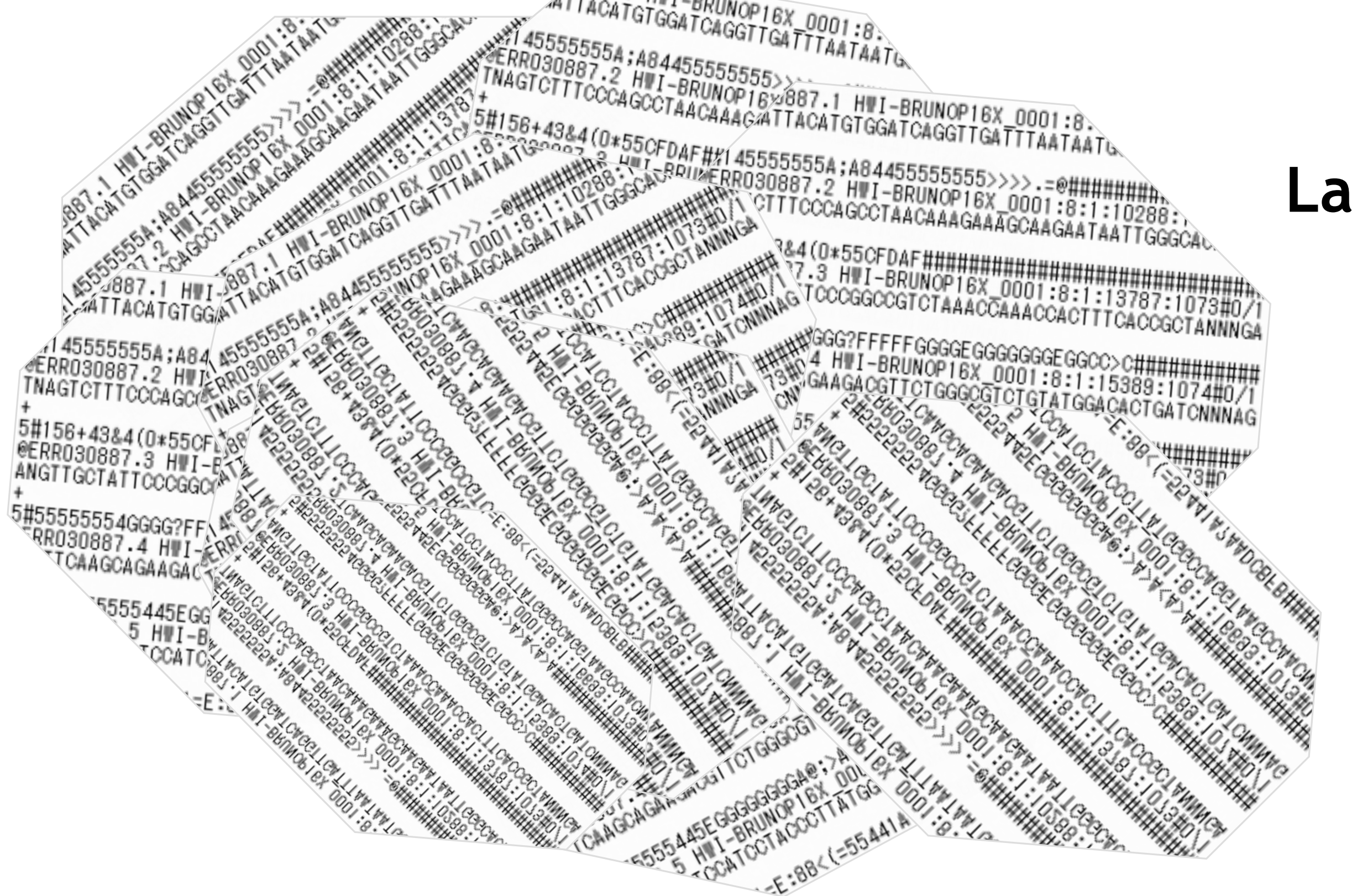
Roadmap for filtering Massively Parallel Sequencing (MPS) datasets

Anne-Laure Ferchaud, Jean-Sébastien Moore, Eric Normandeau, Laura Benestan, Thierry Gosselin, Louis Bernatchez



suggested workflow[1]

Lane quality (*e.g.*, Fastqc)

