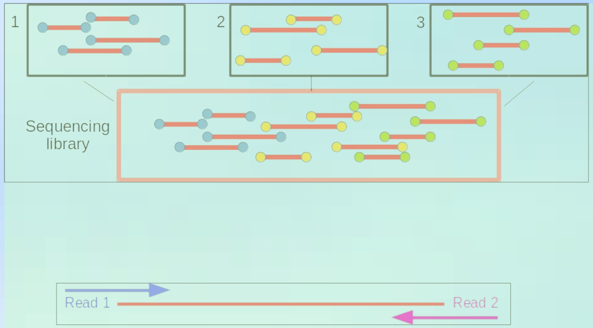
ISGA2015

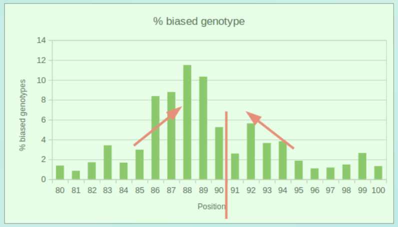
Santiago de Compostela, 21-26 June 2015

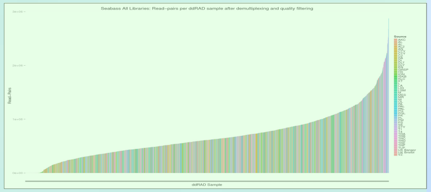
“Analytical power and biases of double digestion RAD (ddRAD) genotyping by sequencing in three european marine aquaculture species”

Maroso, F.[a,b], Hermida, M.[b], Pardo, B. G.[b], Carr, A.[c], Franch, R.[a], Martínez, P.[b], Bargelloni, L.[a]

[a] Dipartimento di Biomedicina Comparata e Alimentazione, Università degli Studi di Padova, 35020, ITALY
[b] Departemento de Genética, Universitade de Santiago de Compostela, Campus de Lugo, SPAIN
[c] Fios Genomics Ltd., Edinburgh BioQuater, Edinburgh EH16 4SB, UK







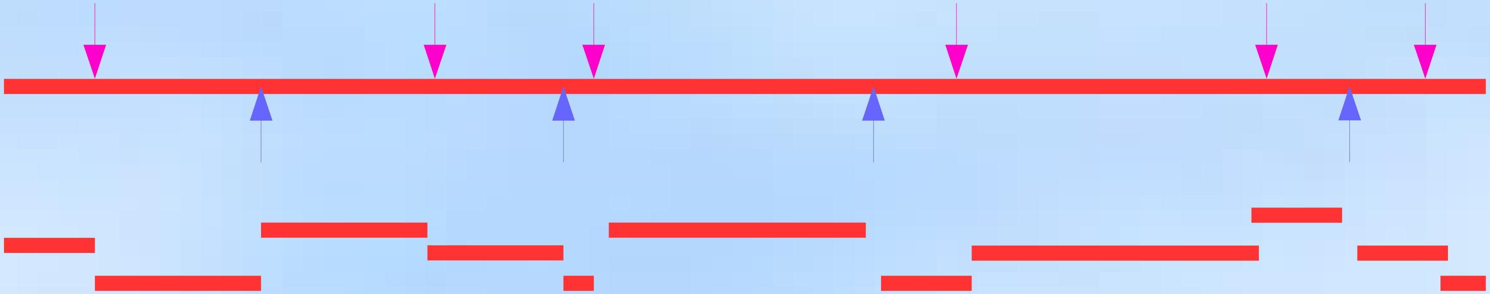
SEVENTH FRAMEWORK PROGRAMME

www.aquatrace.eu

dd-RAD sequencing protocol

Reduce genome complexity (< 1‰):

