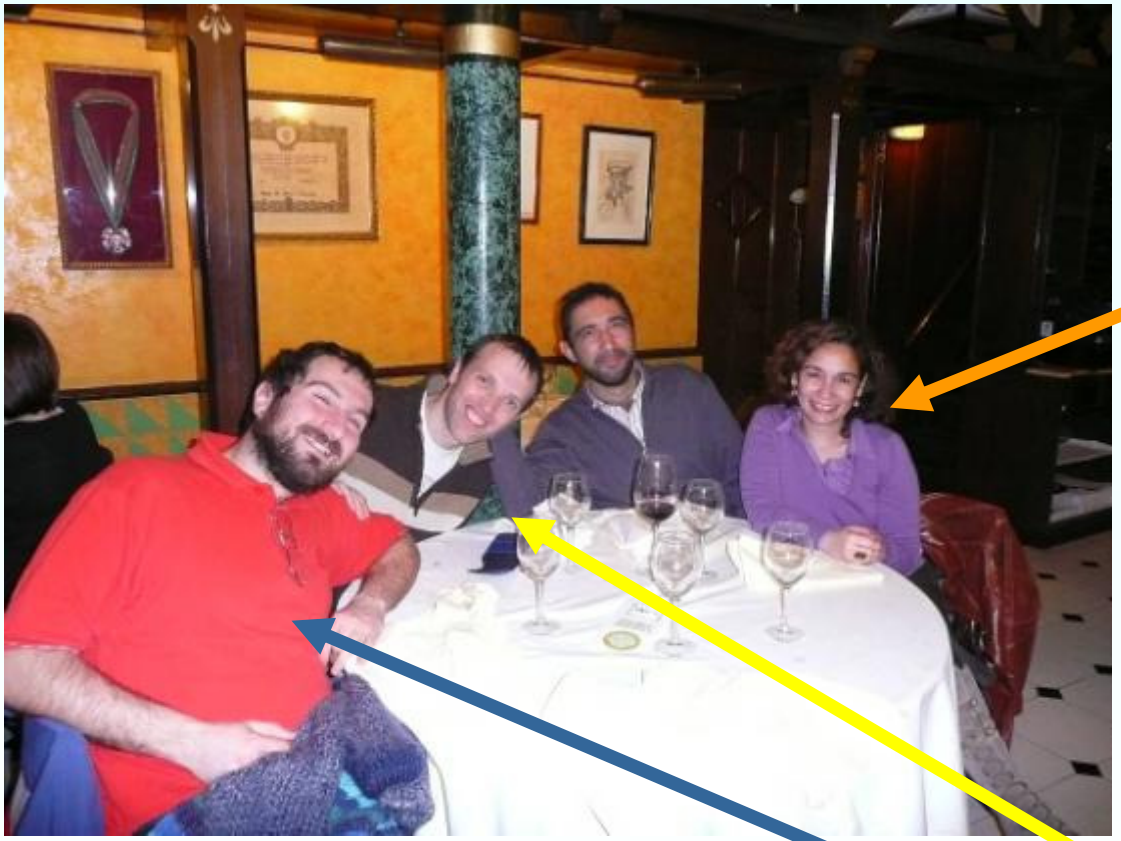
Next generation animal sequencing to meet tomorrow's needs

Miguel Pérez-Enciso

Veterinary School UAB & ICREA – Barcelona – Spain

miguel.perez@uab.es

et al.



INIA-Madrid

M.C. Valdovinos

Wageningen

M. Groenen

H.J. Megens

Andreia Amaral

UIUC – Illinois

L.B. Schook

CRG-Barcelona

H. Himmelbauer

R. Kofler

UAB

Sebas Ramos

Luca Ferretti

A. Esteve

Outline

- ☺, 😐 and ☹ of NGS technologies
- Future applications
- An example: partial resequencing of a highly inbred Iberian pig
- Challenges
- Conclusion

NGS technologies: Pros & cons

- Main advantages:
  - Massive throughput.
  - 100 fold reduction in cost / bp

- Main limitations:
  - Shotgun sequencing (sequence capture but expensive).
  - Preparation of libraries can be expensive and tedious.
  - Short reads (but 454 is now 400 bp).
  - Increased computer power and bioinformatic skills are required.

Future applications

- Quasi complete genomic data for every species
- A complete population polymorphism picture
- True genomewide selection
- Beyond the species: metagenomics