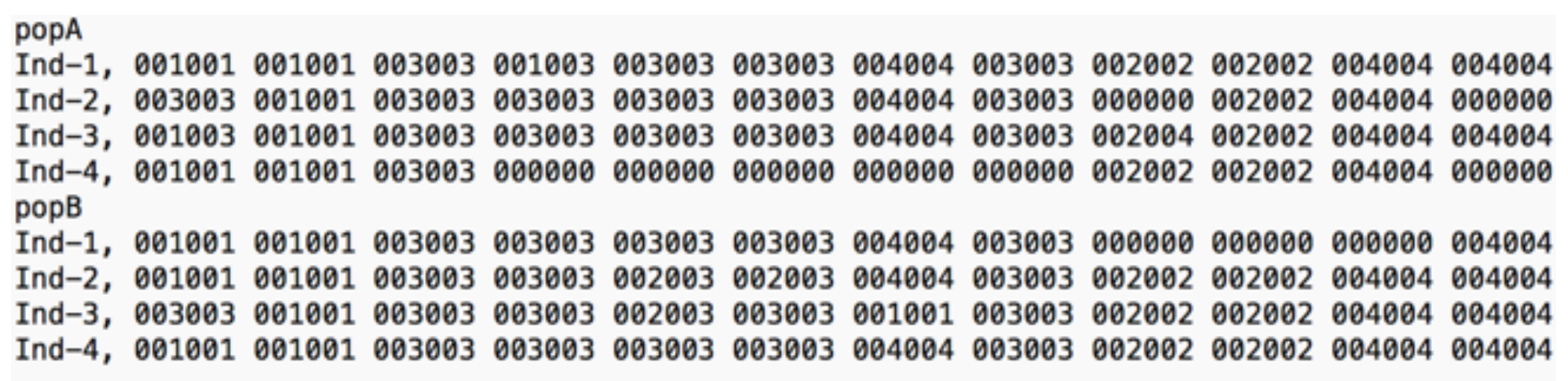
Adaptor removal (*e.g.*, cutadapt)

Demultiplexing and read quality trimming (*e.g.*, *process_radtags*)

Alignment (*de novo* (*e.g.*, *ustacks*) *versus* mapping to a reference genome (*e.g.*, *BWA*))

SNPs calling (*e.g.*, stacks[2])

Filtering (*e.g.*, *see the steps detailed below*)



```
popA
Ind-1, 001001 001001 003003 001003 003003 003003 004004 003003 002002 002002 004004 004004
Ind-2, 003003 001001 003003 003003 003003 003003 004004 003003 000000 002002 004004 000000
Ind-3, 001003 001001 003003 003003 003003 003003 004004 003003 002004 002002 004004 004004
Ind-4, 001001 001001 003003 000000 000000 000000 000000 000000 002002 002002 004004 000000
popB
Ind-1, 001001 001001 003003 003003 003003 003003 004004 003003 000000 000000 000000 004004
Ind-2, 001001 001001 003003 003003 002003 002003 004004 003003 002002 002002 004004 004004
Ind-3, 003003 001001 003003 003003 002003 003003 001001 003003 002002 002002 004004 004004
Ind-4, 001001 001001 003003 003003 003003 003003 004004 003003 002002 002002 004004 004004
```

Missing data***

Sources of missing data

Variation in DNA quality

Variation in DNA concentrations

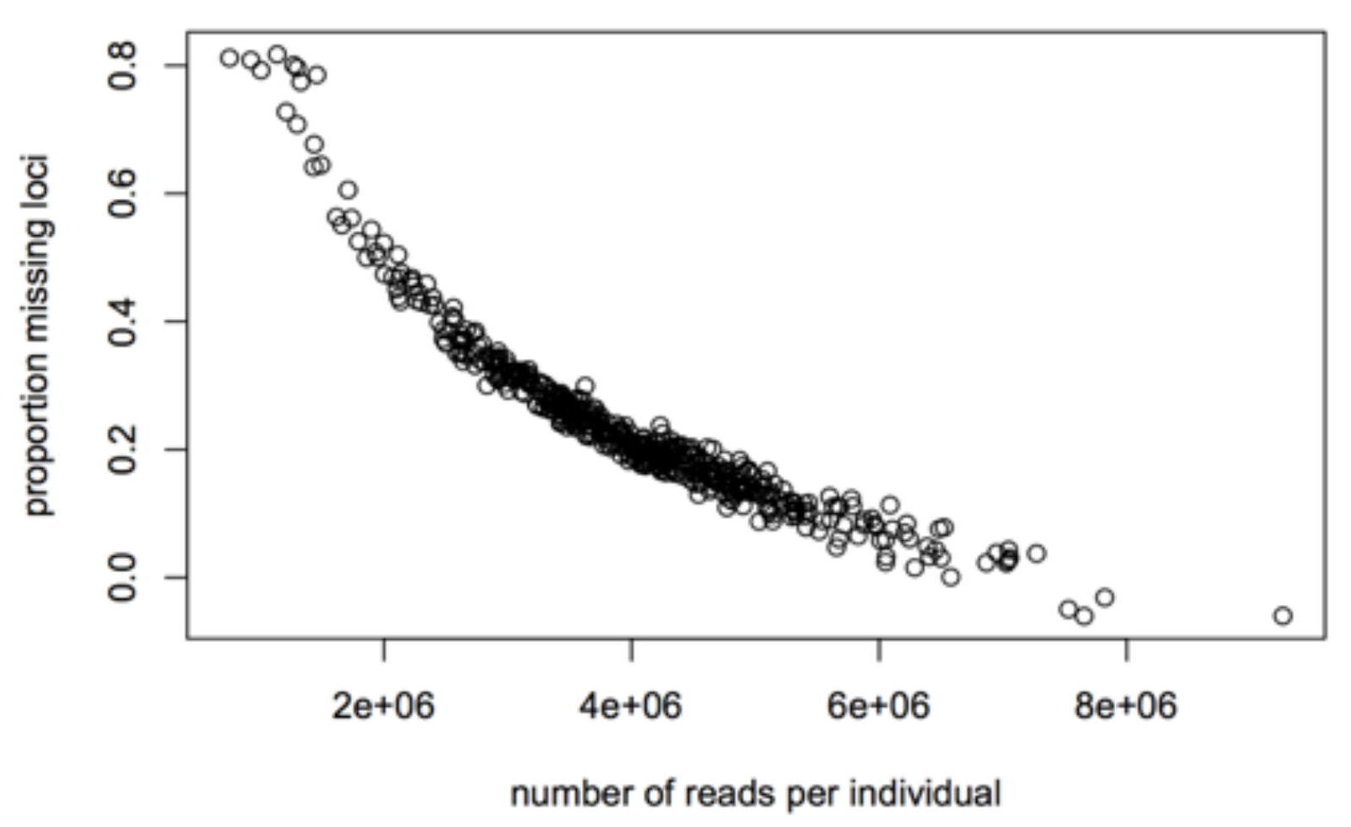
Coverage thresholds[5]

