Haplotype-based somatic mutation calling in heterogeneous cancer samples

Daniel Cooke

University of Oxford

dcooke@well.ox.ac.uk

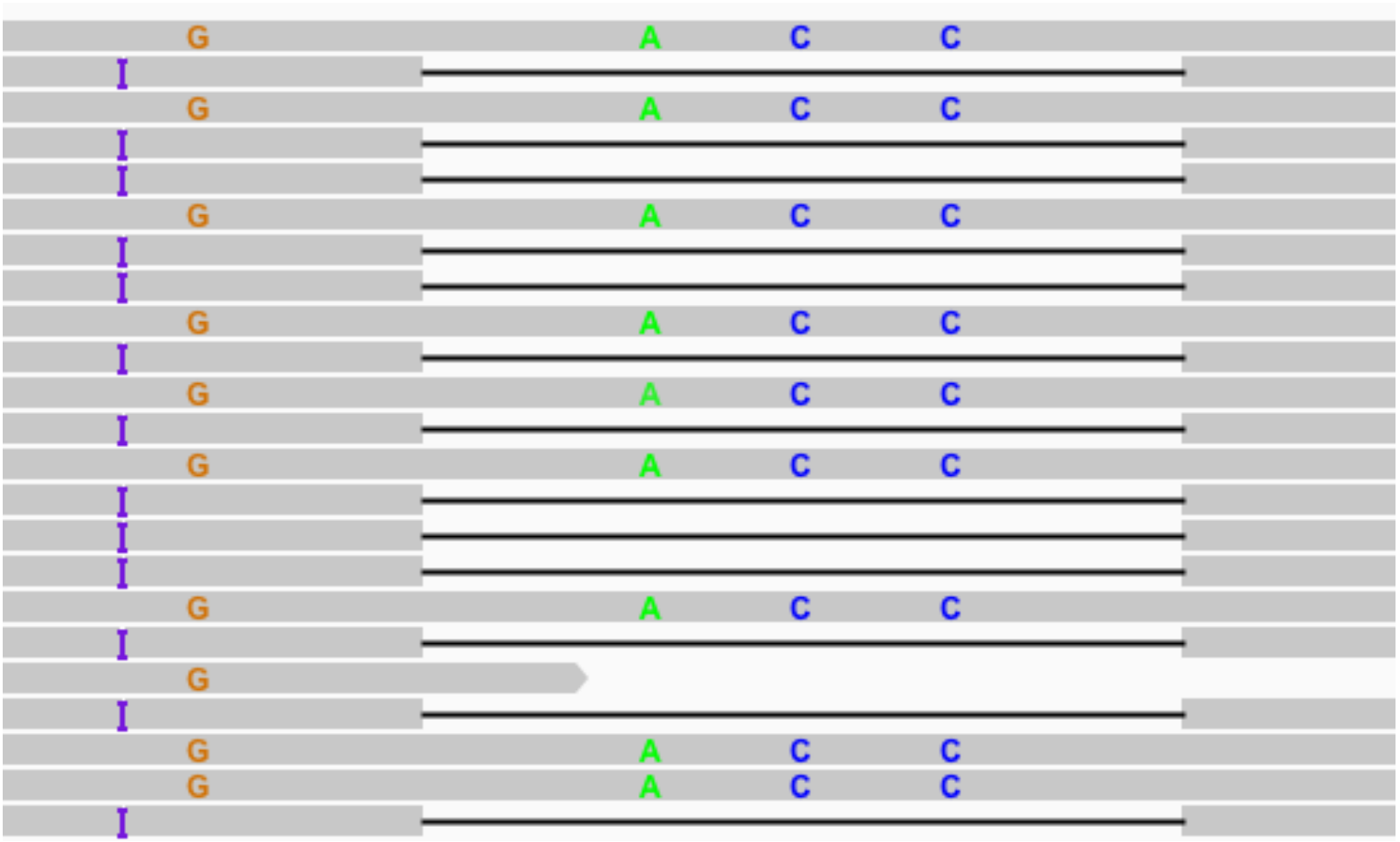
October 26, 2015

Haplotype-based variant calling

