ROSLIN

EAAP 2016 - 67th Annual Meeting of the
European Federation of Animal Science
Belfast UK, 29 Aug – 2 Sept 2016

A family-based imputation algorithm for high- and low-coverage sequence data

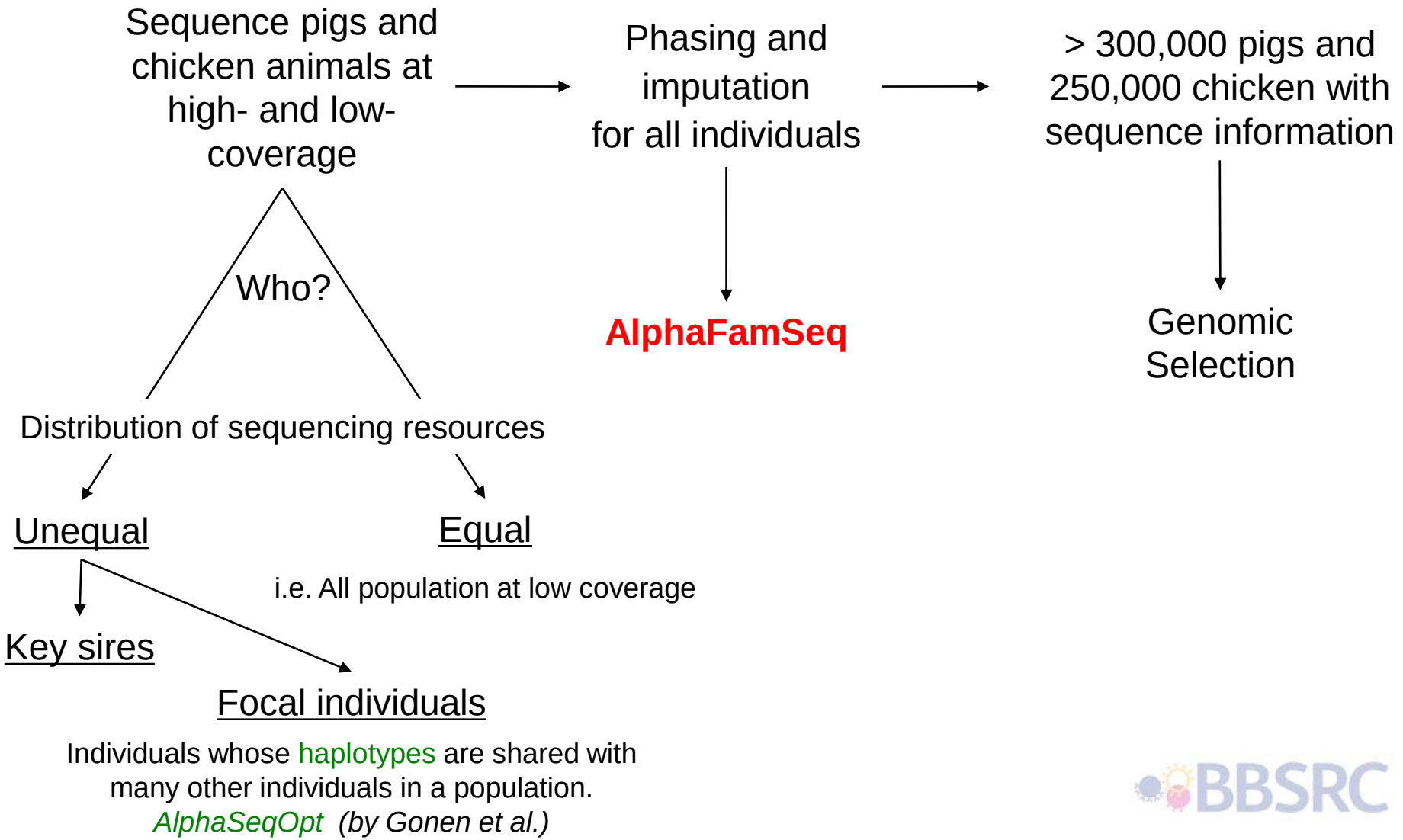
Mara Battagin*
Gregor Gorjanc
Serap Gonen
Roberto Antolin
John Hickey

* mara.battagin@roslin.ed.ac.uk

THE UNIVERSITY of EDINBURGH

BBSRC

Genomic Selection 2.0



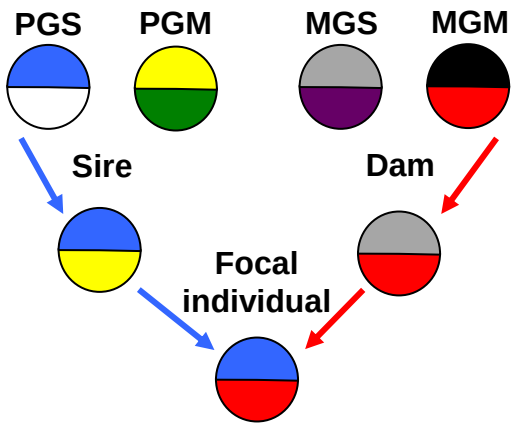
“heuristic phasing algorithm for high- and low-coverage sequence data”

