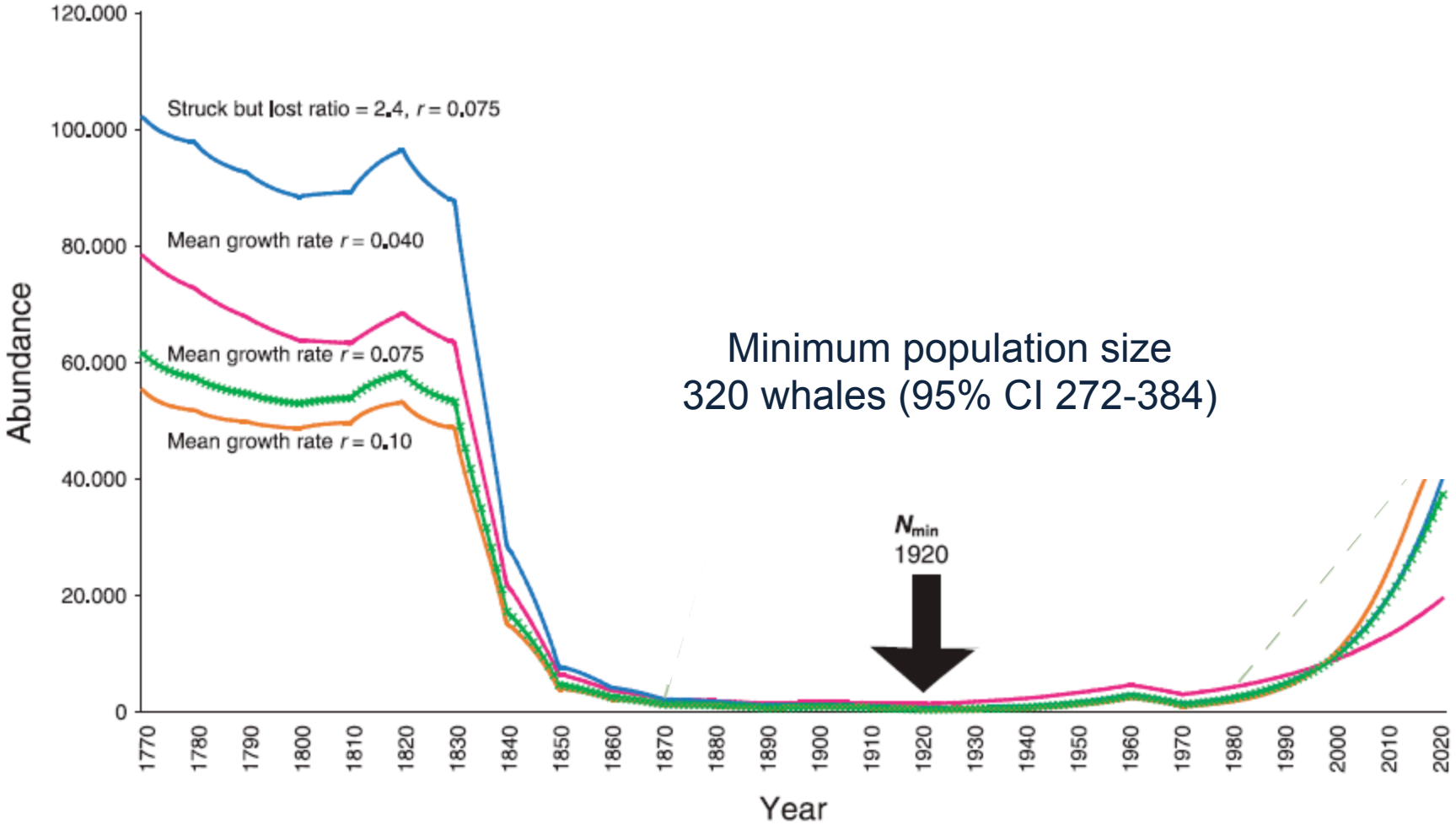
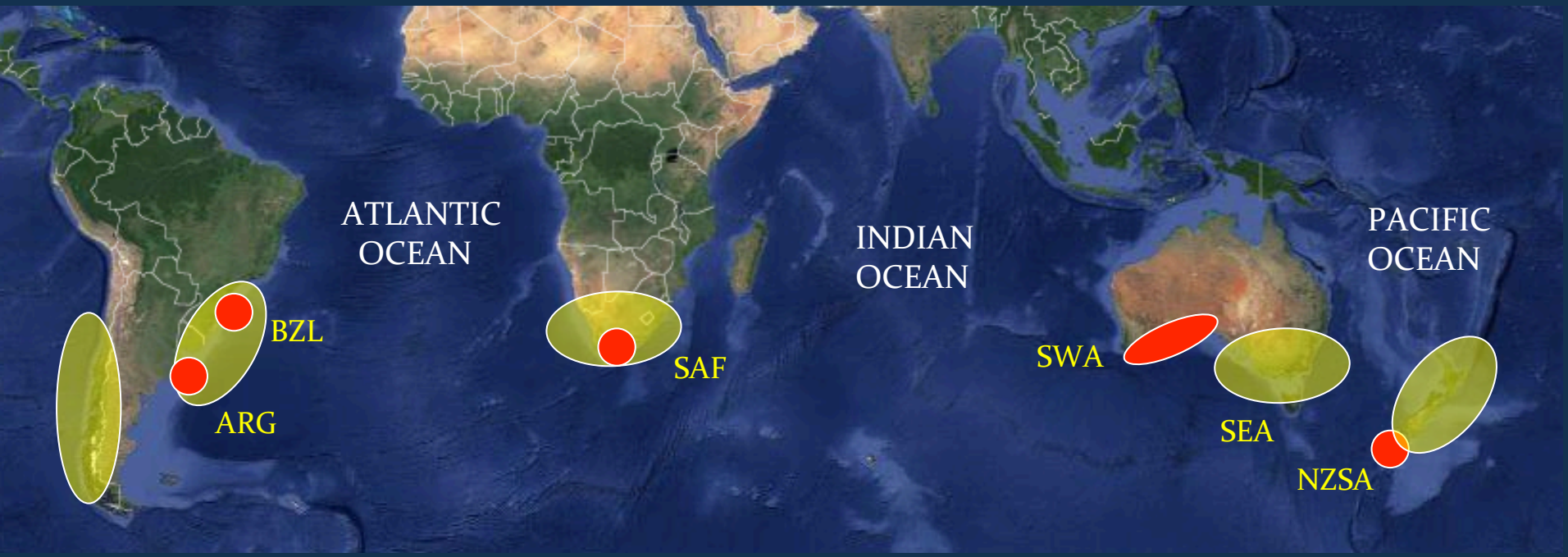


Figure: Jackson et al. 2008

Southern right whales, *now*

Spatially variable recovery

N~12,000



Large aggregations during winter

Sporadic sightings

Right whale: novel model organism

- Well studied
- Females show long-term fidelity to winter calving sites
- Maternally-directed fidelity to migratory destinations – 'migratory culture'
- Highly mobile
- Migrate 1000s km from coastal wintering grounds to offshore summer feeding grounds

→ No obvious barriers to dispersal

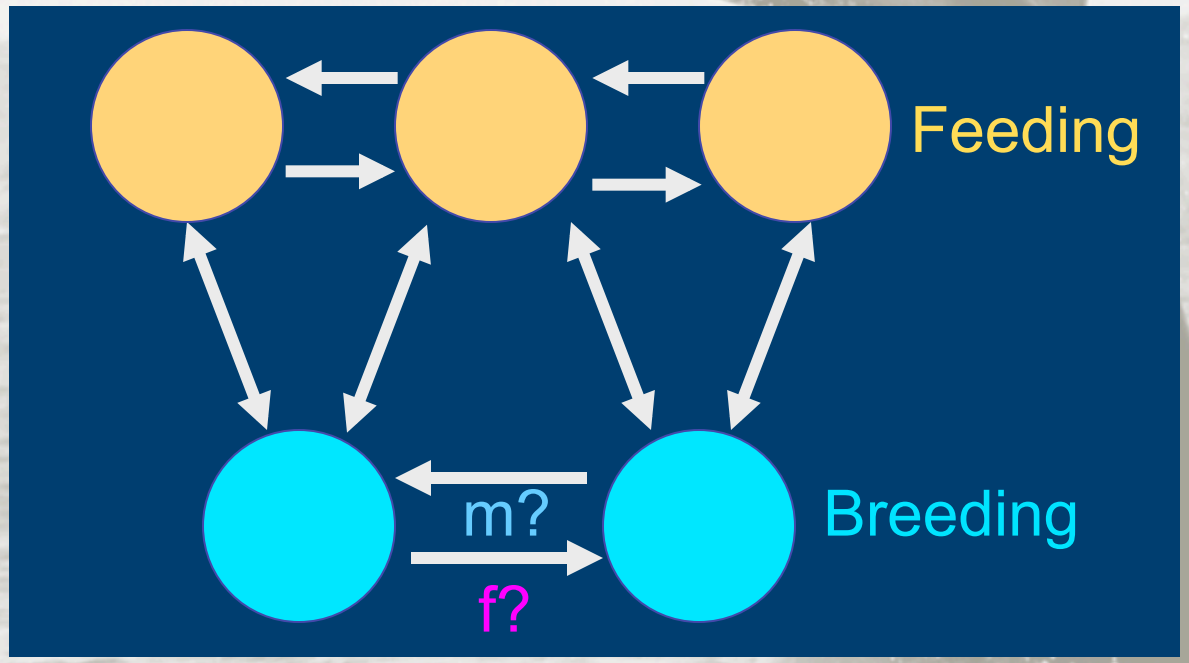
→ Migratory culture influences patterns of connectivity