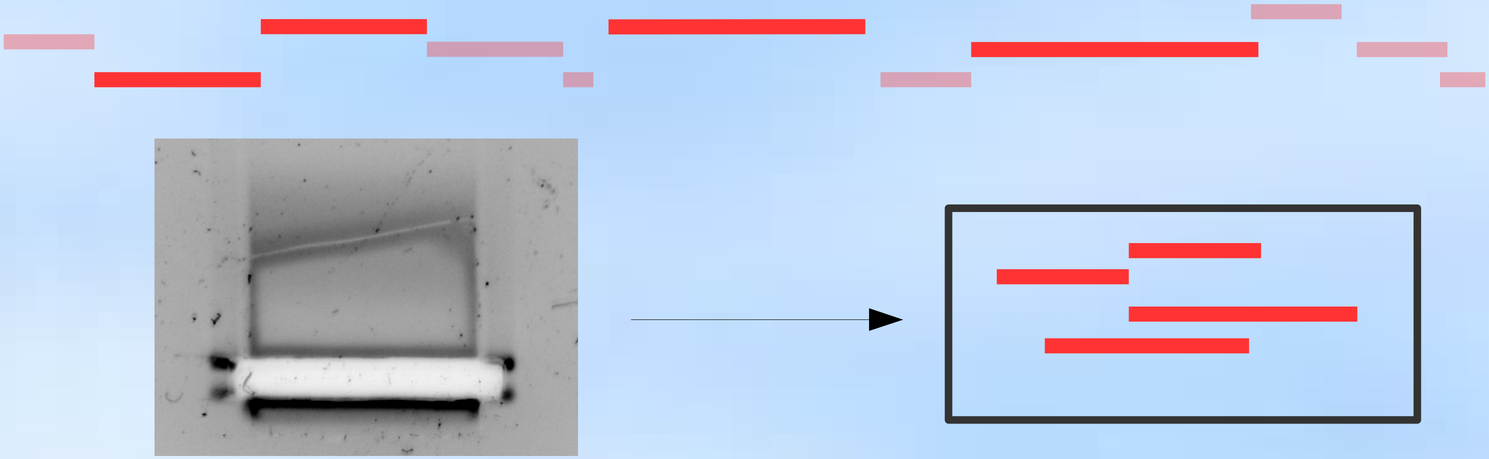
- DNA cut with two enzymes



dd-RAD sequencing protocol

Reduce genome complexity (< 1‰):

- DNA cut with two enzymes
- Fragments selected by size (approx. 300-600 bp)

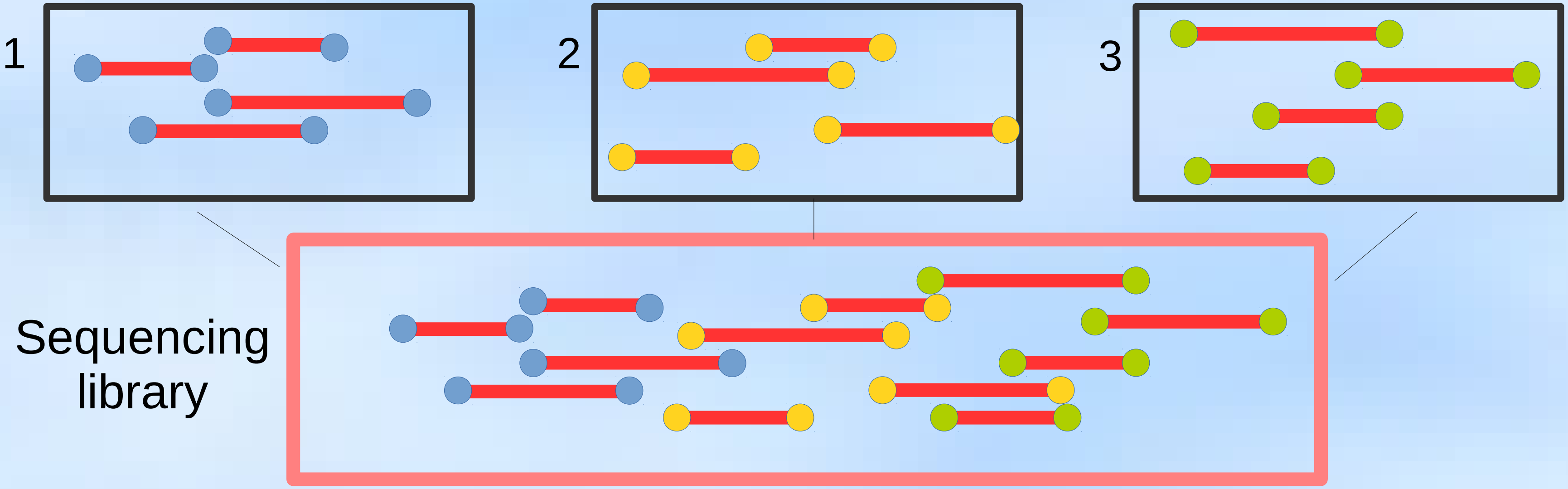


Agarose gel size selection

dd-RAD sequencing protocol

Reduce genome complexity (< 1‰):

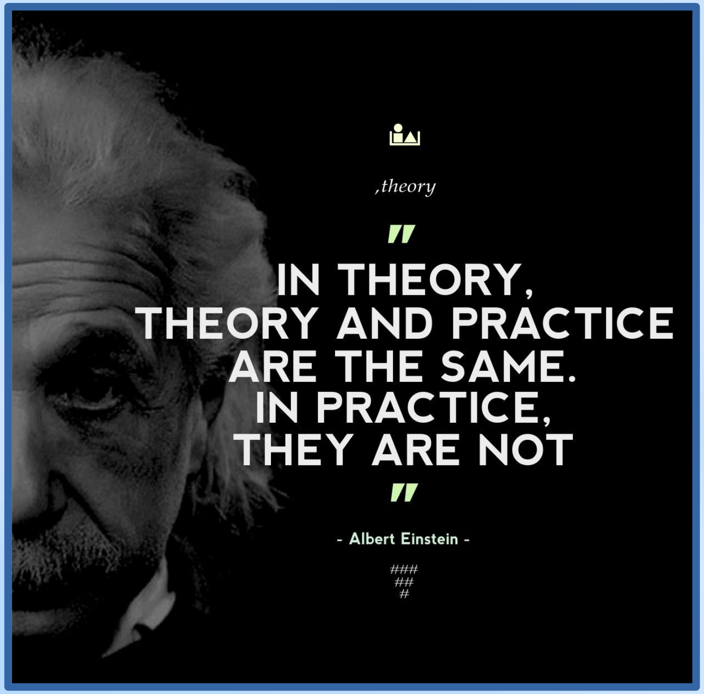
- DNA cut with two enzymes
- Fragments selected by size (approx. 300-600 bp)
- Samples pooled (144). Barcoding to recognize them after sequencing