AlphaFamSeq

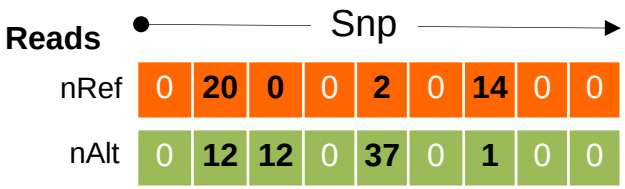
AlphaSuite program

Family-based method

Developed for sequence data



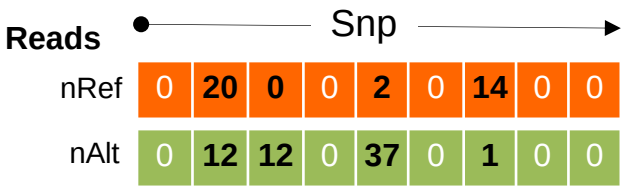
Method
AlphaFamSeq

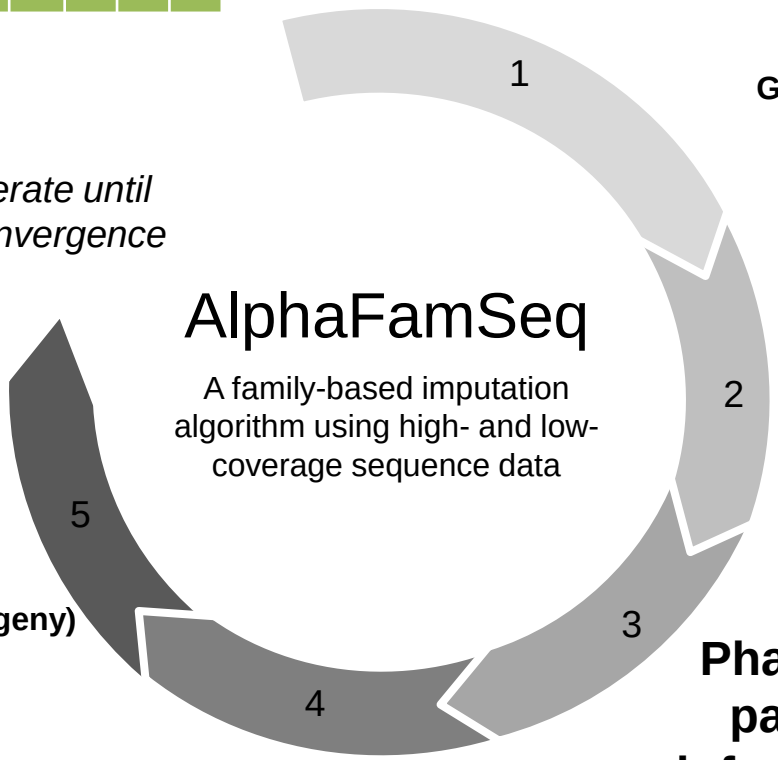


AlphaFamSeq

A family-based imputation algorithm using high- and low-coverage sequence data

Method
AlphaFamSeq





Variant Calling
1% error rate
Genotype threshold >0.99

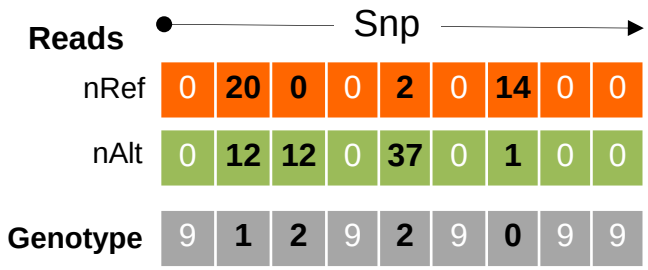
Phasing based on parents informative markers

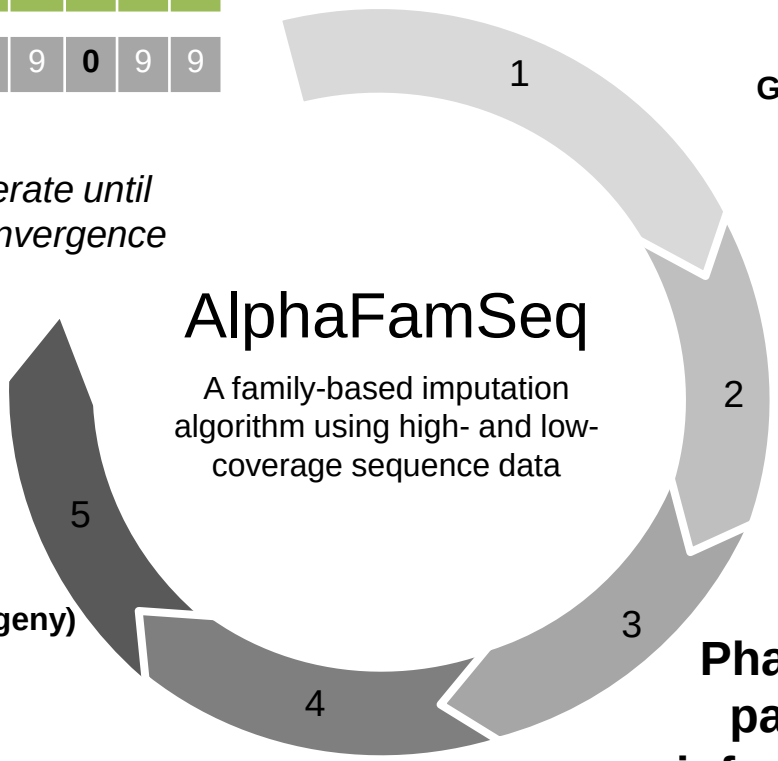
Phasing based on parent-progeny informative markers

Founder Assignment

Build the trio consensus
(grandparent – parent – progeny)

Method
AlphaFamSeq





Variant Calling
1% error rate
Genotype threshold >0.99

Phasing based on parents informative markers

Phasing based on parent-progeny informative markers

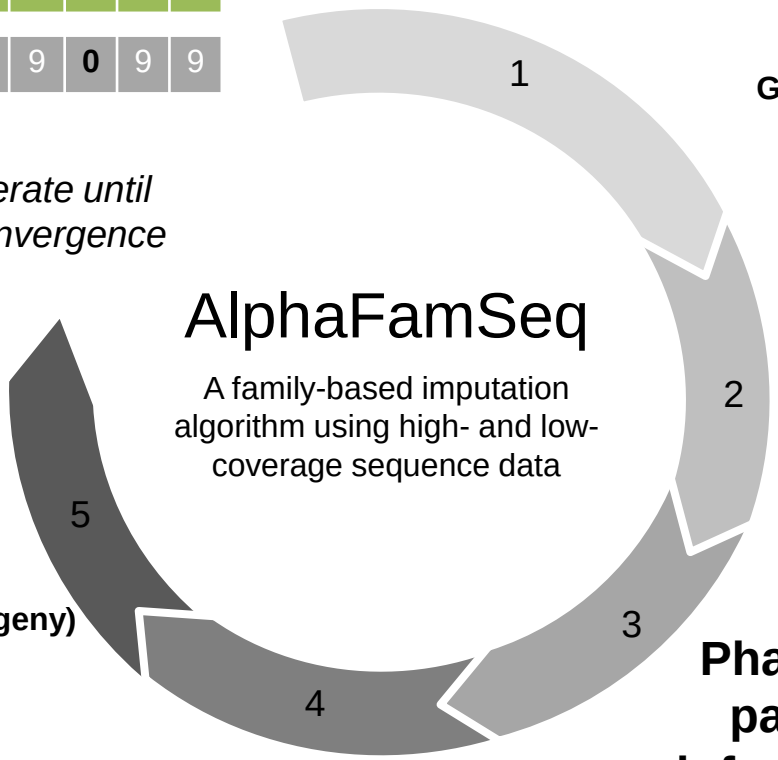
Founder Assignment

Build the trio consensus
(grandparent – parent – progeny)

Iterate until convergence

Method
AlphaFamSeq

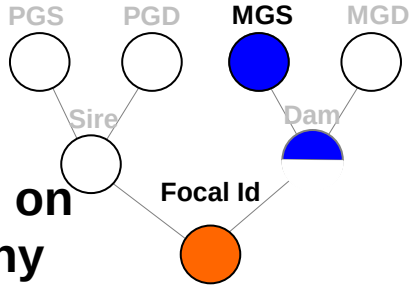
Reads	Snp →								
nRef	0	20	0	0	2	0	14	0	0
nAlt	0	12	12	0	37	0	1	0	0
Genotype	9	1	2	9	2	9	0	9	9



