Storing and analysing a million genomes on a desktop computer

Gregor Gorjanc, Jana Obsteter, Gabriela Mafra Fortuna, Roger Ros-Freixedes, Martin Johnsson, Ivan Pocrnic

InterBull & EAAP

Lyon, 2023-03-24

Where have we been & where are we going?

- Started career with pedigree-based mixed models
- Rode on the excitement of introducing genomic selection
  - 2010s ~10^3 individuals
  - 2015s ~10^5 individuals
  - 2020s ~10^6 individuals
  - 2030s ~10^9 individuals???

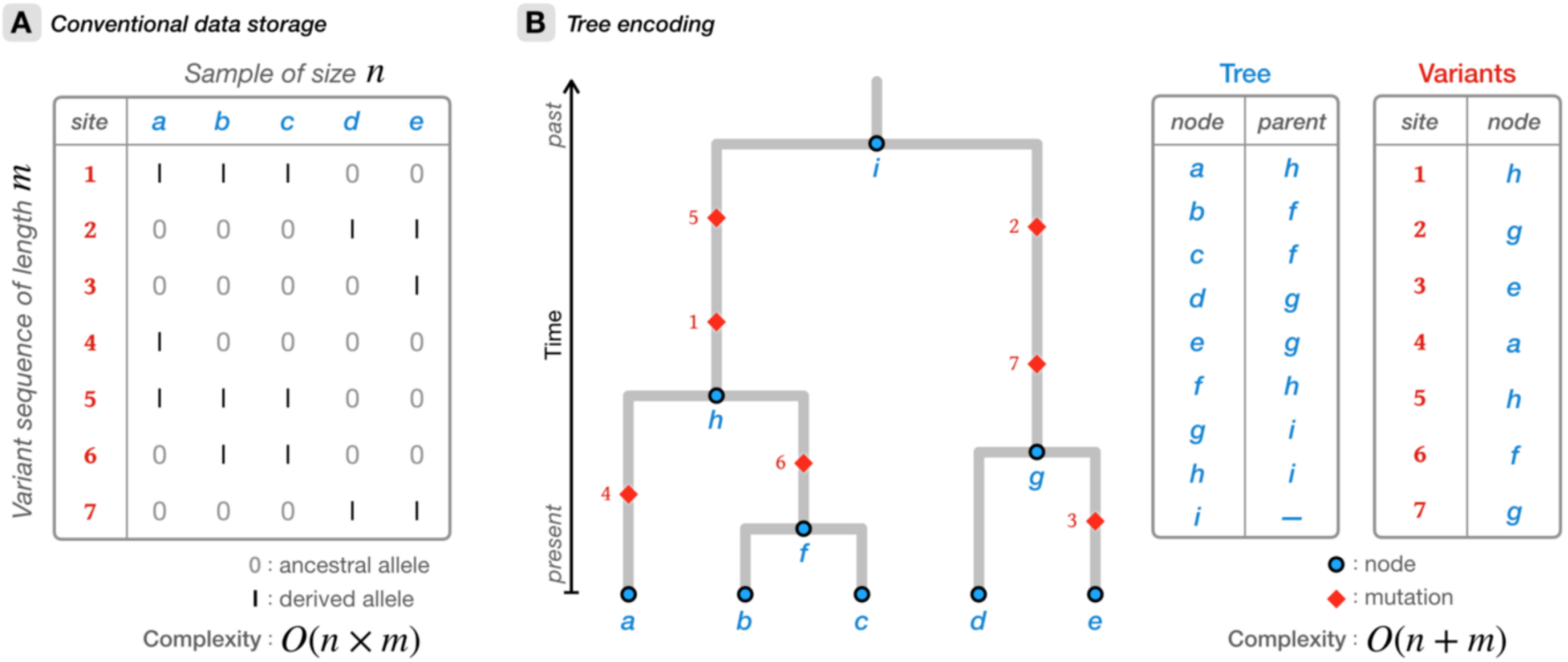
SNP array genotypes

- Contributed to the whole-genome sequencing "craze"
  - 2015s ~10^3 individuals
  - 2020s ~$10^{4\text{-}6}$ individuals
  - 2030s ~$10^{6\text{-}9}$ individuals???

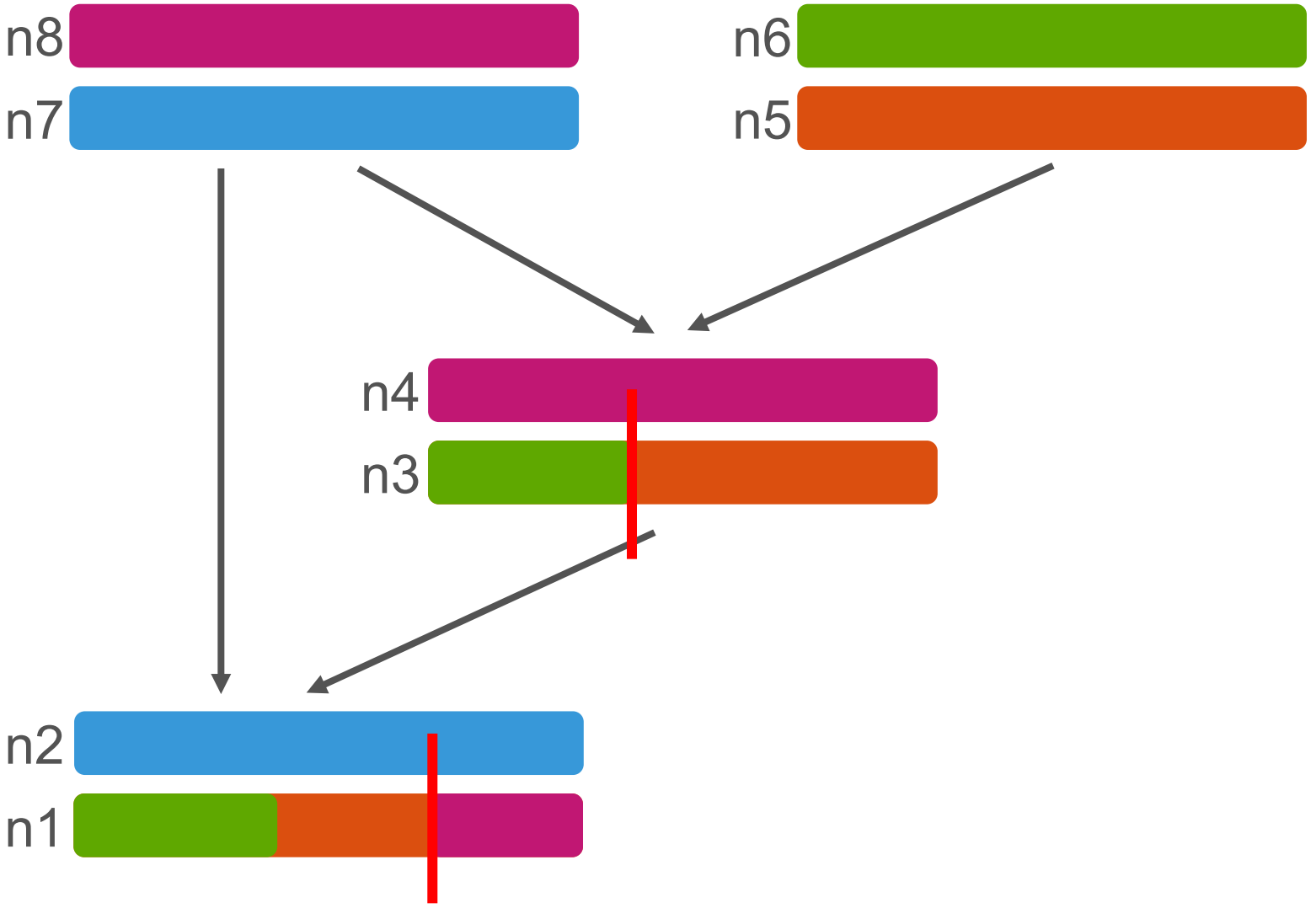
Whole-genome sequences
(with & without imputation)

