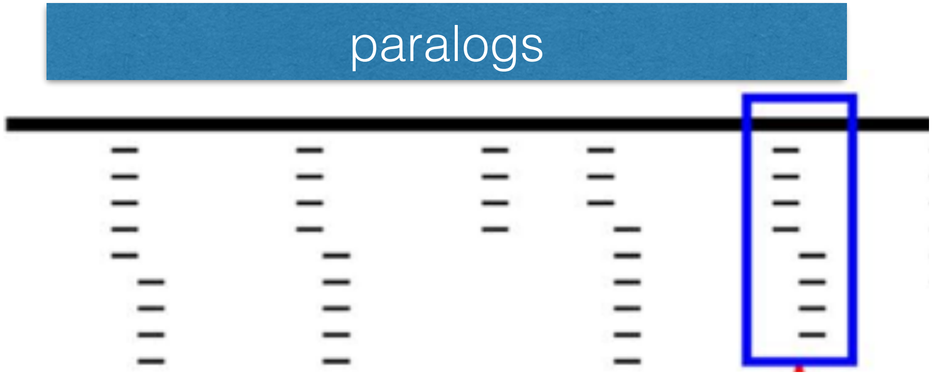
Assembing and using RAD locus catalogs: exploring the parameter space

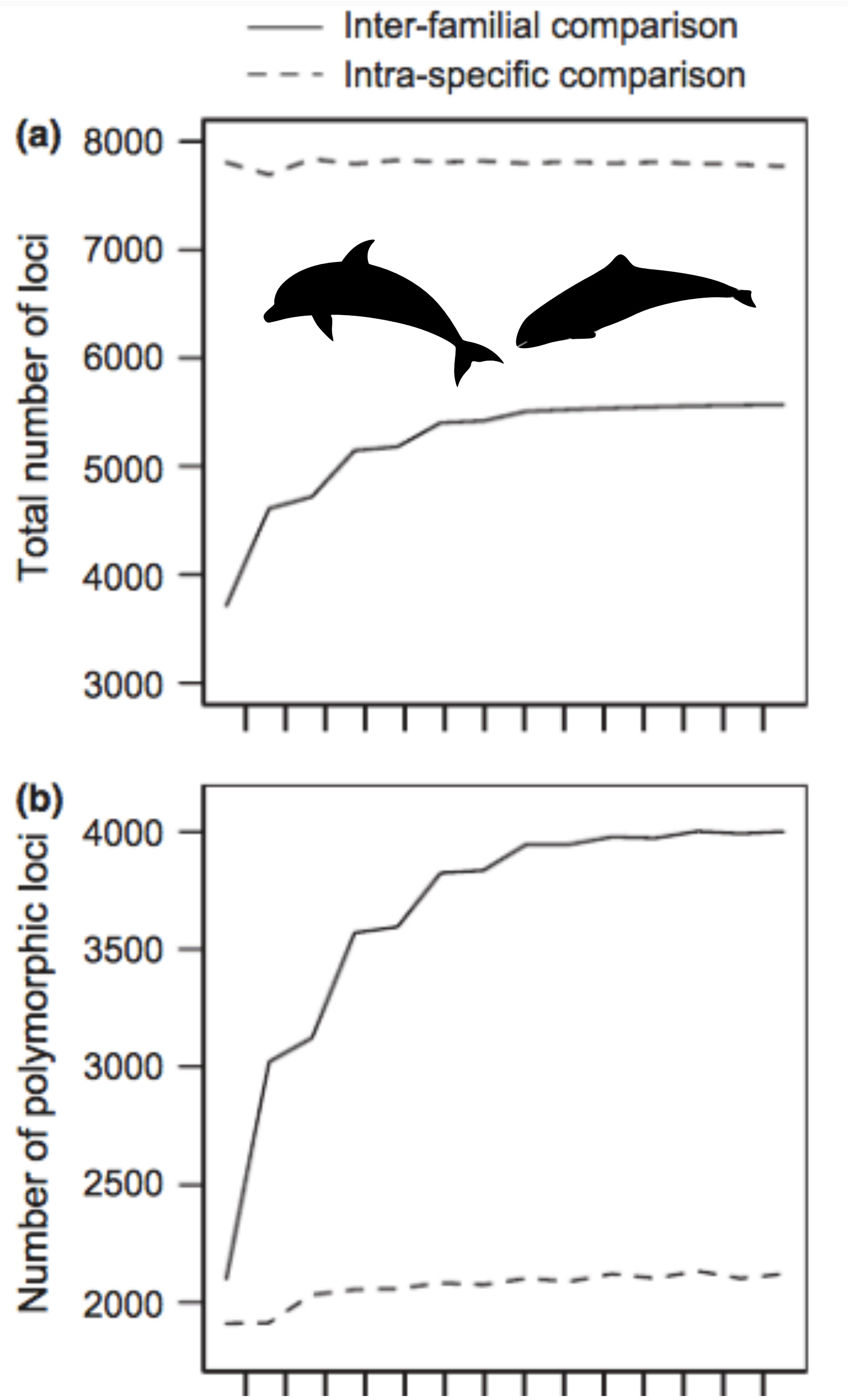
Eric Pante
LIENSs laboratory
UMR 7266 CNRS - La Rochelle University

http://epante.wordpress.com/

Difficulty of building a locus catalog in a nutshell



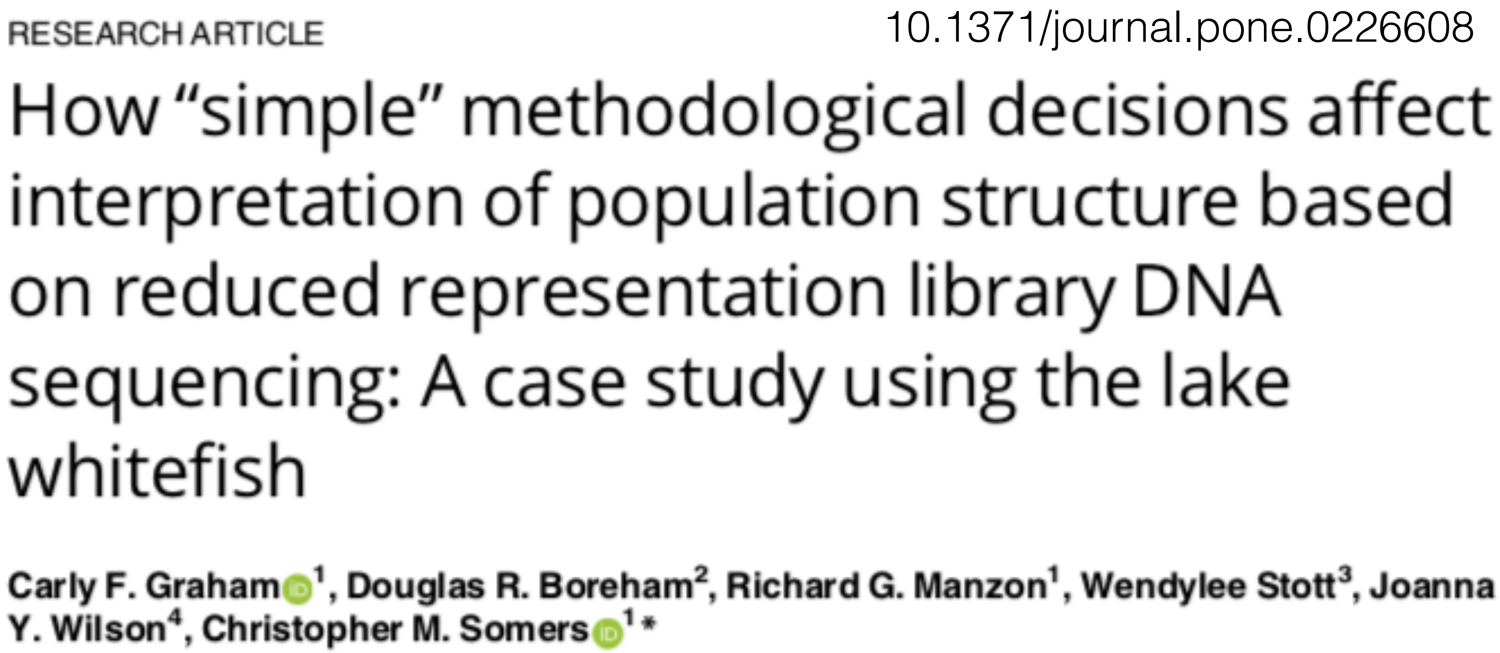
paralogs



intra-locus genetic variation

the trick in building a locus catalog is essentially to find the compromise between assembly of a large number of single-copy loci and few paralogous loci

Viricel et al 2014, MER

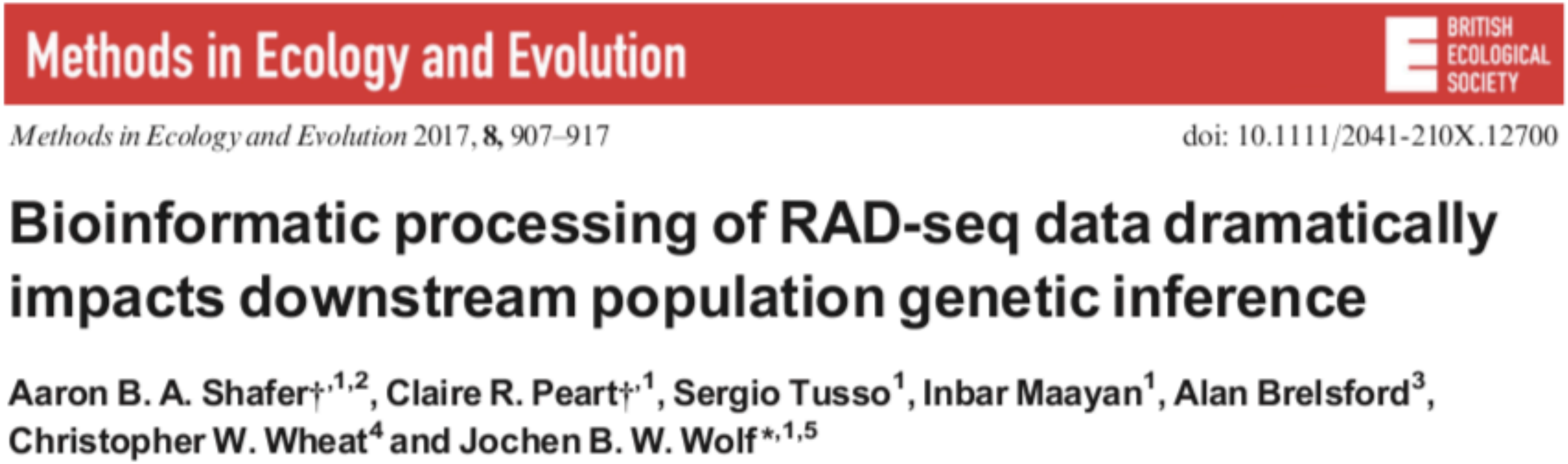


RESEARCH ARTICLE

10.1371/journal.pone.0226608

How “simple” methodological decisions affect interpretation of population structure based on reduced representation library DNA sequencing: A case study using the lake whitefish

Carly F. Graham[1], Douglas R. Boreham[2], Richard G. Manzon[1], Wendylee Stott[3], Joanna Y. Wilson[4], Christopher M. Somers[1]*



Methods in Ecology and Evolution

BRITISH
ECOLOGICAL
SOCIETY

Methods in Ecology and Evolution 2017, **8**, 907–917 doi: 10.1111/2041-210X.12700

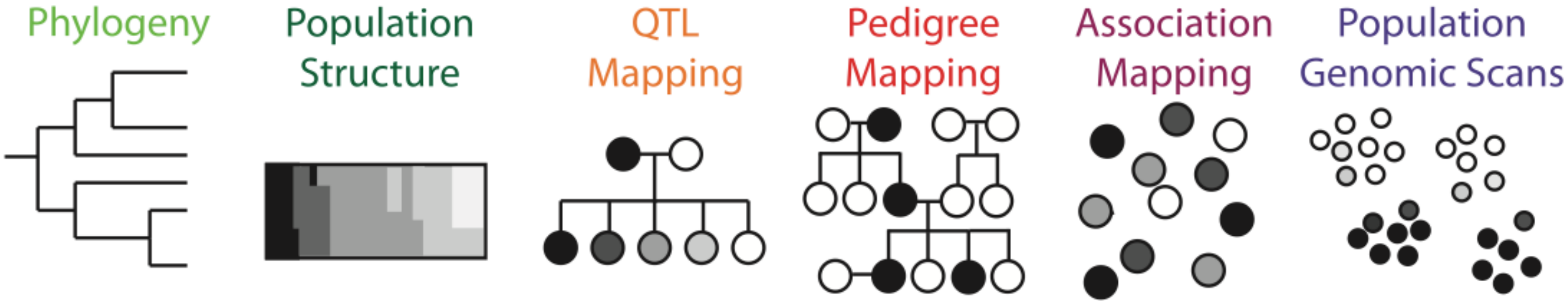
Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference

Aaron B. A. Shafer†[1,2], Claire R. Peart†[1], Sergio Tusso[1], Inbar Maayan[1], Alan Brelsford[3], Christopher W. Wheat[4] and Jochen B. W. Wolf*[1,5]

Applications in evolutionary biology:

- different scales : different problems linked to catalog assembly
  - depth of coverage on SNPs
  - linkage among SNPs
  - type I / II errors for genotyping
  - sequencing of coding vs non-coding regions

Today we will focus on issues linked to estimating population genetics parameters



Peterson et al (2012) PLoS ONE

Plan

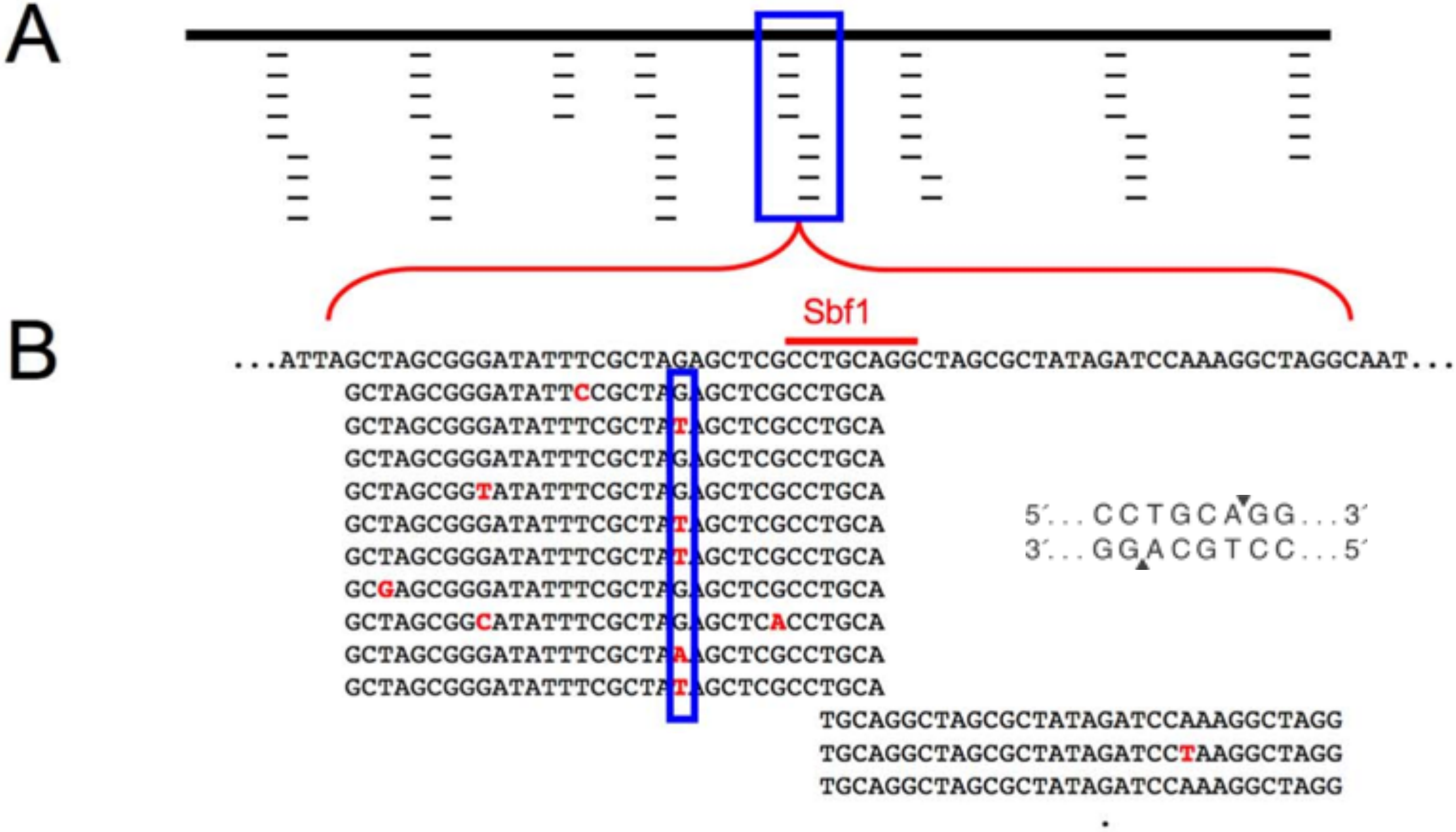
- Setting up your experiment
- Setting up your analysis pipeline
- Setting up a parameter selection strategy

Some difficulties with SNP genotyping

Source	Description	Références (e.g.)
Genome characteristics	GC content, genome size, genome architecture (duplications) polymorphism / methylation on restriction sites (locus dropout or mutation-disruption) …	Roberts et al (2010) Davey et al (2013) Gautier et al (2013)
Laboratory	quality of lab reagents, contamination, pipetting errors, enzyme sensitivity to DNA quality, equi-molarity of purified DNA samples, PCR bias / error / duplicates, library size selection…	Bonin et al (2004) Baird et al (2008), Peterson et al (2012) Hohenlohe et al (2012)
Sequencing	sequencing errors; preferential sequencing of alleles or loci (eg GC content, hairpins…)	Meachan et al (2011) Nielsen et al (2011) Hohenlohe et al (2012) Loman et al (2012)