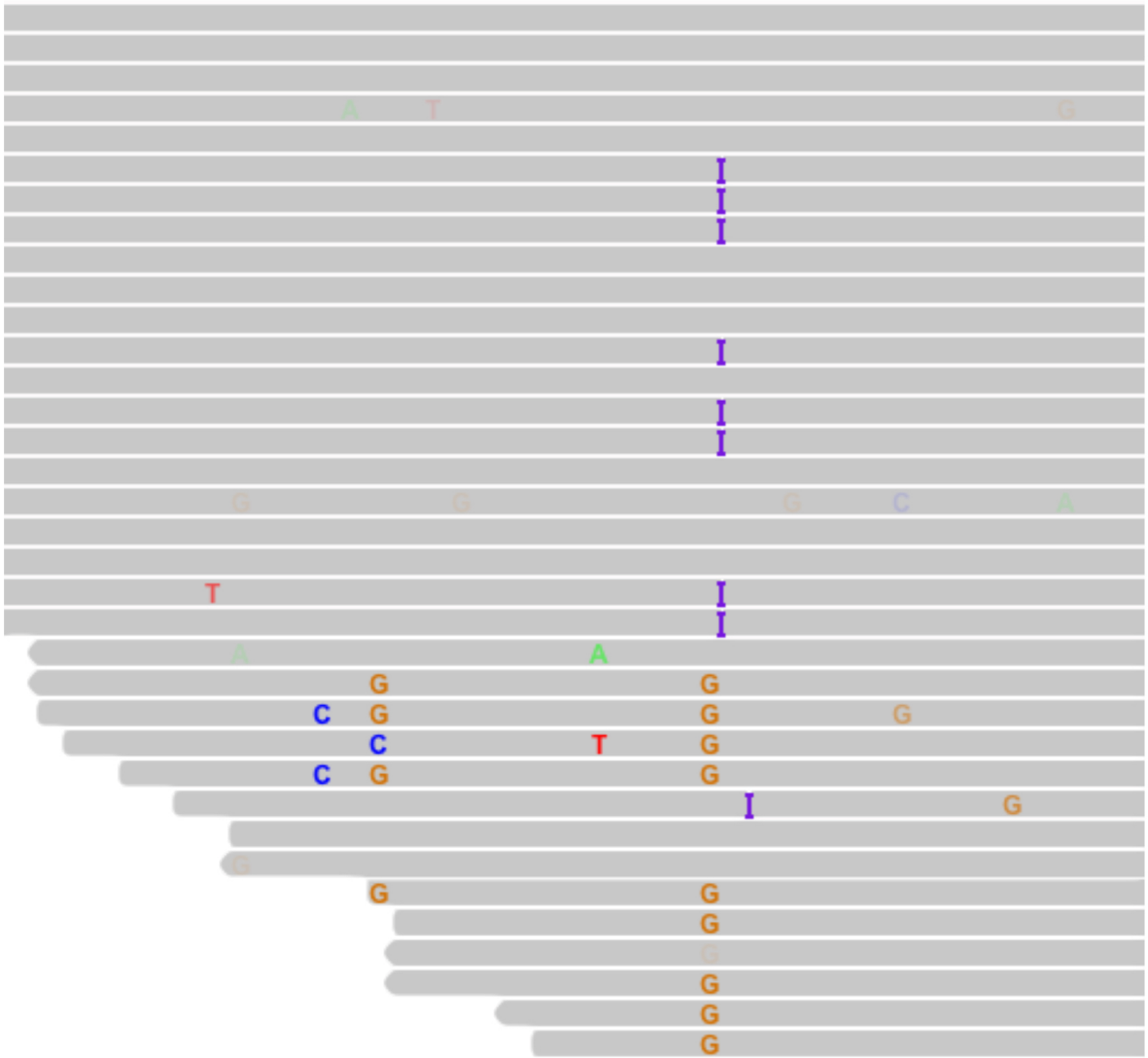
gatk

freebayes

Haplotype methods can resolve alignment errors



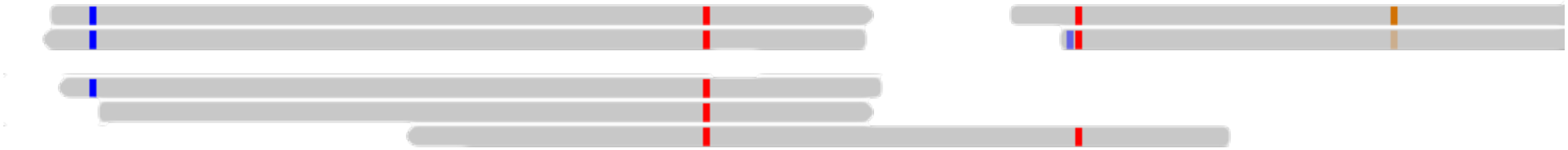
Haplotype methods can resolve alignment errors