Variant Calling
1% error rate
Genotype threshold >0.99

Iterate until convergence

Phasing based on parents informative markers



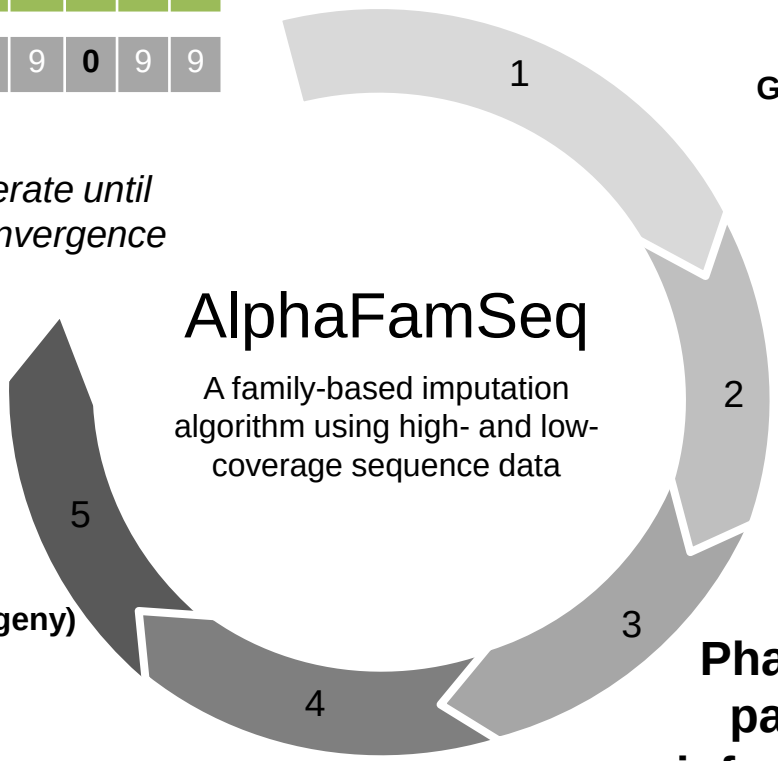
Phasing based on parent-progeny informative markers

Founder Assignment

Build the trio consensus
(grandparent – parent – progeny)

Method
AlphaFamSeq

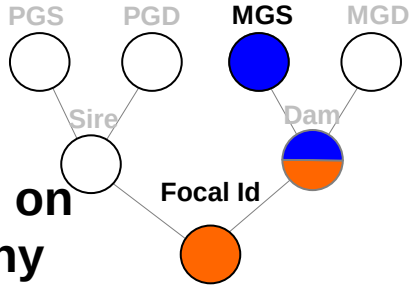
Reads	Snp →								
nRef	0	20	0	0	2	0	14	0	0
nAlt	0	12	12	0	37	0	1	0	0
Genotype	9	1	2	9	2	9	0	9	9



Variant Calling
1% error rate
Genotype threshold >0.99

Iterate until convergence

Phasing based on parents informative markers



Phasing based on parent-progeny informative markers

Founder Assignment

Build the trio consensus
(grandparent – parent – progeny)

Method
AlphaFamSeq

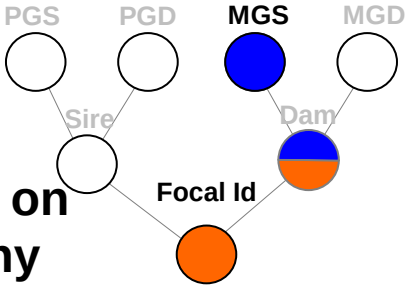
Reads	Snp →								
nRef	0	20	0	0	2	0	14	0	0
nAlt	0	12	12	0	37	0	1	0	0
Genotype	9	1	2	9	2	9	0	9	9

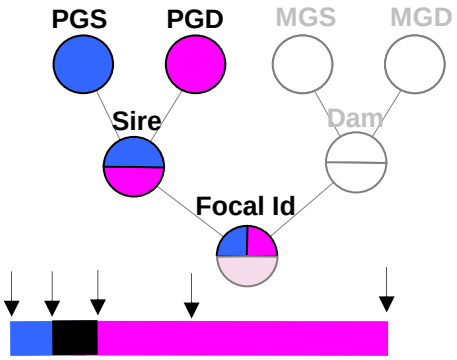
AlphaFamSeq

A family-based imputation algorithm using high- and low-coverage sequence data

Iterate until convergence

1. **Variant Calling**
   1% error rate
   Genotype threshold >0.99
2. **Phasing based on parents informative markers**
3. **Phasing based on parent-progeny informative markers**
4. **Founder Assignment**
5. **Build the trio consensus**
   (grandparent – parent – progeny)





Method
Algorithm tested through Simulation

Simulate **true genotypes (AlphaSim)**

10 chromosomes

1 Morgan length

100,000 segregating sites

2 scenarios

Within Family imputation

Selected 5 families as replicates

Simulate sequence data different scenarios

Number of family members sequenced (from 1 to 7)

Sequencing depth (from 0 to 30x)

Accuracy of phasing the focal individual for 823,543 scenarios

Multiple Families imputation

External **cattle** pedigree (14,814 ids)