Mastretta-Yanes et al (2014) Mol Ecol Res

a key step: **choice of restriction enzyme(s)** will affect the shape of the catalog : *nb of loci, locus depth, level of mutation-disruption*


A
B
Sbf1
...ATTAGCTAGCGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGCAAT...
GCTAGCGGGATATTCCGCTAGAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGTATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCGAGCGGGATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGCATATTTCGCTAGAGCTCACCTGCA
GCTAGCGGGATATTTCGCTAAAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
TGCAGGCTAGCGCTATAGATCCAAAGGCTAGG
TGCAGGCTAGCGCTATAGATCCTAAGGCTAGG
TGCAGGCTAGCGCTATAGATCCAAAGGCTAGG
5′...CCTGCAGG...3′
3′...GGACGTCC...5′

Hohenlohe et al, PLoS Genetics 2010

choice of restriction enzyme(s)

RADtag counter from GenePool, Edinburgh

To use this counter:

1 Enter the GC content of your target genome here: **0.4** proportion GC

2 Enter the size in megabases of your genome here: **2000** taille génome (Mb)

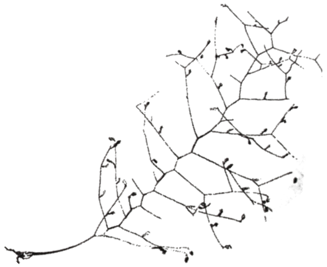
3 Enter the fold coverage of RADtags you require here: **30** couverture

4 Enter the per-pool plexity you plan to use here: **96** plexity

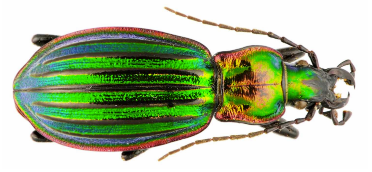
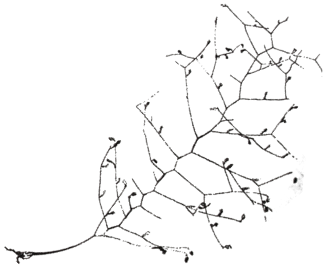
5 Enter number of million reads per lane
(please contact the GenePool for throughput currently achieved on the GAIIx and HiSeq platforms) **80** million reads

Overhang	**TGCA**			**GGCC**			**AATT**	
Enzyme	**SbfI**	**PstI**	**NsiI**	**NotI**	**EaeI**	**EagI**	**EcoRI**	**ApoI**
Site	CCTGCA*GG	CTGCA*G	ATGCA*T	GC*GGCCGC	Y*GGCCR	C*GGCCG	G*AATTC	R*AATTY
Site frequency	5.76E-06	0.000144	0.000324	2.56E-06	0.0004	0.000064	0.000324	0.002025
Sites/Mb	6	144	324	3	400	64	324	2025
Number of sites in genome	**11520**	**288000**	**648000**	**5120**	**800000**	**128000**	**648000**	**4050000**
Number of tags	23040	576000	1296000	10240	1600000	256000	1296000	8100000
Num sequences for coverage	691200	17280000	38880000	307200	48000000	7680000	38880000	243000000
Million sequences per pool	66.4	1658.9	3732.5	29.5	4608.0	737.3	3732.5	23328.0
does your pool fit in one lane?	**YES**	**NO**	**NO**	**YES**	**NO**	**NO**	**NO**	**NO**

choice of restriction enzyme(s)

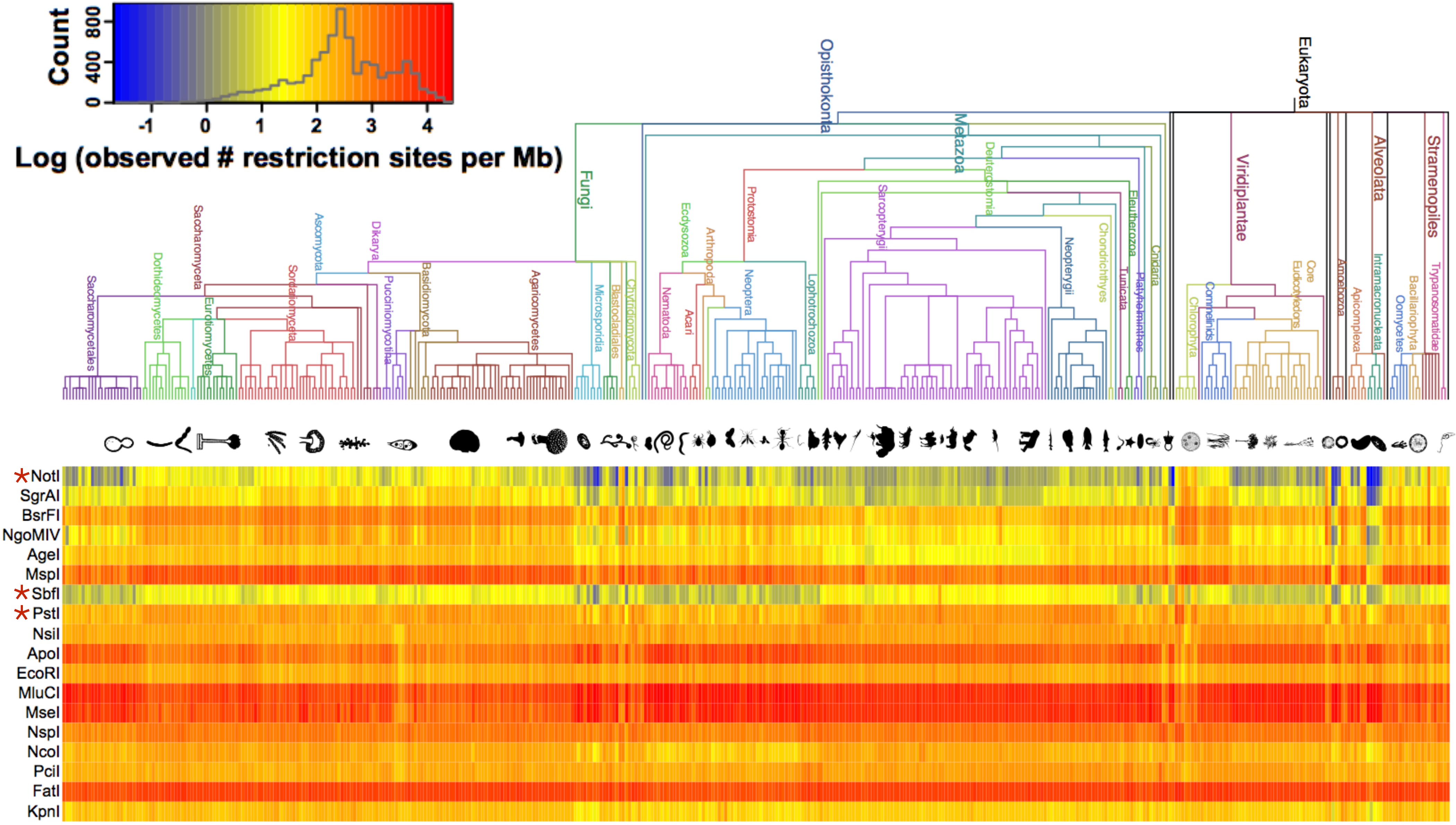
Model	Divergence level	Enzyme	Genome size	Expected coverage	Expected number of restriction sites	Multiplexing (nb. indiv)
	Beetles: Cruaud et al (2014), Mol Biol Evol					
	1-17 MY	*PstI*	300 MB	48x	49 068	31
	Dolphins: Viricel et al (2014), Mol Ecol Res					
	0-19 MY	*NotI*	3 GB	38x	10 714	92
	Corals: Pante et al (2014), Heredity					
	0-17 MY ?	*SbfI*	224 MB - 1.8 TB	30x	23 040	91

choice of restriction enzyme(s)

Model	Divergence level	Enzyme	Genome size	Expected coverage	Expected number of restriction sites	Multiplexing (nb. indiv)
	Beetles: Cruaud et al (2014), Mol Biol Evol					
	1-17 MY	*PstI*	300 MB	48x	49 068	31
	Dolphins: Viricel et al (2014), Mol Ecol Res					
	0-19 MY	*NotI*	3 GB	38x	10 714	92
	Corals: Pante et al (2014), Heredity					
	0-17 MY ?	*SbfI*	224 MB - 1.8 TB	30x	23 040	91

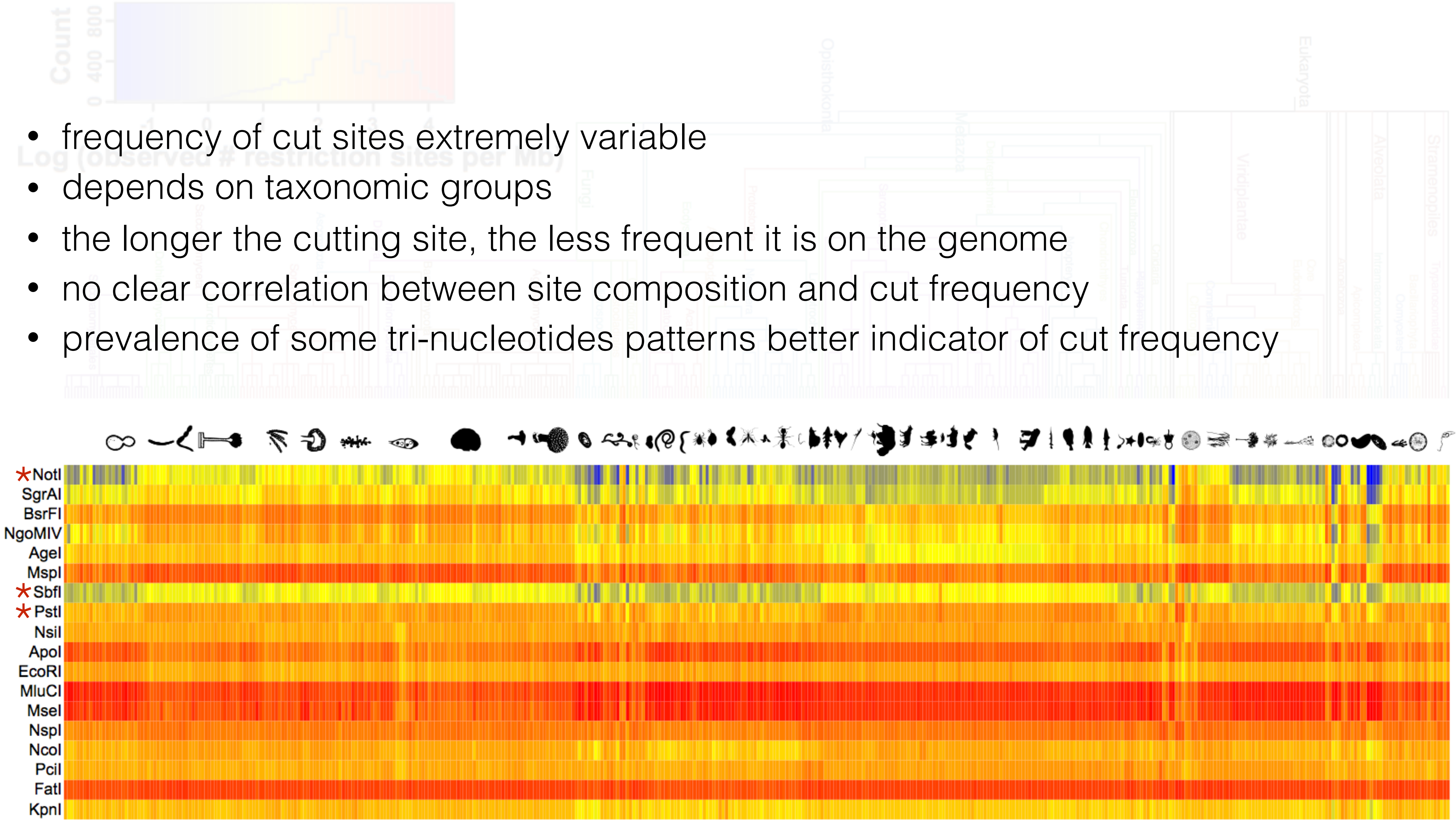
choice of restriction enzyme(s)

PredRAD: Herrera et al (2014) BioRxiv



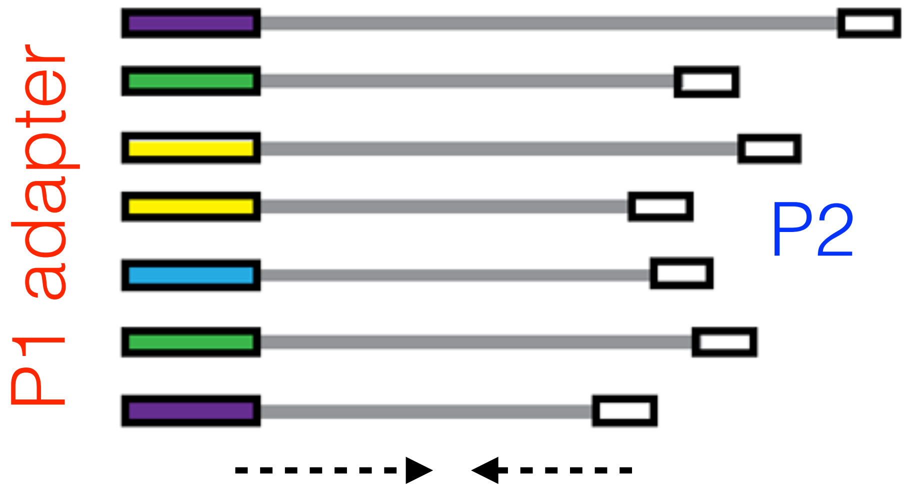
choice of restriction enzyme(s)

- frequency of cut sites extremely variable
- depends on taxonomic groups
- the longer the cutting site, the less frequent it is on the genome
- no clear correlation between site composition and cut frequency
- prevalence of some tri-nucleotides patterns better indicator of cut frequency



choice of sequencing platform and library construction strategy

"now-generation" sequencing of short fragments
(Illumina / SOLiD / Ion Torrent PGM)

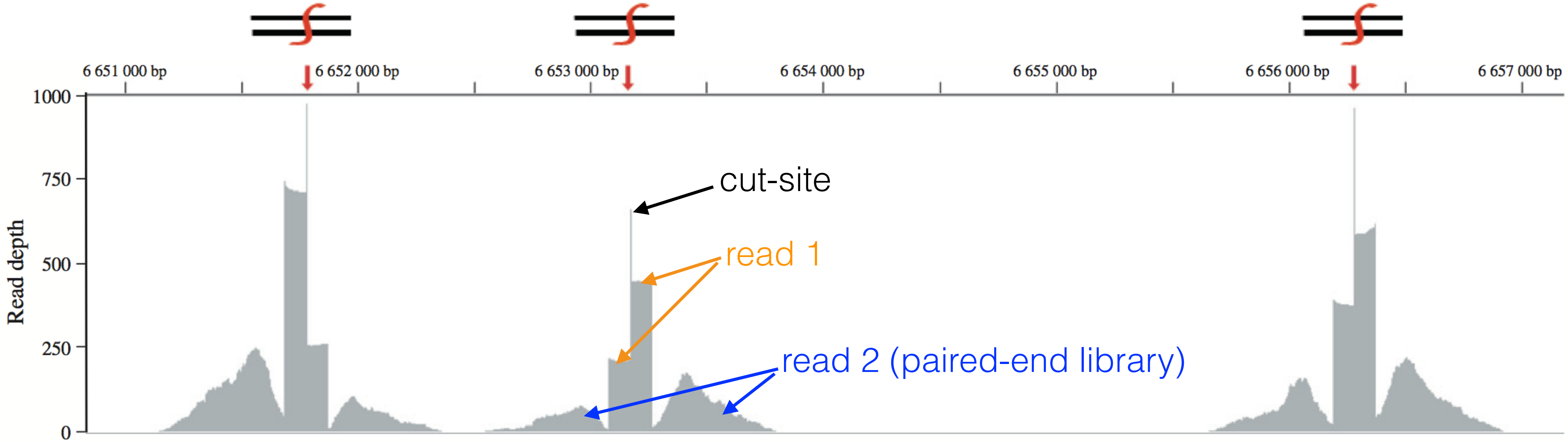


variable length of sequenced fragment:
~ 35 nt to 250 nt or more for contig'ed PE

Baird et al,
PLoS ONE 2008

choice of sequencing platform and library construction strategy

"now-generation" sequencing of short fragments (Illumina / SOLiD / Ion Torrent PGM)



nb RAD-tags = 2 x restriction sites

variable length and position of your R2 in PE experiments

Davey et al (2013) Mol Ecol

choice of sequencing platform and library construction strategy

Method	Strategy	Reference
sdRAD	the original ? use of 1 restriction enzyme, DNA shearing by sonication	Baird et al (2008)
ddRAD	coupling of 2 enzymes differing by their cutting frequencies	Peterson et al (2012)
2b-RAD	use of IIB type enzymes, cut DNA in small (33-36nt) fragments of uniform size	Wang et al (2012)
ezRAD	enzyme nb ≥ 1; simplified prep'; reduced cost (30 librairies < \$10K)	Toonen et al (2013)
BestRAD	uses biotinylated adapters to extract restriction site-adjacent DNA from gDNA early on in library prep	Ali et al (2016)