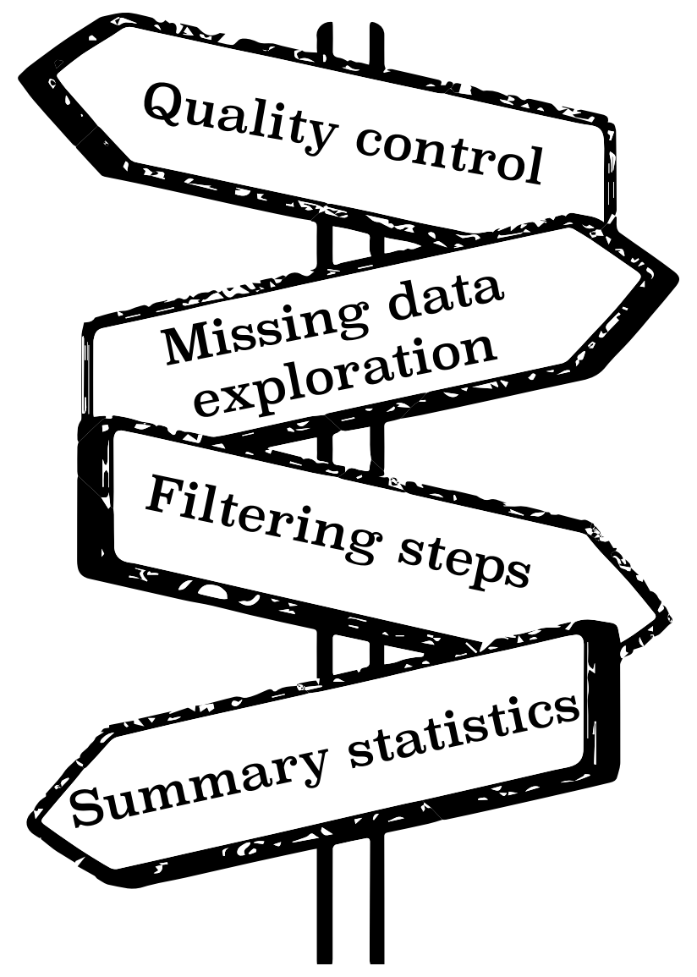
Sequence identity cutoffs[4]