Haplotype methods give local phasing

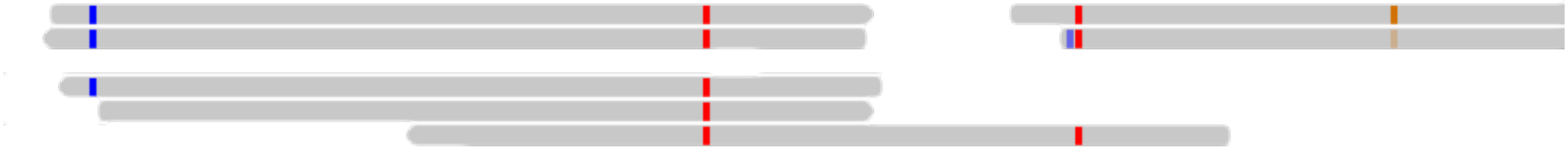
Haplotype methods give local phasing

First haplotype



Haplotype methods give local phasing

First haplotype



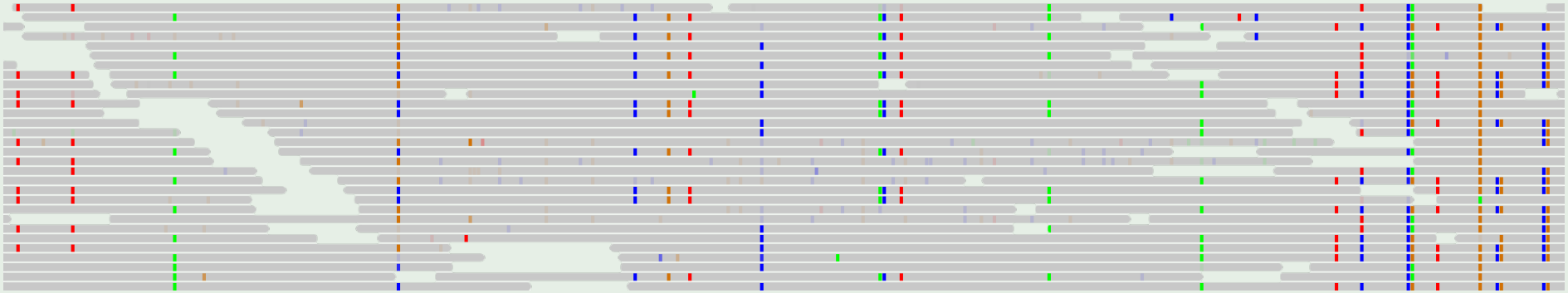
Second haplotype

Phasing is often intractable

$$\#\text{haplotypes} \approx 2^{\#\text{alleles}}$$

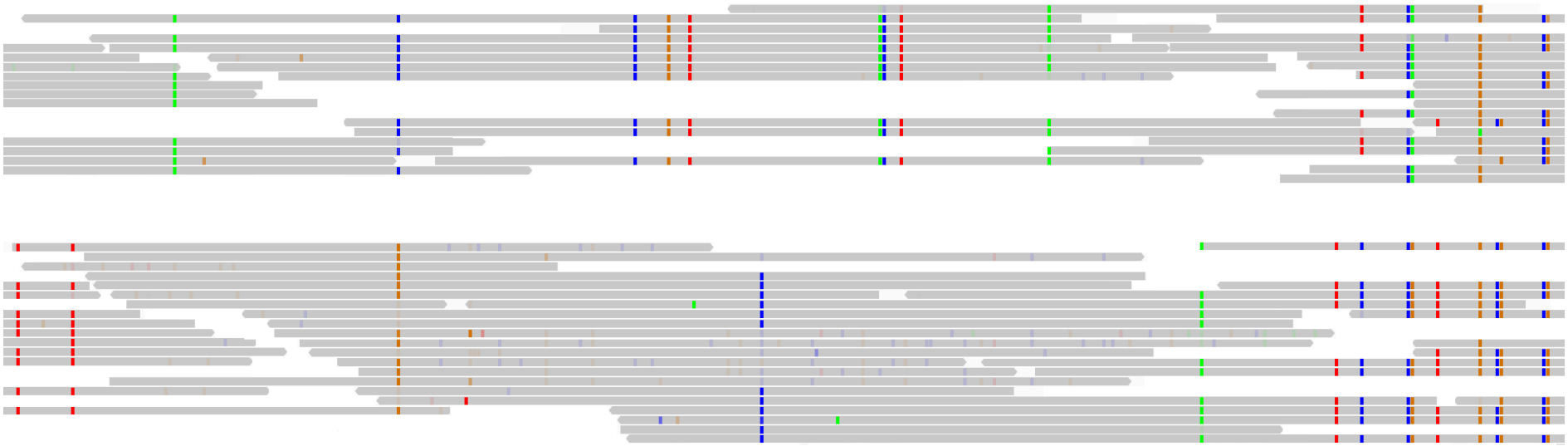
$$\#\text{genotypes} = \binom{\#\text{haplotypes} + \text{ploidy} - 1}{\#\text{haplotypes}}$$