Handling MEGA-SCALE through data generation process

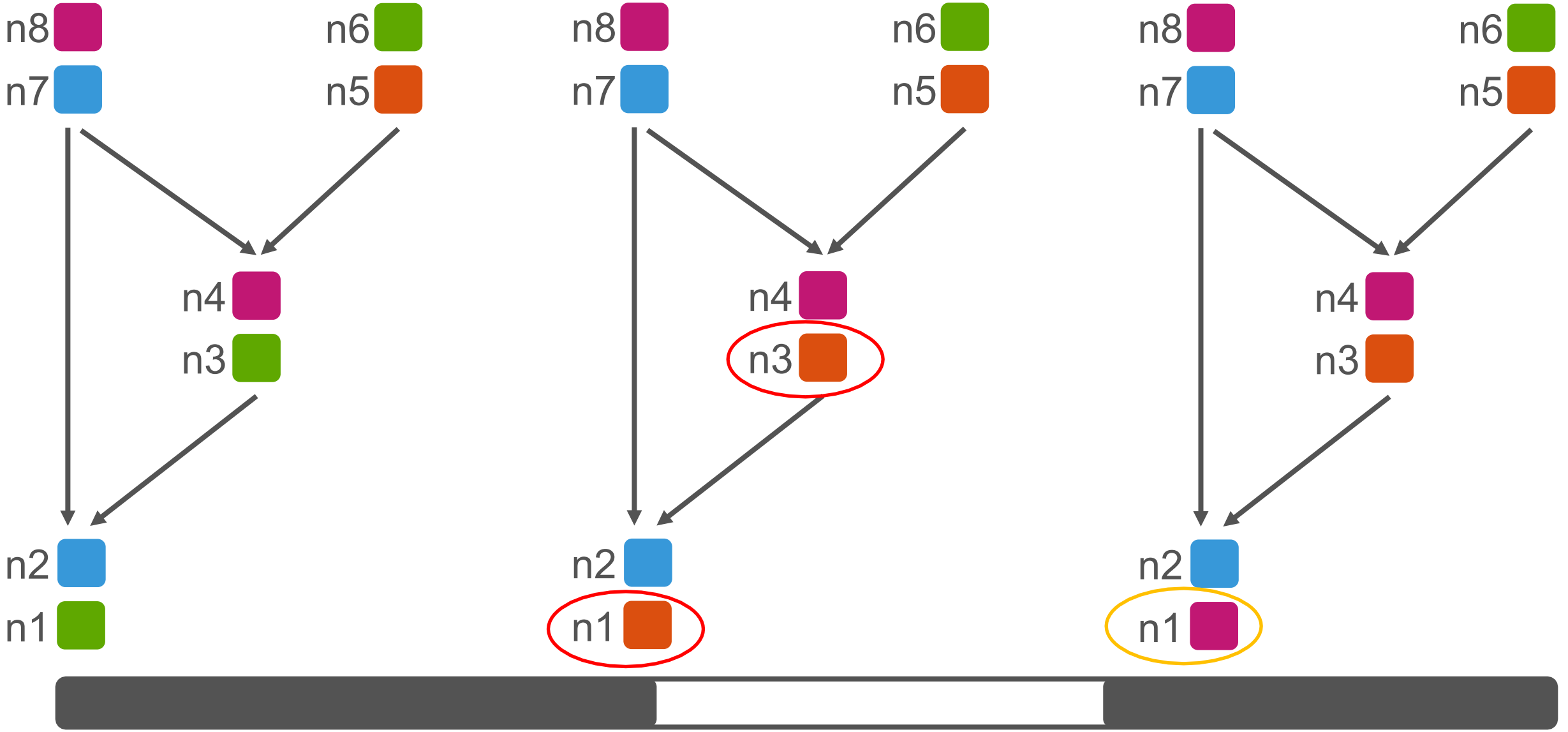
Kelleher et al. (2019, Nature Genetics; …)