Shotgun sequencing

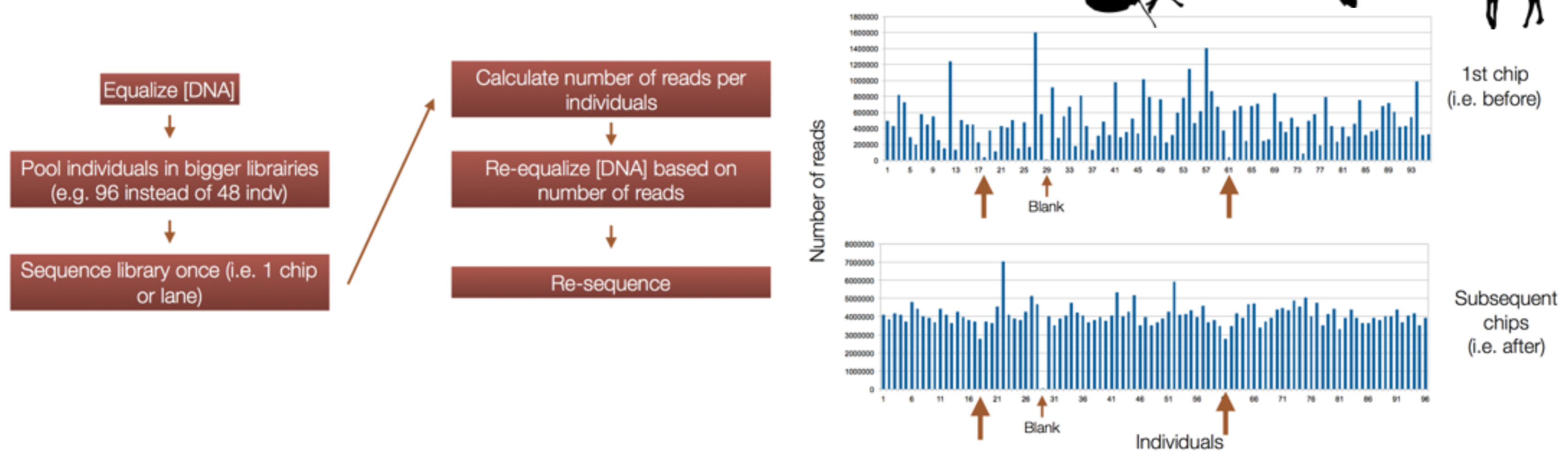
Allelic dropout due to mutations in cut sites[6,7]

Overall number of reads per individual (locus) is probably the most important determinant of missing data





How to reduce the variance in read number:


Equalize [DNA]

Pool individuals in bigger librairies (e.g. 96 instead of 48 indv)

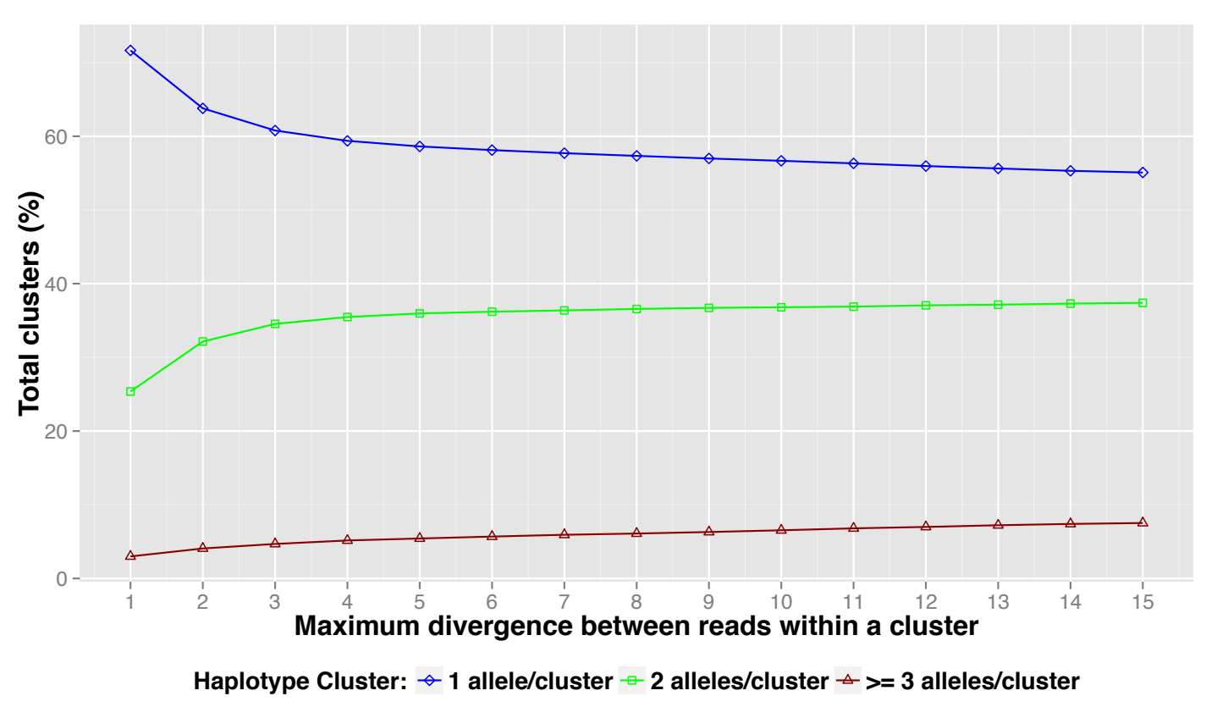
Sequence library once (i.e. 1 chip or lane)