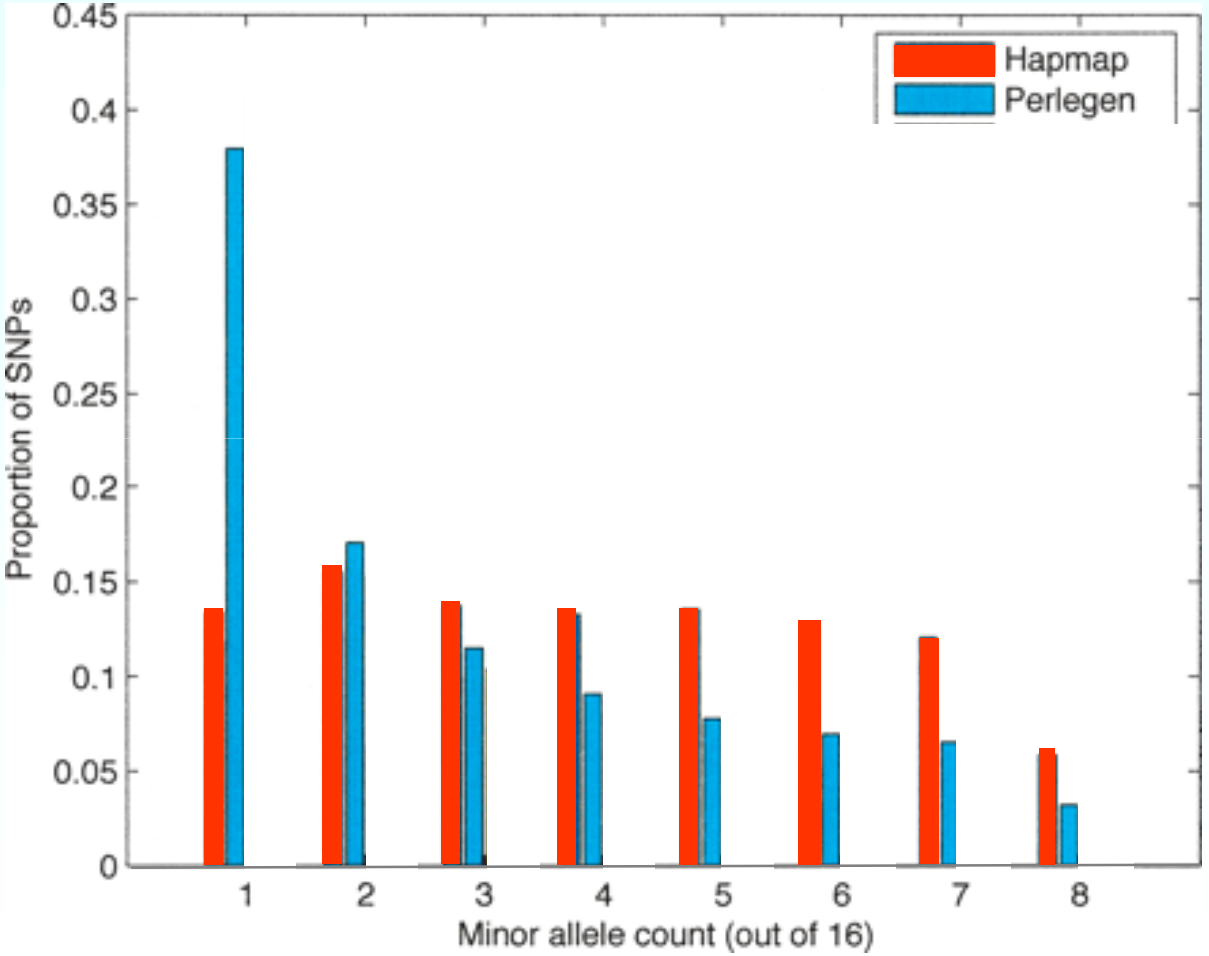
Future applications (i): Quasi complete genomic data for every species

- Thus far, comprehensive genomic data has been available only in a very few species.
- NGS will provide fast and complete data *à la carte* in any species, and will allow to catch up to other model or human species.
- Particularly relevant in aquaculture.
- A single technology may fulfill all needs (e.g., expression microarrays will be replaced by transcriptome sequencing).

Future applications (ii): A complete population polymorphism picture

- Thus far, large scale polymorphism data has been obtained from genotyping SNPs.
- These SNPs are a biased sample of all SNPs that are present in the population studied (SNP ascertainment bias).
- Full or partial resequencing will allow to detect not only all SNPs but also the rest of variants (CNVs, structural variants...).
- Software for structural variants under development.

SNP ascertainment bias: Human population frequency spectra



Clark et al. (2005) Genom Res 15:1496

Future applications (iii): True genomewide selection

- Currently, genomewide assisted selection is beginning to be implemented via massive genotyping.
- Problems: SNP bias and partial characterization of variability (eg, no structural variants).
- I envisage a future where partial shotgun / directed sequencing will be massively carried out via pools where each individual / family will be tagged individually.

Future applications (iv): Beyond the species & metagenomics

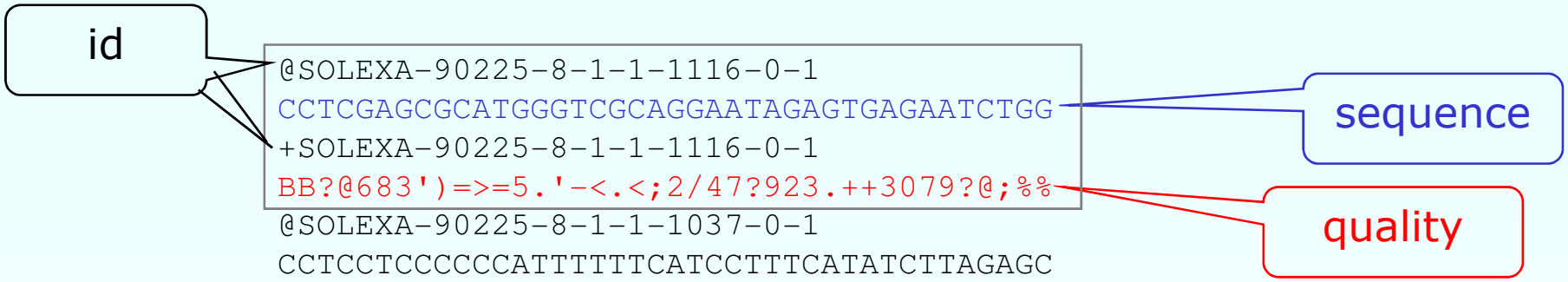
- The animals, including livestock, live in ecosystems.
- Response to selection depends on the environment, an environment that has in turn a genetic component (co-evolution).
- Therefore, a possibility is to study how selection response depends in part from external genetic environment, an immediate target could be to analyze ruminant's bacterian flora.

Example: partial resequencing of an Iberian pig

- We applied Reduced Representation Library Solexa (Genome Analyzer) resequencing to a highly inbred Iberian pig, the Guadyerbas line, which we have used in many diverse experiments.
- It is a very fat black hairless line.
- Known pedigree since its founding in 1950s.
- Digestion with HaeIII, resequencing of 160-200 bp bands.
- 3 lanes ~ 14 million reads after filtering.