→ Provides a good case study for examining the role of behaviour in connectivity

Migratory culture and genetic population structure

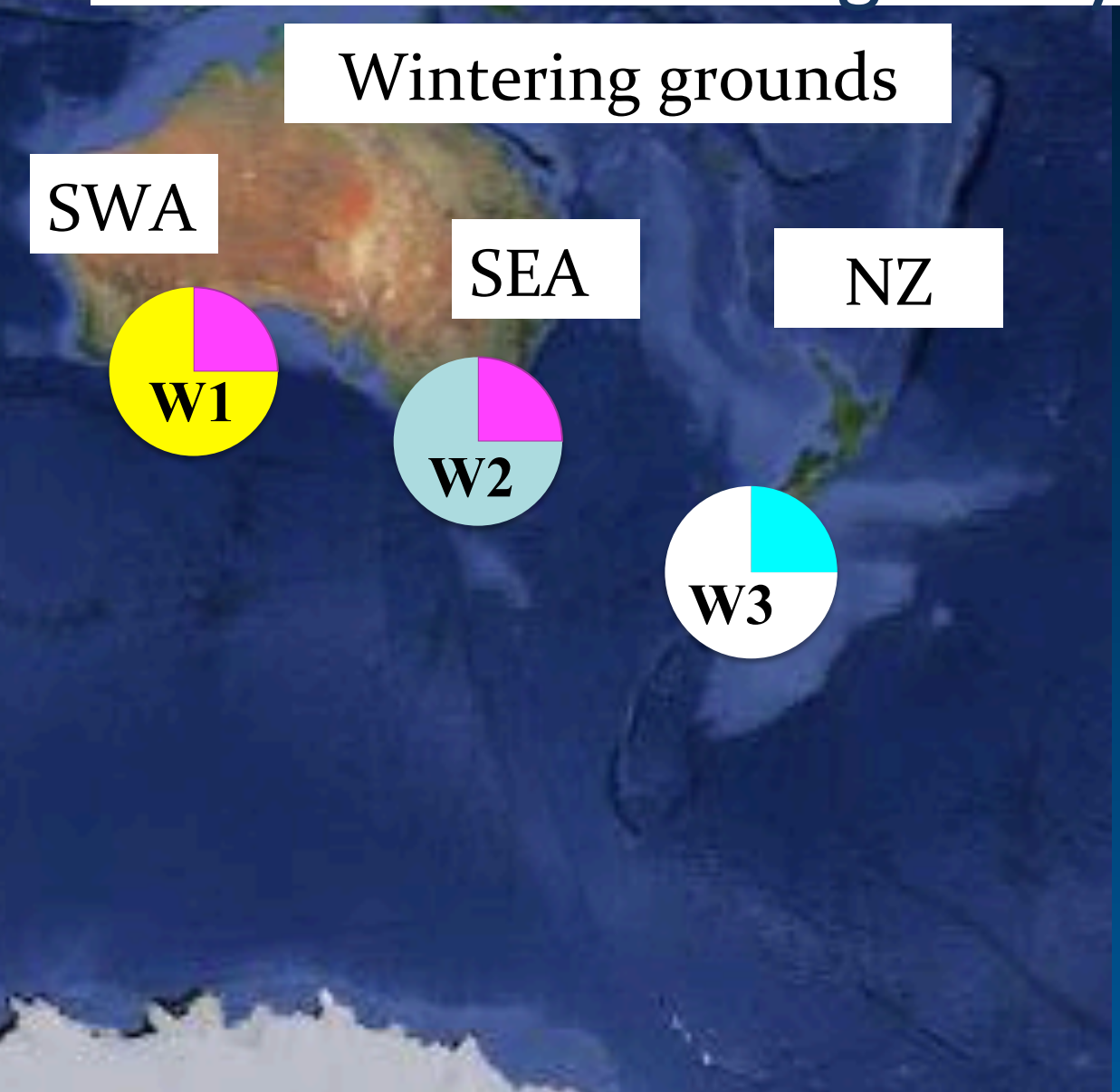
- Does migratory culture persist on an evolutionary time scale and shape patterns of gene flow and connectivity?
  - for both feeding and breeding grounds?
  - for both sexes?

- Or is connectivity sufficient to randomise genetic diversity?
- Role of scale and habitat?
- Recovery and recolonisation?



Archetypes of migratory structure
(courtesy C. Scott Baker, B. Taylor)

Considering migratory culture at both ends of the migratory network

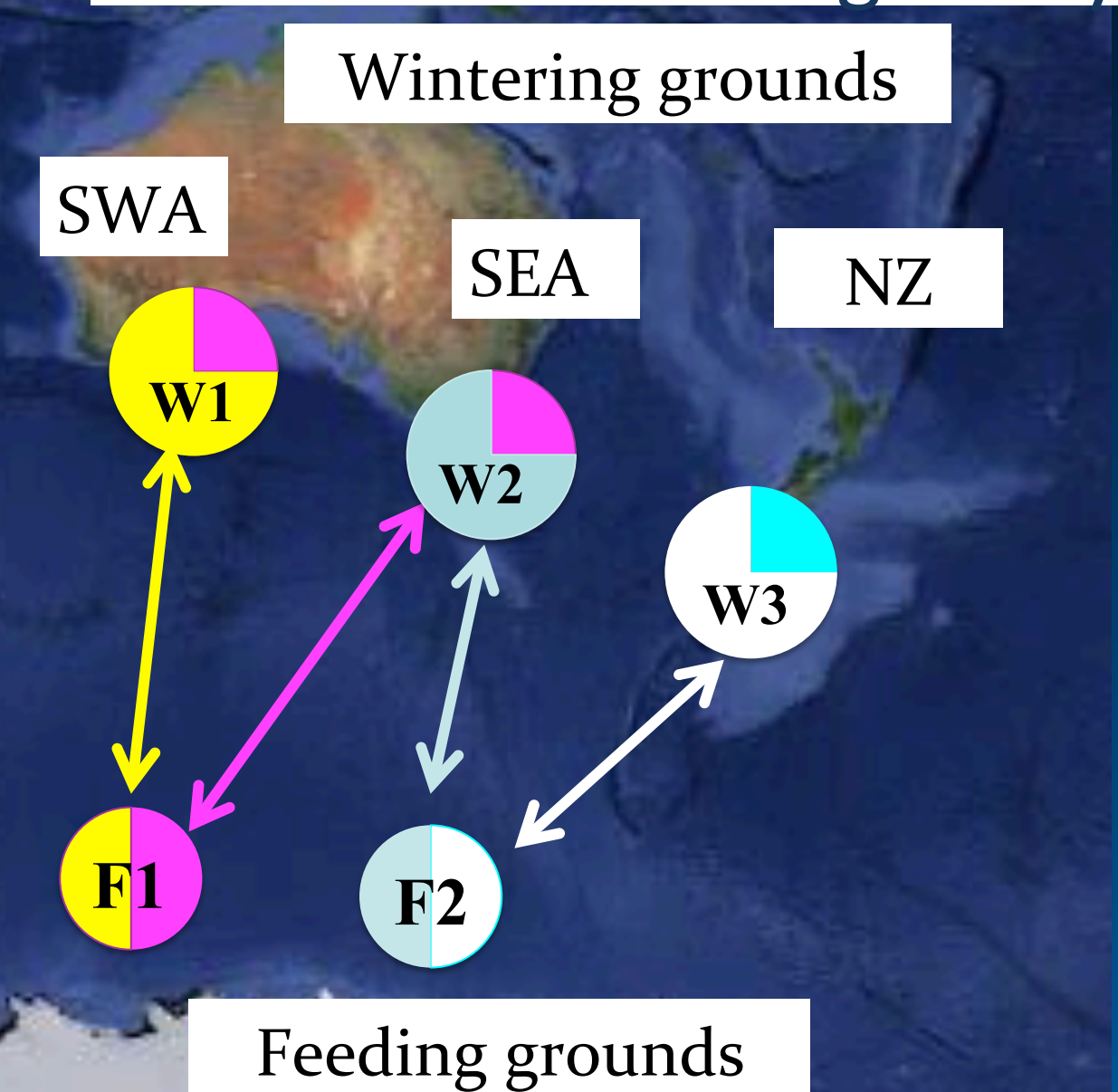


Migratory culture seems to have structured maternal lineages on wintering grounds