Extracted 3 generations for 500 sires

Total 1,614 animals

Key sires

Equal distribution

Focal individuals

Budget

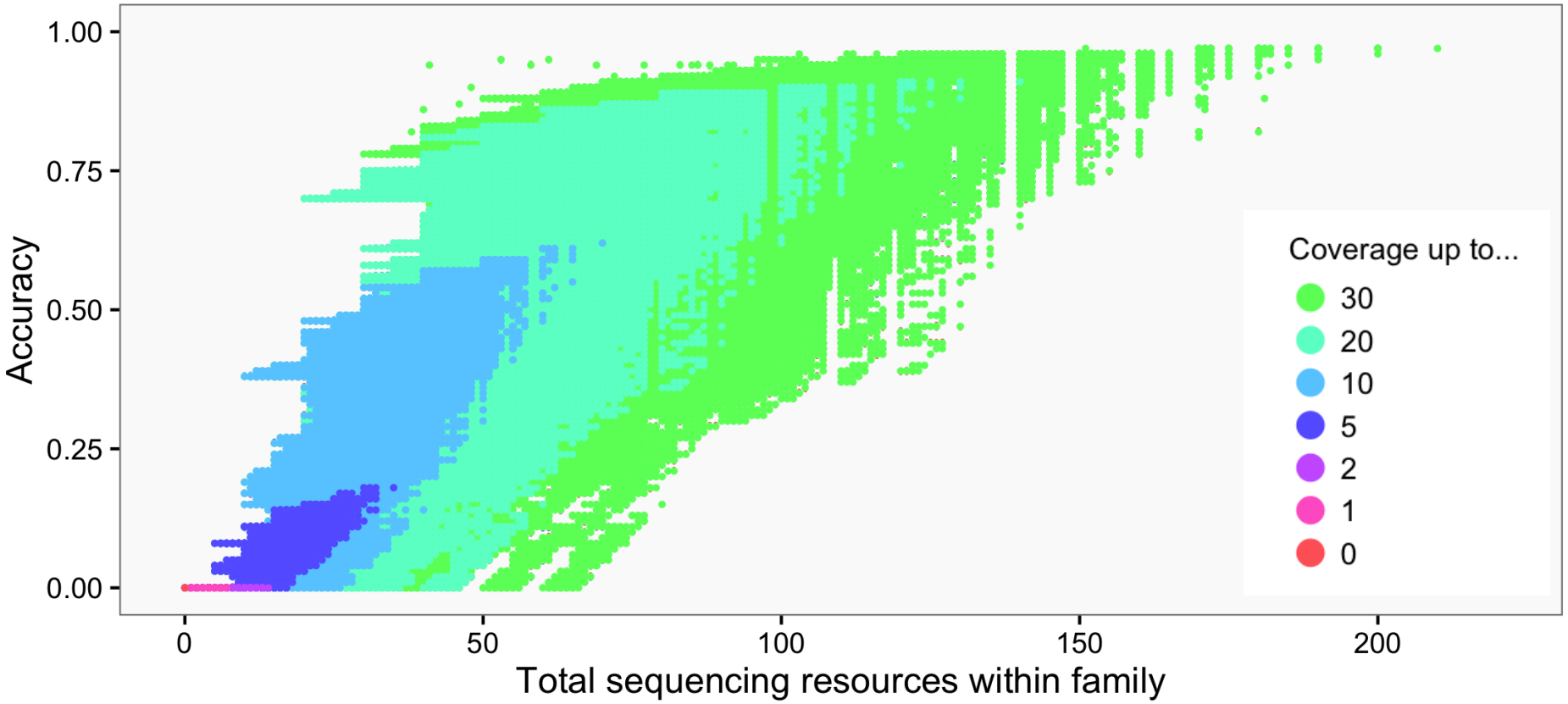
£ 420,000

AlphaSeqOpt

(by Gonen et al.)

Results

Accuracy of phasing within family



Cost of the library/individual = £40

Cost of 1*x* whole genome = £80

Accuracy of phasing multiple families

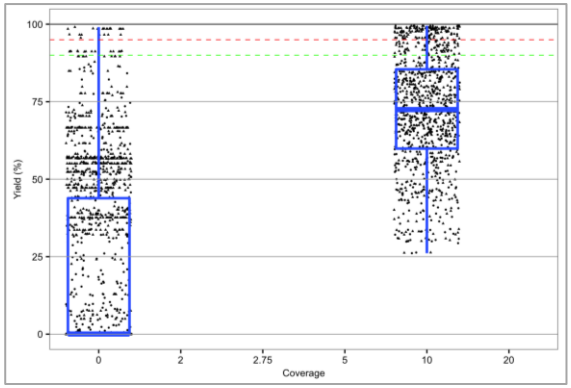
Percentage of **Correct** and **Wrong** phased gamete by read depth

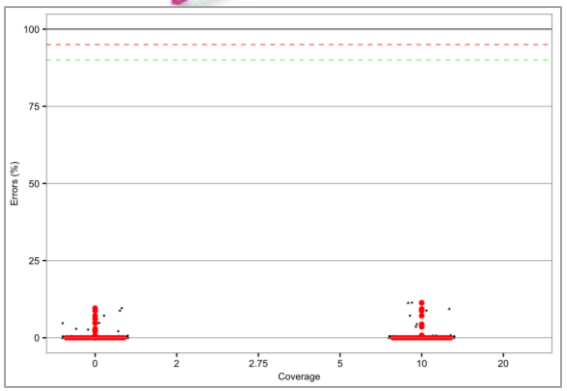
Key Sires

Min 0.00

Mean 35.41

Max 99.50



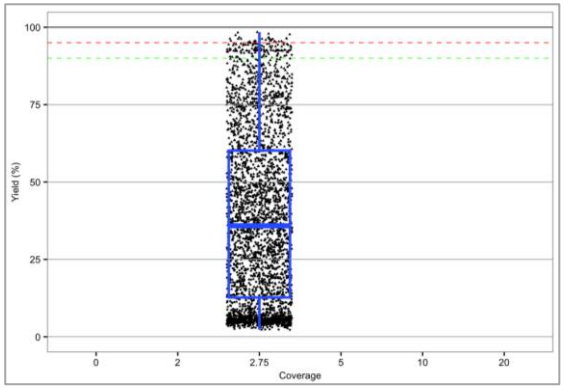


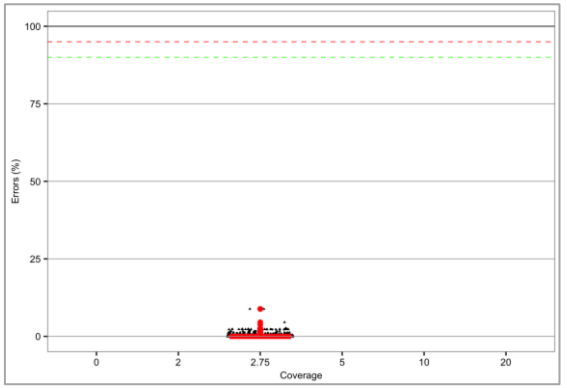
All population

Min 2.25

Mean 39.05

Max 98.60



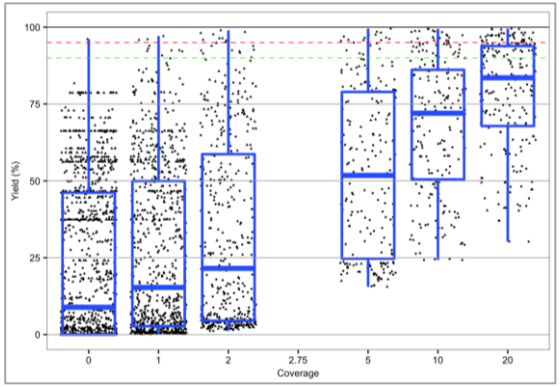


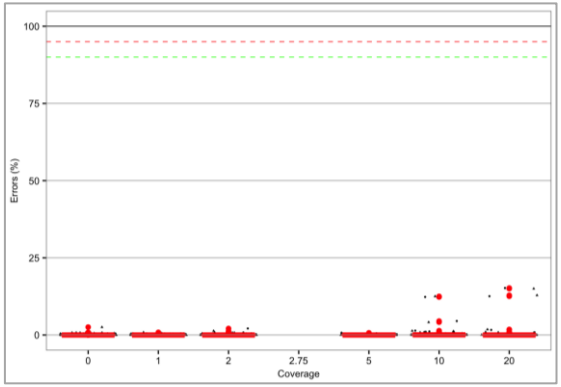
Focal individuals

Min 0.00

Mean 34.28

Max 99.55





Conclusions

Heuristic phasing algorithm for
high- and low-coverage sequence data

Able to increase the **sequence information**
in livestock population

Good accuracy of can be achieved

ToDoList:
Test it with the **real** data!