From Theory to Practice: loss of information and genotyping biases

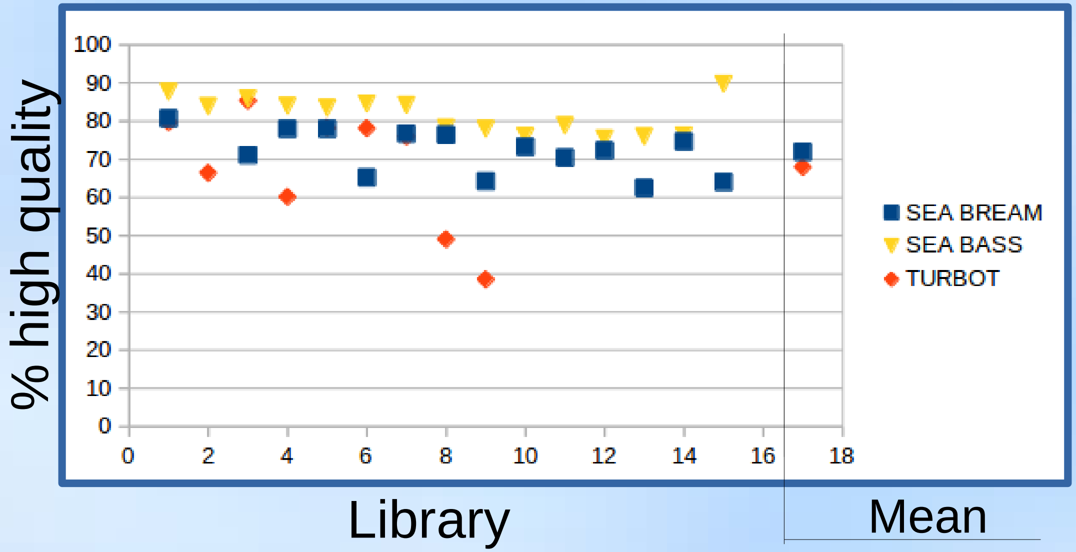
Several steps causing loss of information and analytical power:

- Quality filtering of sequenced reads
- Normalization of samples within library
- Fragment length and coverage
- Genotyping bias



Sequencing and quality filtering

- Illumina HiSeq technology (100 bp, pair end), throughput of 120 M reads
- Quality filter: 10-50% (average 25%) reduction in the number of reads



Quality of normalization

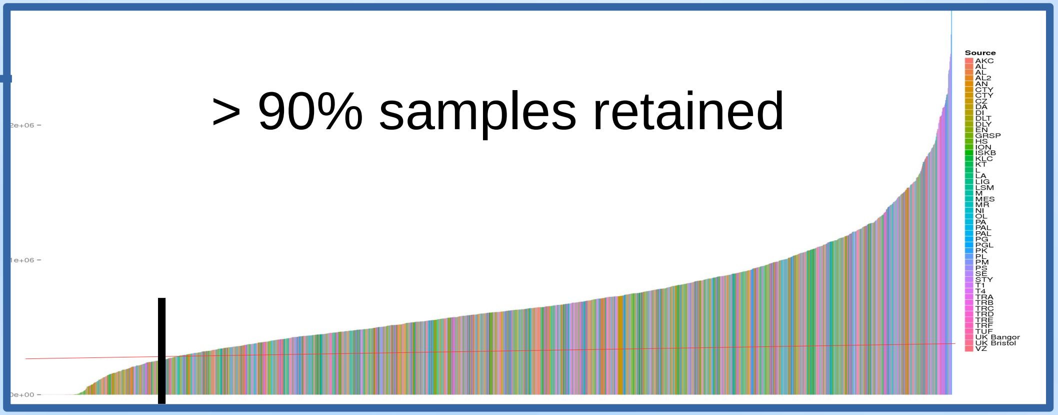
- Not all samples represented by same number of reads
- Threshold of 150'000 reads → samples genotyped for at least 80% loci
- Between 7%-30% samples with less than 150k reads, depending on the libraries