- Significant differences in maternally-inherited mtDNA markers (Patenaude et al 2007)
- Variable difference in bi-parentally inherited markers (Carroll et al 2011, 2015)

Piecharts show simplistic example of maternal allele frequencies

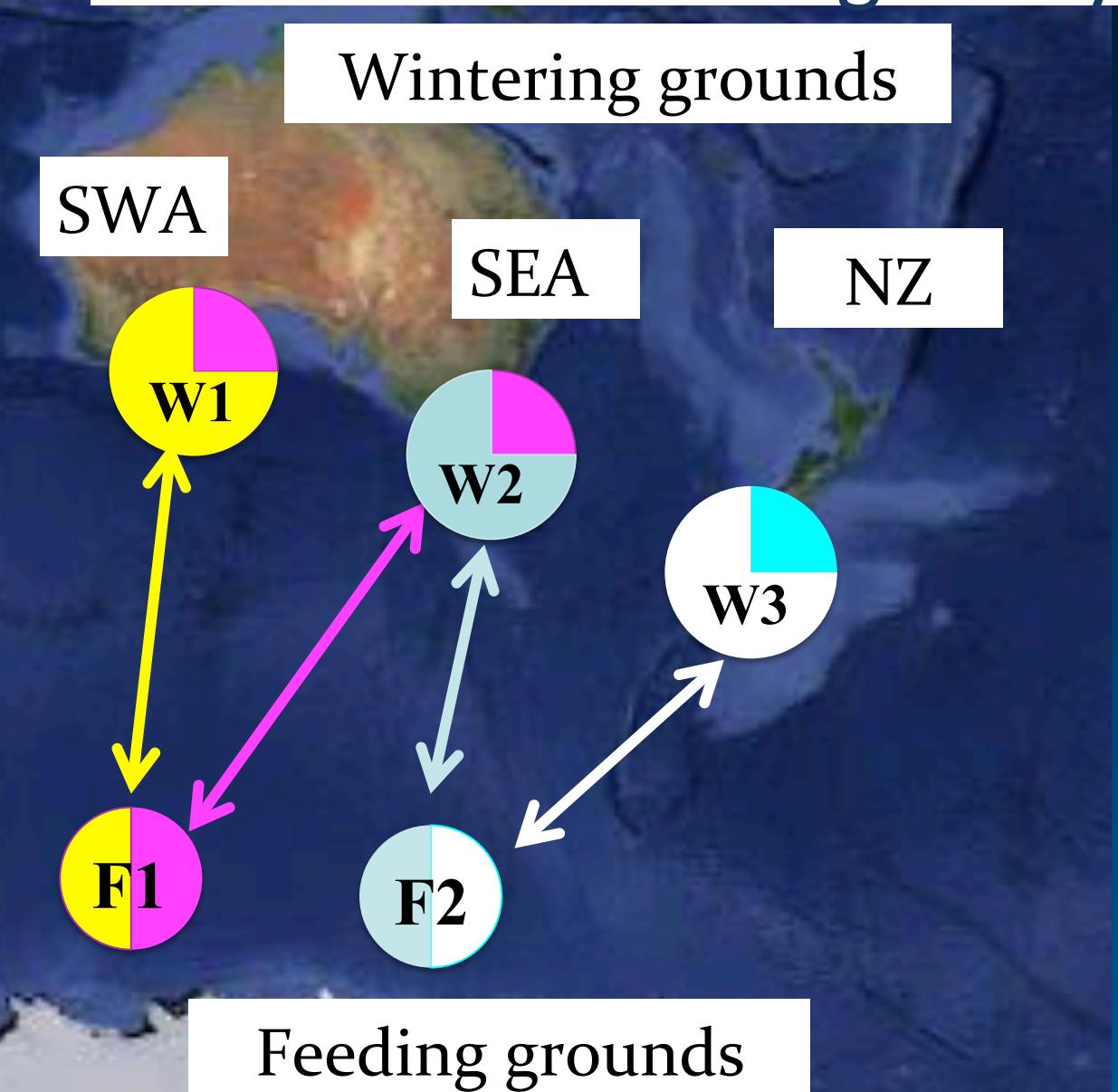
Considering migratory culture at both ends of the migratory network



Whales from different wintering grounds mix on feeding grounds (Baker et al 1999, Mate and Best 2011, Rowntree et al 2001)

Evidence that maternal lineage influences feeding ground choice (Valenzuela et al 2009, Carroll et al 2015)

Considering migratory culture at both ends of the migratory network



Do shared migratory cultural traditions to feeding grounds facilitate gene flow between wintering grounds?

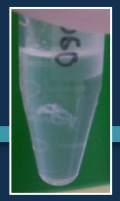
Constructing DNA and stable isotope profiles of living whales

Sample Collection

COLLECT SKIN BIOPSY SAMPLE

TISSUE

DNA



DNA

Mitochondrial DNA

Proxy for maternal lineage – migratory tradition

Stable isotope profile

Proxy for feeding ground tradition

DNA Profile

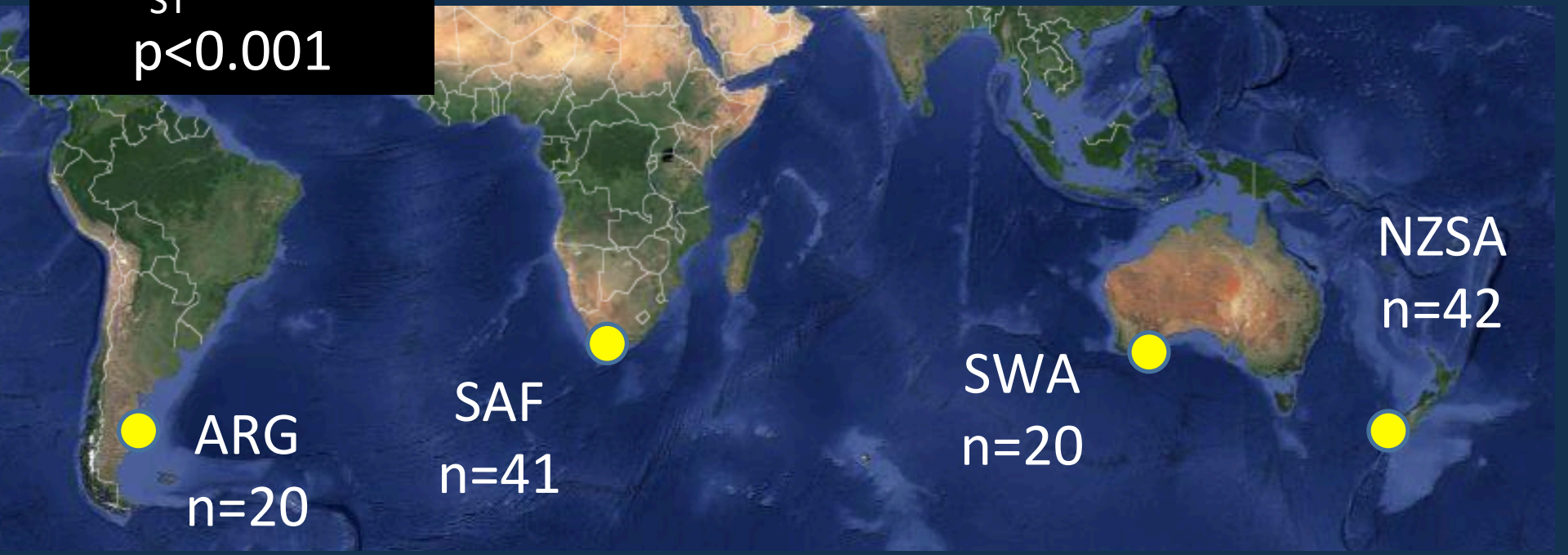
Estimate connectivity or infer population of origin