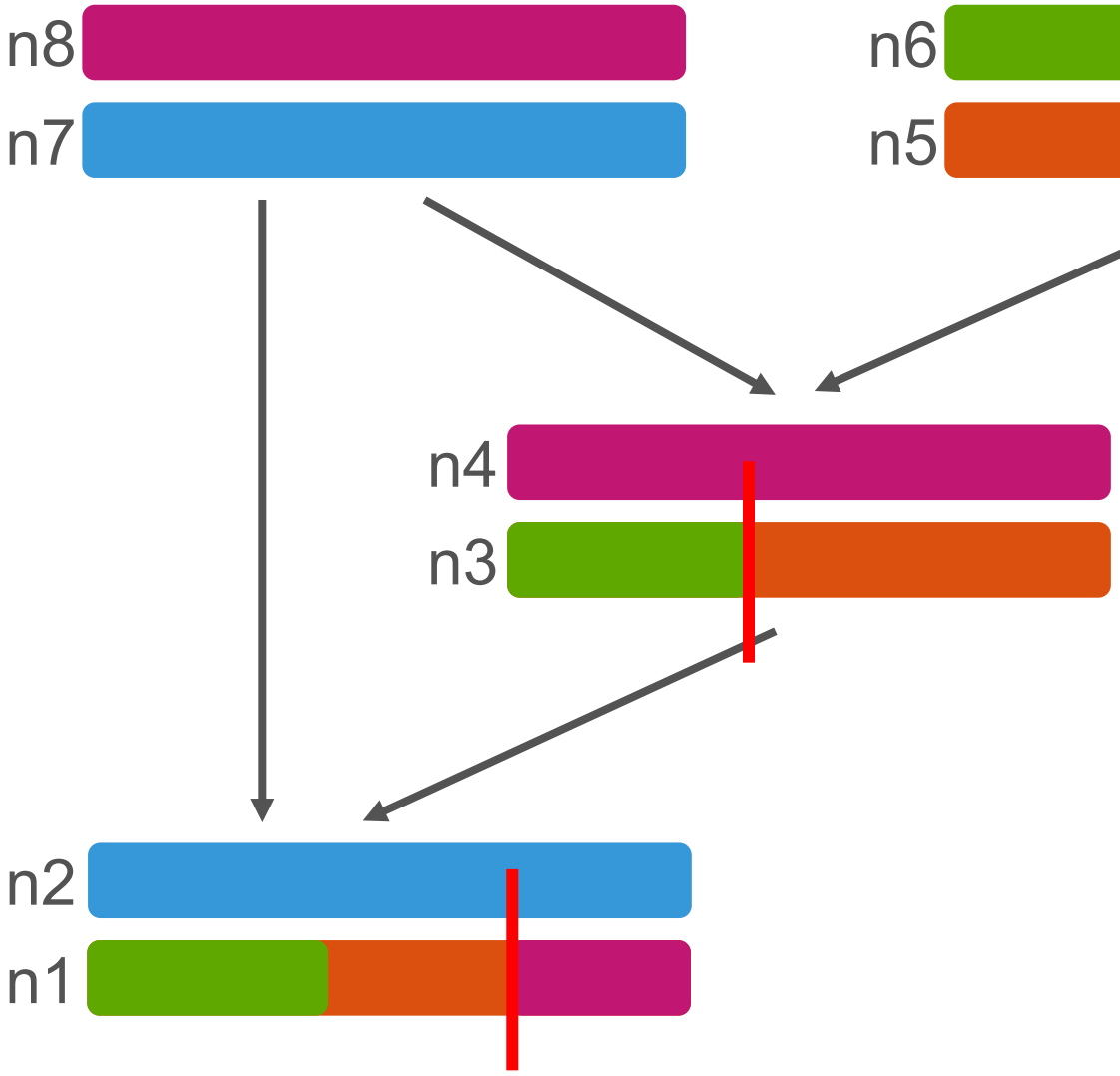
Tracking chromosome segments



Ancestral recombination graph & Local trees

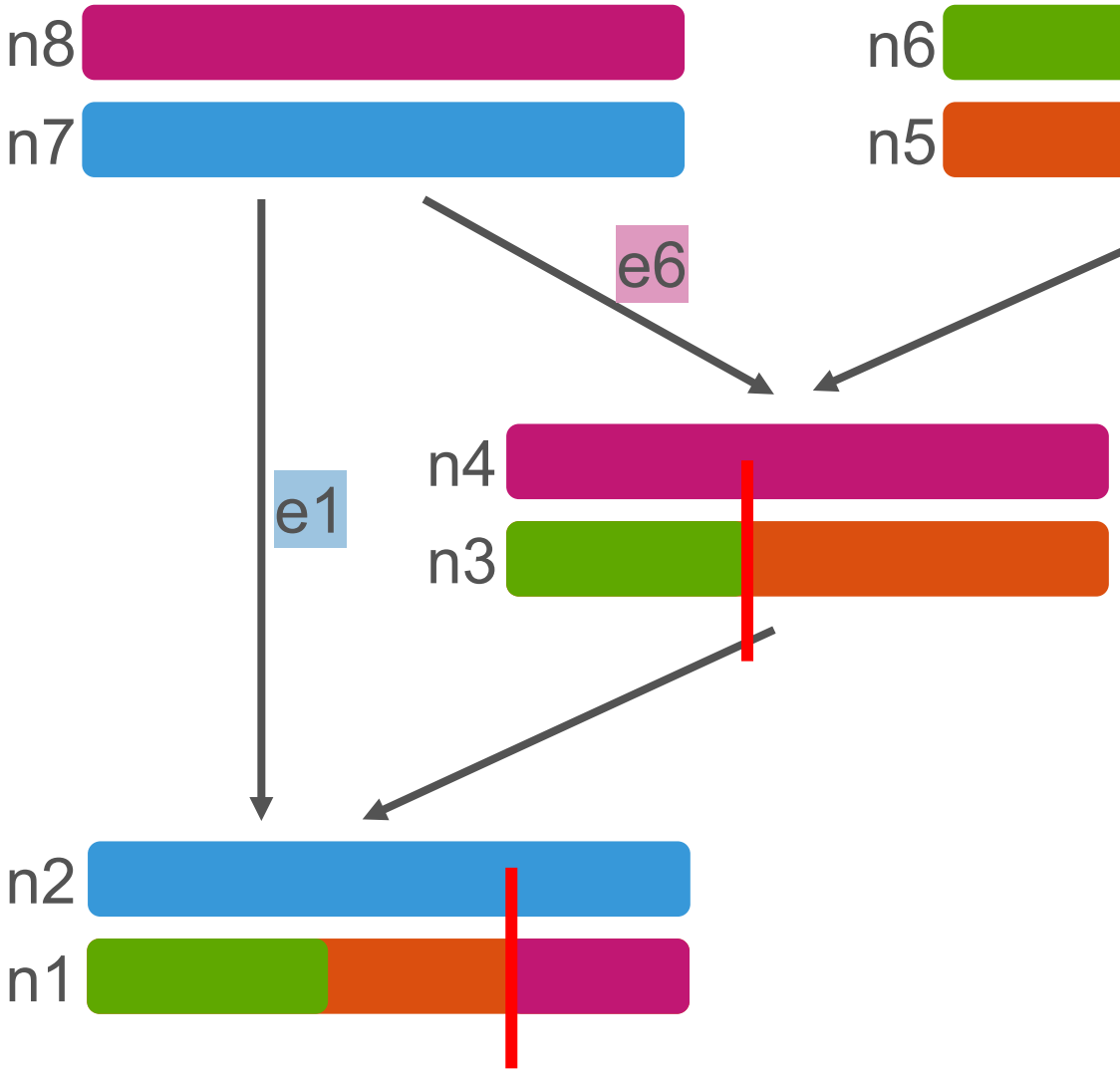


Tree sequence – Nodes



Node	Time
n1	0
n2	0
n3	1
n4	1
n5	2
n6	2
n7	2
n8	2