Quality of normalization

Library 1

Number of reads

150 k reads



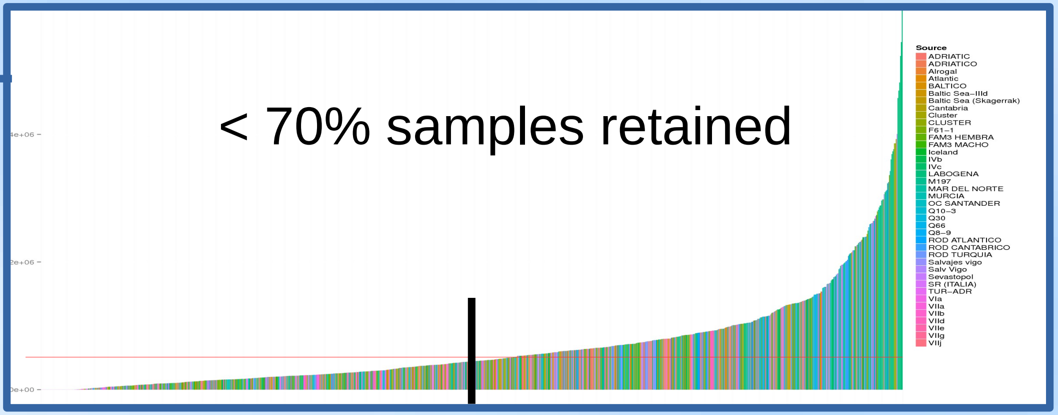
> 90% samples retained

Samples

Library 2

Number of reads

150 k reads

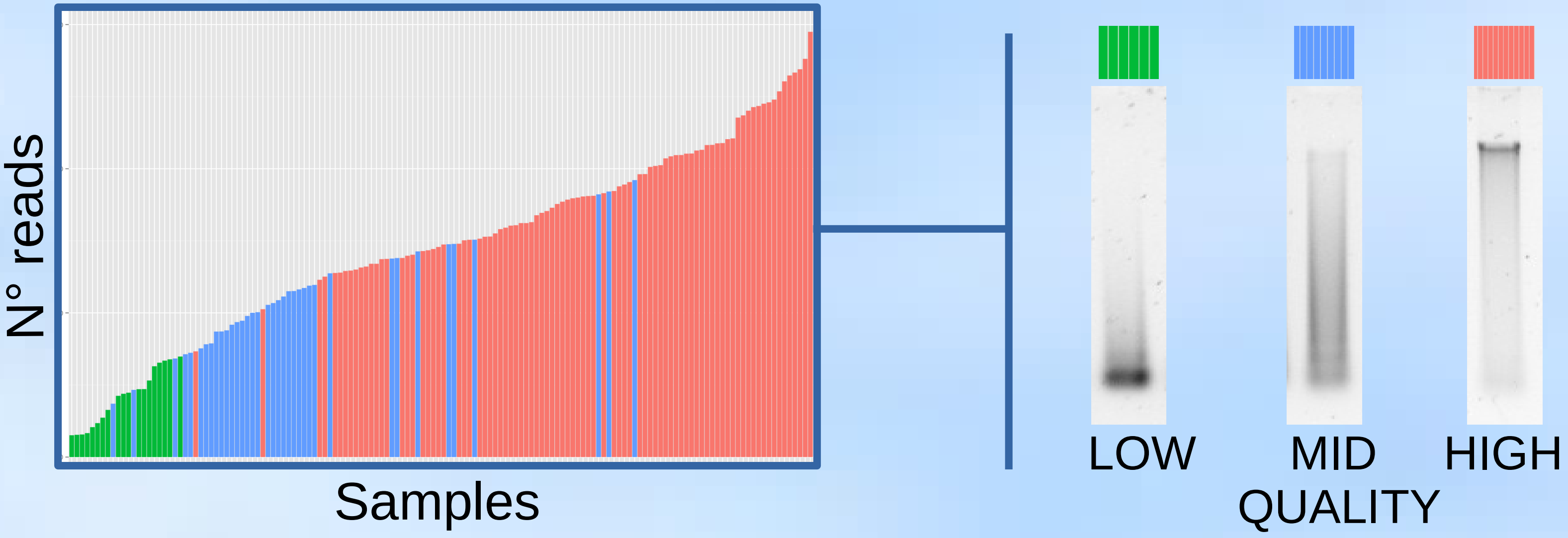


< 70% samples retained

Samples

Quality of normalization

- Laboratory procedures can affect the quality of the normalization (pipetting accuracy, DNA quantification...)
- Correlation between DNA degradation and N° reads



Bioinformatic pipeline (STACKS package)

- 3-4 replicates/species in all libraries (10-14)
- Reads are trimmed → loss of sequenced bp:
  - Last 3 bp removed (lower quality)
  - Barcodes at the beginning of the sequence (7 bp)
  - Enzyme recognition site (5-6 bp) → not variable!

total 15 bp

Bioinformatic pipeline (STACKS package)

- 3-4 replicates/species in all libraries (10-14)
- Reads are trimmed → loss of sequenced bp:
  - Last 3 bp removed (lower quality)
  - Barcodes at the beginning of the sequence (7 bp)
  - Enzyme recognition site (5-6 bp) → not variable!

total 15 bp

- Aspects to take in consideration:

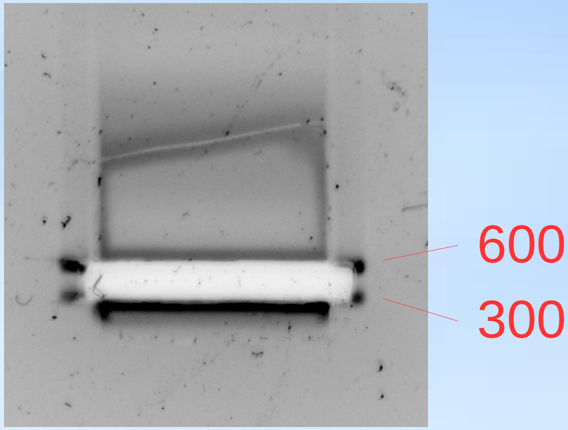
Fragment length

Coverage depth

Errors

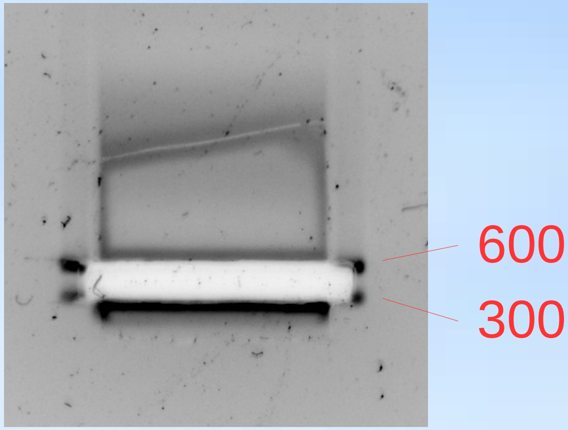
Length of fragments

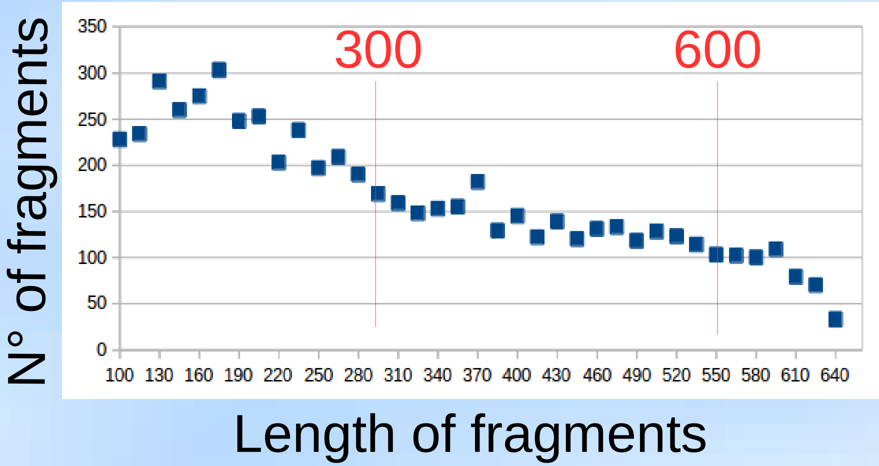
- 300-600 bp agarose gel size selection



Length of fragments

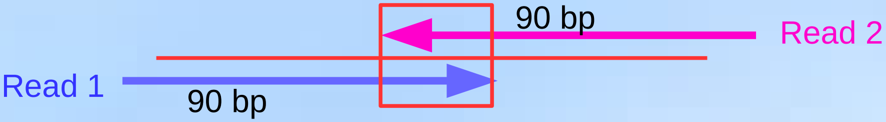
- 300-600 bp agarose gel size selection
- Actual fragment length range (from mapping position) is wider

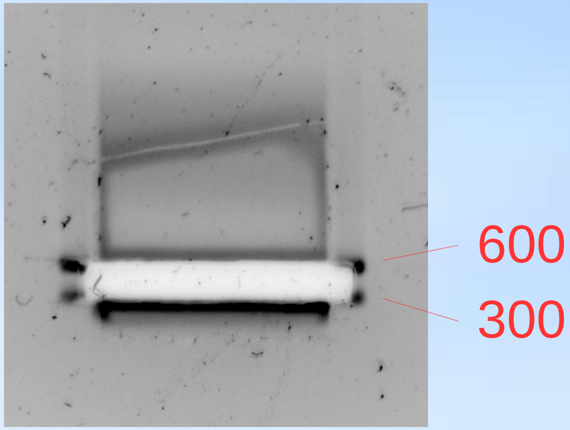


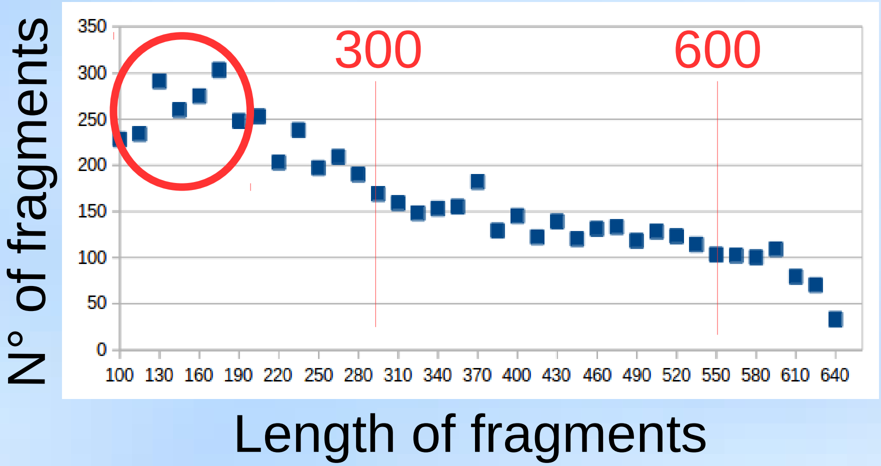


Length of fragments

- 300-600 bp agarose gel size selection
- Actual fragment length range (from mapping position) is wider
- Around 20% of the fragments <180 bp (overlapping)