Previous research:

Patenaude et al (2007) used 275 bp of mtDNA control region to investigate population structure

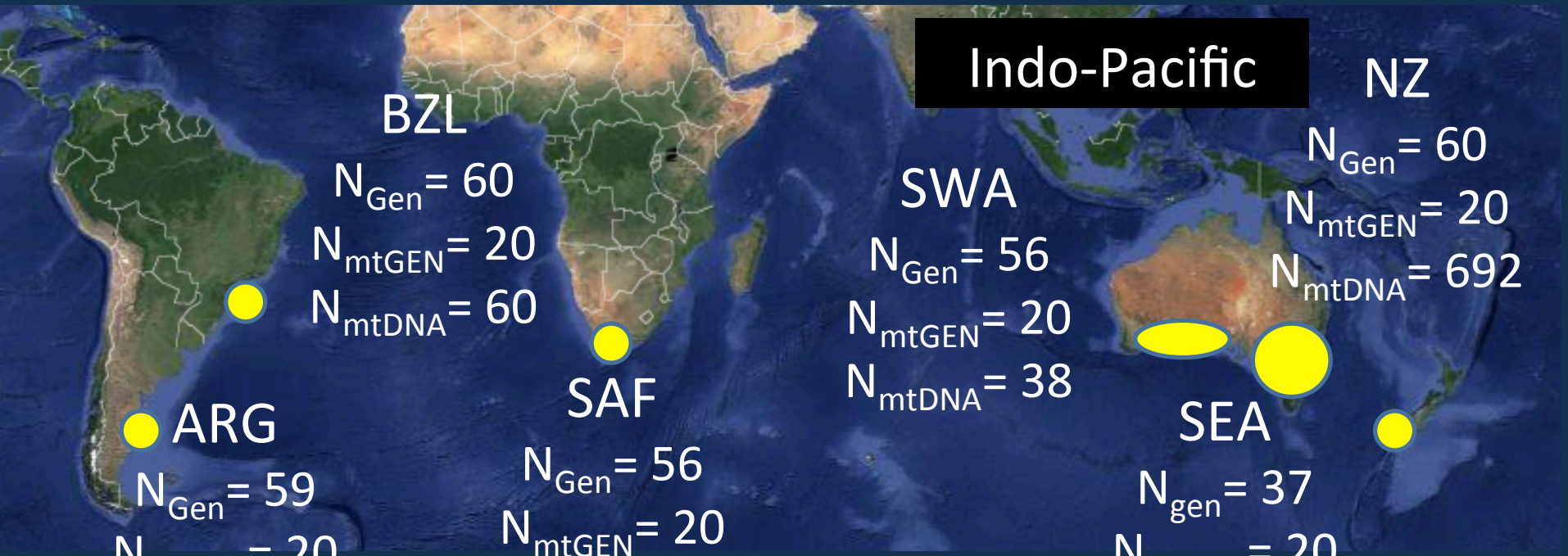
Overall mtDNA
F_{ST}=0.159
Φ_{ST}=0.238
p<0.001

Structuring of mtDNA lineages
Female fidelity to calving grounds
(Patenaude et al 2007)



Current research:

N_{Gen}: ddRAD sequencing
N_{mtGEN}: whole mitogenome sequences
N_{mtDNA}: mtDNA control region sequence
Working with many great collaborators to get a representative circumpolar sample



Global population structure in SRW: current research

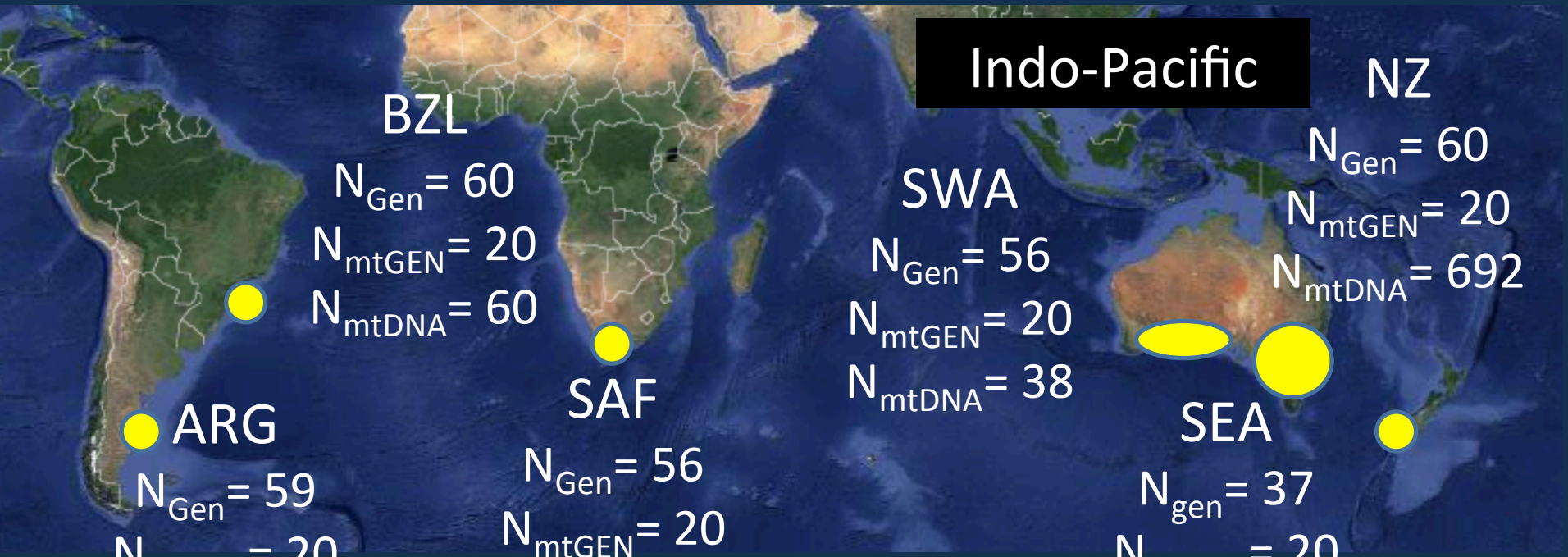
Large network of collaborators contributing samples, data and expertise

Samples/data/ expertise from	Collaborators: Prof O. Gaggiotti, host at St Andrews		
Argentina	J. Seger	V. Rowntree	L. Valenzuela
Brazil	L. Oliveira P. H. Ott	K. Groch P. Flores	T. Frasier B. White
South Africa	P. Palsbøll	M. Bérubé	Late P. Best K. Findlay
Australia & New Zealand	R. Harcourt J. Bannister M. Watson	C. S. Baker N. Patenude	R. Constantine R. Alderman
Genomics/ Stable isotope	K. Andrews M. Bérubé D. R. Grocke	R. Hoelzel C. S. Baker W. Austin	P. Palsbøll D. Steel A. Young

Current research:

Aim (1): estimate migration rates and population assignment probabilities, with conventional population genetics tools

Aim (2): estimate migration rates and population assignment probabilities, given migratory traditions, with novel Bayesian model



Overview of steps in ecological/conservation genomics study

Step	Standard marker e.g. microsats	Next-gen marker e.g. ddRAD
Study design	Pilot study Assess sample quality	
Data generation and processing	Marker quality check e.g.	
	- allele binning - checking for null alleles etc	- depth of coverage - minor allele frequency
	Individual genotype quality check e.g.	
	- excessive homozygosity - repeated failures	- excessive homozygosity - reads recovered
	Replication controls for error, systematic error identification	
Analysis	'Tried and tested' vs newer, faster software	

Steps derived from Morin et al 2010

SRW ddRAD: study design

Aim: 5K SNPs with high coverage across ~350 individuals

Challenges:

- Reproducibility: lab work conducted in four labs in four countries
- Coverage: optimising multiplex scheme for high-throughput sequencing
- Reliability: estimate genotyping error
- Variability in sample DNA quality

SRW ddRAD: pilot study

- ddRAD: 'tuneable' – different combinations of enzymes & size selection windows

- Two different protocols in 4 samples (2 NZ, 1 SEA, 1 SWA)