Calculate number of reads per individuals

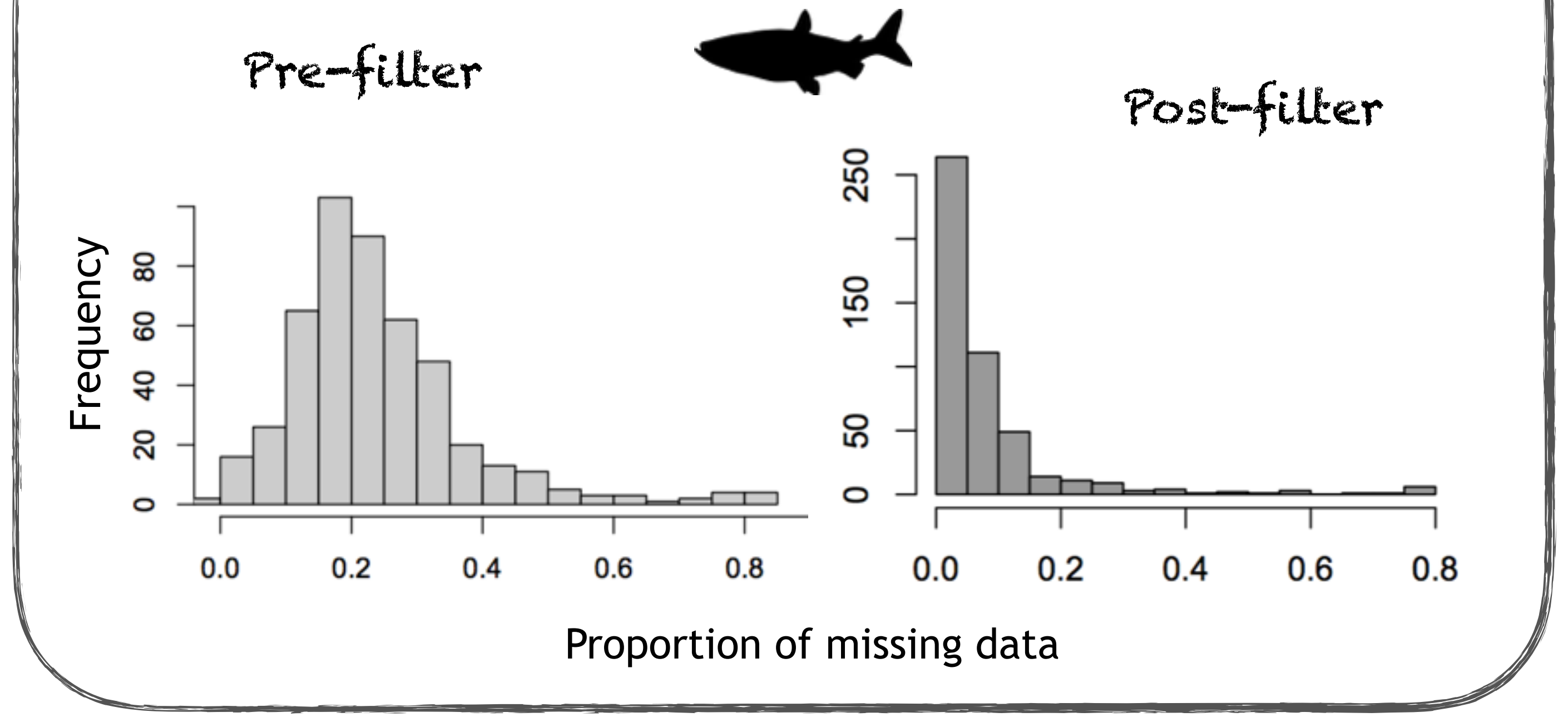
Re-equalize [DNA] based on number of reads