Acknowledgements

Thank you!

mara.battagin@roslin.ed.ac.uk
http://www.alphagenes.roslin.ed.ac.uk/

John Hickey
Gregor Gorjanc
Roberto Antolin
Serap Gonen
Diarmaid De Burca
Andrew Derrington
Chris Gaynor
Paolo Gottardo
Janez Jenko
Daniel Money
Roger Ros Freixedes
Maria Sanchez Perez
Masoud Shirali
Adriana Somavilla

Funding Bodies

Industrial Partners

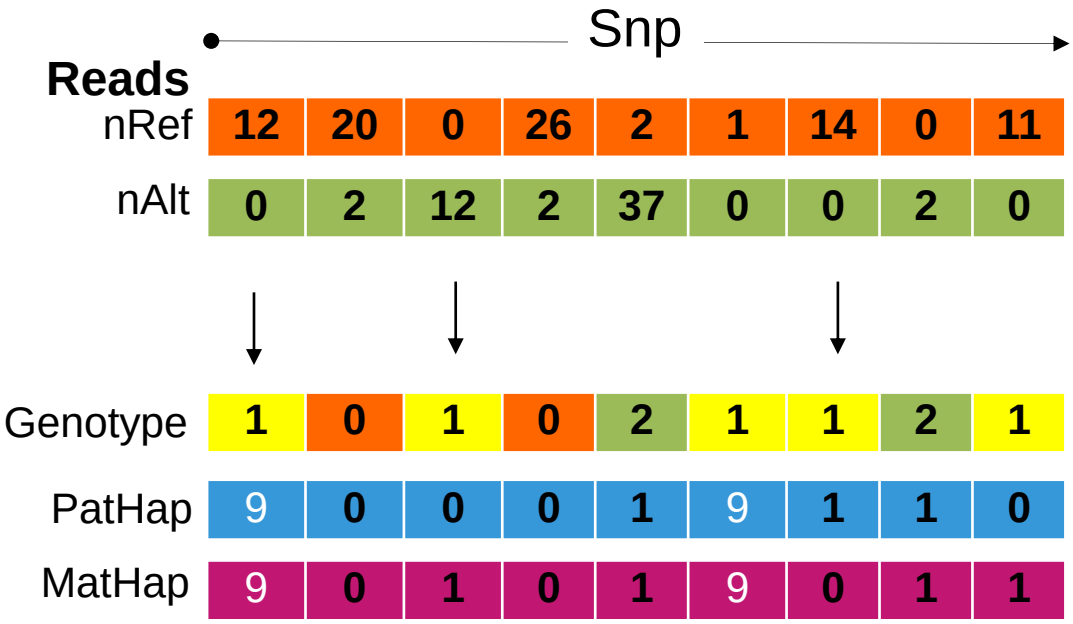
Method
multiple-families sequence data

Coverage	Unequal distribution of resources		Equal distribution of resources
	Key Sires	Focal Individuals	
0x	916	635	
1x		420	
2x		224	
2.75x			1,614
5x		110	
10x	500	116	
20x		109	
30x			
Cost (£)	420,000	419,640	419,640

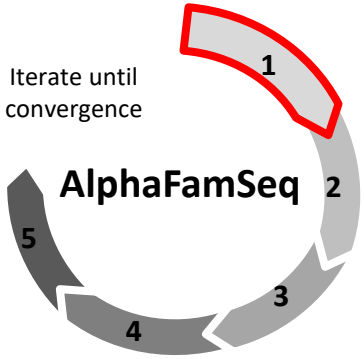
Cost of the library/individual = £40
Cost of 1*x* whole genome = £80

1

Variant calling



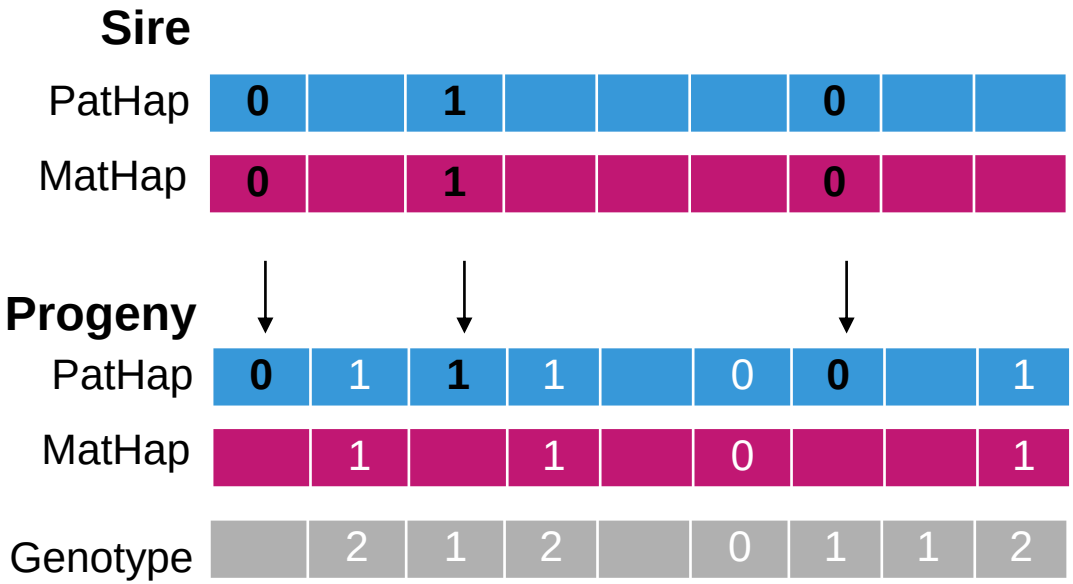
Parameters:
Error Rate = 0.01
Genotype threshold = 0.999