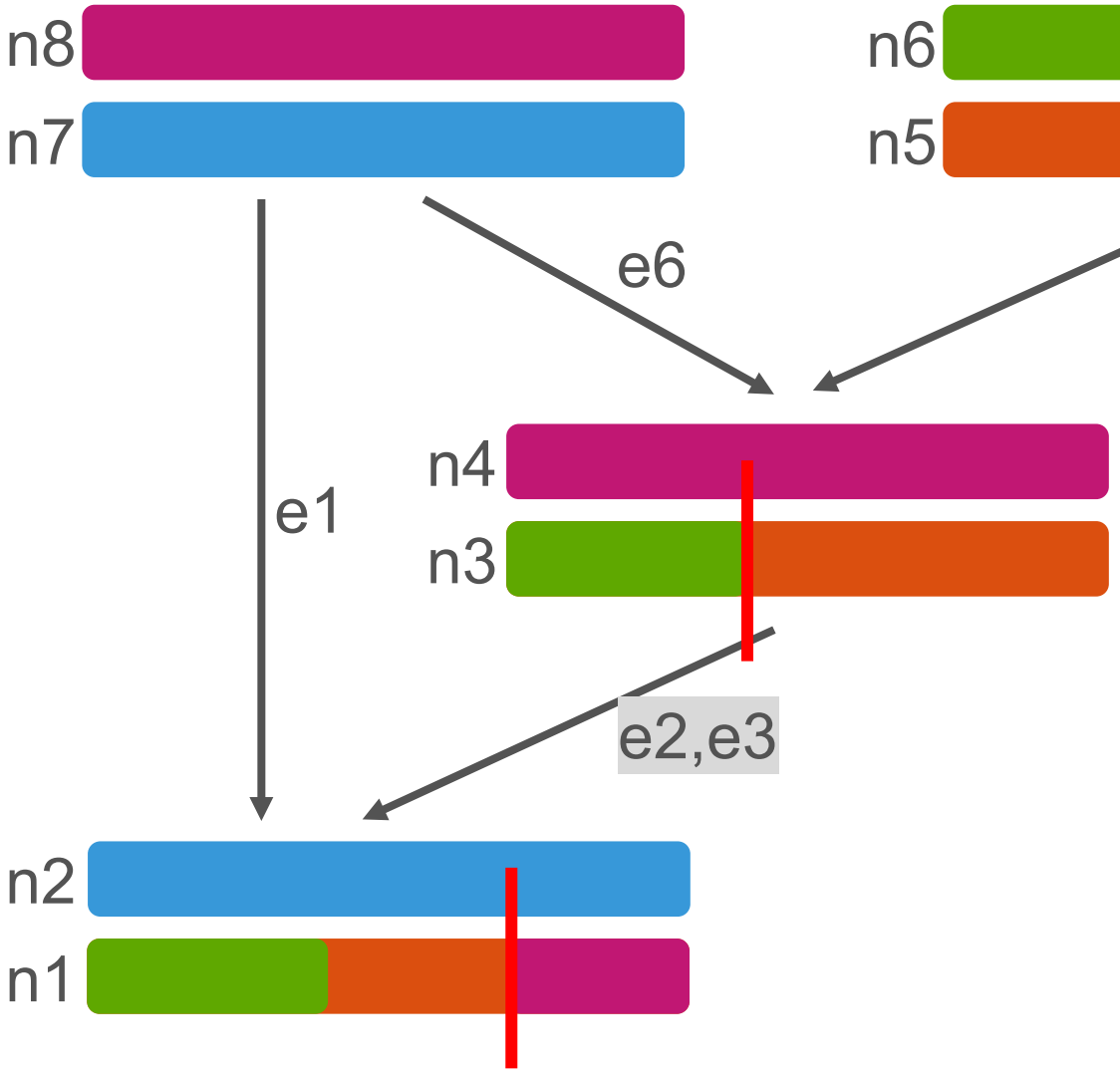
Tree sequence – Nodes & Edges (pedigree)



Node	Time
n1	0
n2	0
n3	1
n4	1
n5	2
n6	2
n7	2
n8	2

Edge	Desc	Anc	Start	Stop
e1	n2	n7	0	100
e6	n4	n8	0	100

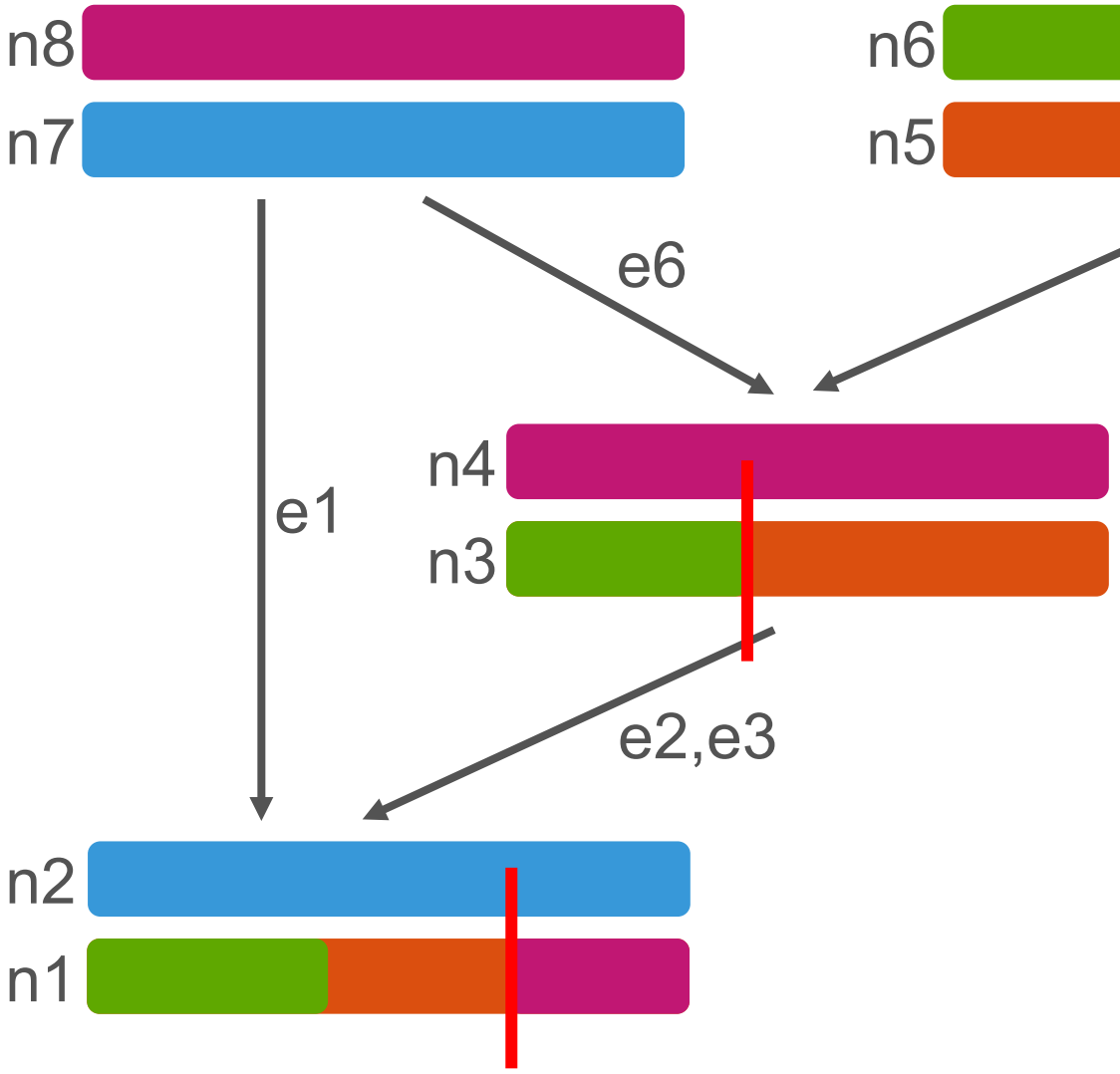
Tree sequence – Nodes & Edges (pedigree)