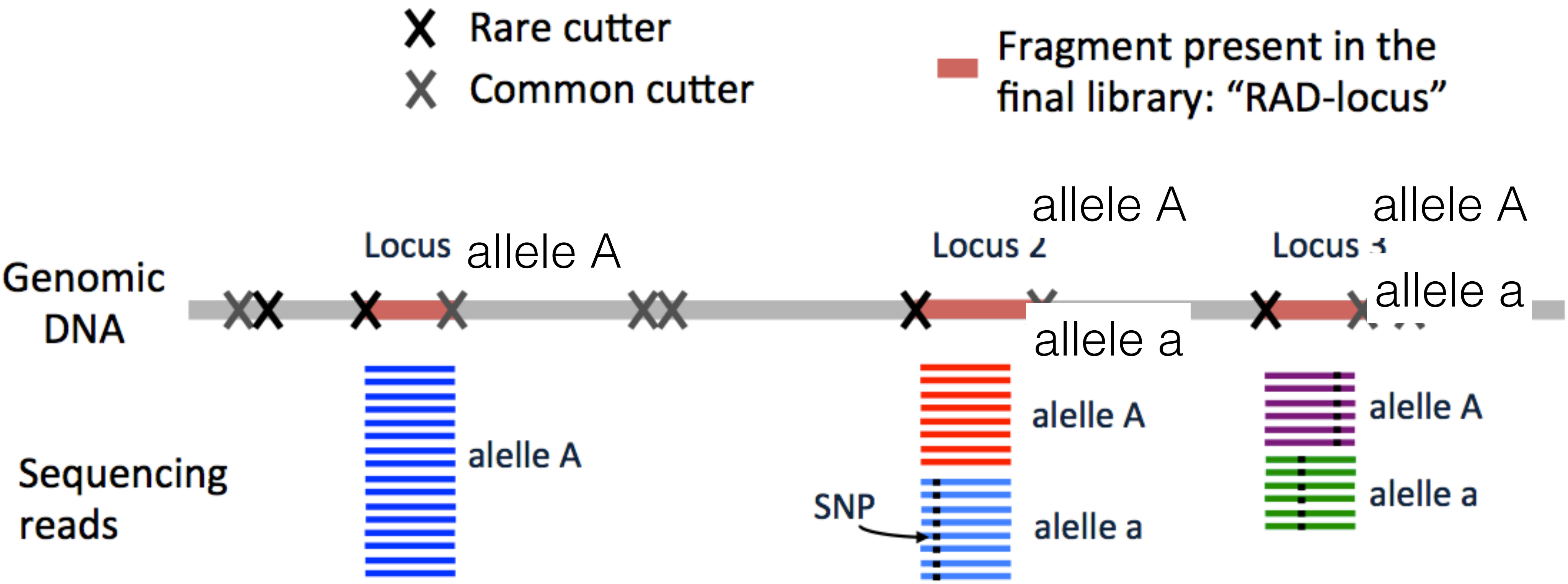
Many others: GBS, teGBS, RESTseq, RRLs, CRoPS, HyRAD …

Comparaison of methods:

Wang et al (2012) Nature Methods, Toonen et al (2013) PeerJ, Lepais et Weir (2014) Mol Ecol Res
Andrews et al (2016) Nat Rev Genet

choice of sequencing platform and library construction strategy

ddRAD (double-digest):
theory says : fewer but better-covered loci, compared to sdRAD

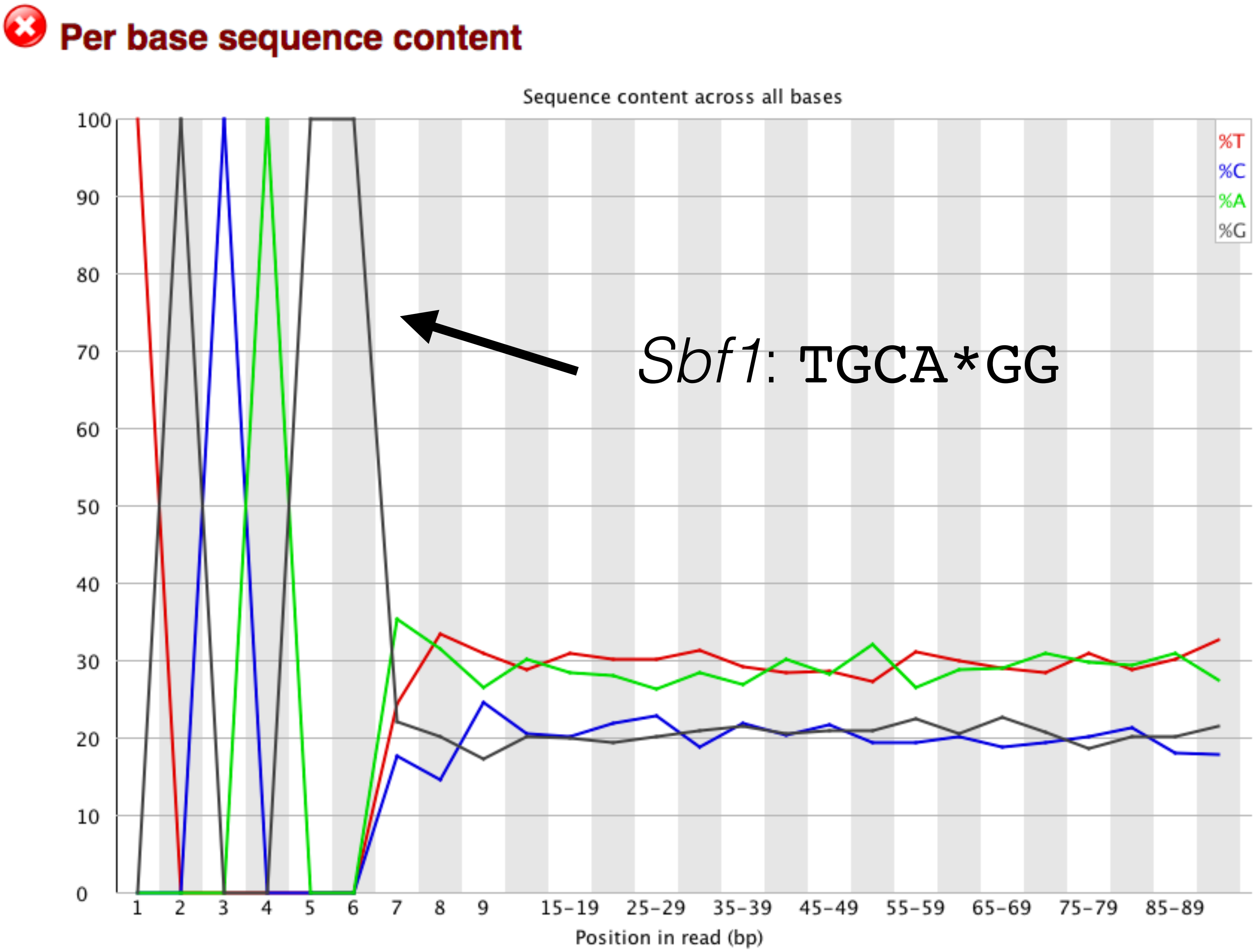


Peterson et al (2012) PLoS ONE (méthode)
Mastretta-Yanes et al (2014) Mol Ecol Res (figure)

Plan

- Setting up your experiment
- Setting up your analysis pipeline
- Setting up a parameter selection strategy

QC is paramount! remember, GIGO :-)



Per base sequence content

Sbf1: `TGCA*GG`

Analysis tools

some redundancy: RADIS relies on STACKS, dDocent relies on Rainbow …

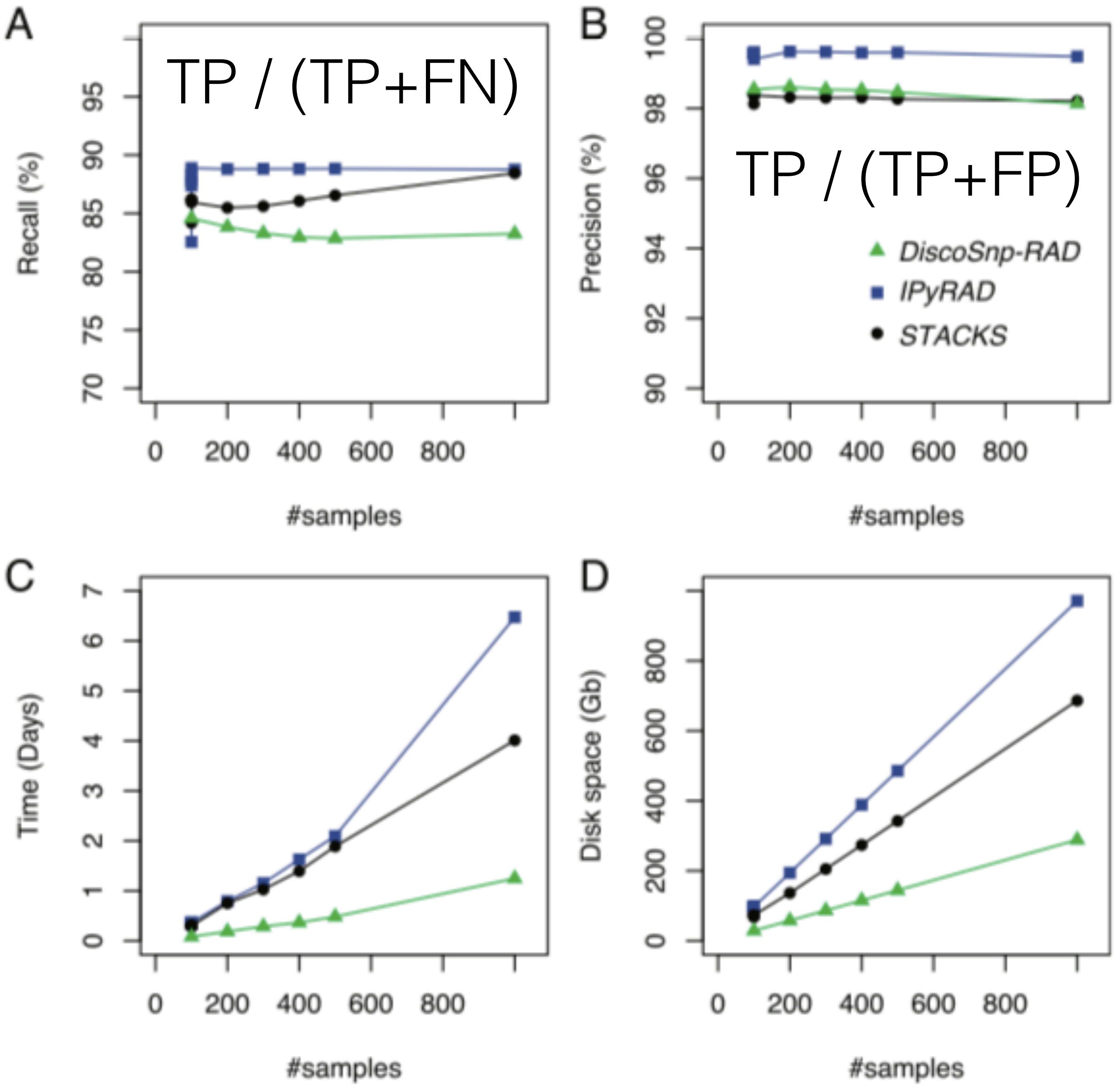
additional “generalist” tools that can be applied to RAD data: GSsnap, GATK, BWA, Stampy, SAMtools, iML …

toul	iouse	authors	pub yr	language	GUI	DOI (or other source)
stacks	RAD pipeline	Catchen et al	2011	C / perl	yes	10.1534/g3.111.000240
RADtools	RAD pipeline	Baxter et al	2011	perl	no	10.1371/journal.pone.0019315
RApiD	RAD pipeline	Willing et al	2011	C / perl	no	10.1093/bioinformatics/btr346
rtd	ddRAD pipeline	Petterson et al	2012	python	no	10.1371/journal.pone.0037135
Rainbow	RAD pipeline	Chong et al	2012	C / perl	no	10.1093/bioinformatics/bts482
RADtyping	linkage maps	Fu et al	2013	perl	no	10.1371/journal.pone.0079960
PyRAD	RAD pipeline	Eaton	2014	python	no	10.1093/bioinformatics/btu121
RADami	RAD tools	Hipp et al	2014	R	no	10.1371/journal.pone.0093975
PredRAD	enzyme choice	Herrera et al	2014	python	no	10.1093/gbe/evv210
dDocent	ddRAD pipeline	Puritz et al	2014	bash	no	10.7717/peerj.431
SimRAD	RAD simulation	Lepais & Weir	2014	R	no	10.1111/1755-0998.12273
aftrRAD	RAD pipeline	Sovic et al	2015			10.1111/1755-0998.12378
HotRAD	RAD pipeline	Assour et al	2015			arXiv:1511.06754
RADIS	RAD wrap-up	Cruaud	2016	perl	no	10.1093/bioinformatics/btw352
simrrls	RAD simulation	Eaton	2016	python	no	github.com/dereneaton/simrrls
RADProc	RAD pipeline	Ravindran et al	2018			10.1111/1755-0998.12954
stacks2	RAD pipeline	Rochette et al	2019			10.1111/mec.15253
ipyrad 😻	RAD pipeline	Eaton & Overcast	2020	python	no	10.1093/bioinformatics/btz966
RADinitio	RAD simulation	Rivera-Colon et al	2020	python		10.1111/1755-0998.13163
DiscoSnp-RAD	RAD pipeline	Gauthier et al	2020		no	10.7717/peerj.9291

When shopping for a pipeline …

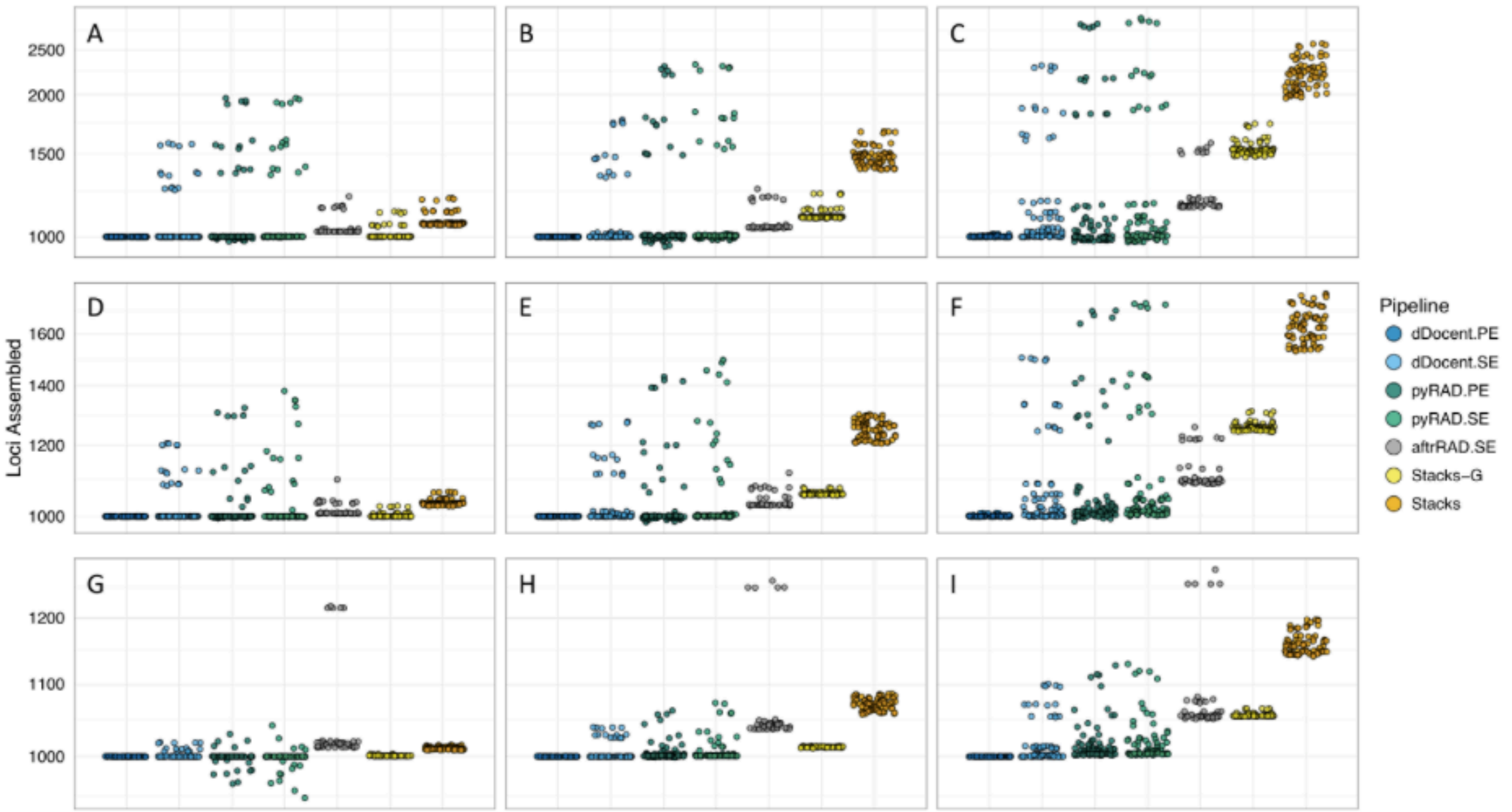
Compare:

1. strategy
2. precision
3. recall
4. time
5. HD space



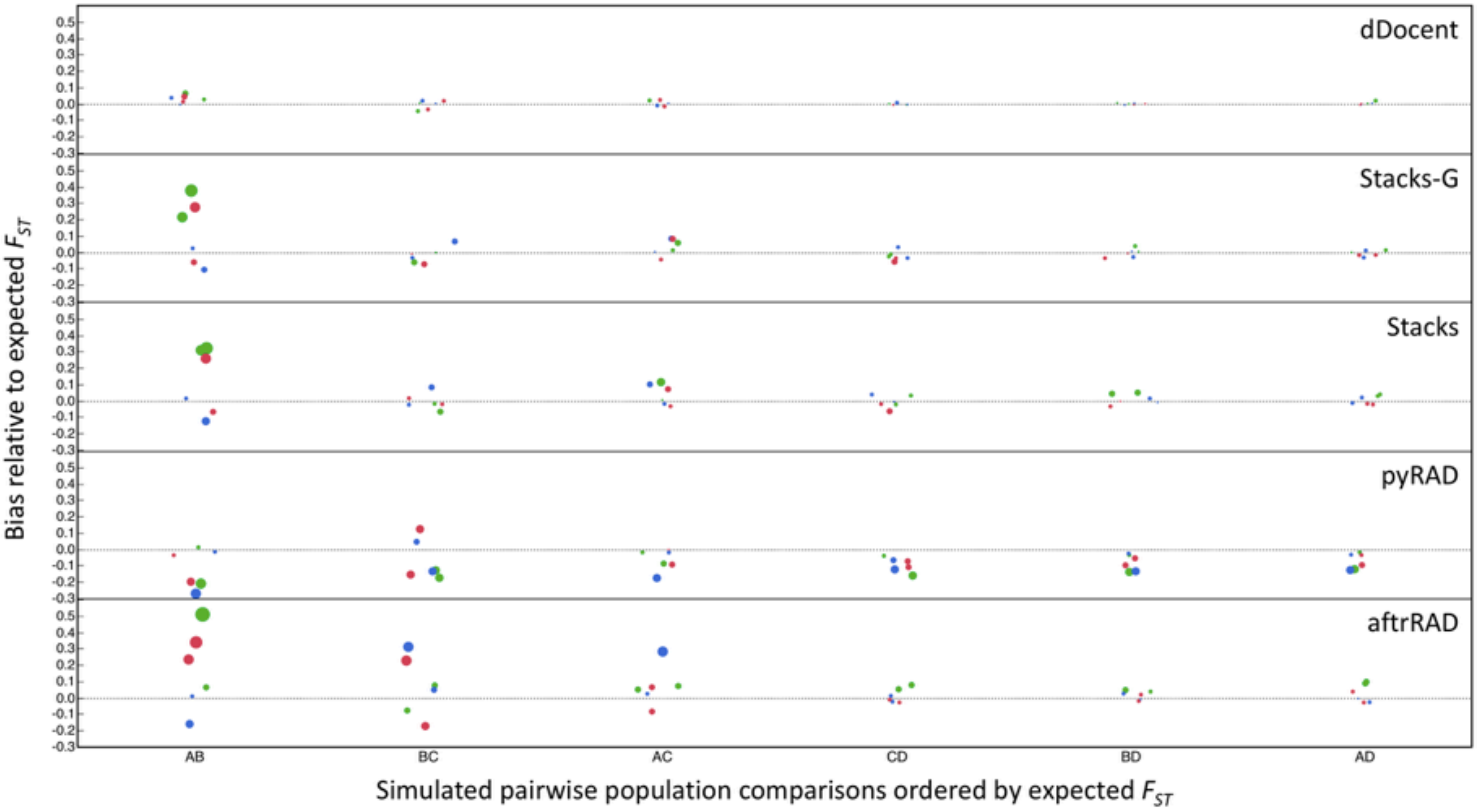
Gauthier et al 2020 PeerJ

When shopping for a pipeline …



https://www.ddocent.com/why/

When shopping for a pipeline …

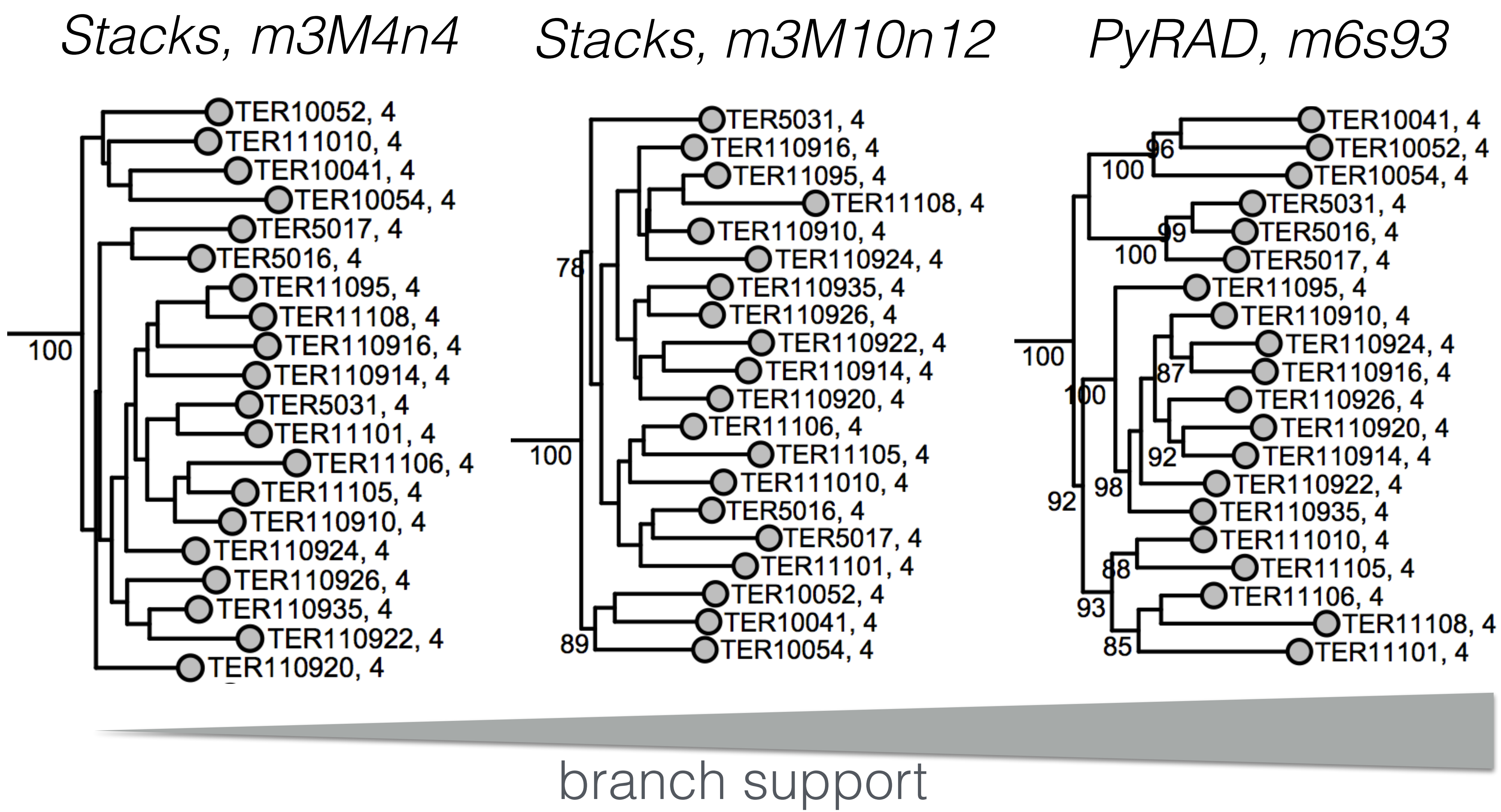


https://www.ddocent.com/why/

Plan

- Setting up your experiment
- Setting up your analysis pipeline
- Setting up a parameter selection strategy

Different filters, different results ?

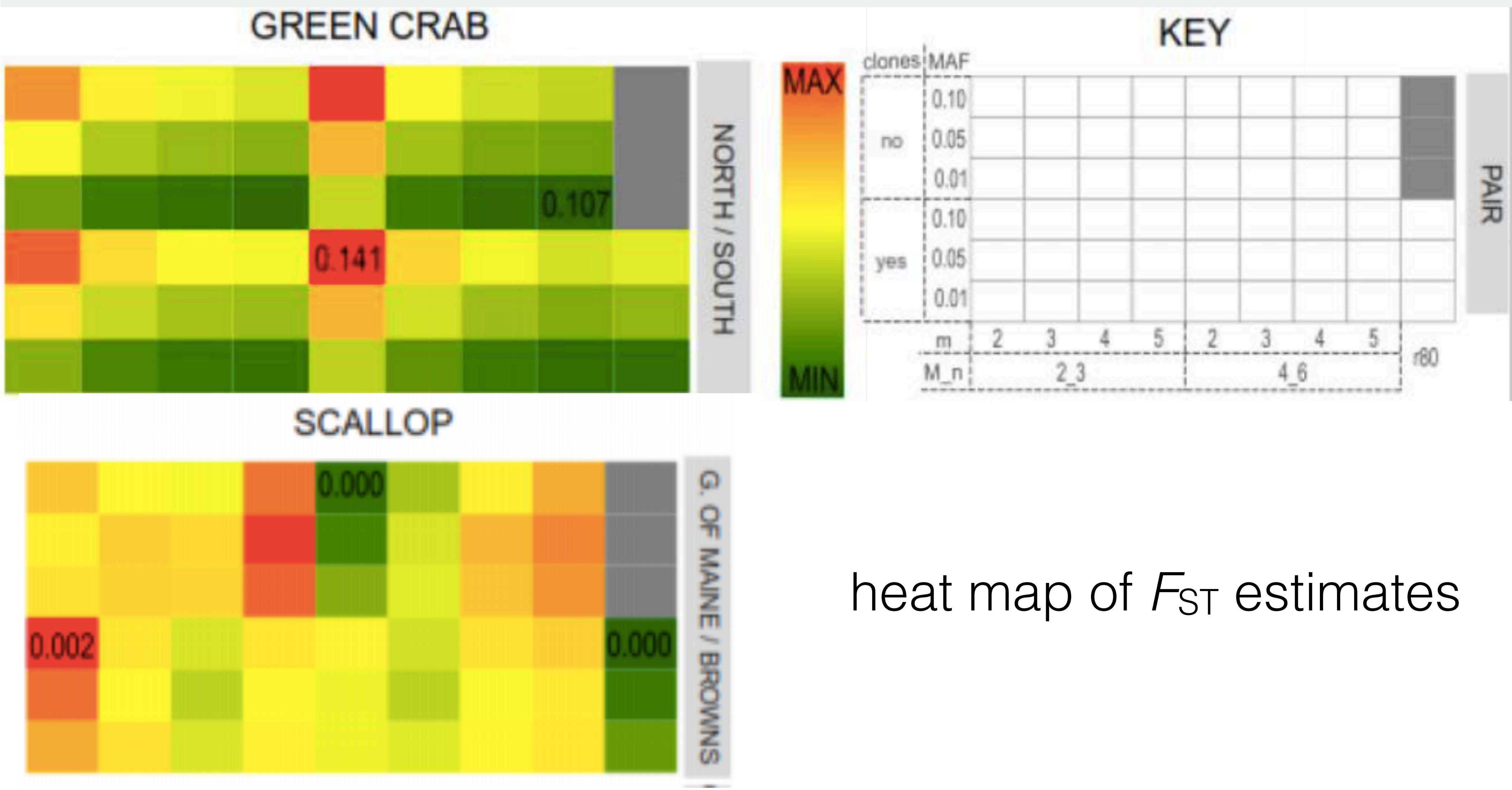


Pante et al (2014) Heredity

Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better?

Natalia Díaz-Arce and Naiara Rodríguez-Ezpeleta*

Marine Research Division, AZTI, Sukarrieta, Spain



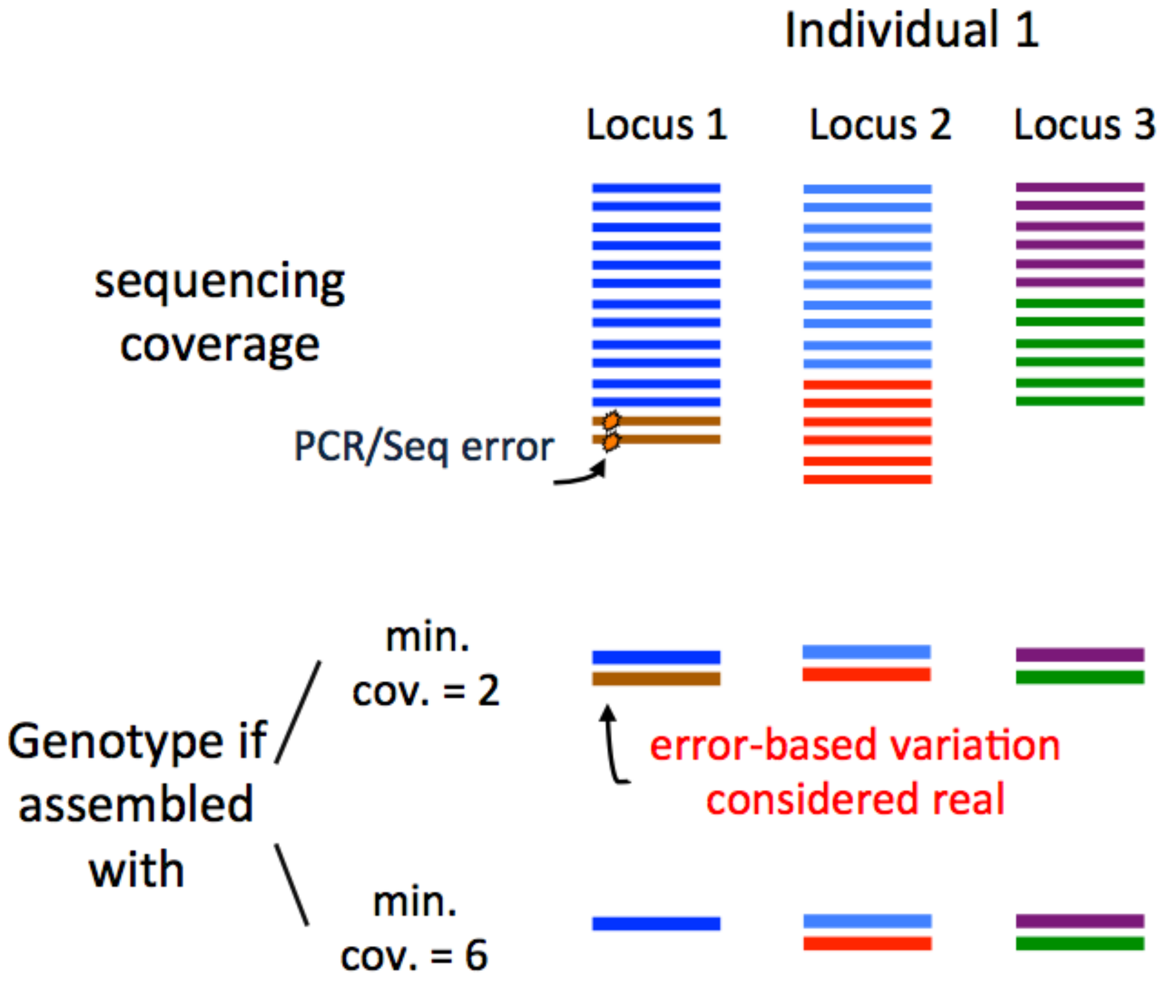
heat map of F_{ST} estimates

Some bioinformatic challenges associated with RAD data

Source	Description	Références (e.g.)
Depth of coverage (DC)	DC threshold too low: genotyping errors DC threshold too high: allele drop-out	Davey et al (2013) Hohenlohe et al (2012) Catchen et al (2013)
PCR duplicates	DC heterogeneous due to overrepresentation of some sequences	Davey et al (2013)
Fragment length	allele / locus drop-out decreases with increasing fragment length	Davey et al (2013)
Repeated regions and paralogs	des séquences similaires, mais non homologues peuvent être assemblées pour former des loci artificiels	Hohenlohe et al (2012) Dou et al (2012)
Indels (insertions / deletions)	some pipelines take them into account (*RApiD*, *PyRAD*), others do not (*Stacks*, *RADtools)*	Peterson et al (2012) Davey et al (2013)
Divergence and reference genome (RG)	the less alleles are divergent from the RG, the more likely they are to be included in the catalog	Pool et al (2010)