A. MluCI & NlaIII restriction enzymes (Peterson et al 2012)
    260 - 310 bp size selection

B. MspI and HindIII restriction enzymes (K. Andrews)
    300 - 400 bp size selection

Pooled and ran on Illumina MiSeq; ~12M PE reads back

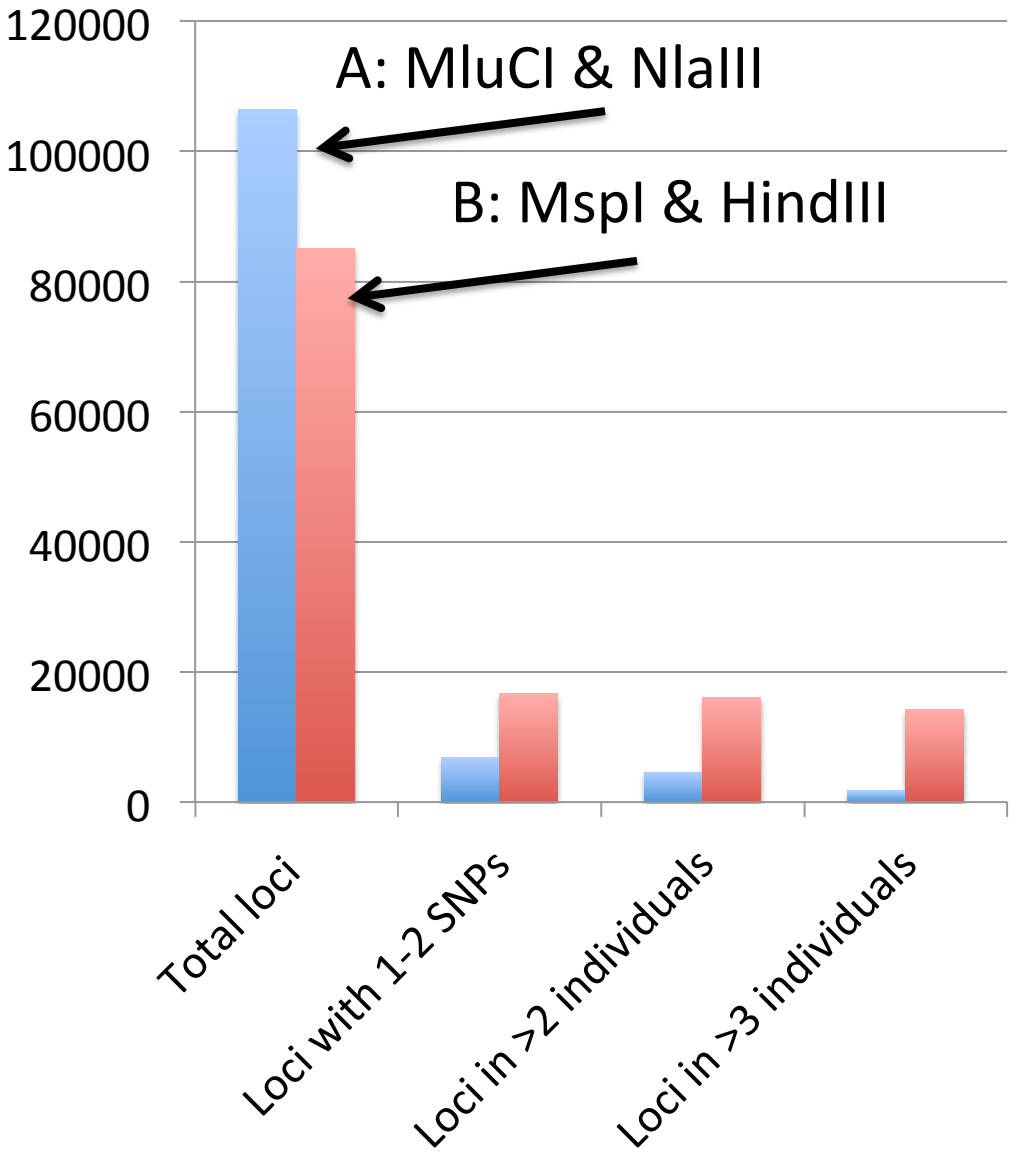
- How many loci are recovered by the different protocols across >2 samples (reproducibility)?

SRW ddRAD: pilot study

- Aligned reads against the bottlenose dolphin genome with bwa and run through the STACKS pipeline