Node	Time
n1	0
n2	0
n3	1
n4	1
n5	2
n6	2
n7	2
n8	2

Edge	Desc	Anc	Start	Stop
e1	n2	n7	0	100
e2	n1	n3	0	70
e3	n1	n4	71	100
e6	n4	n8	0	100

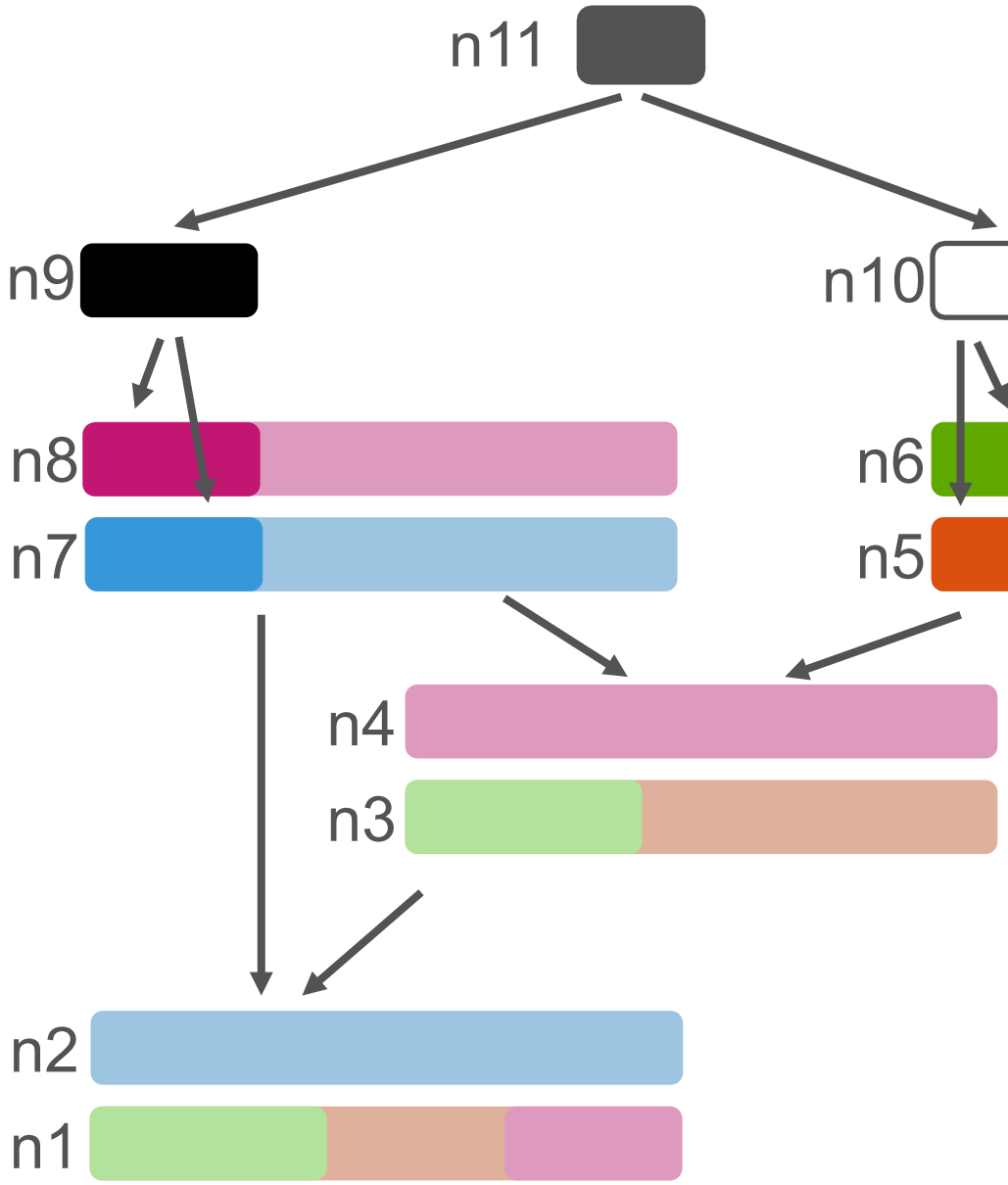
Tree sequence – Nodes & Edges (pedigree)



Node	Time
n1	0
n2	0
n3	1
n4	1
n5	2
n6	2
n7	2
n8	2

Edge	Desc	Anc	Start	Stop
e1	n2	n7	0	100
e2	n1	n3	0	70
e3	n1	n4	71	100
e4	n3	n6	0	40
e5	n3	n5	41	100
e6	n4	n8	0	100