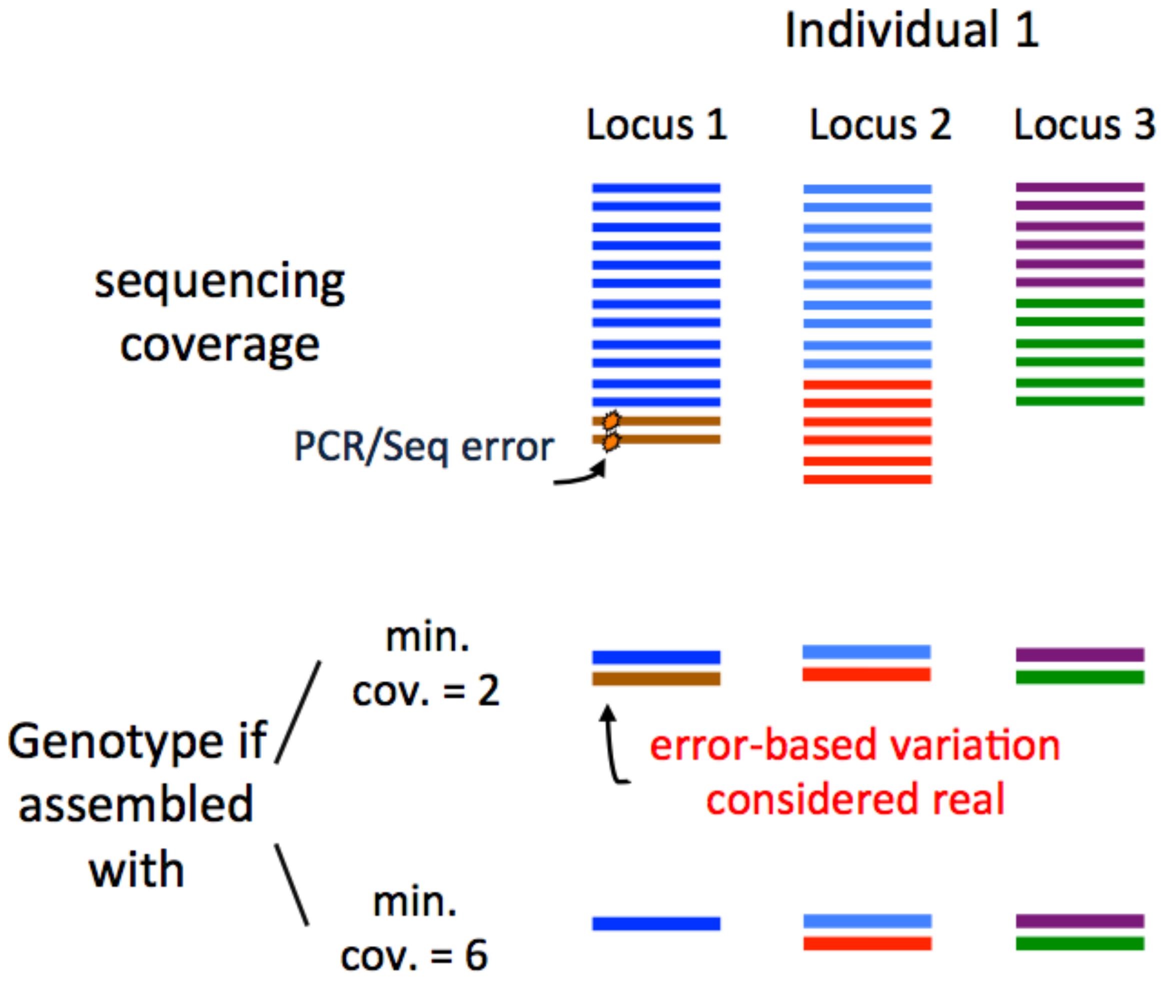
Mastretta-Yanes et al (2014) Mol Ecol Res

Difficultés bioinformatiques : sources d'erreurs de génotypage



Mastretta-Yanes et al (2014) Mol Ecol Res

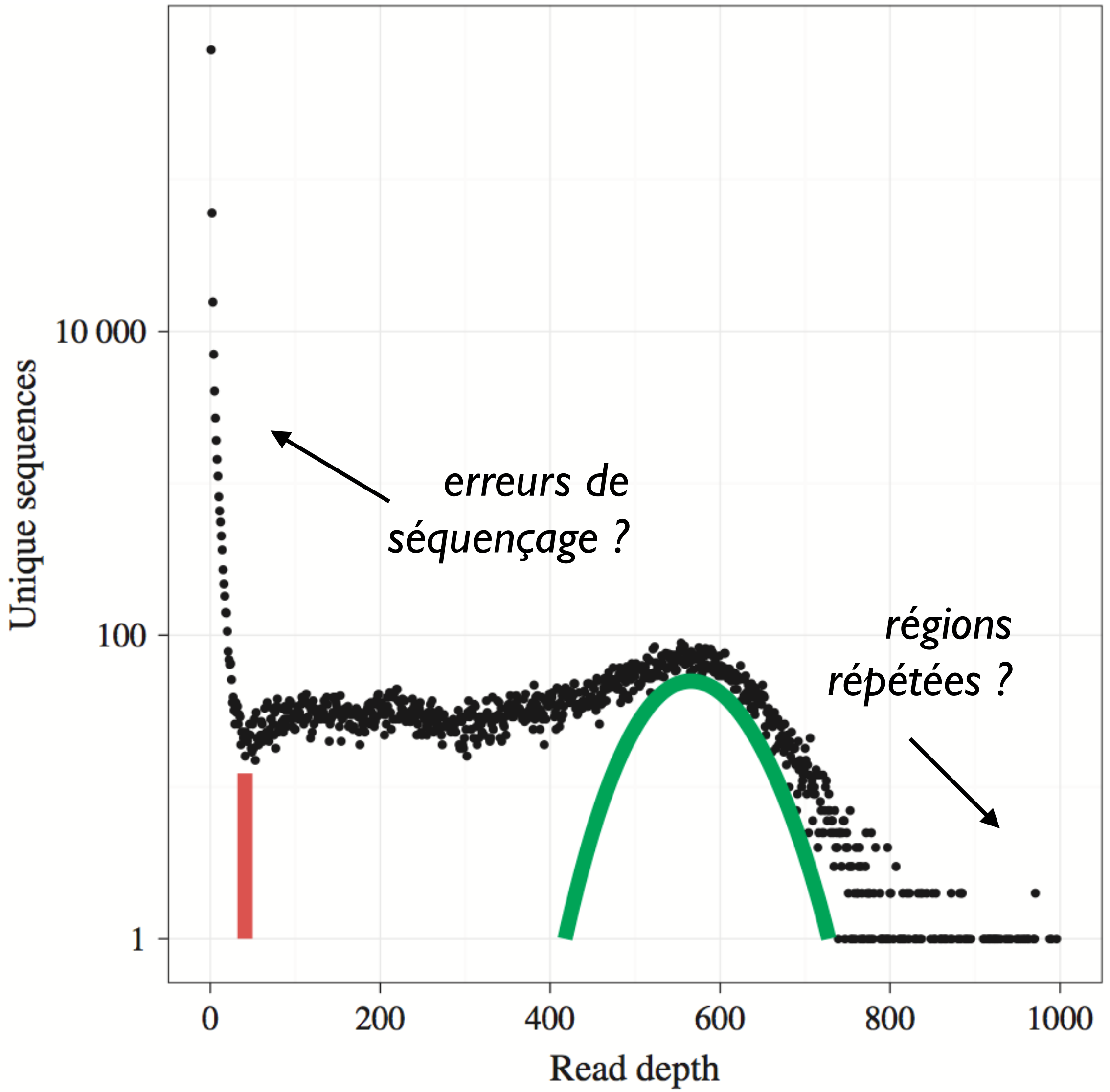
Difficultés bioinformatiques : sources d'erreurs de génotypage



minimum couverture too low inflates error rate

Mastretta-Yanes et al (2014) Mol Ecol Res

Difficultés bioinformatiques : distribution de séquences uniques

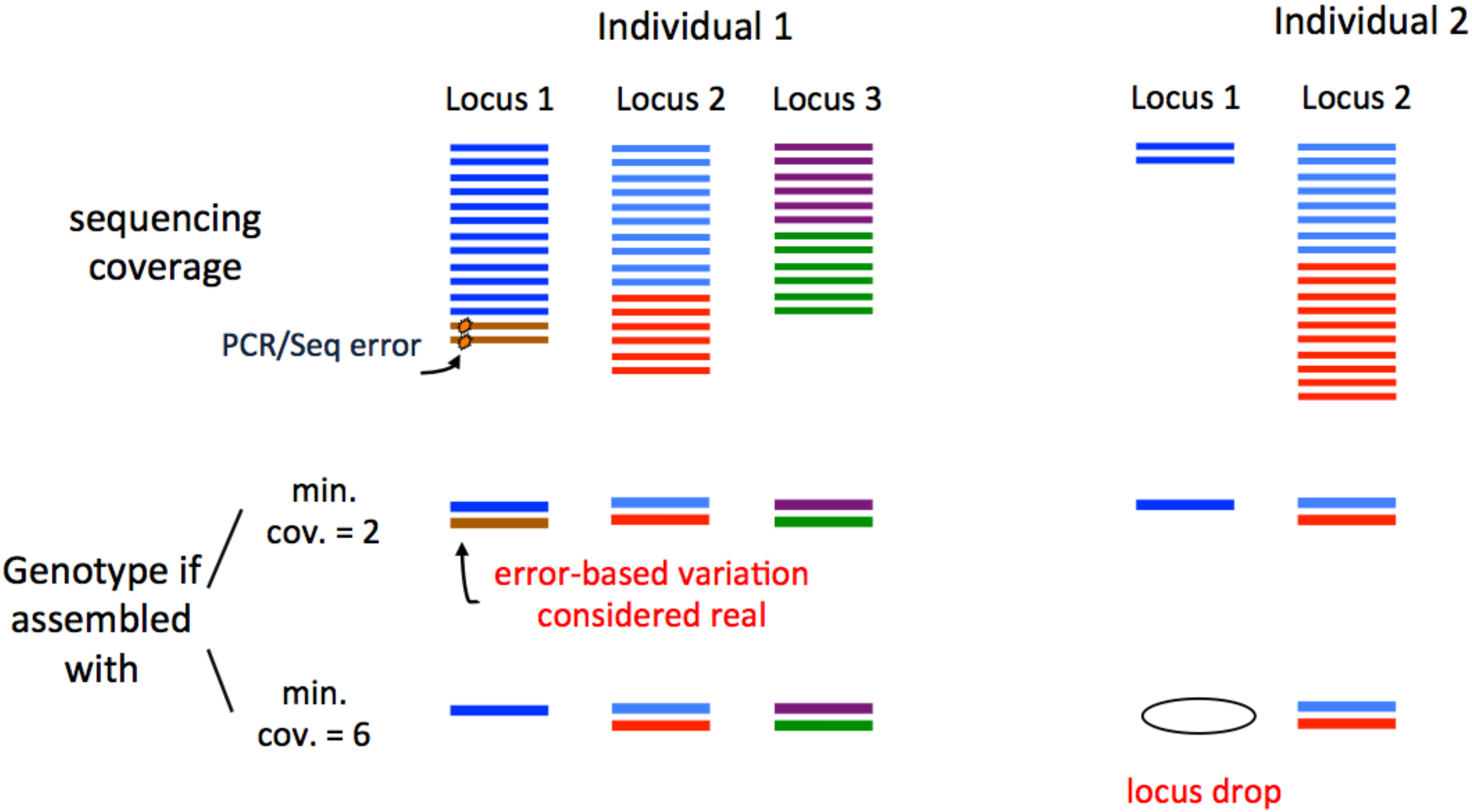


erreurs de séquençage ?

régions répétées ?

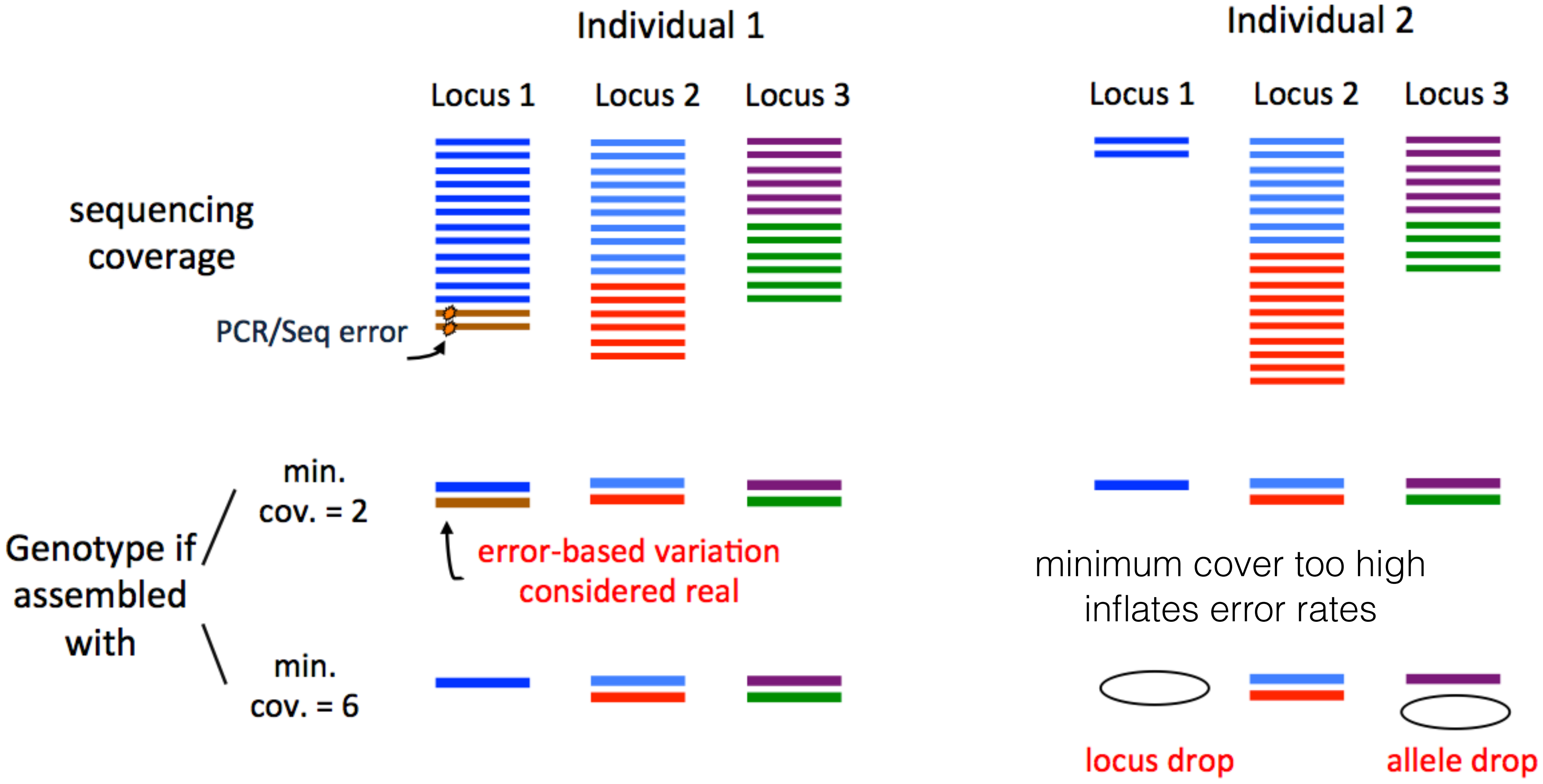
Davey et al (2013) Mol Ecol

Difficultés bioinformatiques : sources d'erreurs de génotypage



Mastretta-Yanes et al (2014) Mol Ecol Res

Difficultés bioinformatiques : sources d'erreurs de génotypage



Mastretta-Yanes et al (2014) Mol Ecol Res

The nitty-gritty of catalog building

Some published recommendations for optimising catalog building

authors	yr	DOI	data	pipeline	take home message(s)
Mastretta-Yanes et al	2015	10.1111/1755-0998.12291	plant	stacks	importance of biological and technical replicates to compute genotyping error rate and optimise parameter values
McCartney-Melstad et al	2017	10.1111/1755-0998.13029	frog	pyrad	"we develop a novel set of metrics to determine sequence similarity thresholds that maximize the correct separation of paralogous regions and minimize oversplitting naturally occurring allelic variation within loci."
Paris et al	2017	10.1111/2041-210X.12775	trout / penguin / earthworms	stacks	the 80% rule as a generally effective method to select the core parameters for STACKS.
Shafer et al	2017	10.1111/2041-210X.12700	sea lions	stacks / pyrad / ddocent	"We recommend that RAD-seq studies employ reference-based approaches to a closely related genome, and due to the high stochasticity associated with the pipeline advocate the use of multiple pipelines to ensure robust population genetic and demographic inferences."
Diaz-Arce et al	2019	10.3389/fgene.2019.00533	crab / mackerel / scallop	stacks	"(i) recovery of higher numbers of polymorphic loci is not necessarily associated with higher genetic differentiation, (ii) that the presence of PCR duplicates, selected loci assembly parameters and selected SNP filtering parameters affect the number of recovered polymorphic loci and degree of genetic differentiation, and (iii) that this effect is different in each dataset, meaning that defining a systematic universal protocol for RAD-seq data analysis may lead to missing relevant information about population differentiation."
Graham et al	2020	10.1371/journal.pone.0226608	lake whitefish	stacks	""simple" methodological decisions with caution, especially when working on non-model species"

y-gritty of catalog building

All are based on empirical datasets, which are intrinsically different

Some pub hed recommendations for optimising catalog building

authors	yr	DOI	ata	pipeline	take home message(s)
Mastretta-Yanes et al	2015	10.1111/1755-0998.12291	plant	stacks	importance of biological and technical replicates to compute genotyping error rate and optimise parameter values
McCartney-Melstad et al	2017	10.1111/1755-0998.13029	frog	pyrad	"we develop a novel set of metrics to determine sequence similarity thresholds that maximize the correct separation of paralogous regions and minimize oversplitting naturally occurring allelic variation within loci."
Paris et al	2017	10.1111/2041-210X.12775	trout / penguin / earthworms	stacks	the 80% rule as a **generally** effective method to select the core parameters for STACKS.
Shafer et al	2017	10.1111/2041-210X.12700	sea lions	stacks / pyrad / ddocent	"We recommend that RAD-seq studies employ reference-based approaches to a closely related genome, and due to the high stochasticity associated with the pipeline advocate the use of multiple pipelines to ensure robust population genetic and demographic inferences."
Diaz-Arce et al	2019	10.3389/fgene.2019.00533	crab / mackerel / scallop	stacks	"(i) recovery of higher numbers of polymorphic loci is not necessarily associated with higher genetic differentiation, (ii) that the presence of PCR duplicates, selected loci assembly parameters and selected SNP filtering parameters affect the number of recovered polymorphic loci and degree of genetic differentiation, and (iii) that this effect is different in each dataset, meaning that defining a systematic universal protocol for RAD-seq data analysis may lead to missing relevant information about population differentiation."
Graham et al	2020	10.1371/journal.pone.0226608	lake whitefish	stacks	""simple" methodological decisions with caution, especially when working on non-model species"