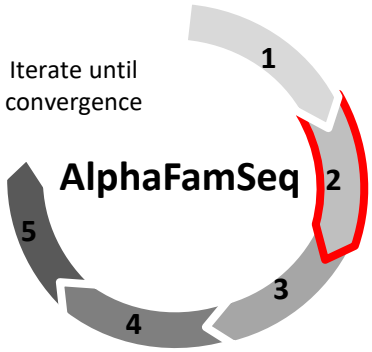


2

Phasing based on parents informative markers

Phase the markers inherited from homozygous parents



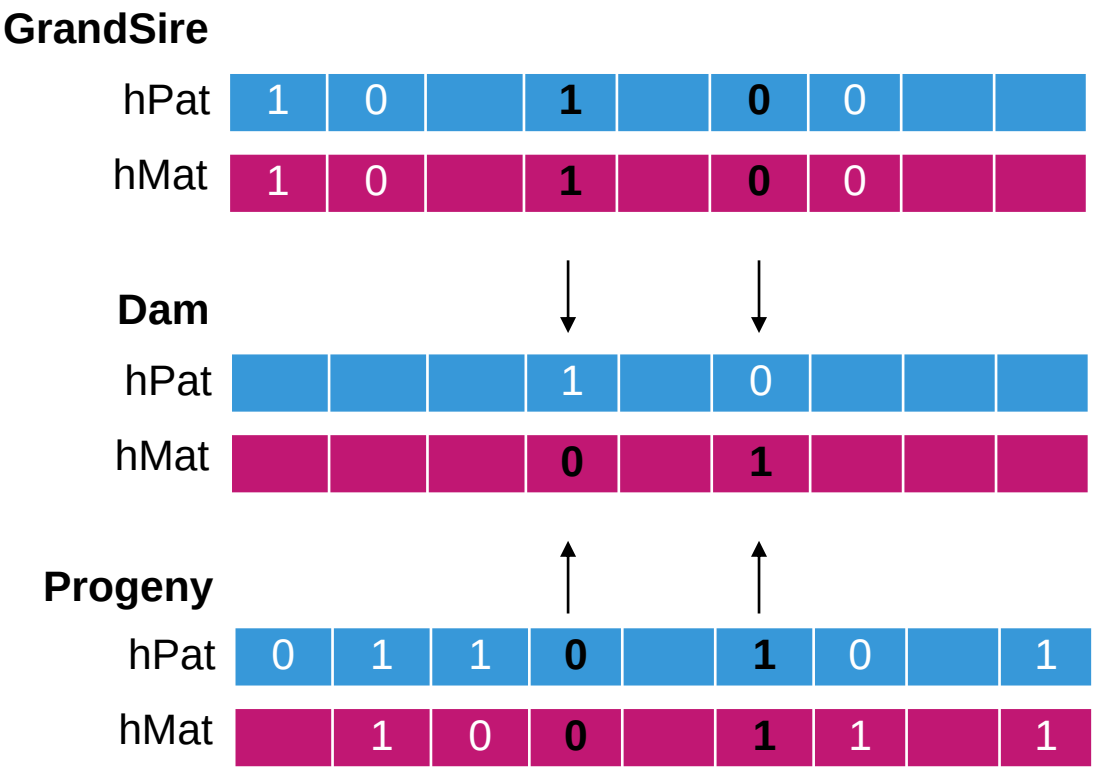


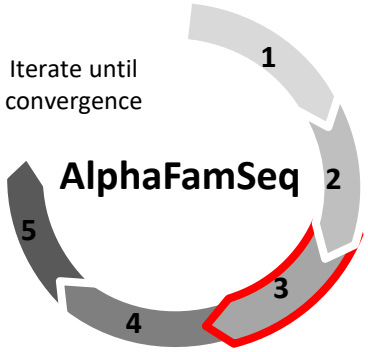
ROSLIN

3

Phasing based on parent-progeny informative markers

Phase the heterozygous markers of individuals that have
a parent and progeny with opposite alleles

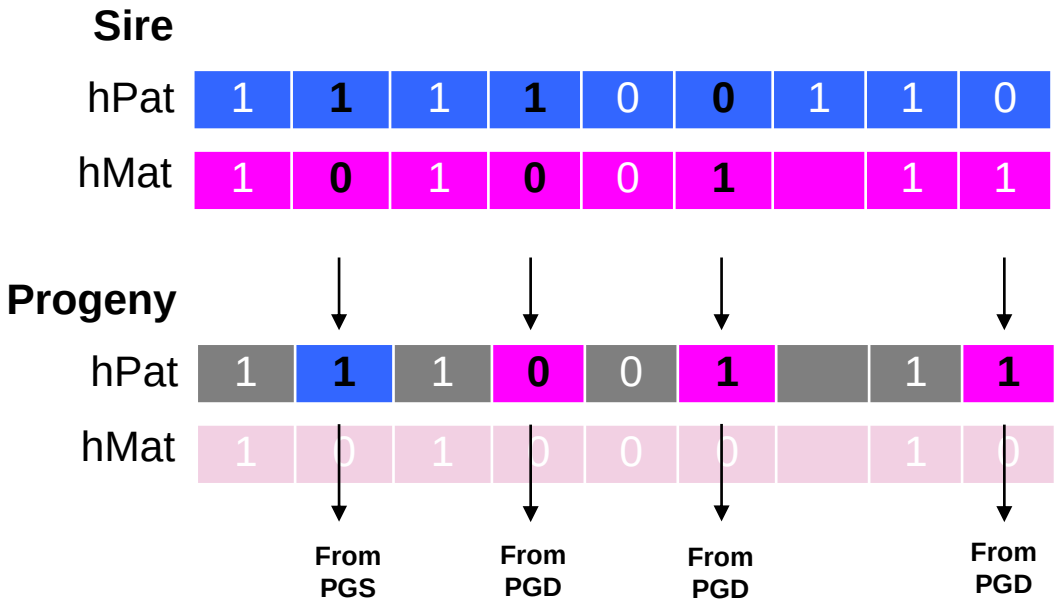


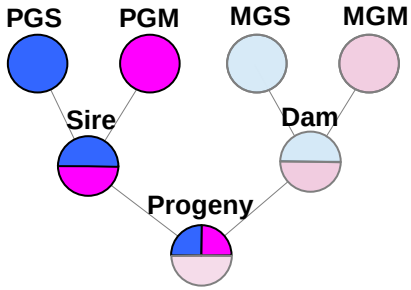


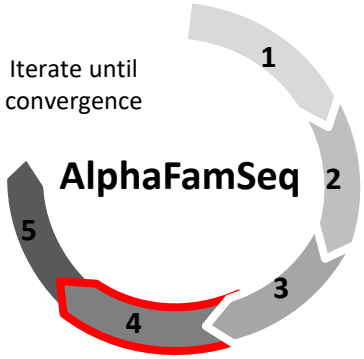
4

Founder Assignment

Identify trios (grandparent, parent, individual) by assigning grandparents to individuals using the heterozygous markers of their parents