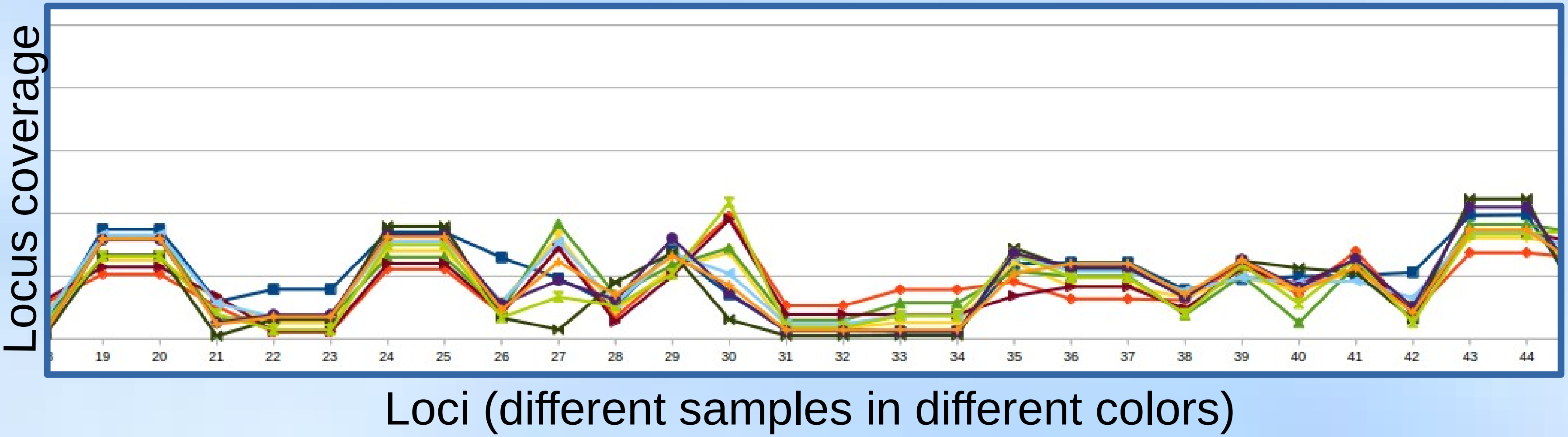






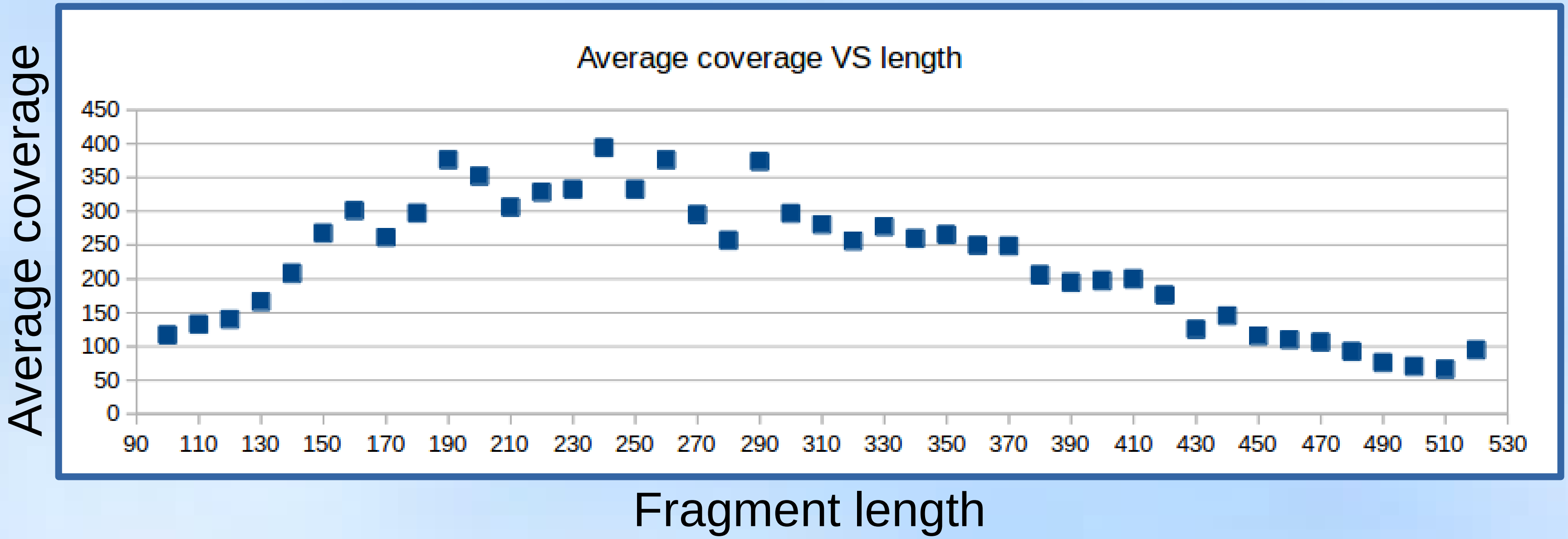
Coverage and fragment length

- Correlation between different loci and coverage per sample
- Depth of coverage is not homogeneous within fragment of different lengths



Coverage and fragment length

- Correlation between different loci and coverage per sample
- Depth of coverage is not homogeneous within fragment of different lengths



Analysis of genotyping errors

- The most frequent genotype was considered as the “correct” one
- 'de-novo' VS 'reference genome based' approaches compared