- Higher proportion of loci variable & typed in more samples with MspI & HindIII

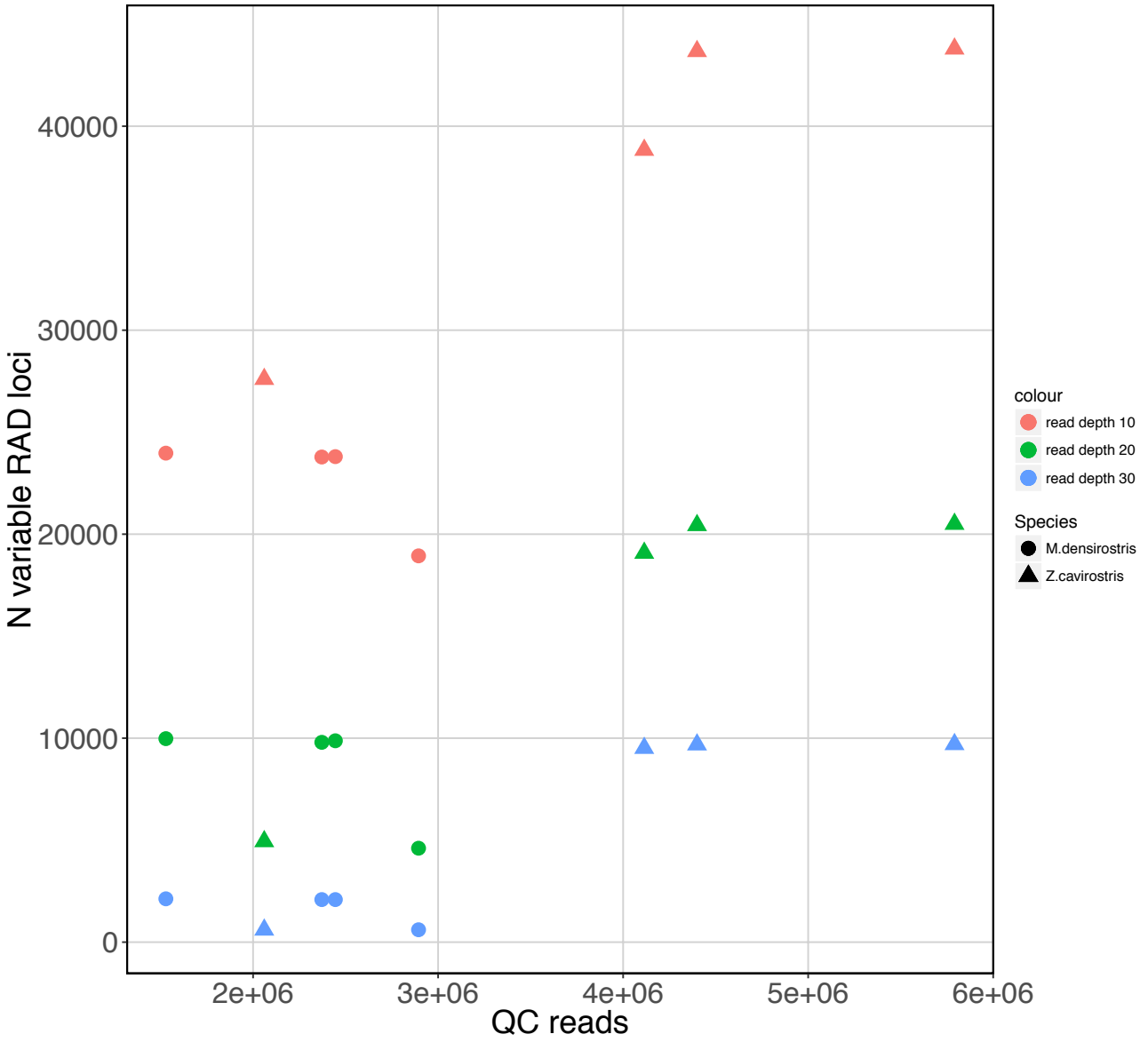
- Multiplexed 60 samples per HiSeq lane



SRW ddRAD: pilot study

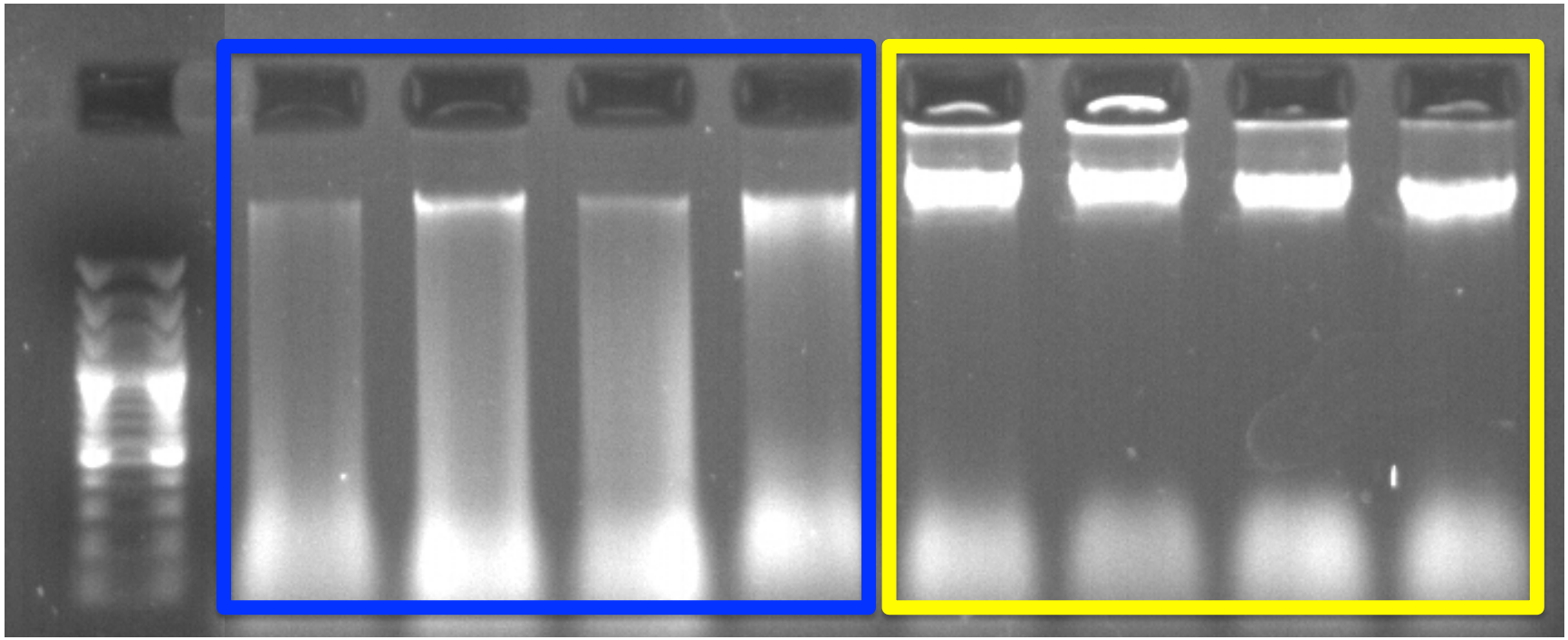
- Plotted number of QC reads versus number of variable loci at different sequencing depths

- Multiplexed 60 samples per HiSeq lane



SRW ddRAD: sample quality

- Variation in sample quality leads to variation in number of reads per samples
- Gel visualisation to check quality of DNA



- Pooled samples of similar quality into same libraries
- Estimate DNA conc. using qubit

SRW ddRAD: study design

So far:

- Run 4 lanes of New Zealand (n=60), Argentinean (n=60) and Australian samples (n=80) ; reran some low-coverage samples
- P. Palsbøll and M. Bérubé at the University of Groningen have run 1 lane of South African samples
- To do: 1 lane of Brazilian samples and additional South African samples from other collaborators

ddRAD data generation & processing: Bioinformatic pipelines

Program (platform)	Reference & citations ✜	Underlying algorithm/process	Benefits, per citation
PyRAD (requires python)	Eaton (2014) 38 citations	U-SEARCH/MUSCLE used to identify loci on a per sample basis and then globally identify orthologous loci	Handles indels; parallel processing; phylogenetics; easier to use
STACKS (windows, linux, mac)	Catchen et al (2011, 2013) 283+158 citations	Identifies loci on a per sample basis ->creates catalog of loci-> genotypes individuals	Faster computationally; both de novo and ref genome
dDocent (linux)	Puritz et al (2014) 7 citations	BASH wrapper that links parts of different bioinformatic pipelines together including STACKS, RAINBOW, GATK	More flexible
AftrRAD (mac and linux)	Sovic et al (2015) 3 citations	ACANA/MAFFT clustering algorithms	Handles indels

✜ Based on google scholar search November 2015

ddRAD data generation & processing: Bioinformatic pipelines

Used STACKS:

- optimised parameters by examining genotyping error rates of replicate samples

- output dataset with optimised parameter set:

-m 2; –M 2; –n 4

loci with 1 to 4 SNPs

depth of coverage 20x

loci found in at least 3 populations in at least 75%

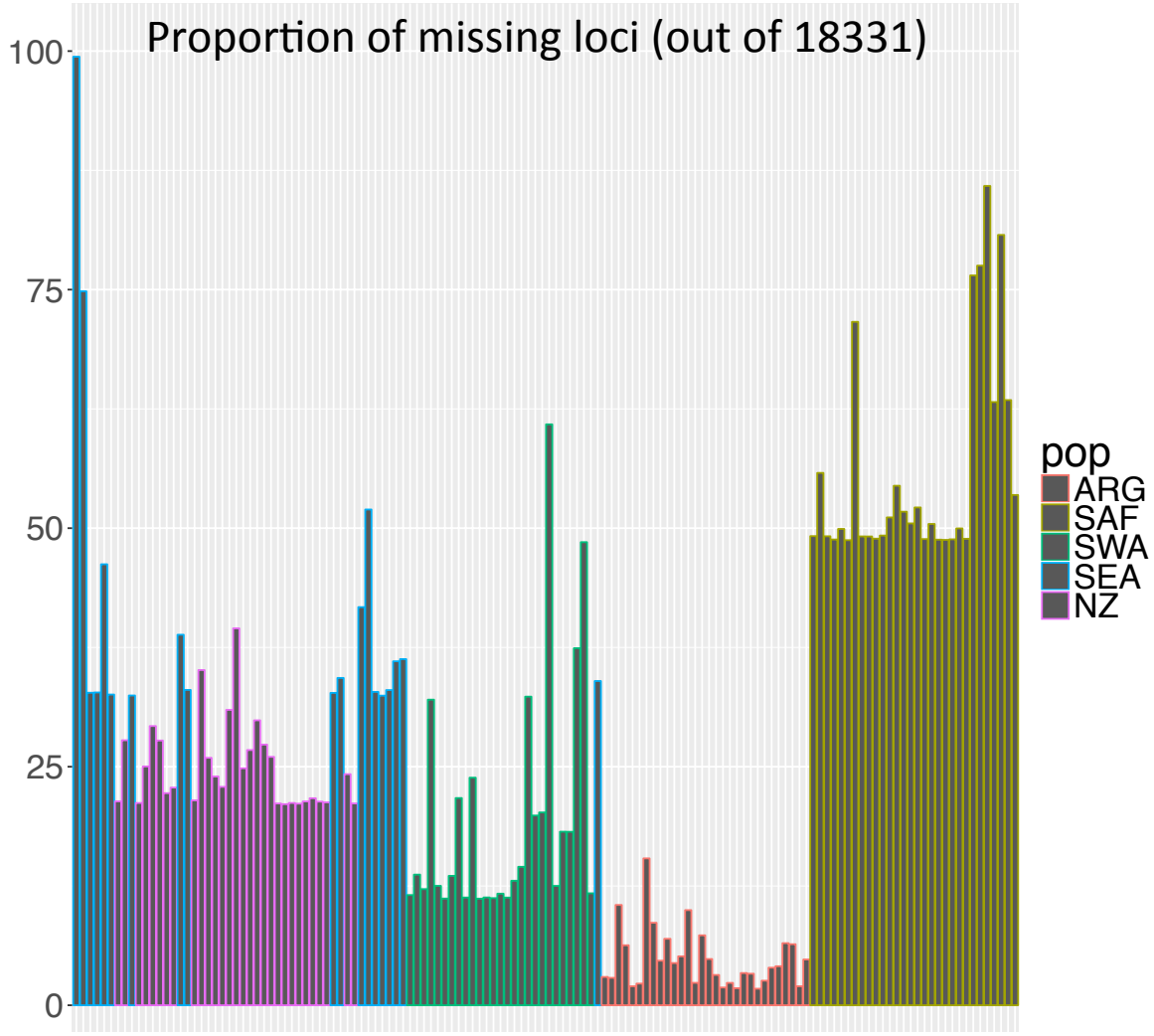
- gave a dataset of 18,331 RAD loci

- other considerations: minor allele frequency; rxstacks; max read depth; etc.

ddRAD data generation & processing: Bioinformatic pipelines

- Problem: the South African dataset had fewer overlapping loci with other populations