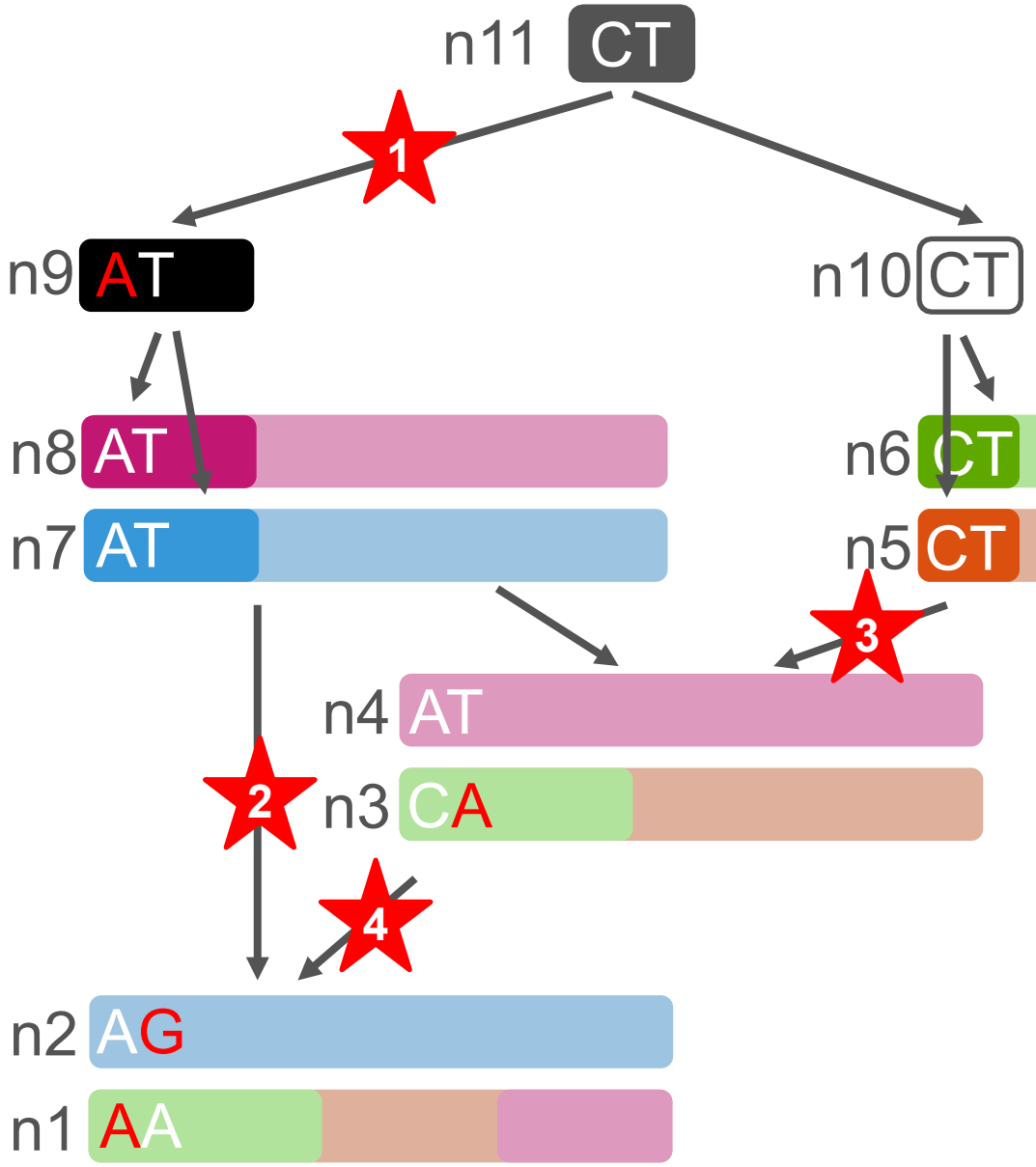
Tree sequence – Nodes & Edges (coalescent)



Node	Time
…	…
n9	100
n10	150
n11	500
…	…

Edge	Desc	Anc	Start	Stop
..	…	…	…	…
e7	n7	n9	0	27
e8	n8	n9	0	27
e9	n5	n10	0	18
e10	n6	n10	0	18
e11	n9	n11	0	27
e12	n10	n11	0	18
…	…	…	…	…