Example: HLA loci

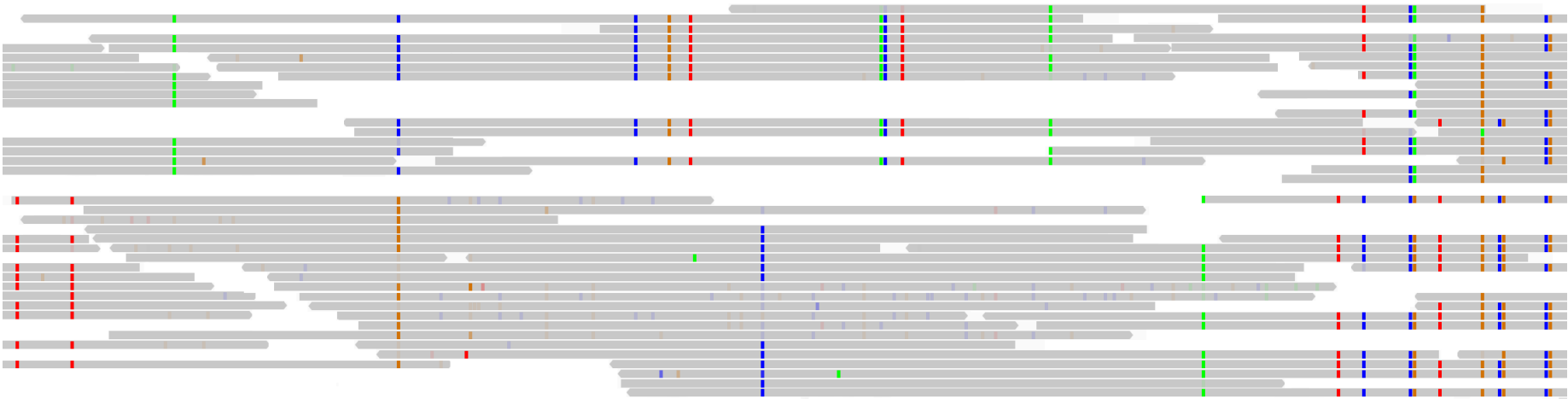


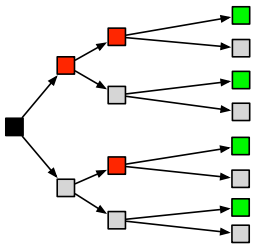
$$\#\text{alleles} \approx 70 \rightarrow \#\text{haplotypes} \approx 2^{70} \rightarrow \#\text{genotypes} \approx \text{lots}$$