How a Guadyerbas look like

How the data look like



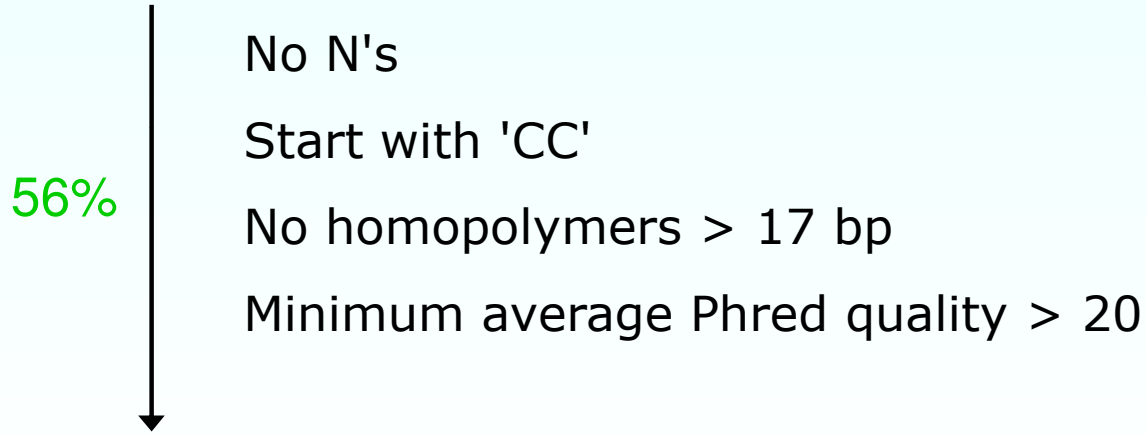

```
@SOLEXA-90225-8-1-1-1116-0-1
CCTCGAGCGCATGGGTCGCAGGAATAGAGTGAGAATCTGG
+SOLEXA-90225-8-1-1-1116-0-1
BB?@683')=>=5.'-<.<;2/47?923.++3079?@;%%
@SOLEXA-90225-8-1-1-1037-0-1
CCTCCTCCCCCCATTTTTTCATCCTTTCATATCTTAGAGC
+SOLEXA-90225-8-1-1-1037-0-1
;@9((7@=)>?5<:?7;::2;(?3<13/<(6@?/(//8/<
@SOLEXA-90225-8-1-1-1133-0-1
CCTCCTGTACCTAACTGTGCACAAGCTGCTTTTACCCACT
+SOLEXA-90225-8-1-1-1133-0-1
BCBCB<18ABB?99?=(8/?9@@;0?<'?@@B=5>A?<?B
@SOLEXA-90225-8-1-1-2027-0-1
CCCCGCCCTTCCCTCTGCTGACTCCATCGTTCCCGCAGCT
+SOLEXA-90225-8-1-1-2027-0-1
BBB>0<?@5':A@7=@<>3*(>@=8''&*,@BA>%%%%%%
@SOLEXA-90225-8-1-2-1861-0-1
CCACGCTGCATGGAATCCACTGCCTTCAATCGAGATCCAT
+SOLEXA-90225-8-1-2-1861-0-1
BCBB7@:/?9>;0:B?BA@A@;?/;;A;>9;%%%%%%%%%
```

Example analysis

- Bioinformatic analysis:
  - Filtering
  - Assembly against reference (assembly 9)
  - SNP detection
- Genetic analysis:
  - Distribution of variability
  - Inference of genetic parameters