Tree sequence – Nodes, Edges, Sites, & Mutations



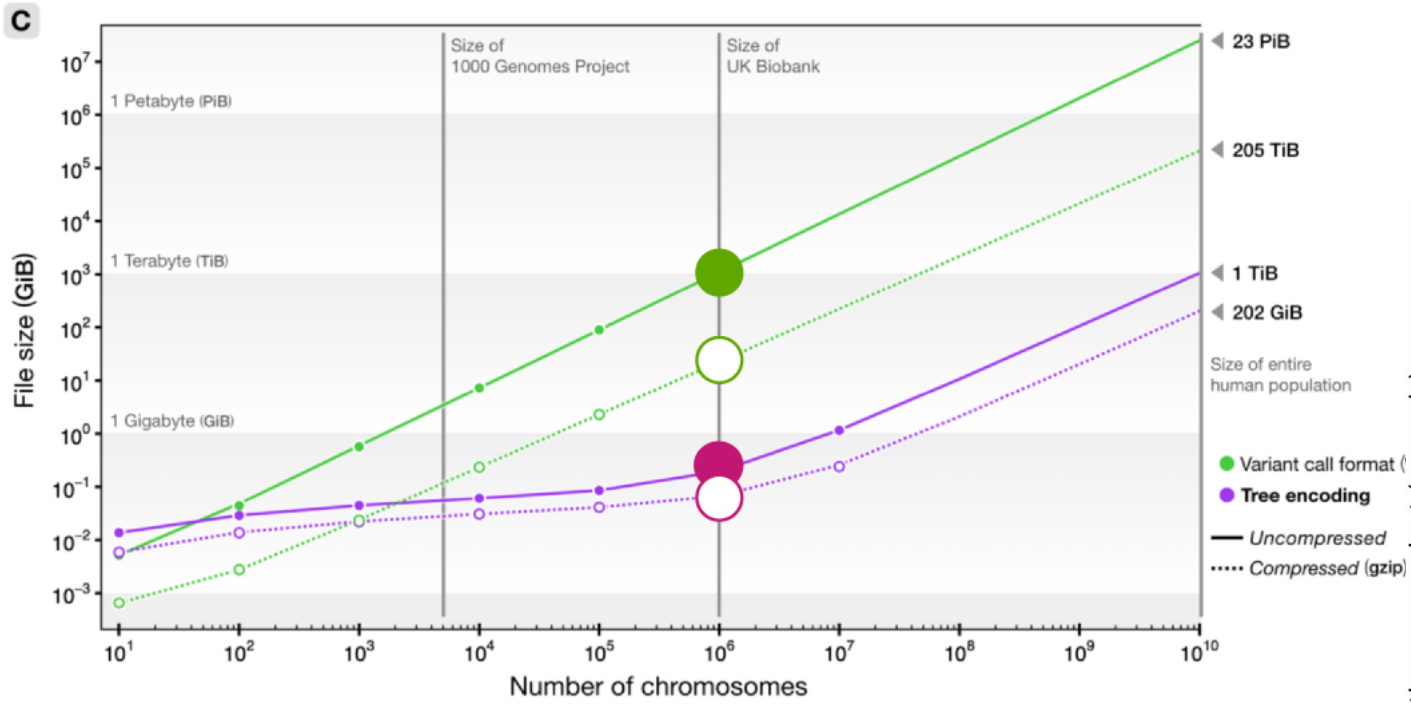
Site	Position	Ancestral
1	5	C
2	17	T

Mutation	Site	Node	Derived
1	1	n9	C→A
2	2	n2	T→G
3	2	n3	T→A
4	1	n1	C→A

Computational power of tree sequences

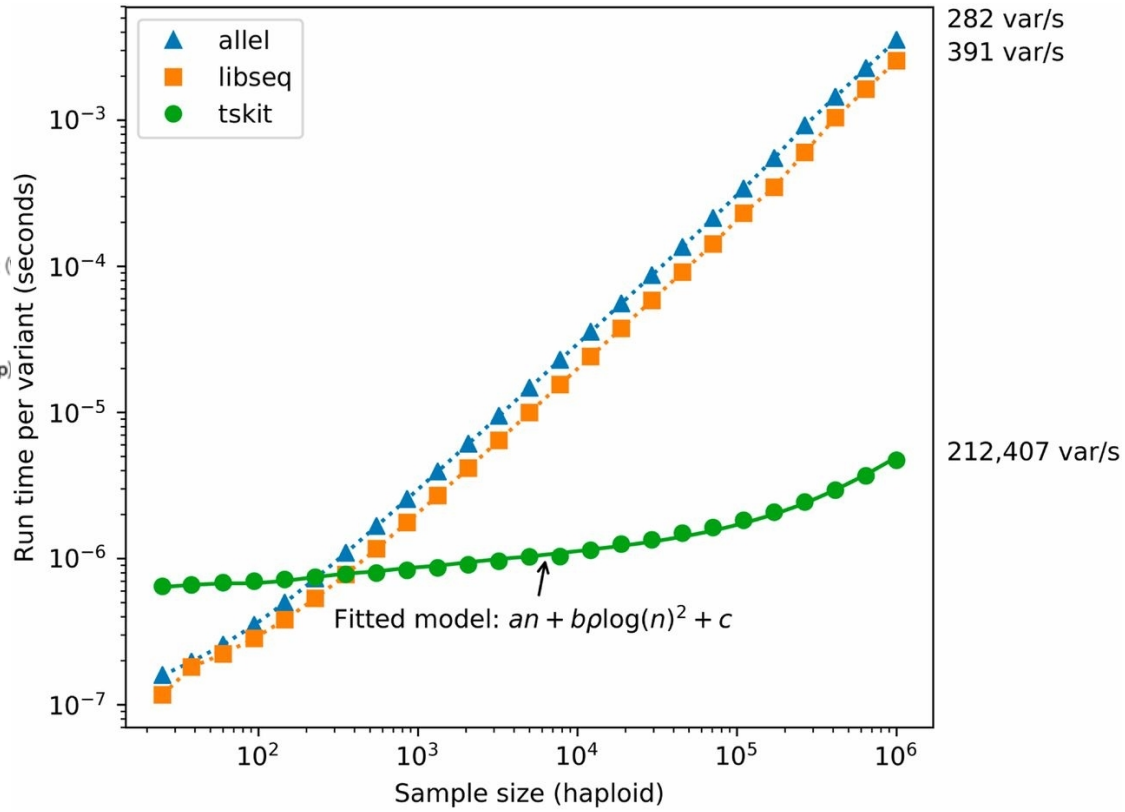
STORAGE

Kelleher et al. (2019, Nature Genetics; …)



COMPUTE

Ralph et al. (2020, Genetics)

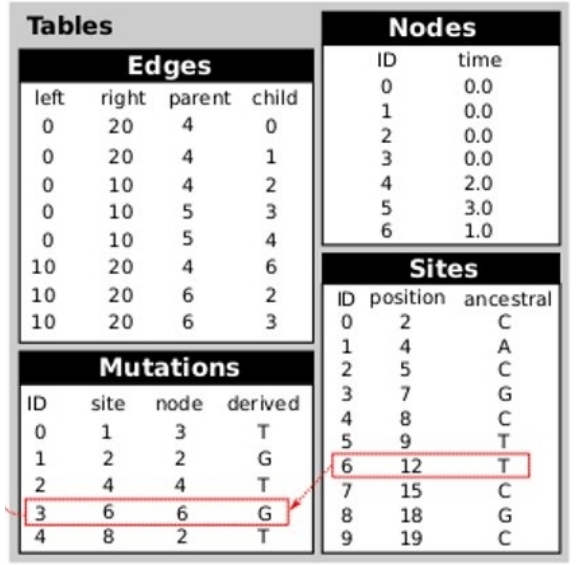


Roslin-Genus/PIC ~1 million pig genomes project

- Pedigree & SNP array data
  - 9 lines with a total of ~450K pigs
  - ~15K-50K markers
- Whole-genome sequence
  - ~8K pigs (a mix of ~1x and ~30x)
  - ~46M variants passed quality control across lines
- Accurate imputation of whole-genomes
  - ~450K diploid pigs * 2 = **~900,000 haploid genomes**
  - ~450K pigs * ~46M sites * 8 bytes / 2^40 = ~152 TiB of memory ☹
    (2-bit storage → ~5TiB of memory)

Roslin-Genus/PIC ~1 million pig genomes project

- Pedigree & SNP array data
  - 9 lines with a total of ~450K pigs
  - ~15K-50K markers
- Whole-genome sequence
  - ~8K pigs (a mix of ~1x and ~30x)
  - ~46M variants passed quality control across lines
- Accurate imputation of whole-genomes
  - ~450K diploid pigs * 2 = **~900,000 haploid genomes**
  - ~450K pigs * ~46M sites * 8 bytes / 2^40 = ~152 TiB of memory ☹
    (2-bit storage → ~5TiB of memory)



Tables

Edges

left	right	parent	child
0	20	4	0
0	20	4	1
0	10	4	2
0	10	5	3
0	10	5	4
10	20	4	6
10	20	6	2
10	20	6	3