The nitty-gritty of catalog building

many focus on stacks but pipelines have intrinsic differences

d recommendations for optimising catalog building

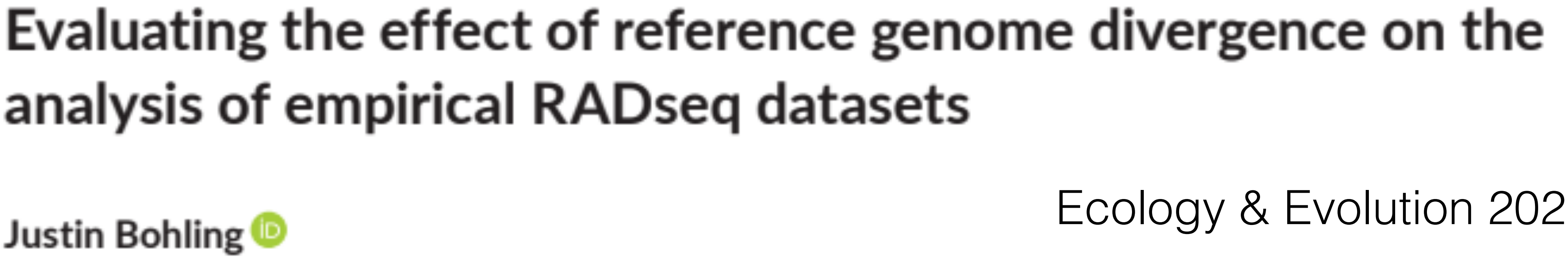
authors	yr	DOI	data	pipeline	take home message(s)
Mastretta-Yanes et al	2015	10.1111/1755-0998.12291	plant	stacks	importance of biological and technical replicates to compute genotyping error rate and optimise parameter values
McCartney-Melstad et al	2017	10.1111/1755-0998.13029	frog	pyrad	"we develop a novel set of metrics to determine sequence similarity thresholds that maximize the correct separation of paralogous regions and minimize oversplitting naturally occurring allelic variation within loci."
Paris et al	2017	10.1111/2041-210X.12775	trout / penguin / earthworms	stacks	the 80% rule as a generally effective method to select the core parameters for STACKS.
Shafer et al	2017	10.1111/2041-210X.12700	sea lions	stacks / pyrad / ddocent	"We recommend that RAD-seq studies employ reference-based approaches to a closely related genome, and due to the high stochasticity associated with the pipeline **advocate the use of multiple pipelines** to ensure robust population genetic and demographic inferences."
Diaz-Arce et al	2019	10.3389/fgene.2019.00533	crab / mackerel / scallop	stacks	"(i) recovery of higher numbers of polymorphic loci is not necessarily associated with higher genetic differentiation, (ii) that the presence of PCR duplicates, selected loci assembly parameters and selected SNP filtering parameters affect the number of recovered polymorphic loci and degree of genetic differentiation, and (iii) that this effect is different in each dataset, meaning that defining a systematic universal protocol for RAD-seq data analysis may lead to missing relevant information about population differentiation."
Graham et al	2020	10.1371/journal.pone.0226608	lake whitefish	stacks	""simple" methodological decisions with caution, especially when working on non-model species"

using a reference genome are not always available, and even so, is not systematically a good thing

gritty of catalog building

ecommendations for optimising catalog building

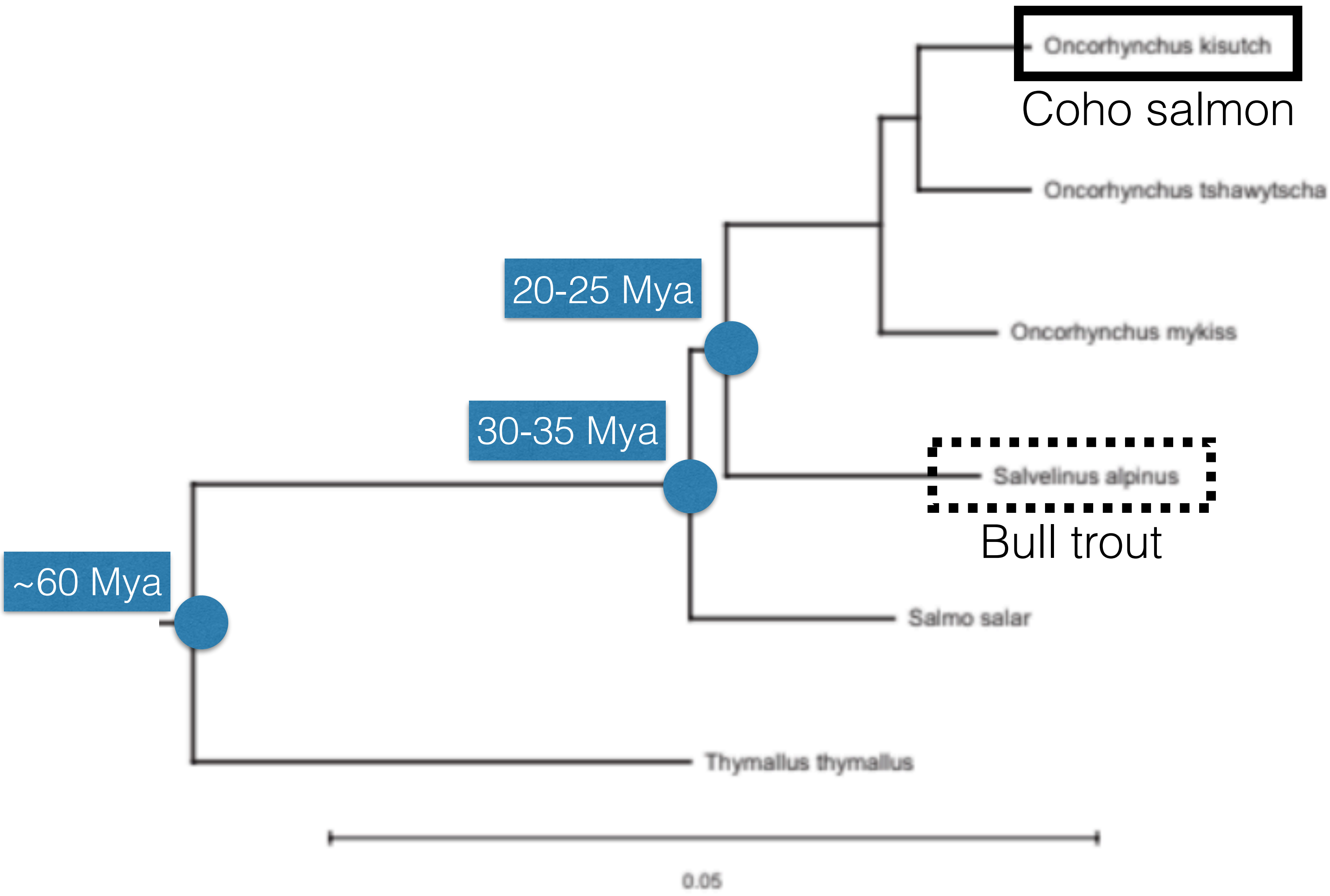
authors	yr	DOI	data	pipeline	take home message(s)
Mastretta-Yanes et al	2015	10.1111/1755-0998.12291	plant	stacks	importance of biological and technical replicates to compute genotyping error rate and optimise parameter values
McCartney-Melstad et al	2017	10.1111/1755-0998.13029	frog	pyrad	"we develop a novel set of metrics to determine sequence similarity thresholds that maximize the correct separation of paralogous regions and minimize oversplitting naturally occurring allelic variation within loci."
Paris et al	2017	10.1111/2041-210X.12775	trout / penguin / earthworms	stacks	the 80% rule as a generally effective method to select the core parameters for STACKS.
Shafer et al	2017	10.1111/2041-210X.12700	sea lions	stacks / pyrad / ddocent	"**We recommend that RAD-seq studies employ reference-based approaches to a closely related genome**, and due to the high stochasticity associated with the pipeline advocate the use of multiple pipelines to ensure robust population genetic and demographic inferences."
Diaz-Arce et al	2019	10.3389/fgene.2019.00533	crab / mackerel / scallop	stacks	"(i) recovery of higher numbers of polymorphic loci is not necessarily associated with higher genetic differentiation, (ii) that the presence of PCR duplicates, selected loci assembly parameters and selected SNP filtering parameters affect the number of recovered polymorphic loci and degree of genetic differentiation, and (iii) that this effect is different in each dataset, meaning that defining a systematic universal protocol for RAD-seq data analysis may lead to missing relevant information about population differentiation."
Graham et al	2020	10.1371/journal.pone.0226608	lake whitefish	stacks	""simple" methodological decisions with caution, especially when working on non-model species"



Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets

Justin Bohling

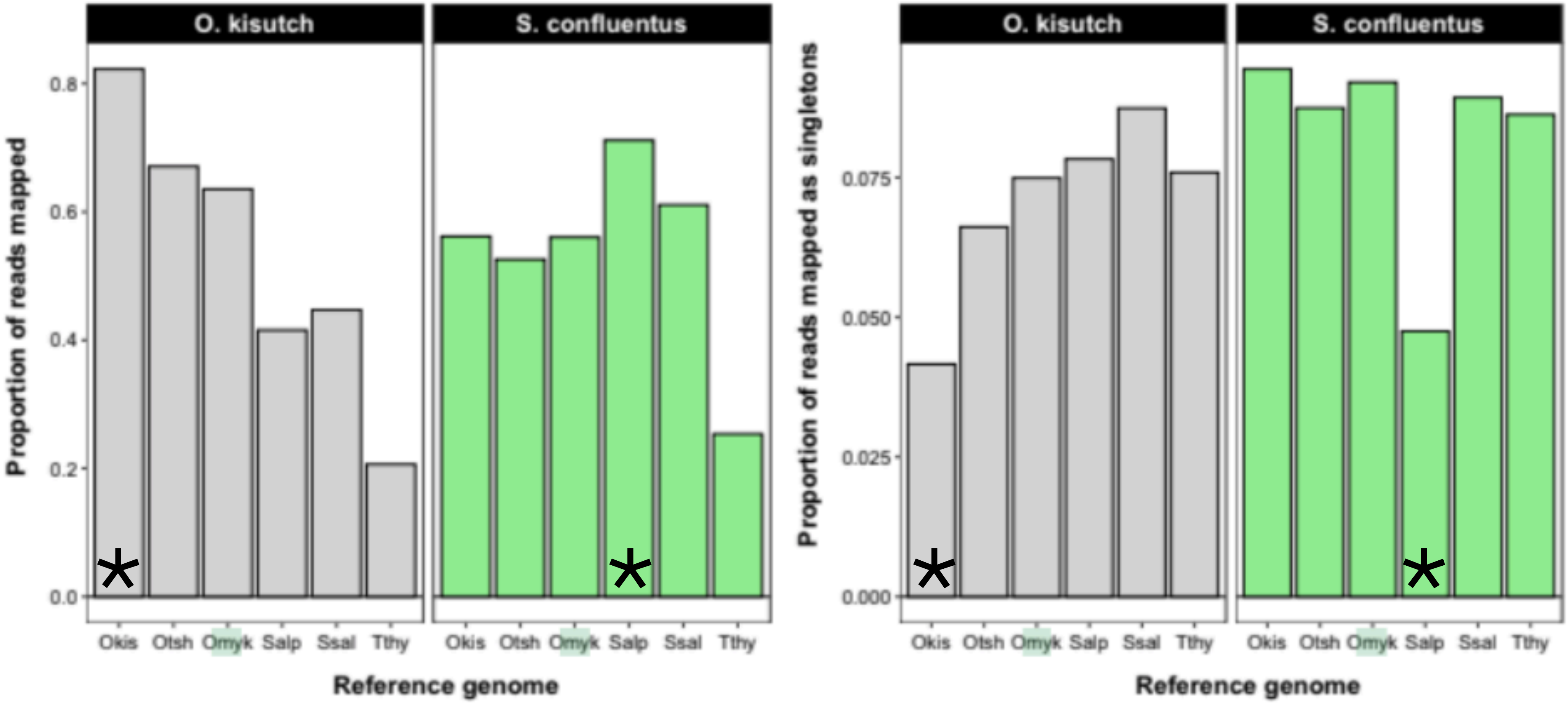
Ecology & Evolution 2020

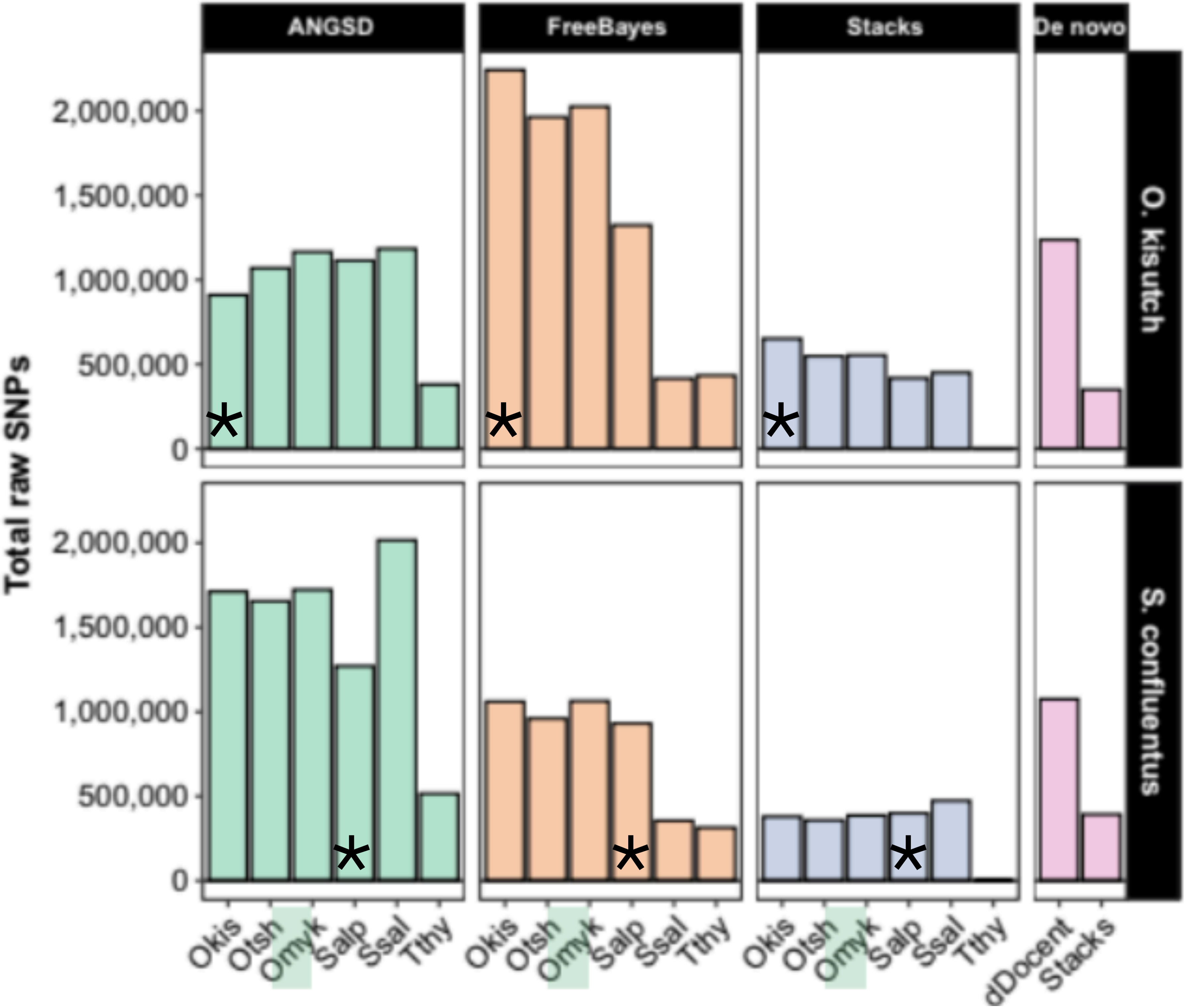


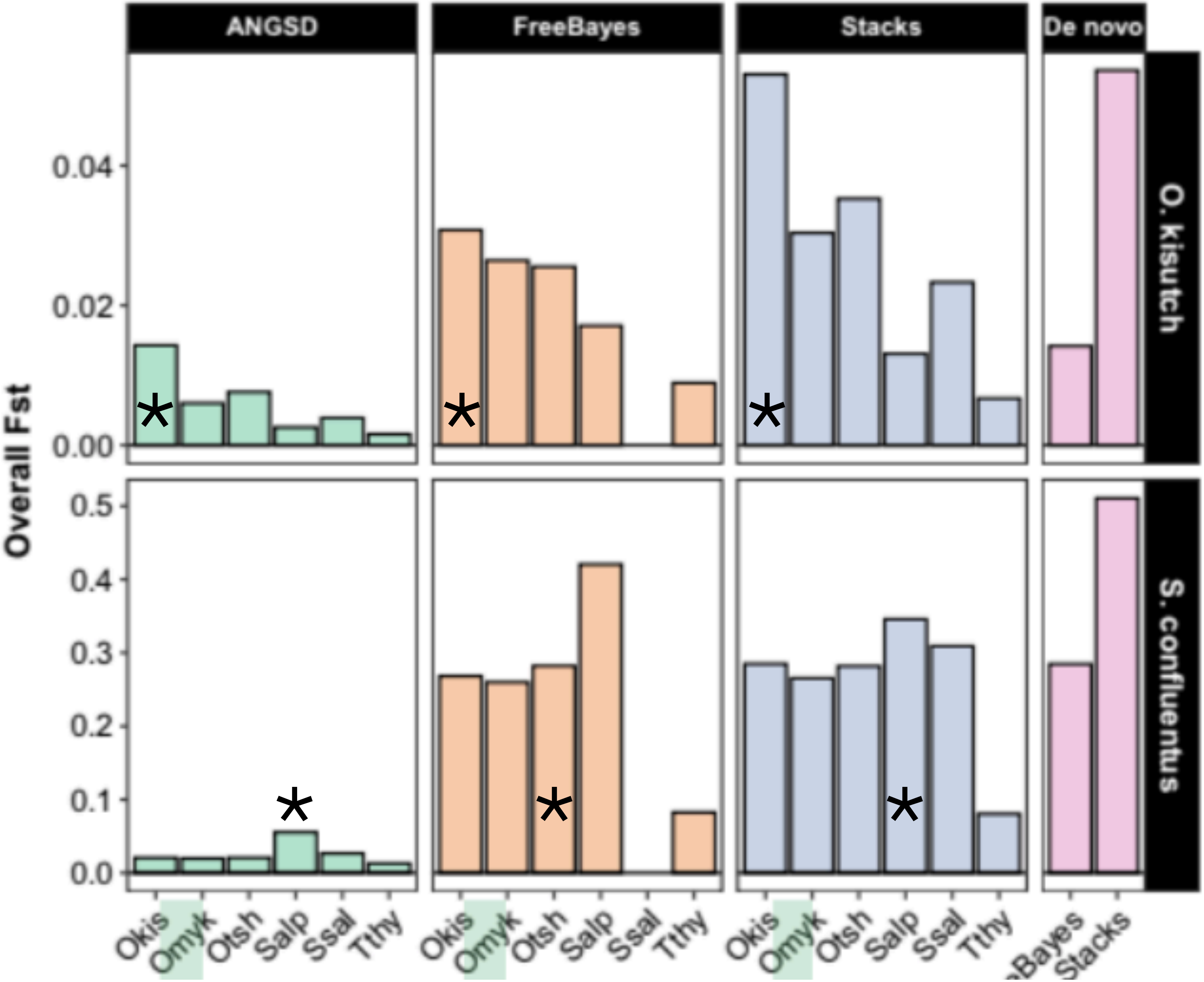
Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets

Justin Bohling

Ecology & Evolution 2020



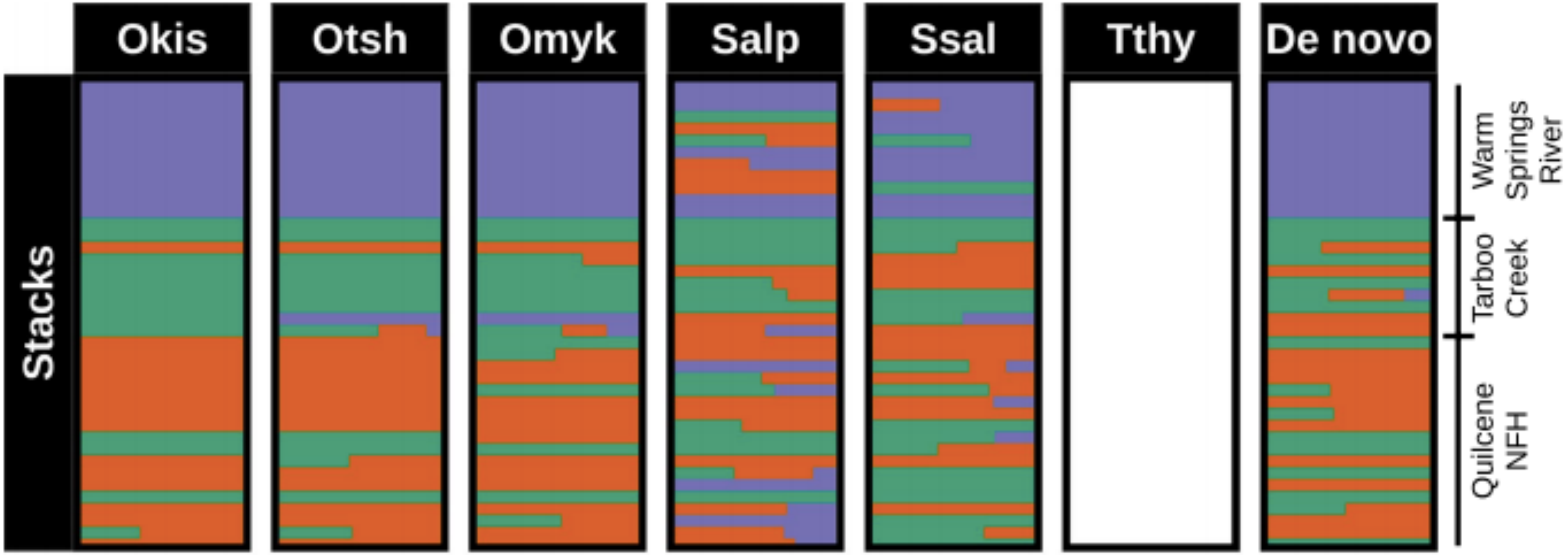

ANGSD
FreeBayes
Stacks
De novo
O. kisutch
S. confluentus
Total raw SNPs
2,000,000
1,500,000
1,000,000
500,000
0
*
Okis
Otsh
Omyk
Salp
Ssal
Tthy
dDocent
Stacks


ANGSD
FreeBayes
Stacks
De novo
O. kisutch
S. confluentus
Overall Fst
0.04
0.02
0.00
0.5
0.4
0.3
0.2
0.1
0.0
*
Okis
Omyk
Otsh
Salp
Ssal
Tthy
eBayes
Stacks

Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets

Ecology & Evolution 2020

Justin Bohling



Admixture results for Coho salmon only, using stacks only.

See also :

Nevado et al 2014 Mol Ecol : human / gorilla

Gopalakrishnan et al 2017 BMC Genomics : dog / wolf

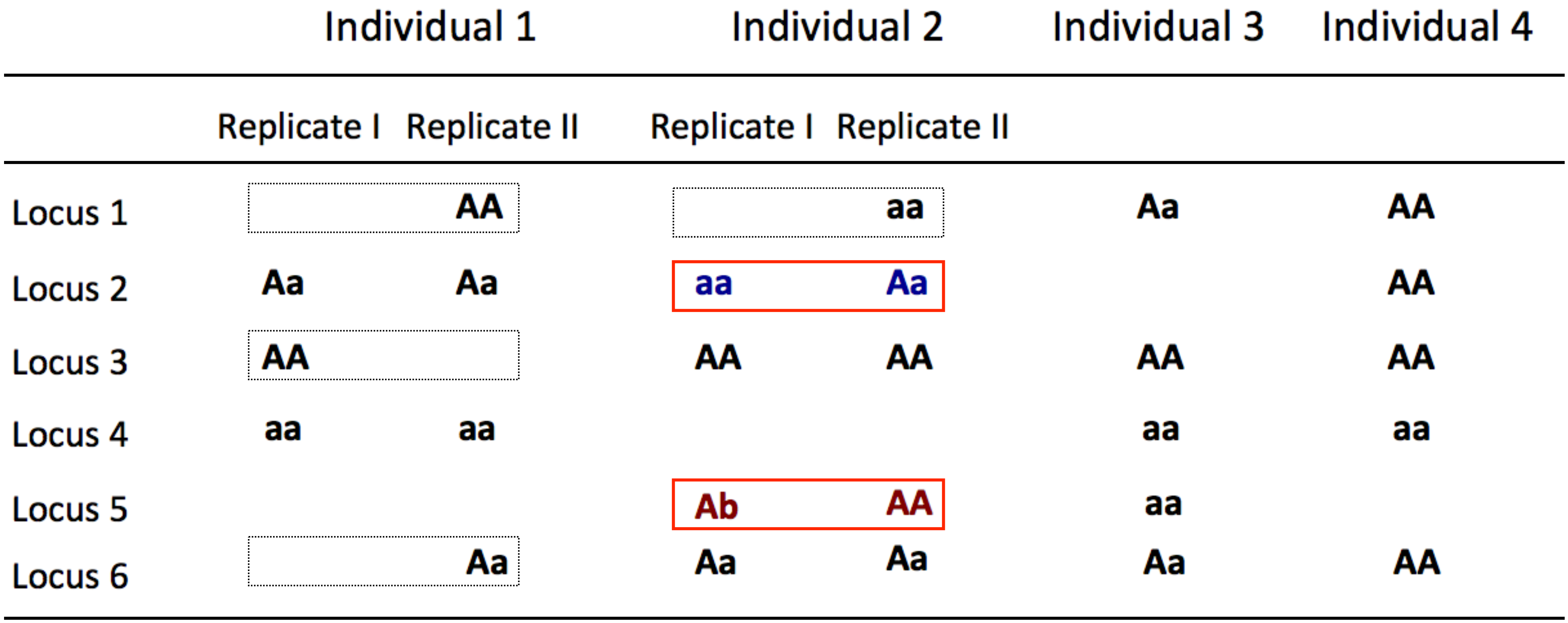
The nitty-gritty of catalog building

replication, on the other hand, is always useful

recommendations for optimising catalog building

authors	yr	DOI	data	pipeline	take home message(s)
Mastretta-Yanes et al	2015	10.1111/1755-0998.12291	plant	stacks	**importance of biological and technical replicates to compute genotyping error** rate and optimise parameter values
McCartney-Melstad et al	2017	10.1111/1755-0998.13029	frog	pyrad	"we develop a novel set of metrics to determine sequence similarity thresholds that maximize the correct separation of paralogous regions and minimize oversplitting naturally occurring allelic variation within loci."
Paris et al	2017	10.1111/2041-210X.12775	trout / penguin / earthworms	stacks	the 80% rule as a generally effective method to select the core parameters for STACKS.
Shafer et al	2017	10.1111/2041-210X.12700	sea lions	stacks / pyrad / ddocent	"We recommend that RAD-seq studies employ reference-based approaches to a closely related genome, and due to the high stochasticity associated with the pipeline advocate the use of multiple pipelines to ensure robust population genetic and demographic inferences."
Diaz-Arce et al	2019	10.3389/fgene.2019.00533	crab / mackerel / scallop	stacks	"(i) recovery of higher numbers of polymorphic loci is not necessarily associated with higher genetic differentiation, (ii) that the presence of PCR duplicates, selected loci assembly parameters and selected SNP filtering parameters affect the number of recovered polymorphic loci and degree of genetic differentiation, and (iii) that this effect is different in each dataset, meaning that defining a systematic universal protocol for RAD-seq data analysis may lead to missing relevant information about population differentiation."
Graham et al	2020	10.1371/journal.pone.0226608	lake whitefish	stacks	"“simple” methodological decisions with caution, especially when working on non-model species"

use of technical replicates to estimate error rate when you do not have a (close-enough) reference genome



	Individual 1		Individual 2		Individual 3	Individual 4
	Replicate I	Replicate II	Replicate I	Replicate II		
Locus 1		AA		aa	Aa	AA
Locus 2	Aa	Aa	aa	Aa		AA
Locus 3	AA		AA	AA	AA	AA
Locus 4	aa	aa			aa	aa
Locus 5			Ab	AA	aa	
Locus 6		Aa	Aa	Aa	Aa	AA

locus dropout

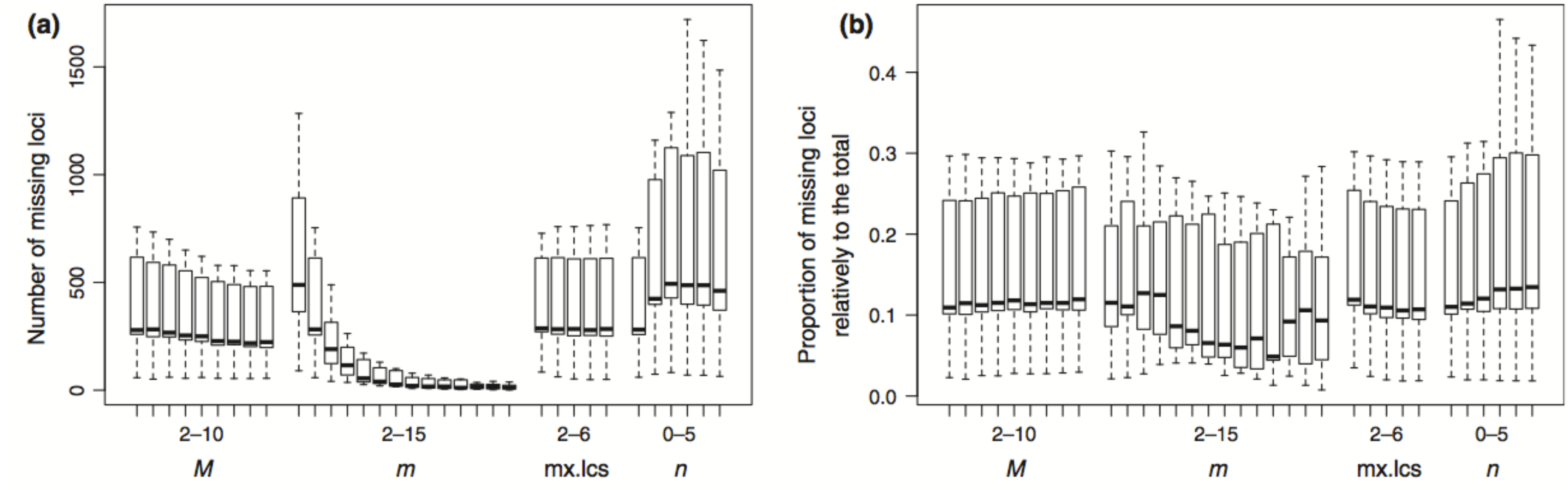
allele dropout ou erreur (PCR ou séquençage)

Mastretta-Yanes et al (2014) Mol Ecol Res

use of technical replicates to estimate error rate when you do not have a (close-enough) reference genome