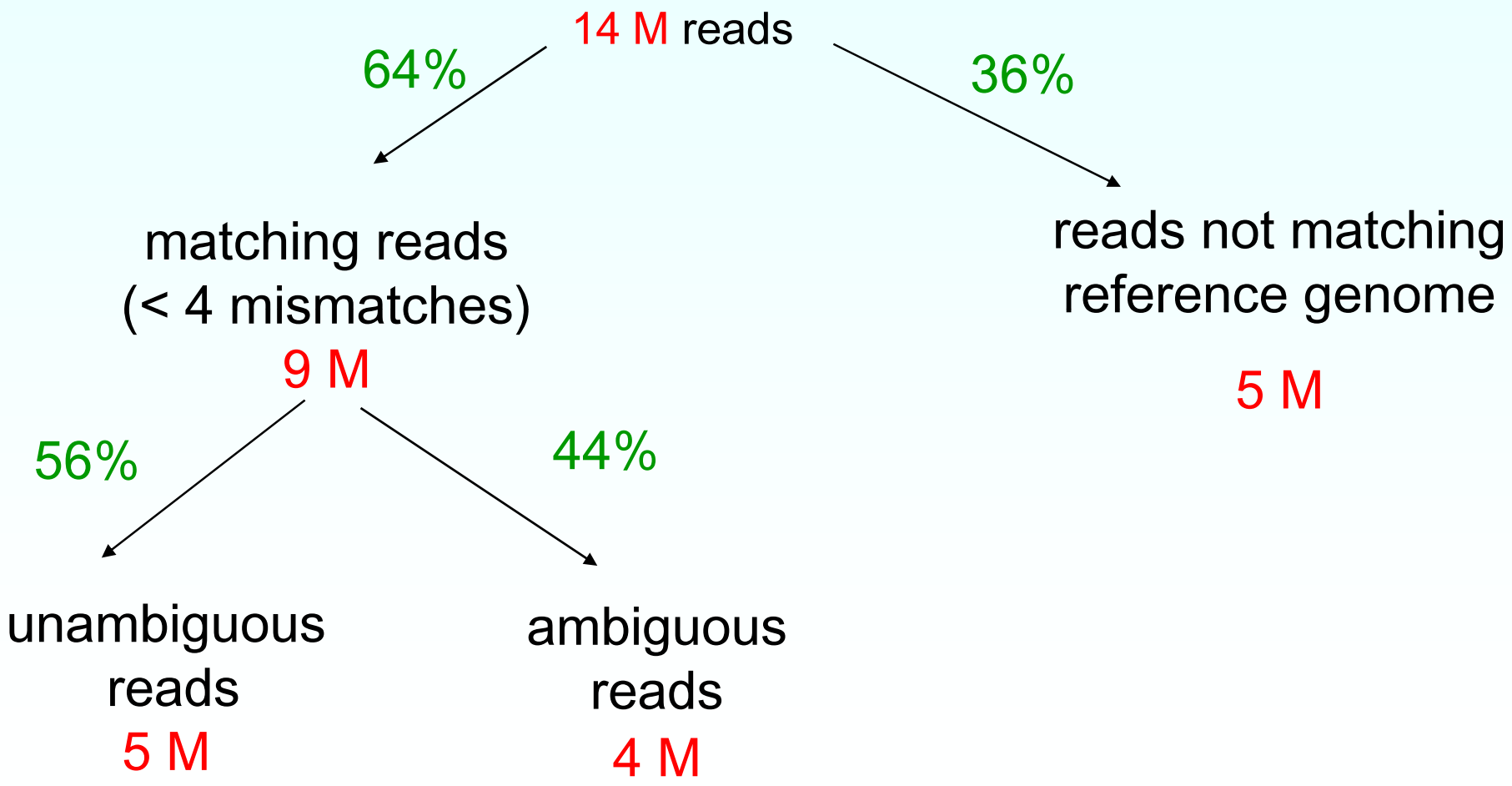
Bioinformatics (i): filtering

25.2 M raw sequences 50 bp



14.1 M filtered sequences 40 bp

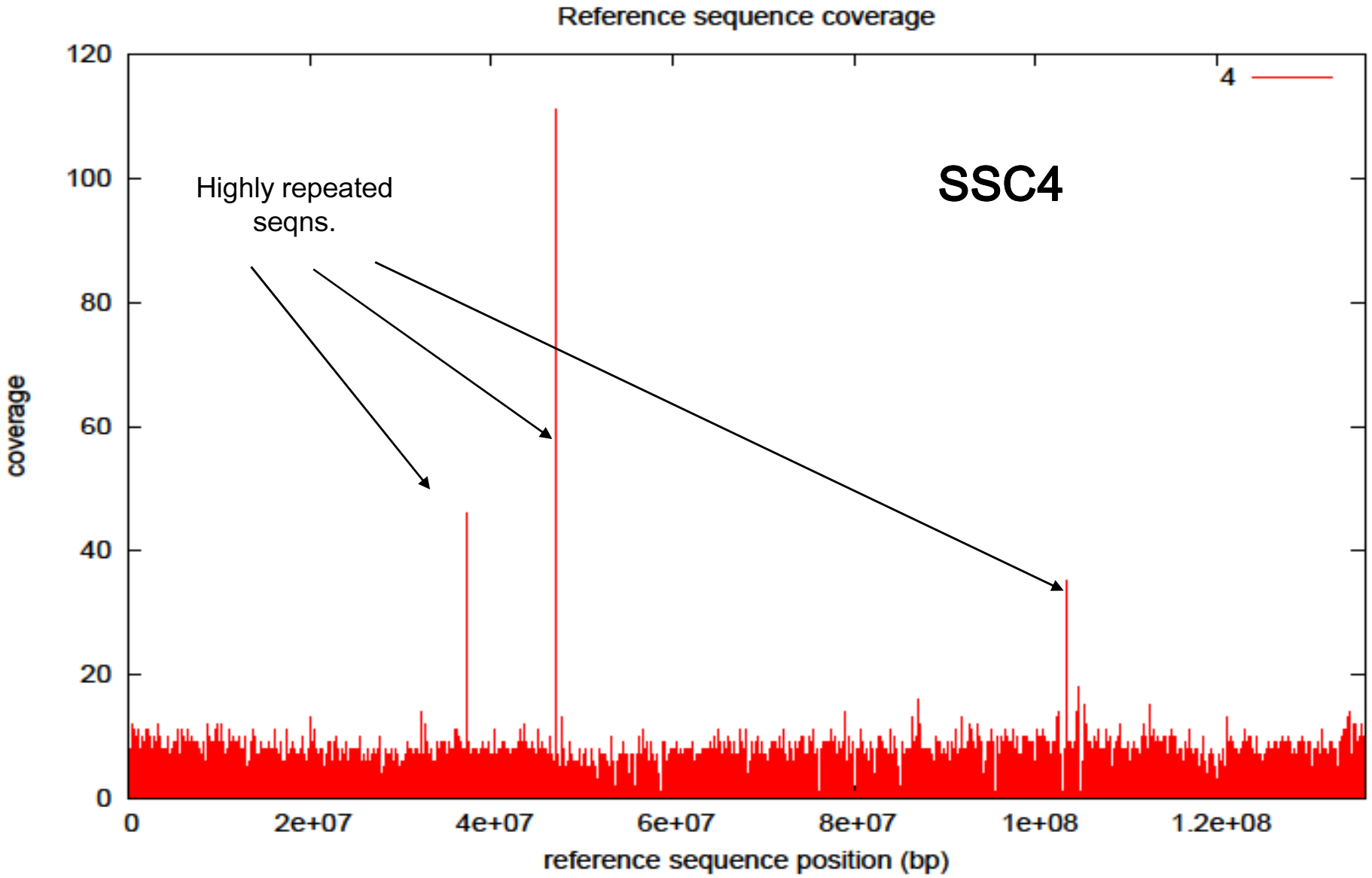
Bioinformatics(ii): Alignment and mapping



Summary

✓ Total assembled	2262.6 Mb
✓ Total sequenced	83.1 Mb
✓ Total sequenced (>2-20x)	25.1 Mb
✓ Average coverage (3-20x)	4.0x

~ 1% genome at 4x


Reference sequence coverage
4
SSC4
Highly repeated
seqns.
120
100
80
60
40
20
0
coverage
0
2e+07
4e+07
6e+07
8e+07
1e+08
1.2e+08
reference sequence position (bp)

Bioinformatics (iii): SNP detection

- ✓ Gem (P. Ribeca, unpublished)
- ✓ MAQ (H. Li et al.)
- ✓ Mosaik suite (G. Marth et al.)