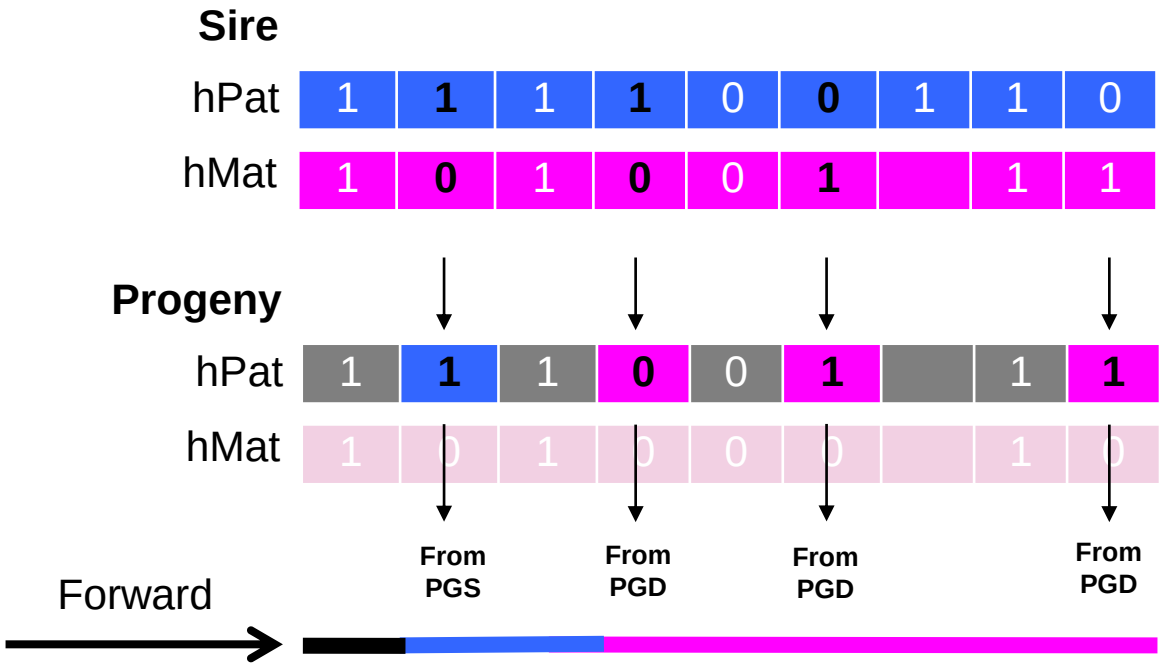


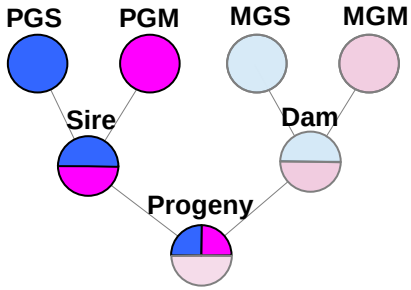


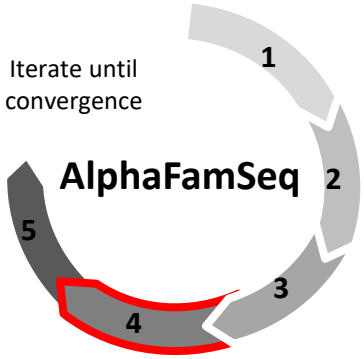
4

Founder Assignment

Identify trios (grandparent, parent, individual) by assigning grandparents to individuals using the heterozygous markers of their parents



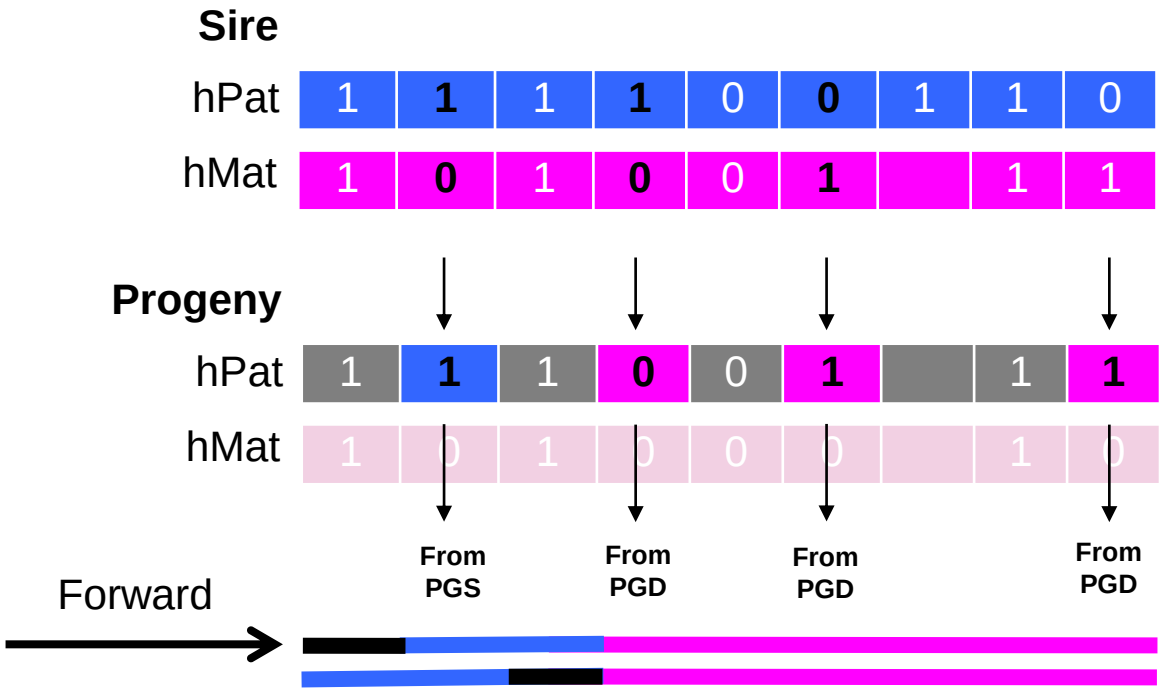


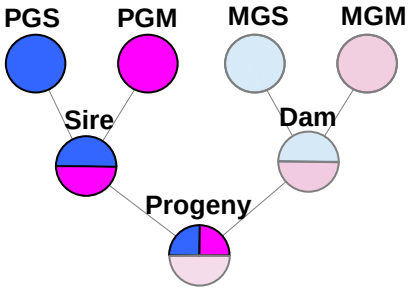


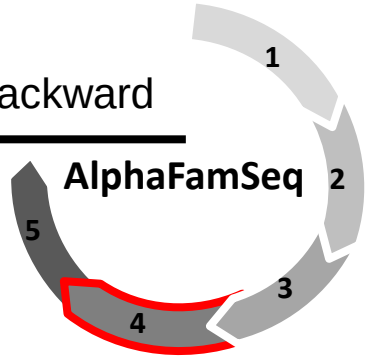
4

Founder Assignment

Identify trios (grandparent, parent, individual) by assigning grandparents to individuals using the heterozygous markers of their parents