Haplotype tree phasing

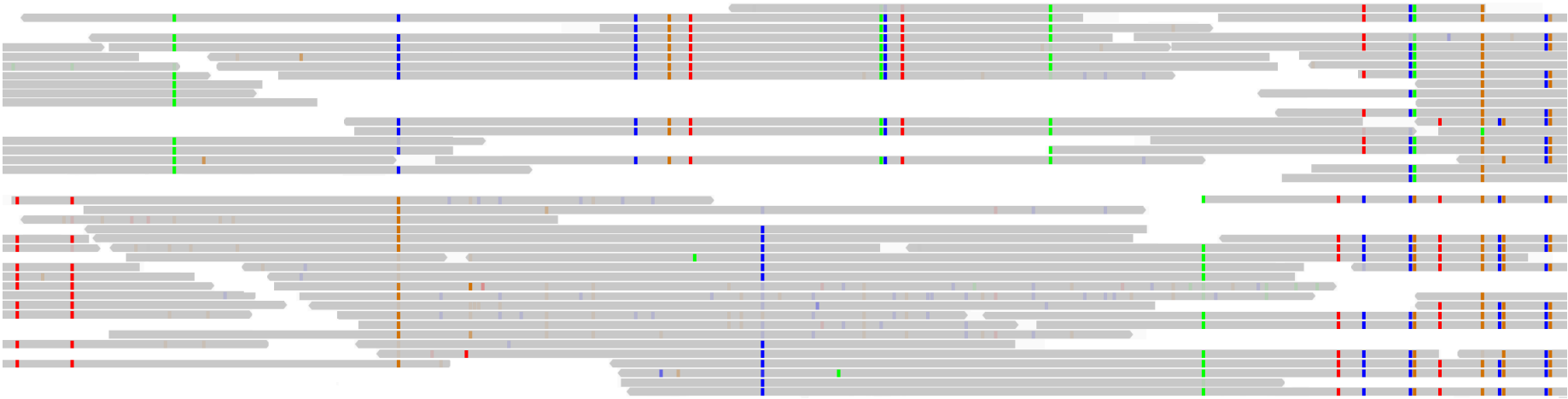


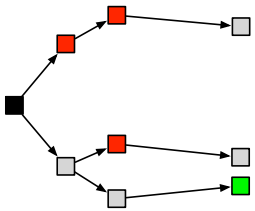
Haplotype tree phasing



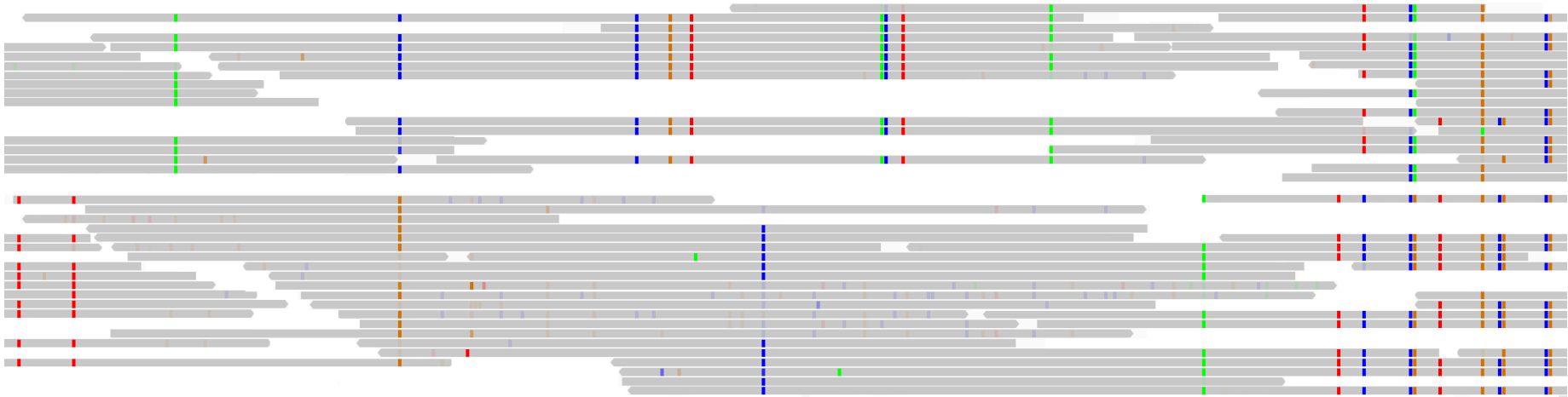