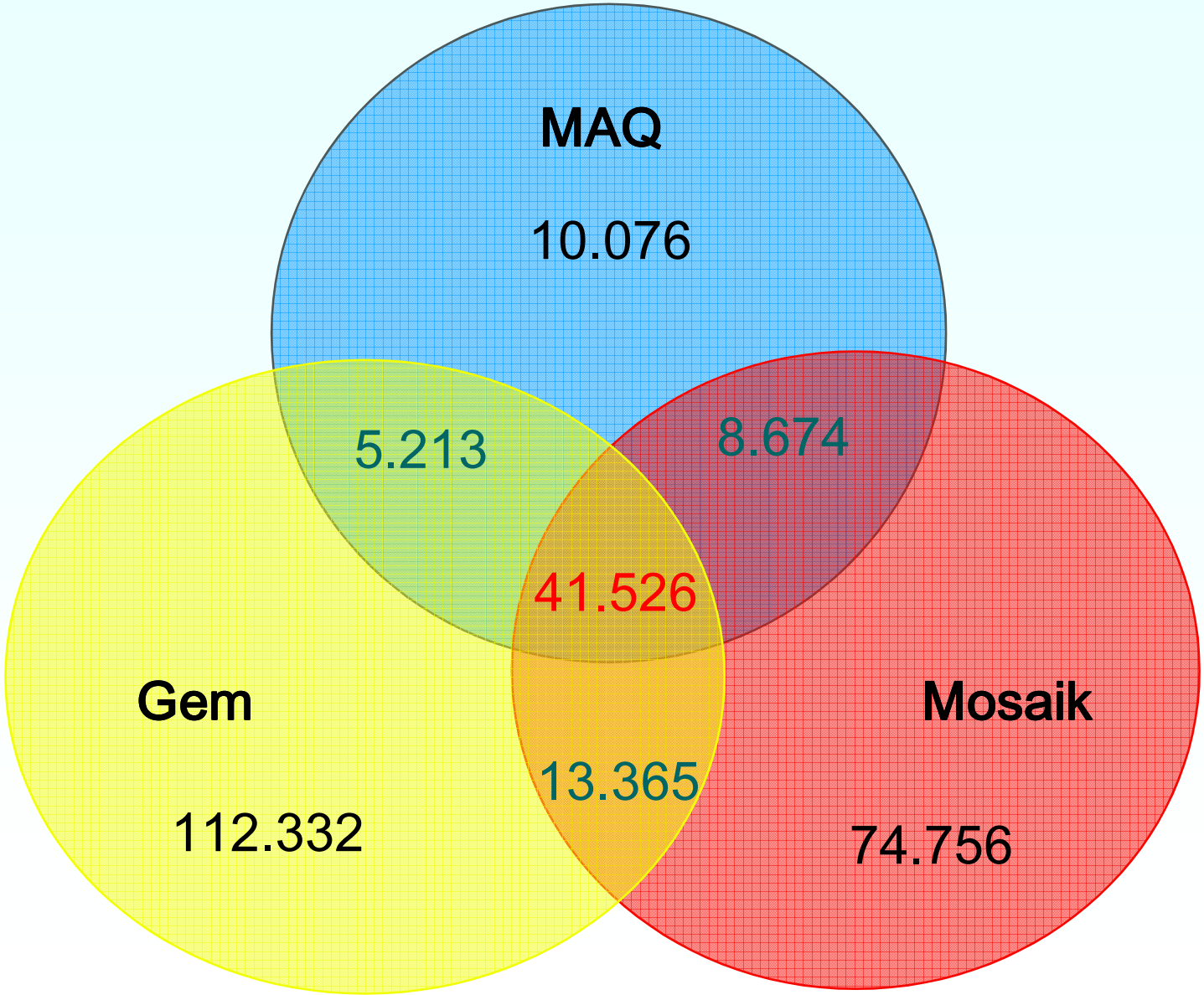
  Mosaik: alignment and assembling

  Gigabayes: short-read SNP and short-INDEL discovery program

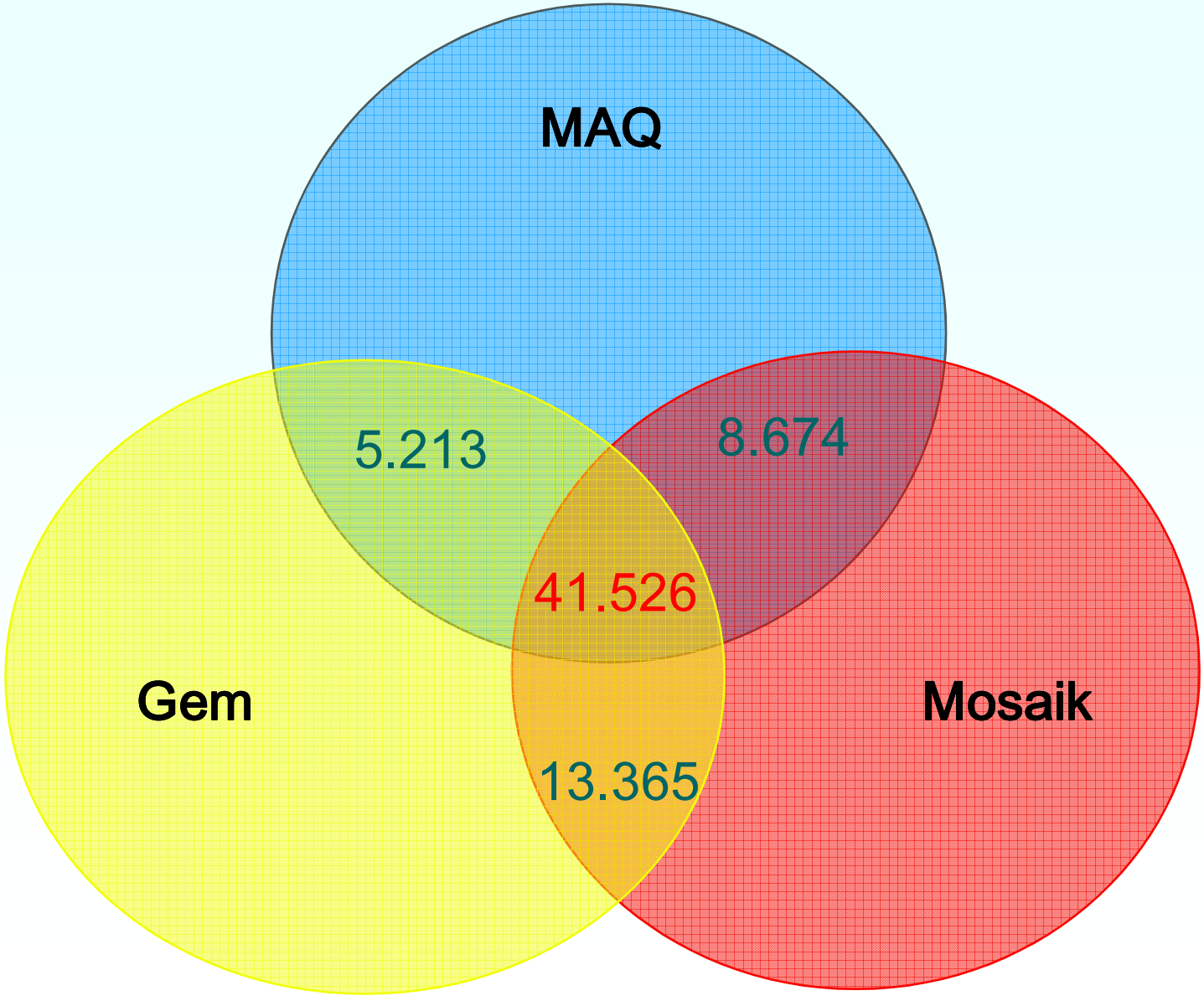
  EagleView: genome assembler viewer

EagleView SNP visualization

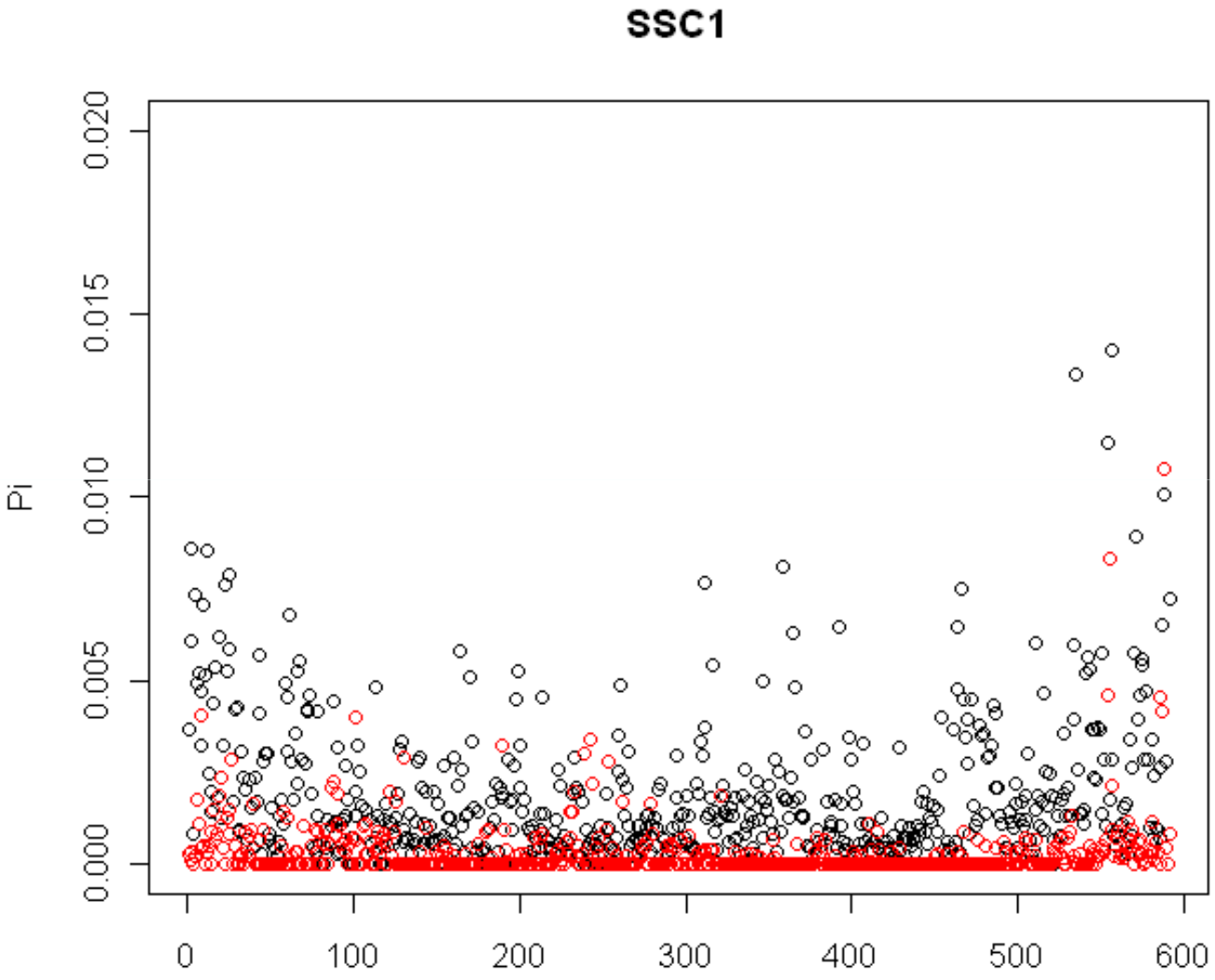