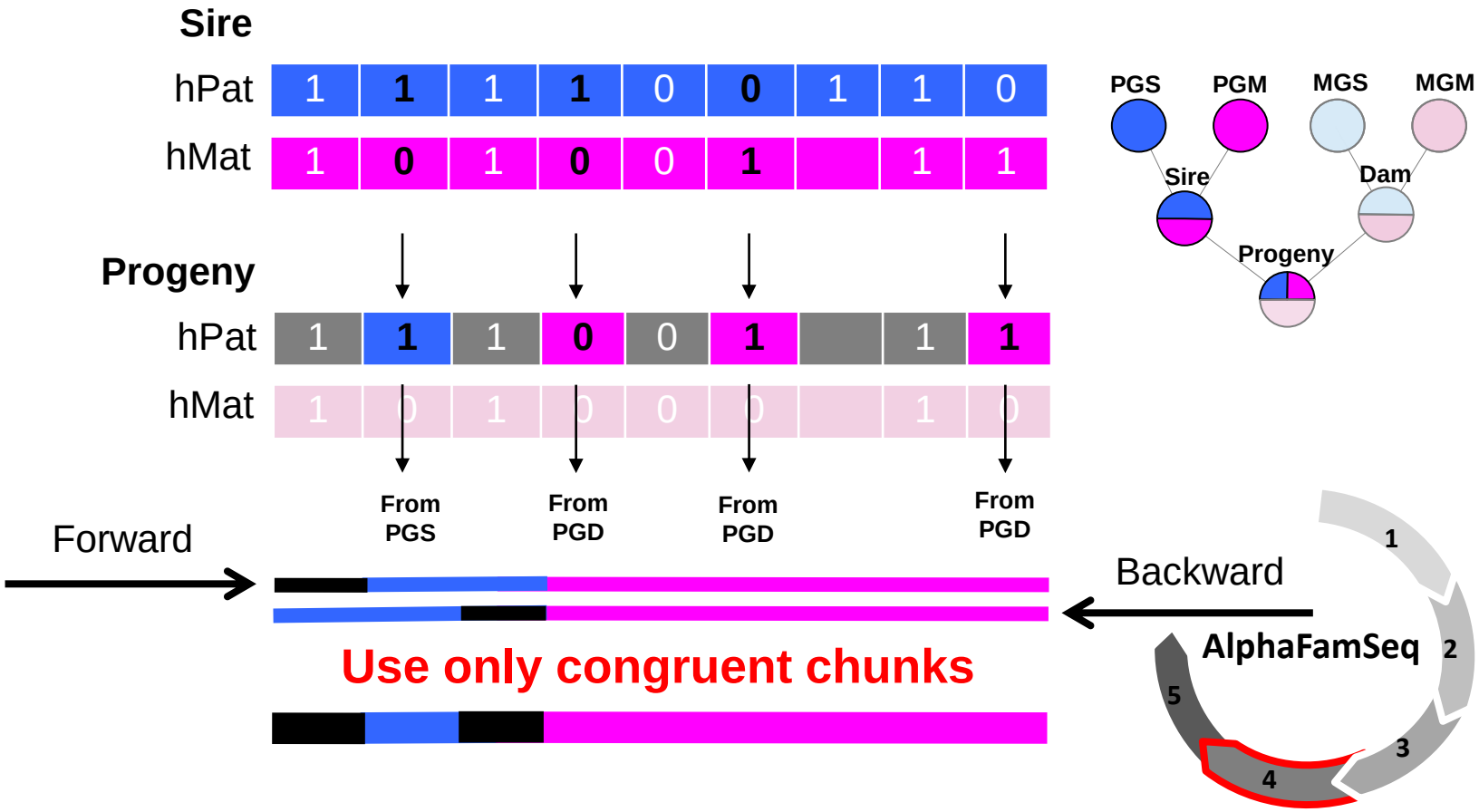




4

Founder Assignment

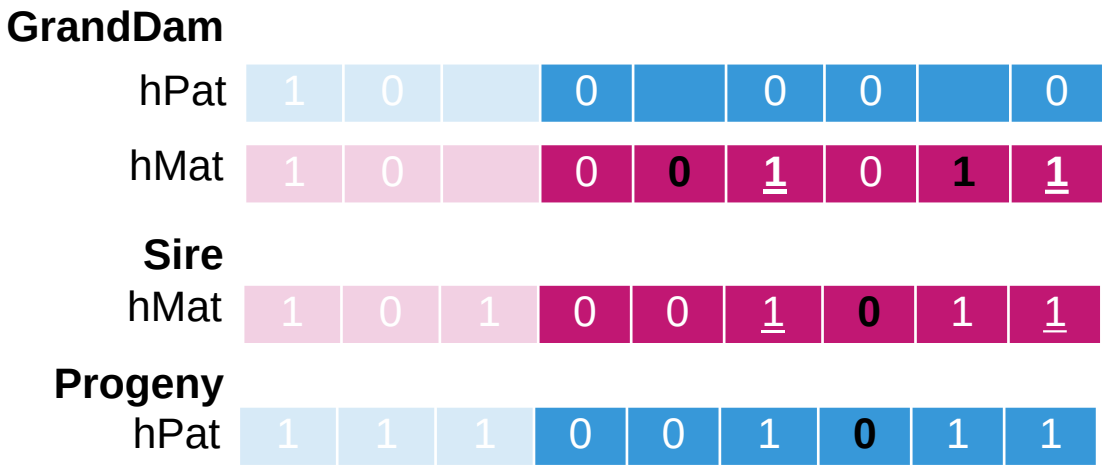
Identify trios (grandparent, parent, individual) by assigning grandparents to individuals using the heterozygous markers of their parents

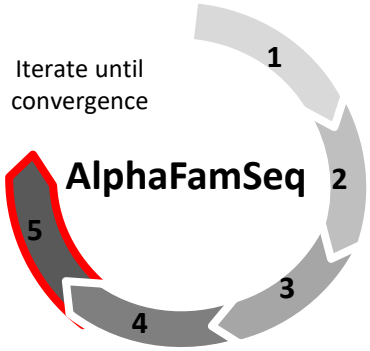


5

Build the trio consensus

Identify chunks of haplotypes within trios and use the chunks to impute missing markers between members of the trio.





Accuracy of phasing multiple families

Chunk of a paternal gamete (100 SNP) for 1,614 individuals sequenced according to three different strategies