- Only library run by collaborators – but followed the same protocol

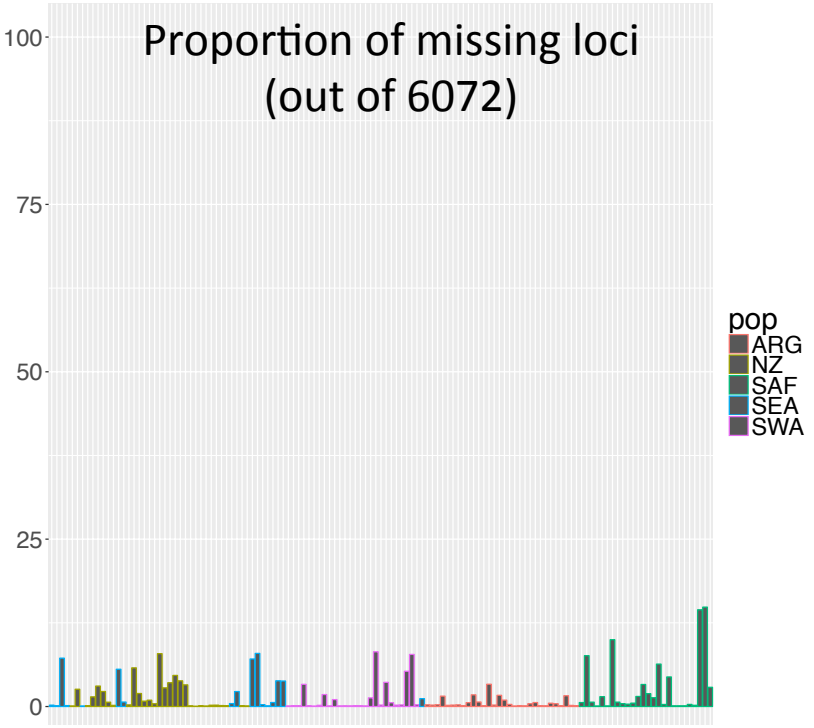


ddRAD data generation & processing: Bioinformatic pipelines

Tweaked export parameters

- RAD loci in all 5 pops

- in >75% individuals



Overall Trial Global Dataset:

6072 RAD loci at 20x depth

Removed 7 low-coverage samples

South Pacific: NZ n = 30; SWA n=26; SEA n=17

South Atlantic: ARG n = 30 ; SAF n = 26

ddRAD data analysis:

Wide range of analyses options for population structure next-gen data

- Classical genetic tests of differentiation, e.g. estimating F_{ST}
- Multivariate statistical analyses (e.g., Principal coordinates analysis)
- Clustering methods
- Estimating migration rates (e.g. BayesAss, TreeMix)
- Approximate Bayesian computing (e.g. FastSimCoal)
- many others.....

ddRAD data analysis: Multivariate statistics

What they do

- Represent genetic distances using a small number of synthetic variables
- Assumption-free i.e no HWE assumed
- Variety of methods (see Jombart et al 2009)
- Variety of packages e.g. in R ade4, adegenet e.g. GenAlEx

PCA and DAPC

- Principal coordinates analysis (PCA) using ade4 in R; missing values interpolated; scaled to reduce bias.
- Discriminant analysis of principal components using adegenet
- takes output of PCA and clusters data to maximise variation between groups (k-means clustering)
- finds principal components maximising differences between groups

ddRAD data analysis: (Bayesian) clustering programs

What they do

- estimate individual assignment probabilities given a number of clusters K
- some estimate the proportion of an individual's genome that originates from different ancestral genepools - *ancestry coefficients*

Why (Bayesian) ?

- the most commonly used program estimates assignment probabilities with Bayesian algorithms (STRUCTURE, BAPS, TESS)
- often implemented using computationally greedy MCMC
- these methods don't scale well to next-gen datasets
- newer programs have been switching to different algorithms to increase computing efficiency (faster to run)

ddRAD data analysis: Some clustering programs

Program	Reference & citations	Data	Underlying algorithm	Benefits
STRUCTURE	Pritchard et al (2000) 15,528 citations	SNP/ microsat	Model based; Bayesian MCMC implementation	Tried & tested
Fast STRUCTURE	Raj et al (2014) 35 citations	Bi-allelic only	Model-free; Fast	Fast
BAPS	Corander et al (2006) 421 citations	SNP/ microsat/ haplotype	Model-based; Spatial & non-spatial, Bayesian stochastic optimization methods	GUI
TESS/ TESS2.3	Chen et al (2007) 431 citations	Bi- / multi-allelic	Model-based; Spatial; Bayesian MCMC implementation	Spatial
TESS3	Caye et al (2016)	Bi- /multi-allelic	Model-free; Spatial; Least squares optimisation	Spatial
sNMF	Frichot et al (2014)	Bi-allelic	Model-free; PCA-like method	Fast

Citations: Based on google scholar search November 2015

ddRAD data analysis: Some clustering programs

Program	Reference & citations	Data	Underlying algorithm	Benefits
STRUCTURE	Pritchard et al (2000) 15,528 citations	SNP/ microsat	Model based; Bayesian MCMC implementation	Tried & tested
Fast STRUCTURE	Raj et al (2014) 35 citations	Bi-allelic only	Model-free; Fast	Fast
BAPS	Corander et al (2006) 421 citations	SNP/ microsat/ haplotype	Model-based; Spatial & non-spatial, Bayesian stochastic optimization methods	GUI
TESS/ TESS2.3	Chen et al (2007) 431 citations	Bi- / multi-allelic	Model-based; Spatial; Bayesian MCMC implementation	Spatial
TESS3	Caye et al (2016)	Bi- /multi-allelic	Model-free; Spatial; Least squares optimisation	Spatial
sNMF	Frichot et al (2014)	Bi-allelic	Model-free; PCA-like method	Fast

ddRAD data analysis: Some clustering programs

Program	Reference & citations	Data	Underlying algorithm	Benefits
STRUCTURE	Pritchard et al (2000) 15,528 citations	SNP/ microsat	Model based; Bayesian MCMC implementation	Tried & tested
Fast STRUCTURE	Raj et al (2014) 35 citations	Bi-allelic only	Model-free; Fast	Fast
BAPS	Corander et al (2006) 421 citations	SNP/ microsat/ haplotype	Model-based; Spatial & non-spatial, Bayesian stochastic optimization methods	GUI
TESS/ TESS2.3	Chen et al (2007) 431 citations	Bi- / multi-allelic	Model-based; Spatial; Bayesian MCMC implementation	Spatial
TESS3	Caye et al (2016)	Bi- /multi-allelic	Model-free; Spatial; Least squares optimisation	Spatial
sNMF	Frichot et al (2014)	Bi-allelic	Model-free; PCA-like method	Fast

SRW ddRAD data analysis: fastSTRUCTURE

Tries to replicate the underlying model of STRUCTURE but with a PCA-like statistical approach

Different models: simple, hierarchical and logistic
- model population-specific allele frequencies differently
- simple: flat prior for population-specific allele frequencies
- hierarchical: based on a demographic model that allows allele frequencies to have shared underlying pattern at all loci (F-model, Falush et al 2003)
- logistic: is a computationally more efficient version of F-model

Ran four iterations of K 1 to 10 under simple model

SRW ddRAD data analysis: fastSTRUCTURE

Methods for inferring K

- K^*_{ε}: value of K that maximizes the log marginal likelihood lower bound of the entire dataset
- best at identifying strong structure
- severely underestimates K with weak population structure

- $K^*_{\varnothing C}$: the smallest number of model components that accounts for almost all of the ancestry in the sample
- overestimates K with weak population structure

Rep		simple
1	K^*_{ε}	2
	$K^*_{\varnothing C}$	2
2	K^*_{ε}	2
	$K^*_{\varnothing C}$	2
3	K^*_{ε}	2
	$K^*_{\varnothing C}$	2
4	K^*_{ε}	2
	$K^*_{\varnothing C}$	2