265,842 SNPs detected



68,778 SNPs retained



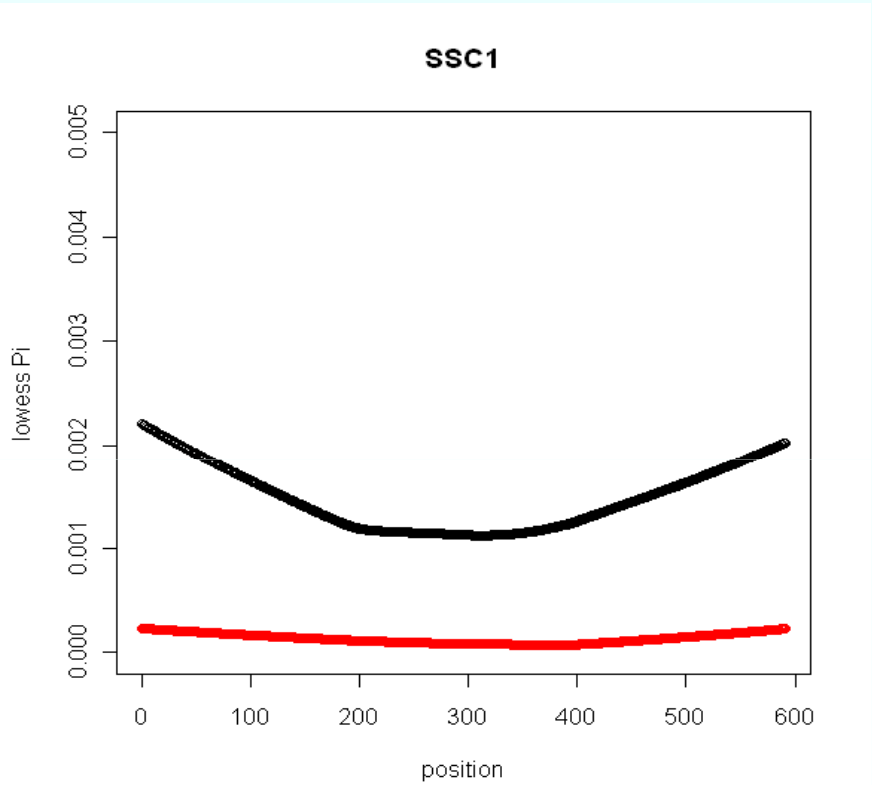
Genetic analysis (i): variability

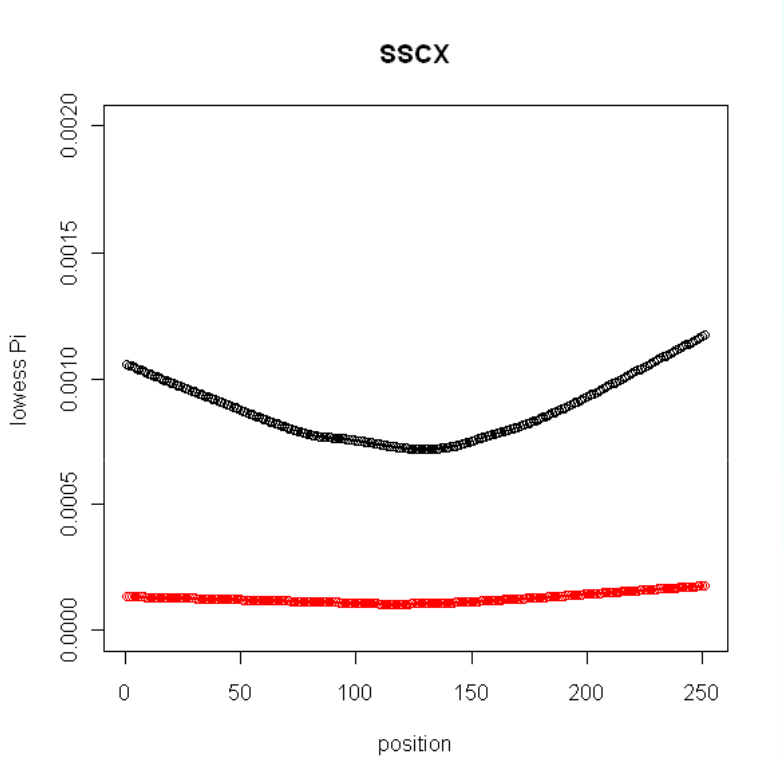


black: fixed differences with assembly

red: SNPs Iberian

Genetic analysis (i): variability





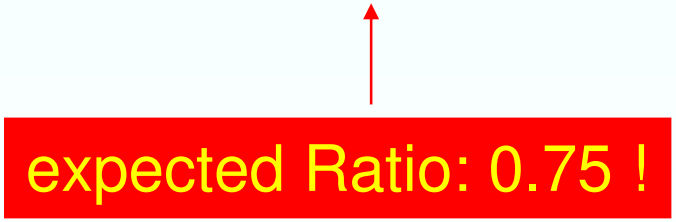
lowess adjusted curves

black: fixed differences with assembly

red: heterozygous Iberian

Genetic analysis (i): nucleotide diversity

	Autosomes	SSCX	Ratio
Fixed Differences	2.3×10^{-3}	1.2×10^{-3}	0.52
SNP	5.3×10^{-4}	2.8×10^{-4}	0.53