	Analysis	rxstacks	Tags	SNPS	Markers	% ERROR
Sea bream	DENOVO	N	3913	2970	1263	1,17
	DENOVO	Y	2353	1175	557	2,43
	REF	N	4753	1943	1341	0,30
	REF	Y	3729	1363	960	0,13
Sea bass	DENOVO	N	1673	639	389	2,83
	DENOVO	Y	1631	546	349	2,80
	REF	N	3162	1012	780	2,07
	REF	Y	3118	952	747	1,82

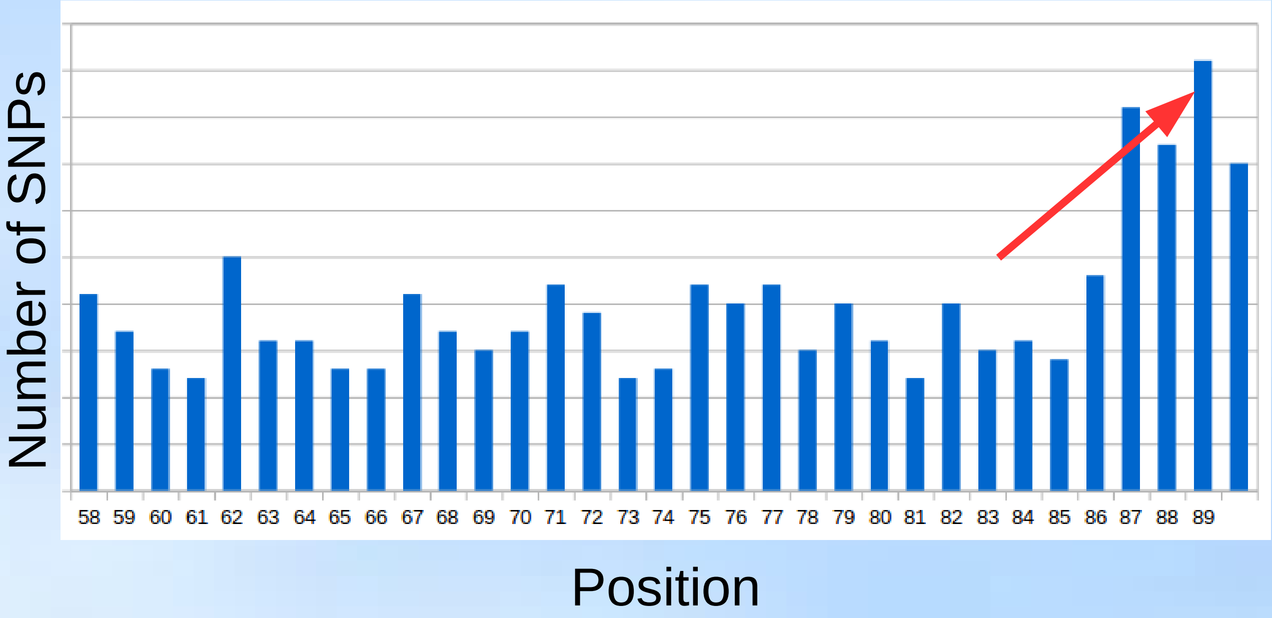
Analysis of genotyping errors

- The most frequent genotype was considered as the “correct” one
- 'de-novo' VS 'reference genome based' approaches compared

	Analysis	*rxstacks*	Tags	SNPS	Markers	% ERROR
Sea bream	DENOVO	N	3913	2970	1263	1,17
	DENOVO	Y	2353	1175	557	2,43
	REF	N	4753	1943	1341	0,30
	REF	Y	3729	1363	960	0,13
Sea bass	DENOVO	N	1673	639	389	2,83
	DENOVO	Y	1631	546	349	2,80
	REF	N	3162	1012	780	2,07
	REF	Y	3118	952	747	1,82

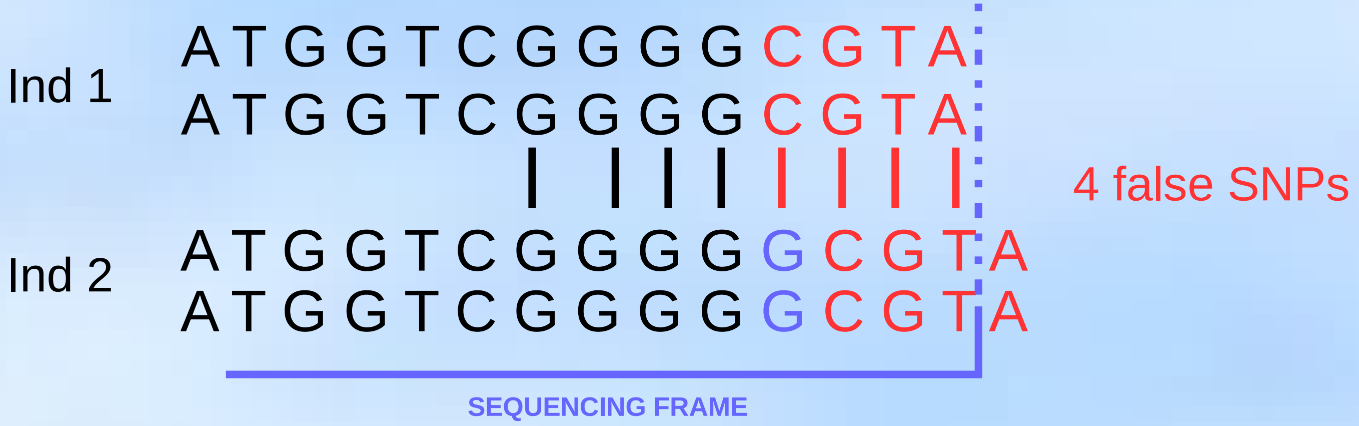
Position of SNPs

- The number of biased genotypes increased toward the end of the reads, in particular at the very last 4 bp