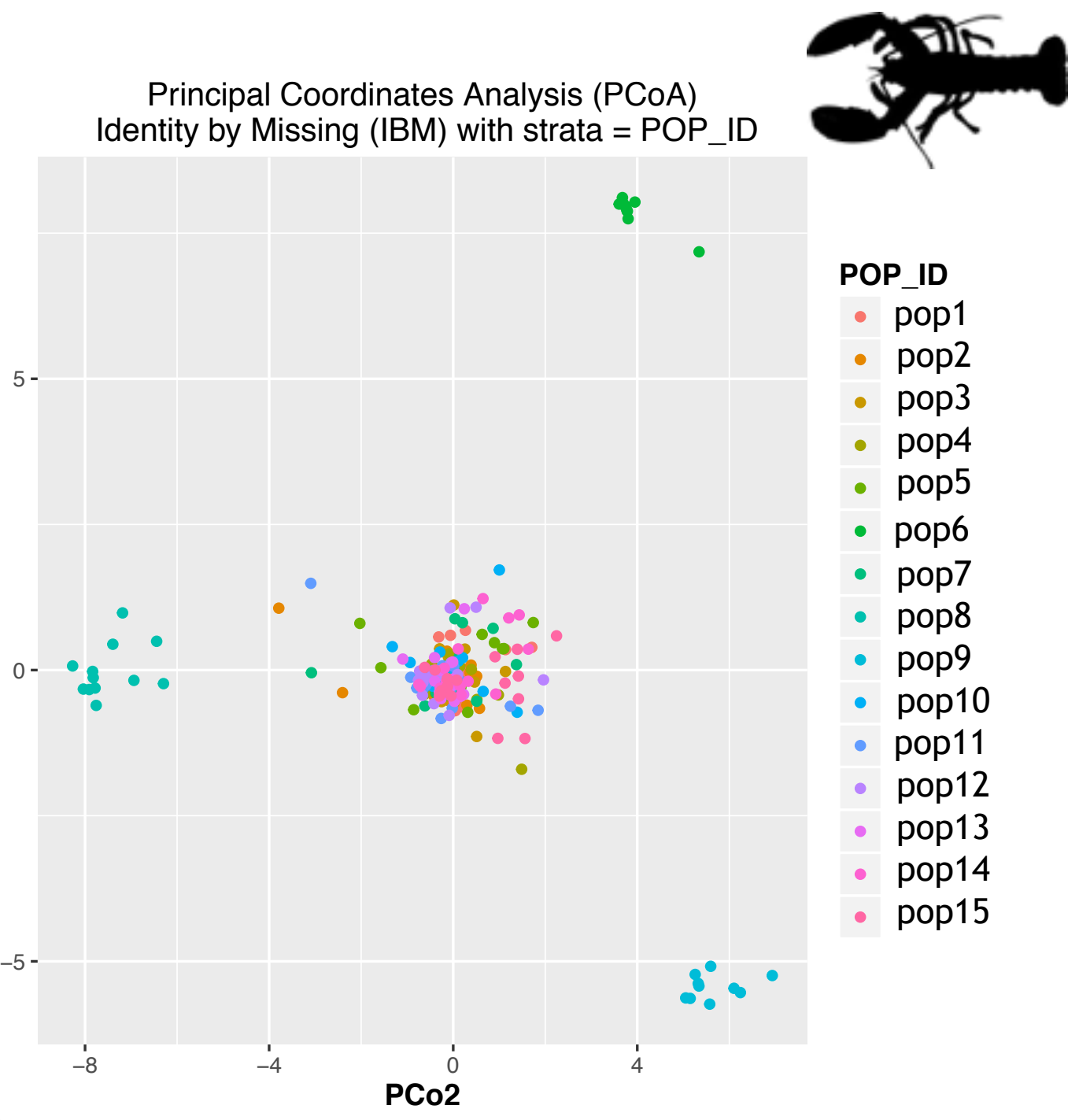
Re-sequence

Optimization of assembly[5]



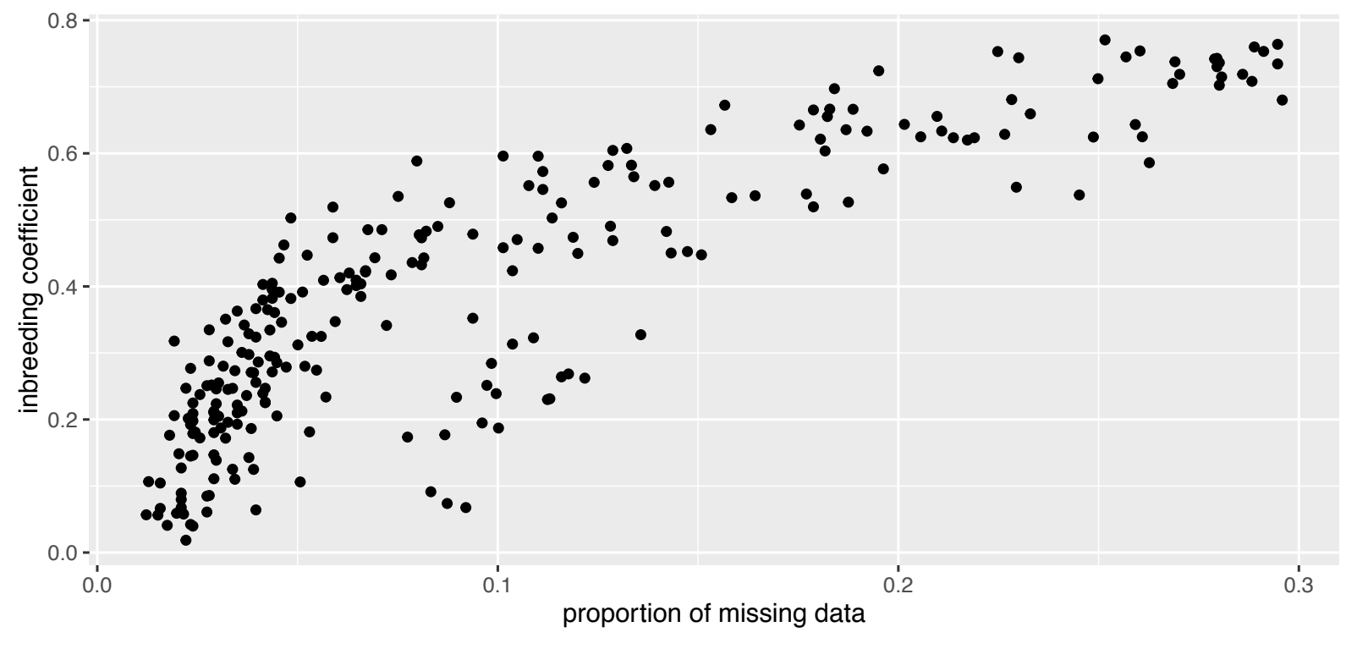
Histogram of proportion of missing loci per individual