In Amaral et al (2009): 0.36

Genetic analysis (ii): HKA

$$T = \frac{\sum_i F_i}{\sum_i S_i} - 1$$

Estimated divergence time (2N units)

Autosomes: 3.16

SSCX: 3.26

$$\hat{\theta}_i = \frac{F_i + S_i}{T + 2} - 1$$

Estimated theta (2N units)

Autosomes: 0.69×10^{-3}

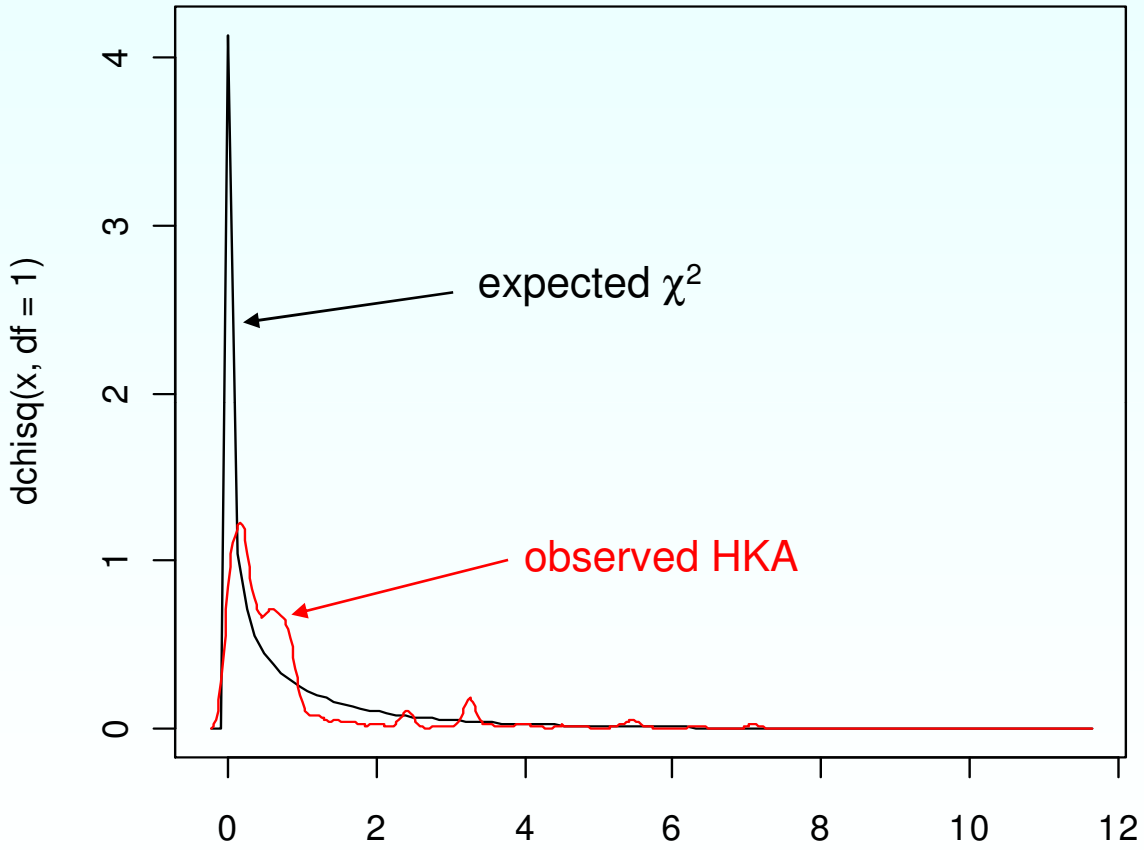
SSCX: 0.28×10^{-3}

$$\chi^2 = \sum_i \left[\frac{[S_i - E(S_i)]^2}{Var(S_i)} + \frac{[F_i - E(F_i)]^2}{Var(F_i)} \right]$$

Fi = Fixed variants between Iberian and assembly (Duroc)

Si = Segregating sites (SNPs) within Iberian

Genetic analysis (ii): HKA



Expected (χ^2, 1df) black line vs. observed HKA statistics

In summary,

- NGS technology is rapidly being developed.
- Main constraints are targeted sequencing and multiple tagging.
- Applications are in its infancy, so far mainly to obtain cheaply thousands of SNPs.
- But I have shown how can we easily go beyond that to grasp how selection and demography has shaped nucleotide variability.
- Many more applications ...

Challenges

- Hardware: Distributed storage and computing.
- Bioinformatics: It is currently the bottleneck.
- Population genomics: How to infer population genetic parameters, accurately but feasibly?
- Simulation tools: To study association and selection schemes.
- Statistics: How to improve upon current association techniques?
- Animal breeding: How to implement True Genome Selection?

CONCLUSION

1. NGS is changing dramatically how genomics research is carried out.
2. It should have an impact also in how funding is optimally allocated and in future PhD education.
3. I envisage that most dynamic research will be carried out by relatively small labs / centers rather than big consortia.
4. Bioinformatics will continue to be for a while an important bottleneck, but not serious enough provided minimum numeric skills.