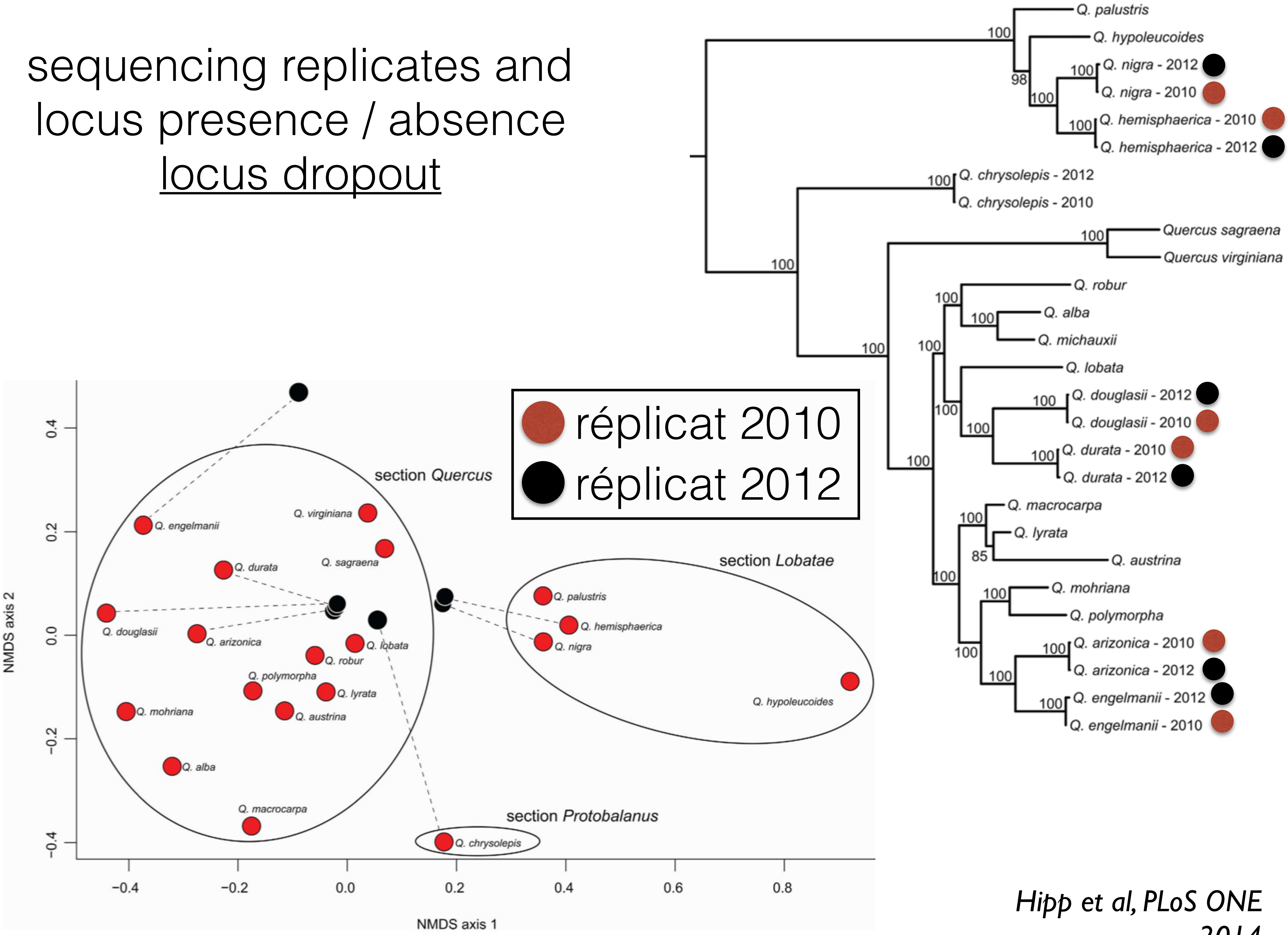
locus presence / absence

<u>locus dropout</u>



Mastretta-Yanes et al (2014) Mol Ecol Res

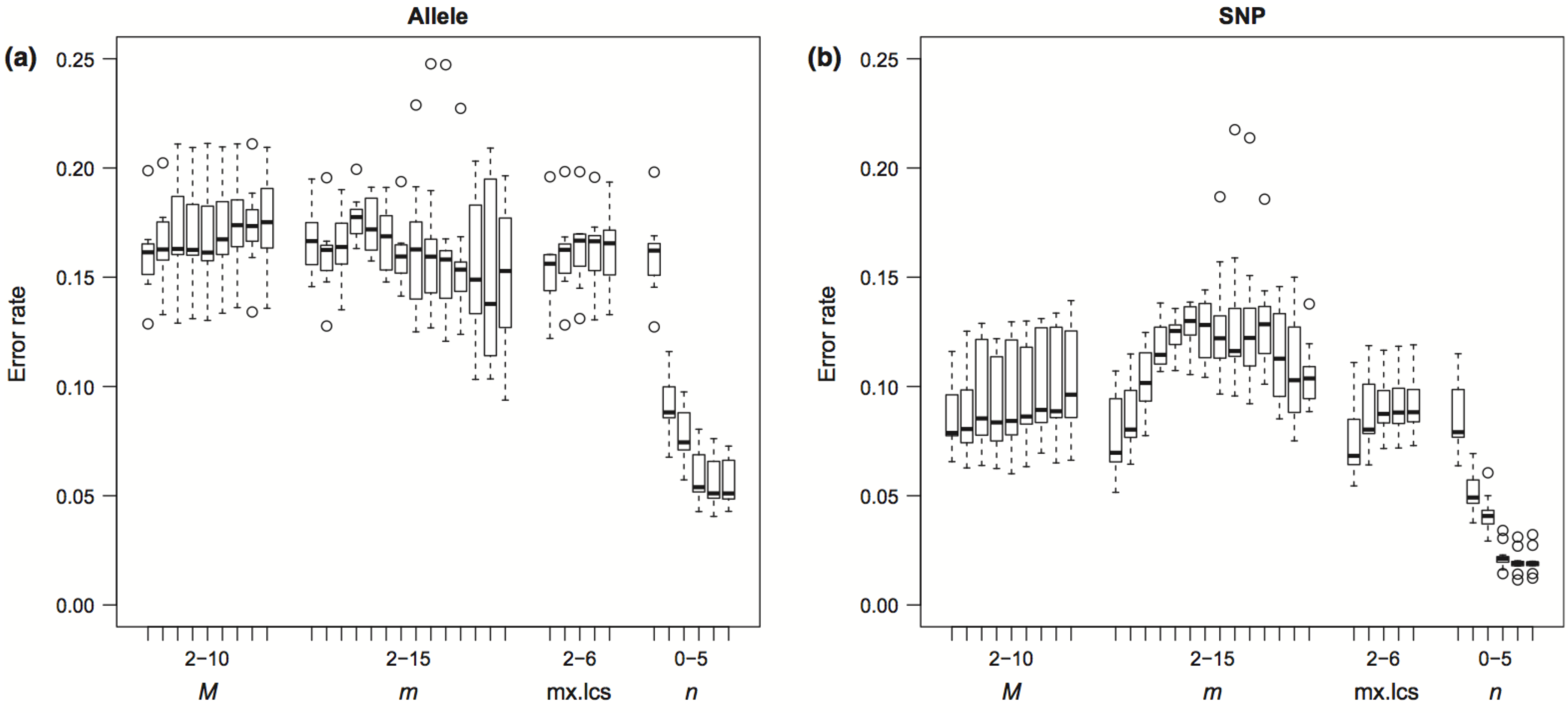
sequencing replicates and locus presence / absence <u>locus dropout</u>



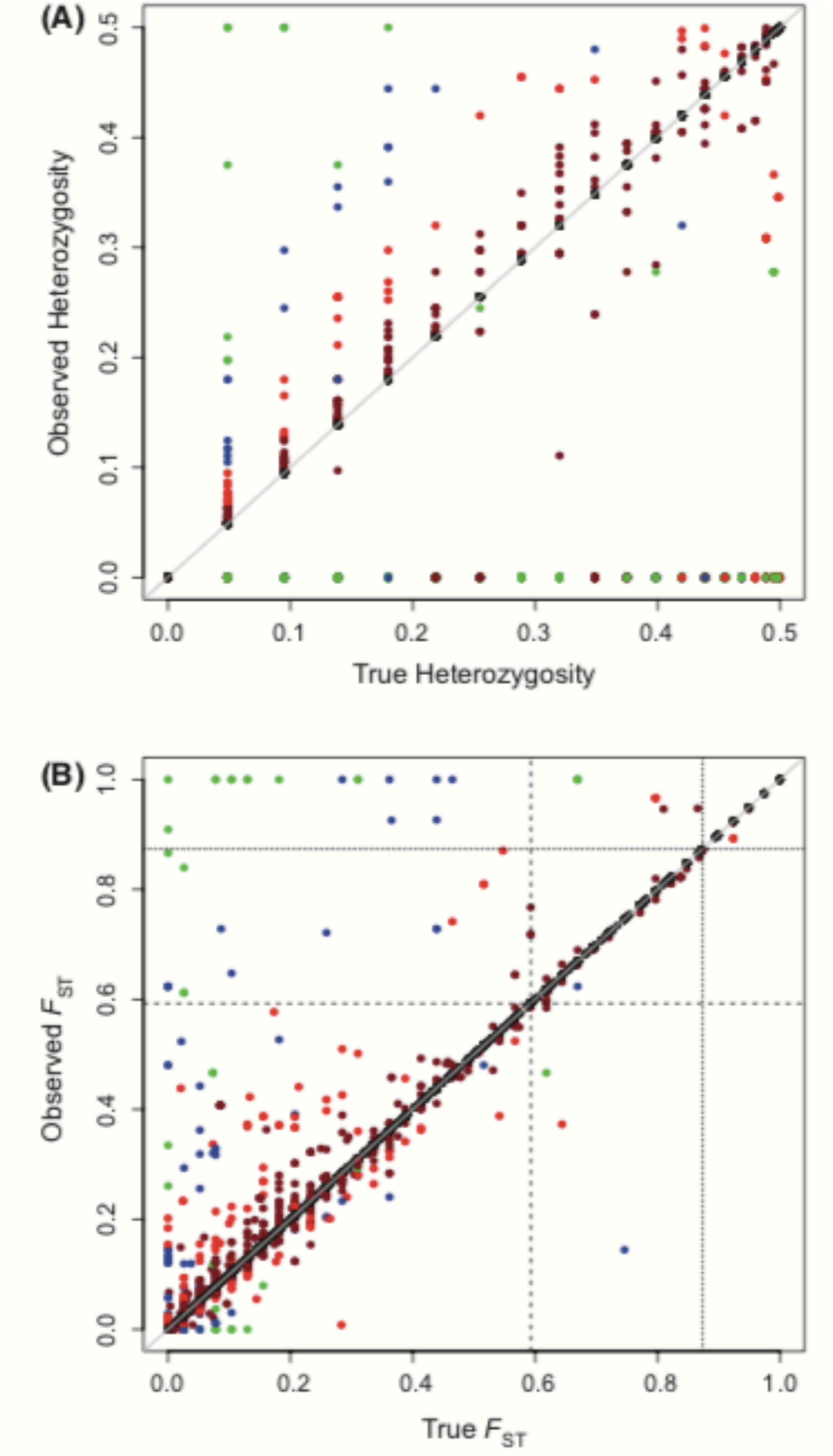
Hipp et al, PLoS ONE 2014

use of technical replicates to estimate error rate when you do not have a (close-enough) reference genome

error detection for alleles and SNPs



Mastretta-Yanes et al (2014) Mol Ecol Res



Another source of allele dropout: polymorphism on restriction sites

allele dropout leads to overestimates of

- genetic variation within and between populations
- heterozygosity
- F_{ST} proportion of F_{ST} outliers

using the distribution of read coverage values over loci to detect markers with a large excess of null alleles

Gautier et al (2013) Mol Ecol

Différents filtres, différents résultats ?

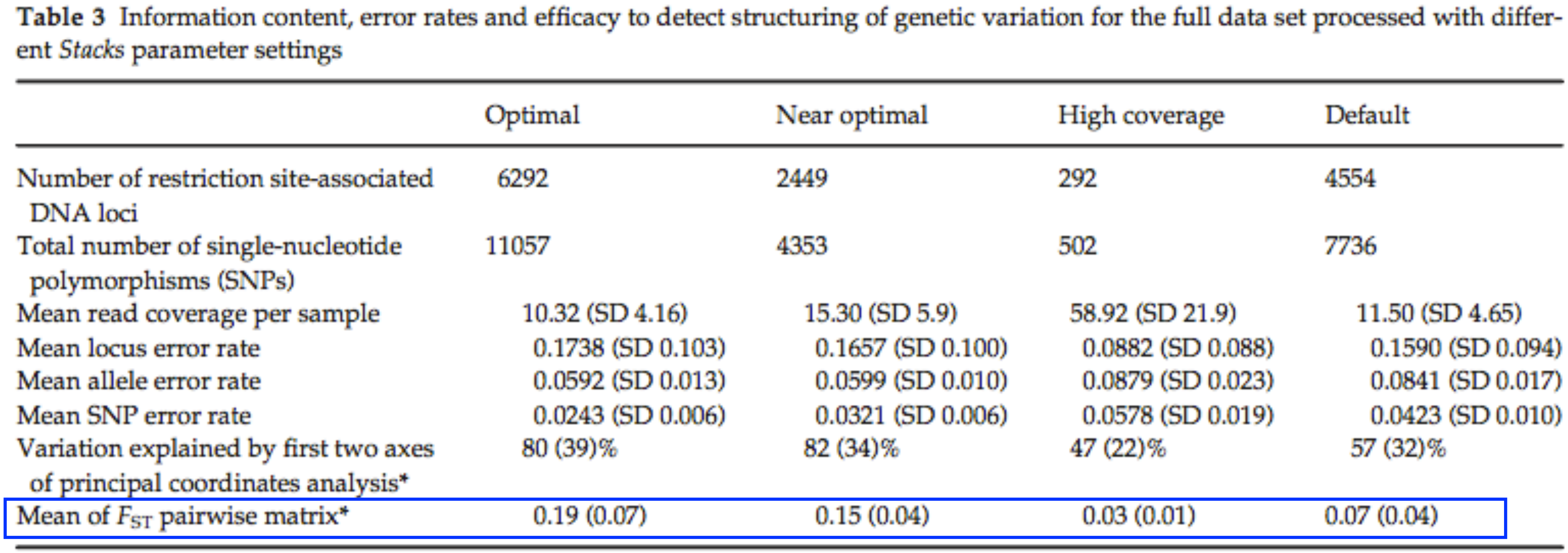


Table 3 Information content, error rates and efficacy to detect structuring of genetic variation for the full data set processed with different *Stacks* parameter settings

	Optimal	Near optimal	High coverage	Default
Number of restriction site-associated DNA loci	6292	2449	292	4554
Total number of single-nucleotide polymorphisms (SNPs)	11057	4353	502	7736
Mean read coverage per sample	10.32 (SD 4.16)	15.30 (SD 5.9)	58.92 (SD 21.9)	11.50 (SD 4.65)
Mean locus error rate	0.1738 (SD 0.103)	0.1657 (SD 0.100)	0.0882 (SD 0.088)	0.1590 (SD 0.094)
Mean allele error rate	0.0592 (SD 0.013)	0.0599 (SD 0.010)	0.0879 (SD 0.023)	0.0841 (SD 0.017)
Mean SNP error rate	0.0243 (SD 0.006)	0.0321 (SD 0.006)	0.0578 (SD 0.019)	0.0423 (SD 0.010)
Variation explained by first two axes of principal coordinates analysis*	80 (39)%	82 (34)%	47 (22)%	57 (32)%
Mean of F_{ST} pairwise matrix*	0.19 (0.07)	0.15 (0.04)	0.03 (0.01)	0.07 (0.04)

“optimal = parameter profile that performed better in experiment 1 optimal parameter values will vary for other RADseq data.”

Mastretta-Yanes et al (2014) Mol Ecol Res

The nitty-gritty of catalog building

r80, for STACKS: selecting the m, M, and n parameter values that provide the maximum number of polymorphic loci present in at least the 80% of the individuals

mmendations for optimising catalog building

auth				peline	take home message(s)
Mas Yane				acks	importance of biological and technical replicates to compute genotyping error rate and optimise parameter values
McC Mels				rad	"we develop a novel set of metrics to determine sequence similarity thresholds that maximize the correct separation of paralogous regions and minimize oversplitting naturally occurring allelic variation within loci."
Paris et al	2017	10.1111/2041-210X.12775	trout / penguin / earthworms	stacks	the 80% rule as a generally effective method to select the core parameters for STACKS.
Shafer et al	2017	10.1111/2041-210X.12700	sea lions	stacks / pyrad / ddocent	"We recommend that RAD-seq studies employ reference-based approaches to a closely related genome, and due to the high stochasticity associated with the pipeline advocate the use of multiple pipelines to ensure robust population genetic and demographic inferences."
Diaz-Arce et al	2019	10.3389/fgene.2019.00533	crab / mackerel / scallop	stacks	"(i) recovery of higher numbers of polymorphic loci is not necessarily associated with higher genetic differentiation, (ii) that the presence of PCR duplicates, selected loci assembly parameters and selected SNP filtering parameters affect the number of recovered polymorphic loci and degree of genetic differentiation, and (iii) that this effect is different in each dataset, meaning that defining a systematic universal protocol for RAD-seq data analysis may lead to missing relevant information about population differentiation."
Graham et al	2020	10.1371/journal.pone.0226608	lake whitefish	stacks	""simple" methodological decisions with caution, especially when working on non-model species"

The nitty-gritty of catalog building

Some published recommendations for optimising catalog building

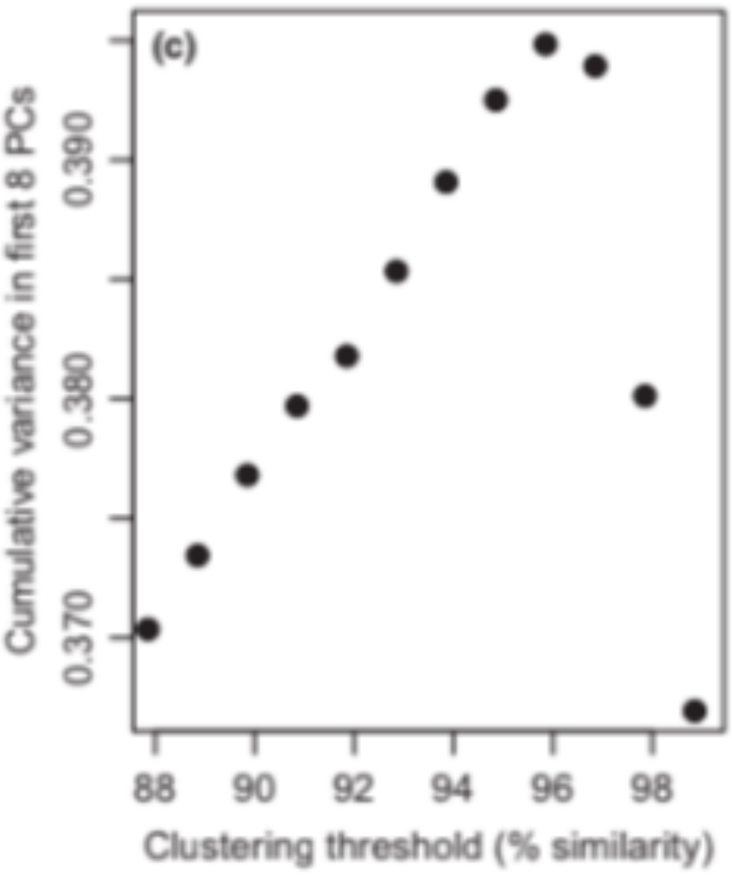
authors	yr	DOI	data	pipeline	take home message(s)
Mastretta-Yanes et al	2015	10.1111/1755-0998.12291	plant	stacks	importance of biological and technical replicates to compute genotyping error rate and optimise parameter values
McCartney-Melstad et al	2017	10.1111/1755-0998.13029	frog	pyrad	"we develop a novel set of metrics to determine sequence similarity thresholds that maximize the correct separation of paralogous regions and minimize oversplitting naturally occurring allelic variation within loci."
Paris et al	2017	10.1111/2041-210X.12775	trout / penguin / earthworms	stacks	the 80% le as a generally effective method to select the core parameters for STACKS.
Shafer et al	2017	10.1111/2041-210X.12700	sea lions	stacks / pyrad / ddocent	"We recom a closely rela pipeline ac tic and demog
Diaz-Arce et al	2019	10.3389/fgene.2019.00533	crab / mackerel / scallop	stacks	"(i) recover with higher ted loci assem per of recovere this effect is different in each dataset, meaning that defining a systematic universal protocol for RAD-seq data analysis may lead to missing relevant information about population differentiation.”
Graham et al	2020	10.1371/journal.pone.0226608	lake whitefish	stacks	““simple” methodological decisions with caution, especially when working on non-model species”

7 metrics to identify that maximises correct separation of paralogs and minimises over-splitting

GitHub repo with scripts to compute metrics from VCF files

Seven metrics to optimise catalog construction

- Fraction of inferred paralogs & diversity measures (metrics 1-4)
- Relationship between missingness and genetic divergence, and slope of isolation by distance (metrics 5-6)
- Phylogenetic resolution (metric 7)



example metric:
#4: cumulative variance explained by first 8 PCA axes

McCartney -Melstad et al (2017) Mol Ecol Res

Mastretta-Yanes et al (2014) Mol Ecol Res

conclusions

- many things affect catalog assembly
  - experimental strategy (sampling, enzyme(s) …)
  - lab work (library constr. sequencing plateform …)
  - bioinformatic pipelines (catalog assembly strategy)
  - pipeline set-up (parameter selection)

conclusions

- practical considerations :
  - consider in-silico enzyme selection
  - consider using biological and technical replicates
  - evaluate the usefulness of reference genome
  - try several pipelines
  - estimate optimal clustering metrics

thanks for your attention!

- thanks to the workshop organisers!
- RAD funding:
  - GDR GE & APEGE (InEE, CNRS)
  - MNHN, Institut ISYEB
  - cluster de calcul YMIR, Université de La Rochelle
  - GenoToul bioinformatics cluster
- RAD buddies, Amélia Viricel (LIENSs), Jawad Abdelkrim (MNHN)