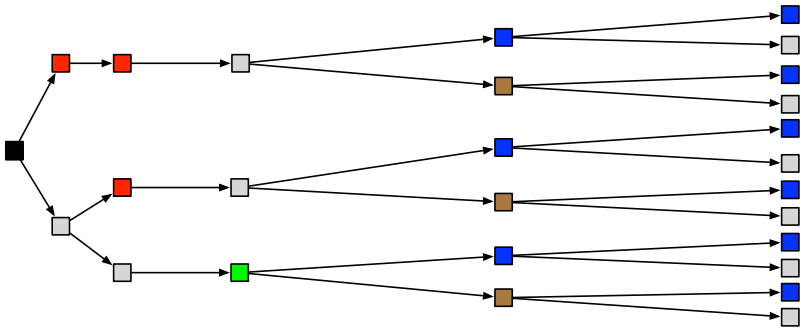
Haplotype tree phasing



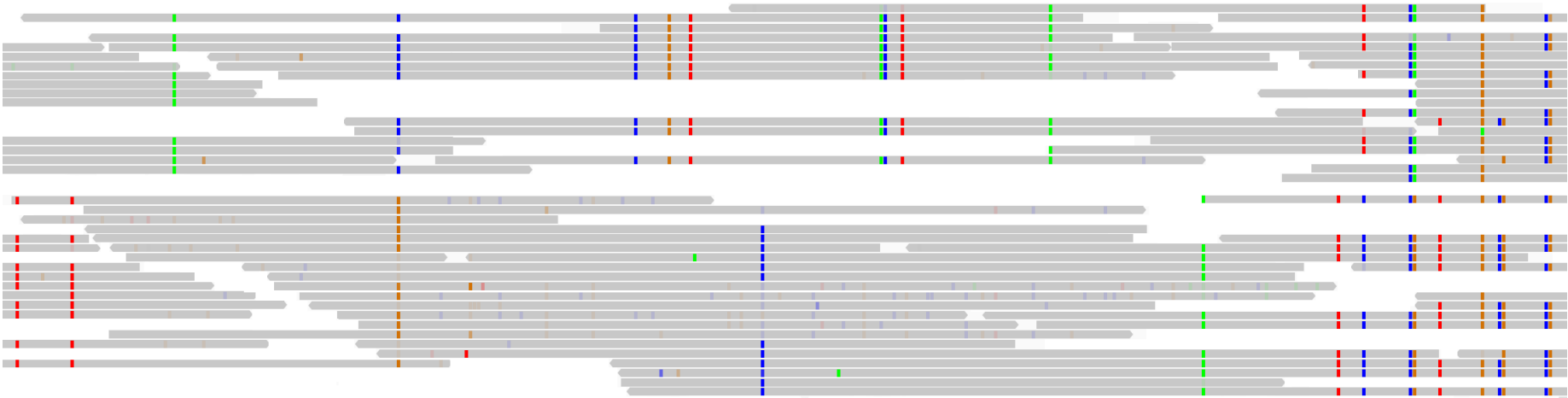


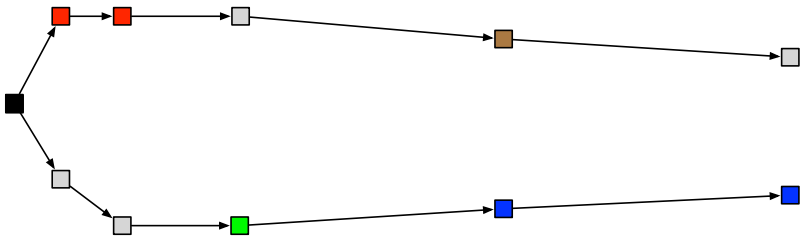