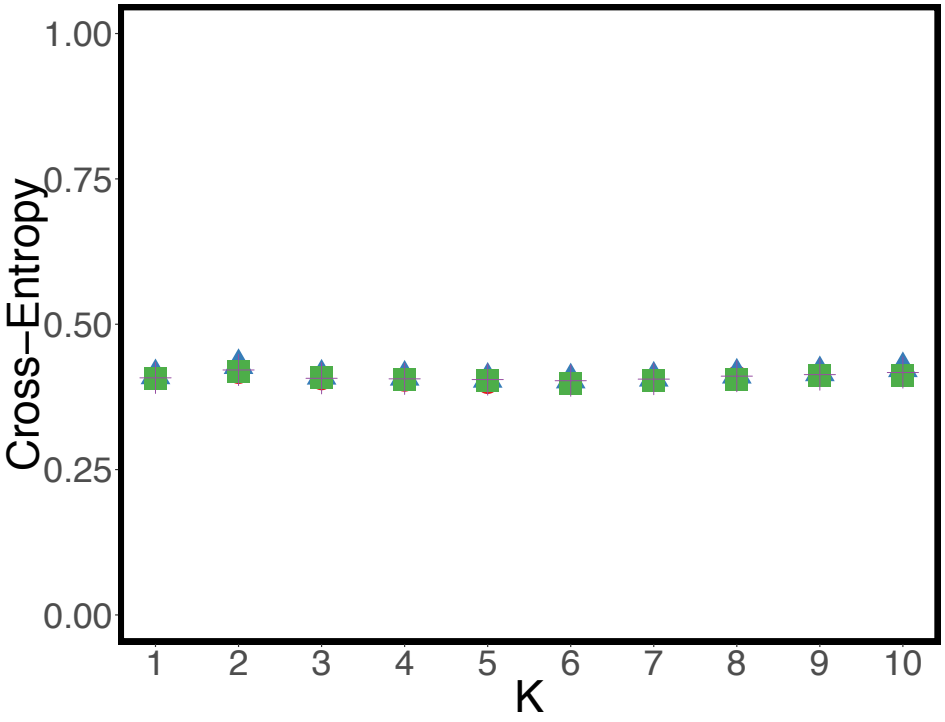
SRW ddRAD data analysis: sNMF

Estimates individual assignment probabilities using PCA-like approach

- Rapid: K 1 to 10 took a few minutes to run
- Run several times to ensure starting values don't influence outcome

Model selection in sNMF

- 'The cross-entropy criterion'
- Compares the genotypic frequencies predicted from the training set to those computed from the test set at each locus (leave 5% data out).
-Smaller values /a plateau in the cross-entropy plot suggest best fit (Frichot & Francois, 2015)



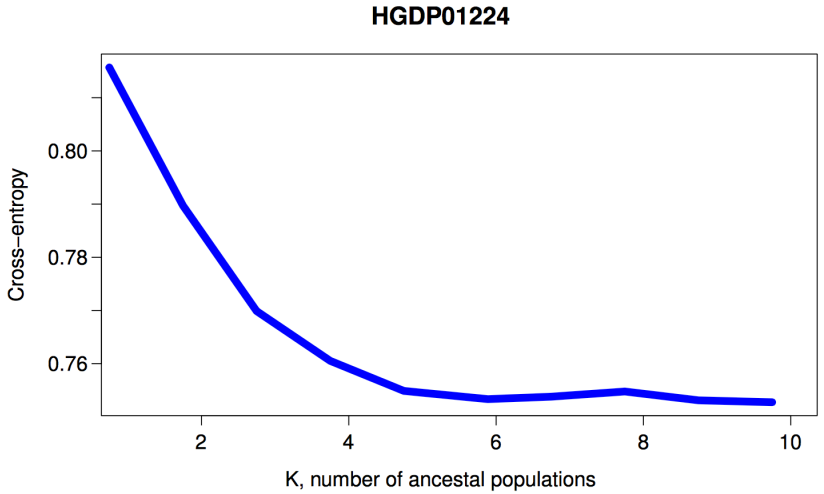
Consistent results across 4 runs

SRW ddRAD data analysis: sNMF

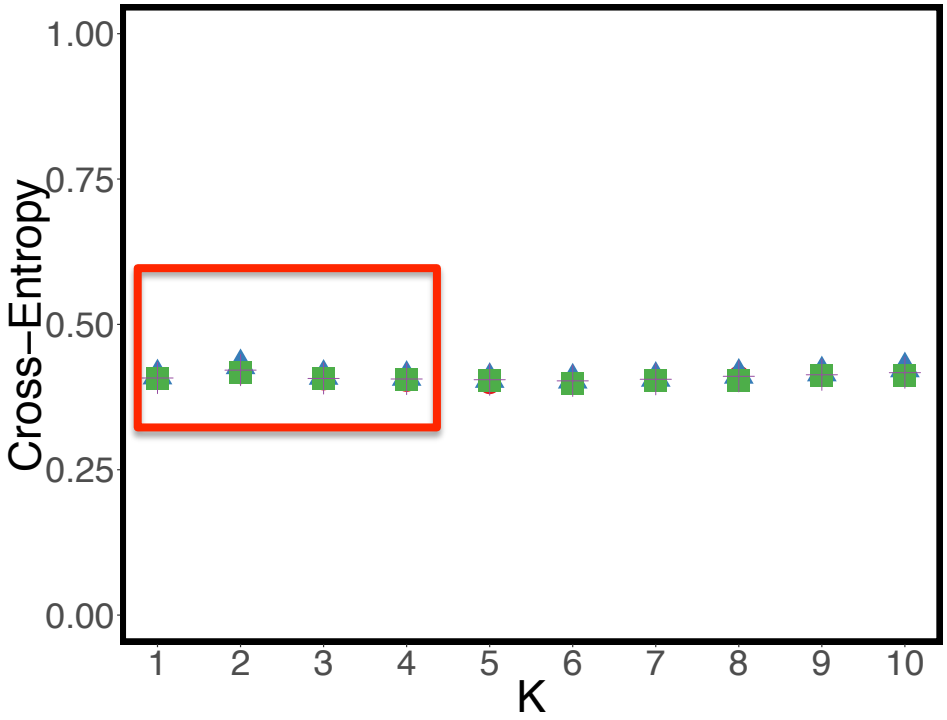
Estimates individual assignment probabilities using PCA-like approach

- Rapid: K 1 to 10 took a few minutes to run
- Run several times to ensure starting values don't influence outcome

Model selection in sNMF



cross-entropy plot suggest best fit (Frichot & Francois, 2015)



Consistent results across 4 runs

SRW ddRAD project: summary

Study design & aim: initial analysis

- Achieved Aim: 5K SNPs with high coverage across initial dataset

Challenges:

- Reproducibility: lab work conducted in four labs in four countries
- Coverage: optimising multiplex scheme for high-throughput sequencing
- Reliability: estimate genotyping error
- Variability in sample DNA quality

SRW ddRAD project: summary

Study design & aim: initial analysis

- Achieved Aim: 5K SNPs with high coverage across initial dataset

Challenges:

- Reproducibility: lab work conducted in four labs in four countries
- Coverage: optimising multiplex scheme for high-throughput sequencing
- Reliability: estimate genotyping error
- Variability in sample DNA quality

Data generation & processing

- Customised STACKS pipeline by finding combinations of parameters that minimised genotyping error
- Identified problem with overlapping loci between library preparations done by different labs

SRW ddRAD project: summary

Global population structure across southern right whale wintering grounds: standard genomics tools

Strong structuring of mtDNA haplotypes suggesting restricted female gene flow
- consistent with maternal fidelity to wintering grounds

Hierarchical population structure: greater differentiation between that within ocean basins
- classical differentiation indices suggest weak but significant structuring between ocean basins
- PCA and DAPC showed samples cluster by ocean basin
- STRUCTURE and sNMF indicate clustering by ocean basin
- simple fastSTRUCTURE model did not detect structure

SRW ddRAD project: future directions

Lab work

- finish ddRAD and mitogenome sequencing by late 2016

Analysis

- consider different datasets for different analyses e.g. make use of larger number of overlapping loci across NZ and Australia

- Historical demography:

Test hypotheses generated by mtDNA phylogeny: high levels of gene flow vs isolation and secondary contact

Lower diversity in South Pacific? Glaciation and the amount of suitable calving habitat; edge effects; whaling?

- Continuing work on novel Bayesian model with Prof O. Gaggiotti

Acknowledgements

ELC supported by Marie Curie from EU and Newton Fellowship from Royal Society
Genotyping costs supported by British Ecological Society grant & MASTS starting funding to OEG

Sample/data holders	Collaborators at St Andrews: Prof O. Gaggiotti
Argentina	Prof J. Seger, Drs. L. Valenzuela, V. Rowntree, University of Utah
South Africa	Prof P. Palsbøll and Dr M. Bérubé, University of Groningen, Dr. Ken Findlay, late Dr. P. Best, University of Pretoria
Australia & New Zealand	Prof R. Harcourt, Macquarie University, Prof C. S. Baker, Oregon State University, Dr R. Constantine, University of Auckland, Dr J. Bannister, Western Australia Museum, Dr N. Patenude, Collégial International Sainte-Anne