Exploring patterns of missing is an important step to decide whether or not some populations or individuals should be excluded.

Individuals with high proportion of missing data could have **elevated homozygosity (probably suggesting allelic dropout)**



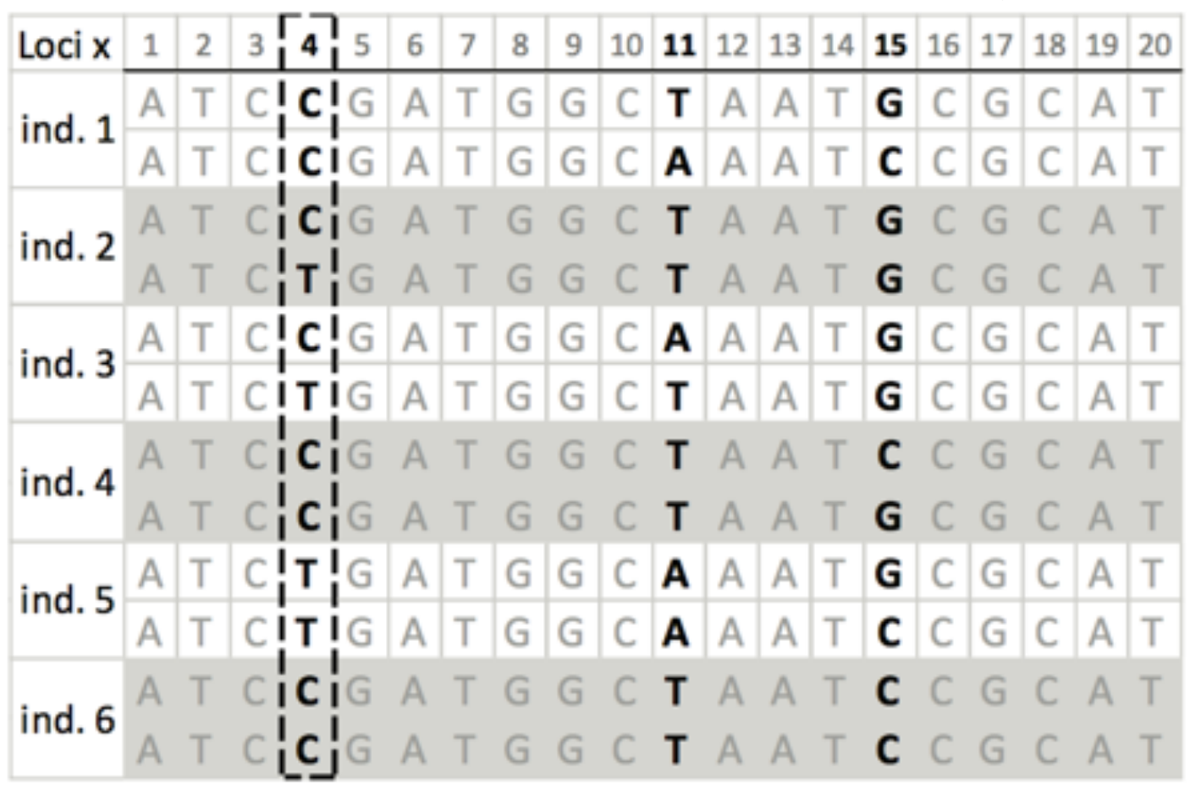
Quality control and filtering MPS data[3]

	Parameter	Library/ Sequencing Lane	Allele	Genotype	Individual	Marker	Sampling site	Population	Globally
⚠	Quality	X			X				
⚠	Assembly and genotyping	X							
⚠	Outliers		X	X	X	X			
⚠	Missingness***	X	X	X	X	X	X	X	X
⚠	Coverage		X	X		X			
⚠	Genotype likehood			X					
⚠	Proportion Genotyped				X	X	X	X	X
▼	He & Fis & HWE				X	X		X	
▼	Minor Allele Frequency					X	X	X	X
⚠	Missingness***	X	X	X	X	X	X	X	X