Haplotype tree phasing



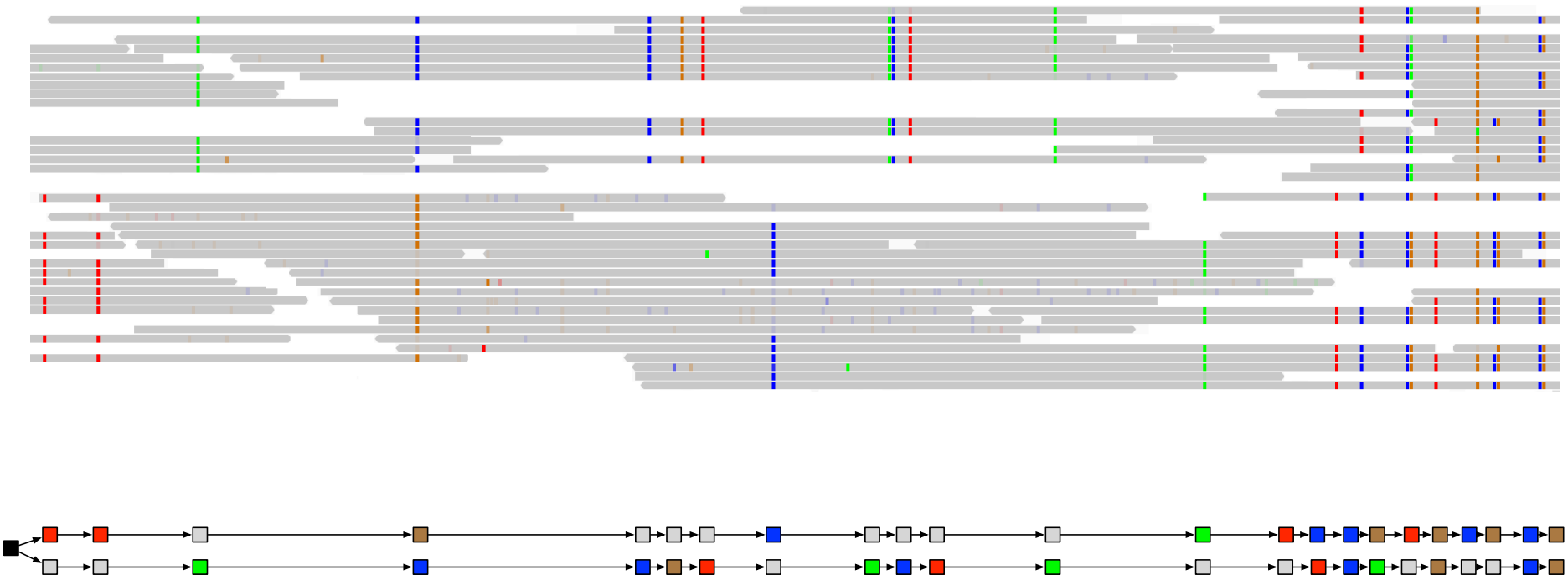


Haplotype tree phasing

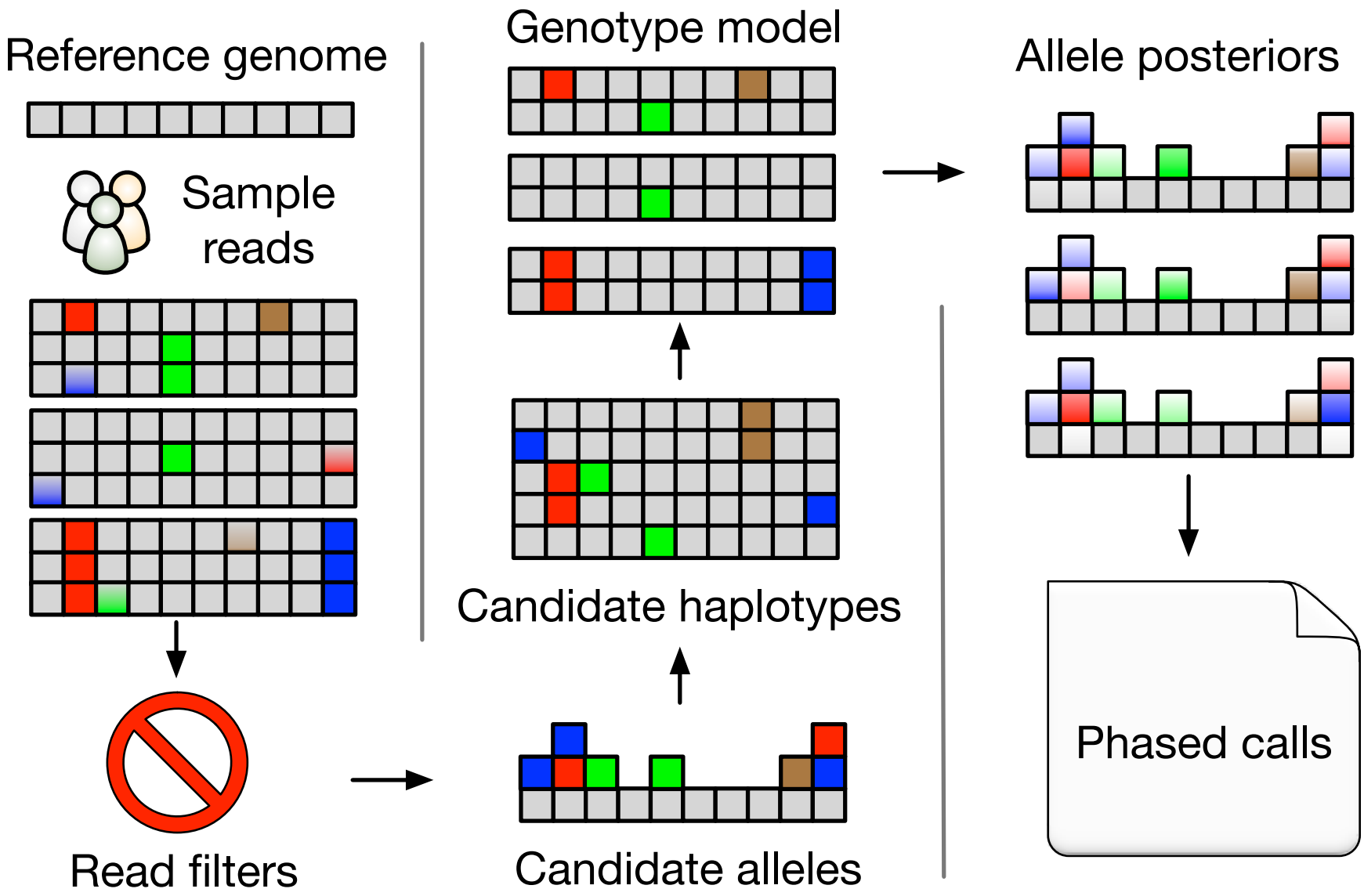




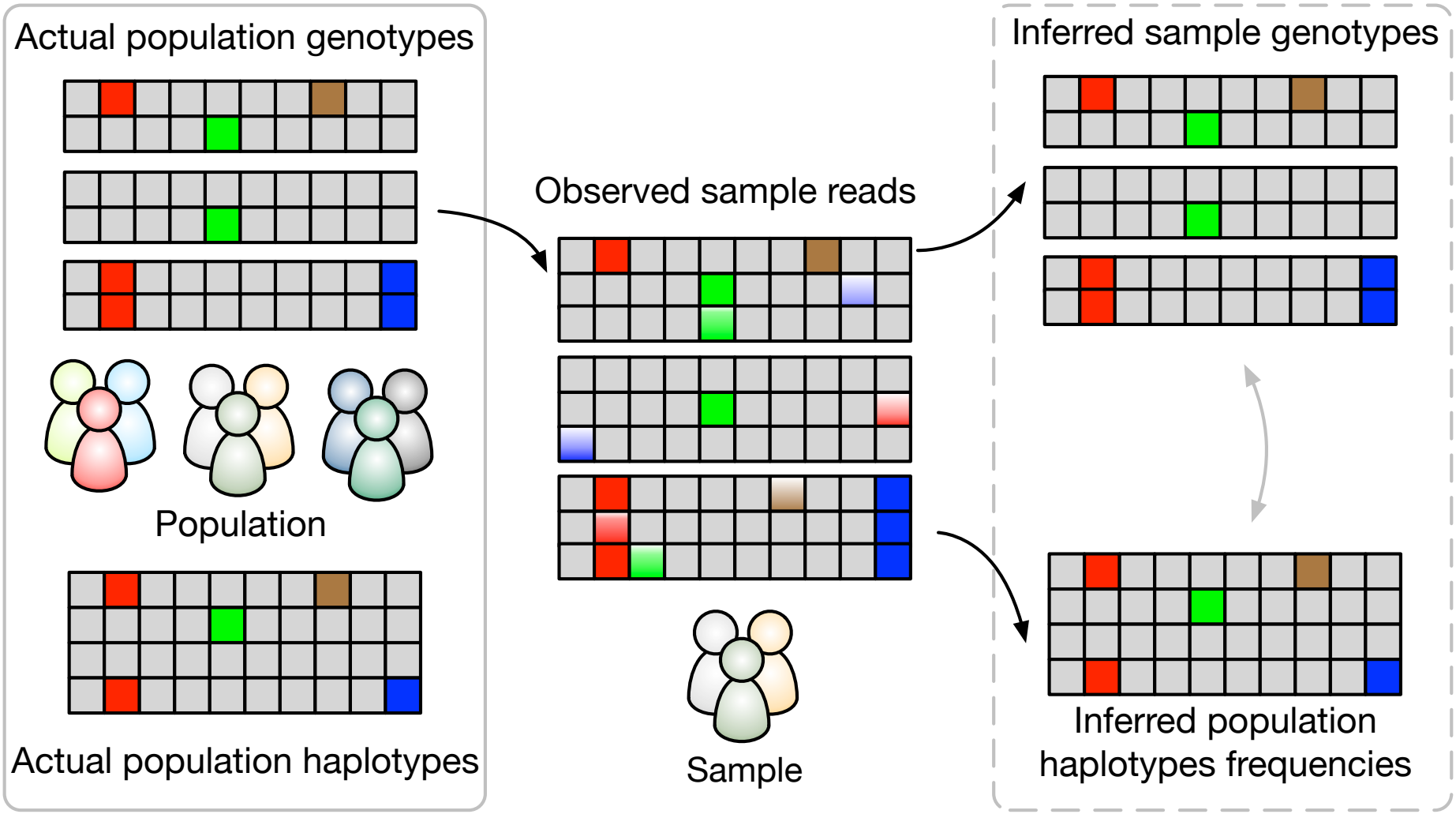