Position of SNPs

- Repetitions cause a shift in the sequences and may introduce false SNPs calling
- To a lesser extent, can be due to PCR/bridge amplification errors



Ind 1

ATGGTCGGGGCGTA

ATGGTCGGGGCGTA

4 false SNPs

Ind 2

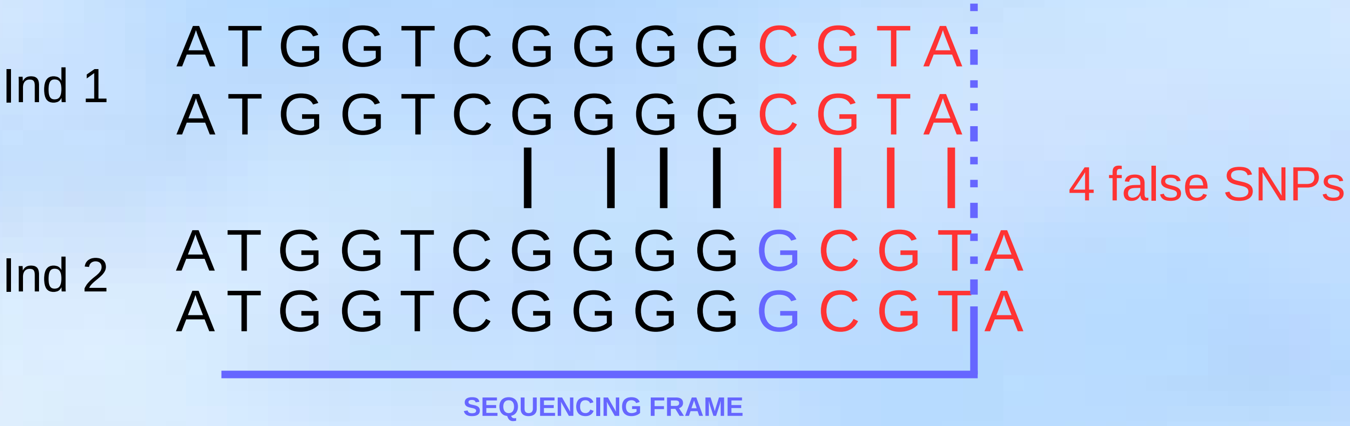
ATGGTCGGGGGCGTA

ATGGTCGGGGGCGTA

SEQUENCING FRAME

Position of SNPs

- Repetitions cause a shift in the sequences and may introduce false SNPs calling
- Biased tags can be identified and eliminated from final analysis (e.g Stacks' 'markers blacklist')



Ind 1

A T G G T C G G G G C G T A

A T G G T C G G G G C G T A

4 false SNPs

Ind 2

A T G G T C G G G G G C G T A

A T G G T C G G G G G C G T A

SEQUENCING FRAME

Take home...

New genotyping technologies allow faster, cheaper and more accurate analysis than ever before; continue improvement...

From theory to practice

↓

On average 55% of the total RAW information are actually used in the analysis

- More accuracy with reference genome
- 'blacklist' loci with repeats to reduce error rate

ISGA2015

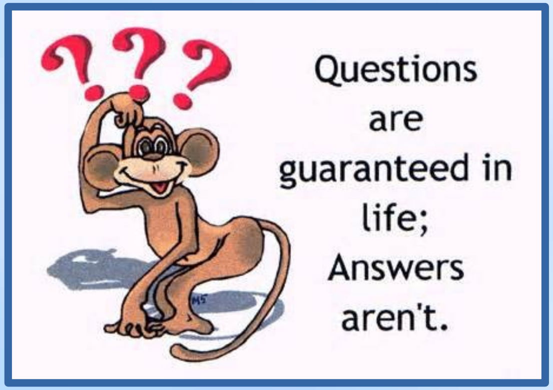
Santiago de Compostela, 21-26 June 2015

“Analytical power and biases of double digestion RAD (ddRAD) genotyping by sequencing in three european marine aquaculture species”

Maroso, F.[a,b], Hermida, M.[b], Pardo, B. G.[b], Carr, A.[c], Franch, R.[a], Martínez, P.[b], Bargelloni, L.[a]

[a] Dipartimento di Biomedicina Comparata e Alimentazione, Università degli Studi di Padova, 35020, ITALY
[b] Departemento de Genética, Universitade de Santiago de Compostela, Campus de Lugo, SPAIN
[c] Fios Genomics Ltd., Edinburgh BioQuater, Edinburgh EH16 4SB, UK

THANK YOU FOR LISTENING!



Questions are guaranteed in life; Answers aren't.

SEVENTH FRAMEWORK PROGRAMME

www.aquatrace.eu

fios GENOMICS