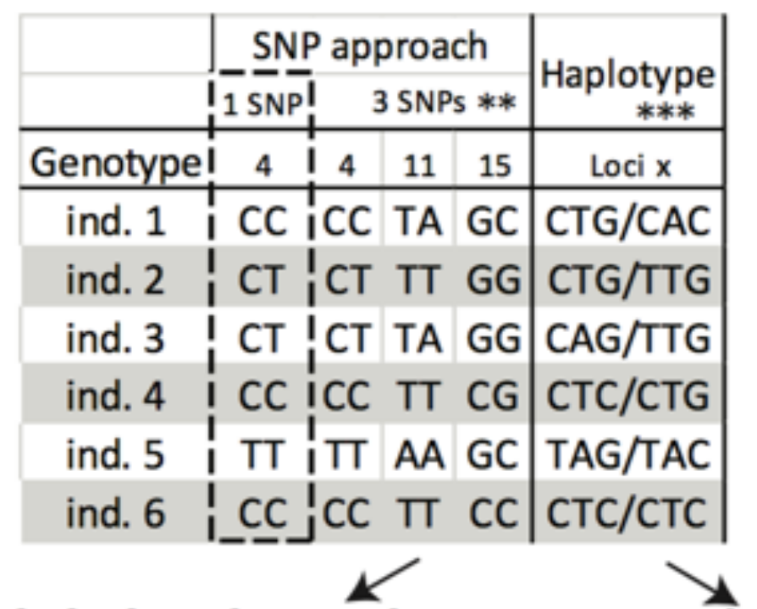
⚠ Quality insurance steps crucial to remove artifactual and uninformative markers to have reliable of genetic parameters

▼ Filtering steps that should not necessarily done, it would depend on the subsequent analysis that need to be done

Will you consider haplotype or SNP level statistics ?[8]



Loci x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ind. 1	A	T	C	C	G	A	T	G	G	C	T	A	A	T	G	C	G	C	A	T
	A	T	C	C	G	A	T	G	G	C	A	A	A	T	C	C	G	C	A	T
ind. 2	A	T	C	C	G	A	T	G	G	C	T	A	A	T	G	C	G	C	A	T
	A	T	C	T	G	A	T	G	G	C	T	A	A	T	G	C	G	C	A	T
ind. 3	A	T	C	C	G	A	T	G	G	C	A	A	A	T	G	C	G	C	A	T
	A	T	C	T	G	A	T	G	G	C	T	A	A	T	G	C	G	C	A	T
ind. 4	A	T	C	C	G	A	T	G	G	C	T	A	A	T	C	C	G	C	A	T
	A	T	C	C	G	A	T	G	G	C	T	A	A	T	G	C	G	C	A	T
ind. 5	A	T	C	T	G	A	T	G	G	C	A	A	A	T	G	C	G	C	A	T
	A	T	C	T	G	A	T	G	G	C	A	A	A	T	C	C	G	C	A	T
ind. 6	A	T	C	C	G	A	T	G	G	C	T	A	A	T	C	C	G	C	A	T
	A	T	C	C	G	A	T	G	G	C	T	A	A	T	C	C	G	C	A	T



	SNP approach				Haplotype ***
	1 SNP	3 SNPs **			
Genotype	4	4	11	15	Loci x
ind. 1	CC	CC	TA	GC	CTG/CAC
ind. 2	CT	CT	TT	GG	CTG/TTG
ind. 3	CT	CT	TA	GG	CAG/TTG
ind. 4	CC	CC	TT	CG	CTC/CTG
ind. 5	TT	TT	AA	GC	TAG/TAC
ind. 6	CC	CC	TT	CC	CTC/CTC

3 linked markers with maximum **3 different genotypes** each **

a multi-SNP locus with a maximum of **6 different haplotypes observed** ***

Here is an example of 6 diploid individuals (ind.) genotyped at loci x, 20 bp long. Among this subset of individuals, 3 SNPs were discovered and accurately called (*), at nucleotide positions 4, 11 and 15. These 3 SNPs could be treated as three different markers (**). Several classic analysis would treat these 3 markers as independent whereas they are physically linked. To counteract this problem, researchers often retain only one SNP, for example the first one, here SNP 4 (see dashed line). However, in order to make use of all the 3 SNPs, the haplotype approach (combining the 3 SNPs in a single haplotype) could be used (***) when filtering and genotyping.

References:

[1] Stacks workflow used and proposed by the Bernatchez ‘s lab : https://github.com/enormandeau/stacks_workflow.
[2] Catchen *et al.*, (2013) Molecular Ecology,22 (11):3142-3140.
[3] Gosselin & Bernatchez L, (2016) https://github.com/thierrygosselin/stackr.
[4] Harvey *et al.*, PeerJ. (2015);3: e895.

[5] Ilut *et al.*, BioMed Research International (2014): 1–9.
[6] Gautier *et al.*, (2013) Molecular Ecology, 22, 3165–3178.
[7] Arnold *et al.*, (2013) Molecular Ecology, 22, 3179- 3190.
[8] Benestan *et al.*, (2016) Molecular Ecology

@AL_Ferchaud