Nodes

ID	time
0	0.0
1	0.0
2	0.0
3	0.0
4	2.0
5	3.0
6	1.0

Mutations

ID	site	node	derived
0	1	3	T
1	2	2	G
2	4	4	T
3	6	6	G
4	8	2	T

Sites

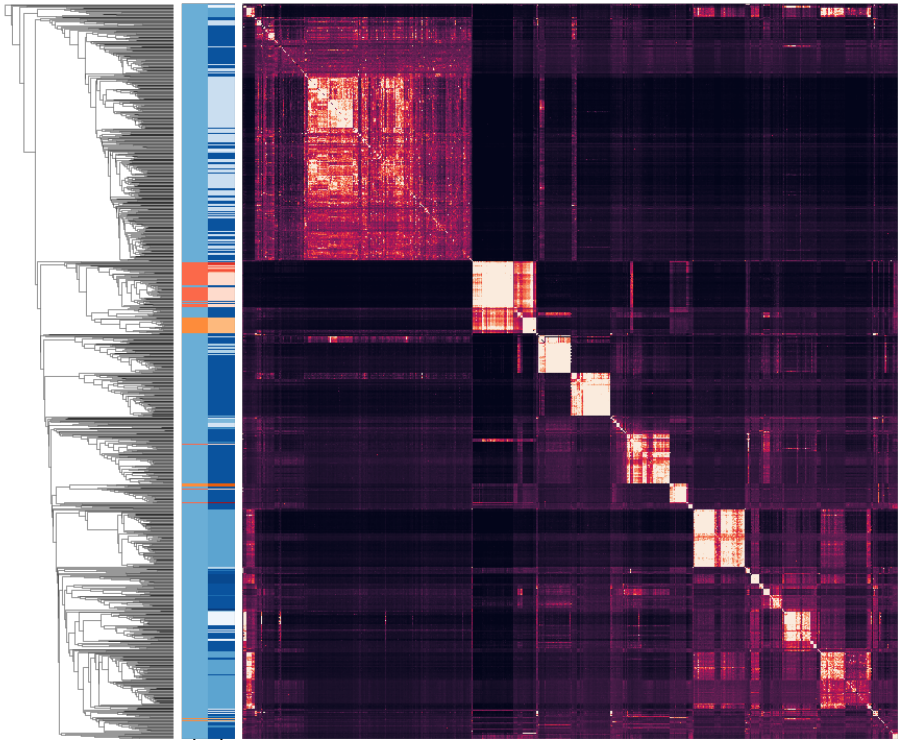
ID	position	ancestral
0	2	C
1	4	A
2	5	C
3	7	G
4	8	C
5	9	T
6	12	T
7	15	C
8	18	G
9	19	C

~15+ GiB vs. ~152 TiB

~99.9% “compression”!!!

Real application: 1000 Bull Genomes data

- 2,716 samples & 157 groups
  - Bos taurus
  - Bos indicus
  - Crossbred & African
  - Bos taurus coreanae (nat. Korean)
  - Bos primigenius (auroch)
  - Bos grunniens (yak)
- 29 autosomal chromosomes with ~116M variants



Session 83
Genetic diversity
Thursday
~ 16h 30m

- ShapeIt (phase), tsinfer (infer tree sequence), & tskit (analyse)
- VCFs: HUGE SIZE → Tree sequences: DECENT SIZE → ”Compression”: LARGE → Analysis: FAST

Conclusion

- MEGA-SCALE genomic datasets are here & growing
- Tree sequence data format to the rescue!?
  - PROS
    - Succinctly encodes the inheritance process
    - Combines pedigree, coalescent, phylogenetics (gene & species trees), segregation, recombination, gene conversion, mutations, IBS, IBD, …
    - Significant storage reduction & fast analyses
    - Novel insights & modelling
  - CONS (on-going work)
    - Novel way of thinking
    - Ancestral alleles, inference from real data, and inputs still HUGE
    - Need to develop specialised algorithms

Join us @ THE UNIVERSITY of EDINBURGH ROSLIN HLab

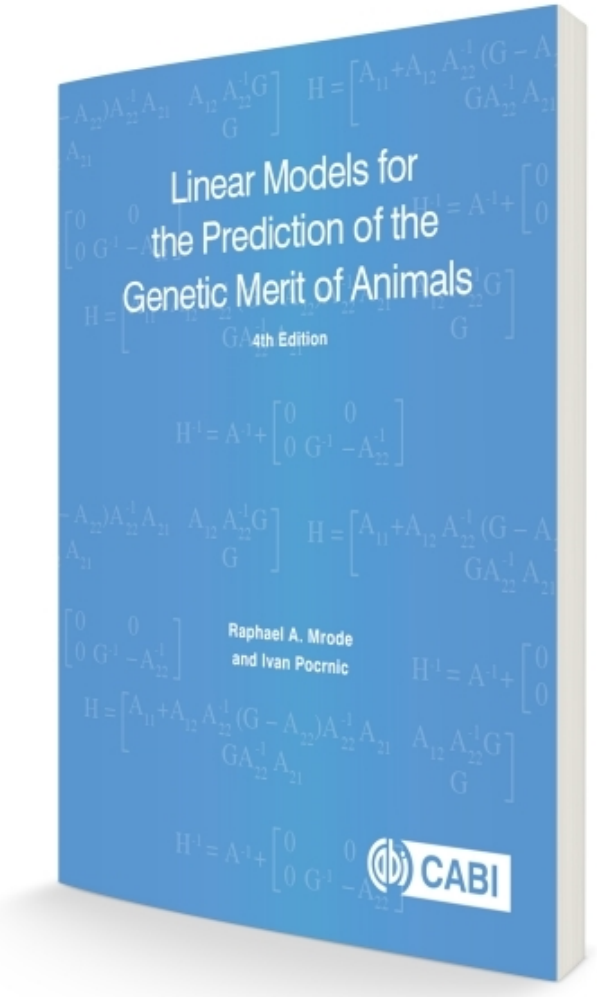
4-year PhD studentship available (start in autumn 2024)!!!

Breeding Programme Modelling with AlphaSimR

THE UNIVERSITY of EDINBURGH

THE DATA LAB
value from data

DDI Data-Driven Innovation
Part of the Edinburgh & South East Scotland City Region Deal



Storing and analysing a million genomes on a desktop computer

Gregor Gorjanc, Jana Obsteter, Gabriela Mafra Fortuna, Roger Ros-Freixedes, Martin Johnsson, Ivan Pocrnic

InterBull & EAAP

Lyon, 2023-03-24