Haplotype tree phasing



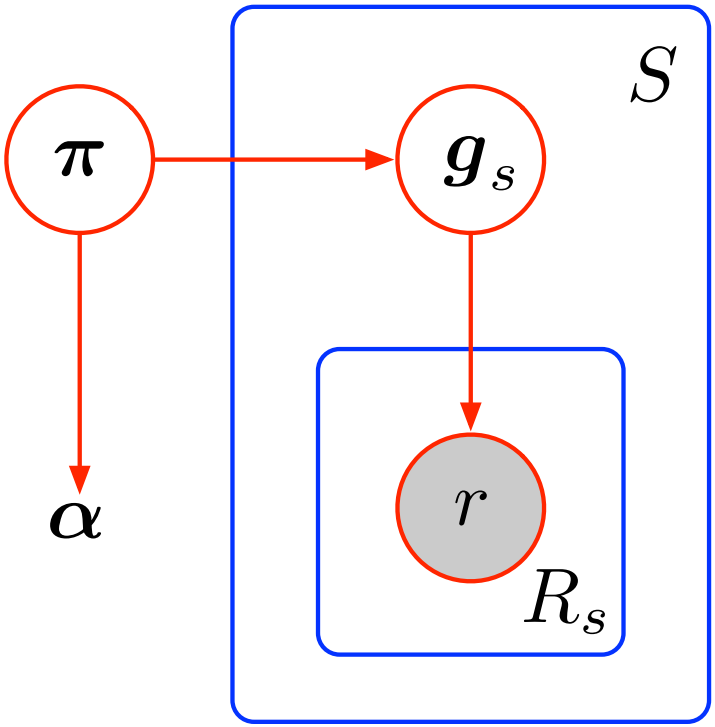
Algorithm overview



Population genotype model: overview



Population genotype model: maths

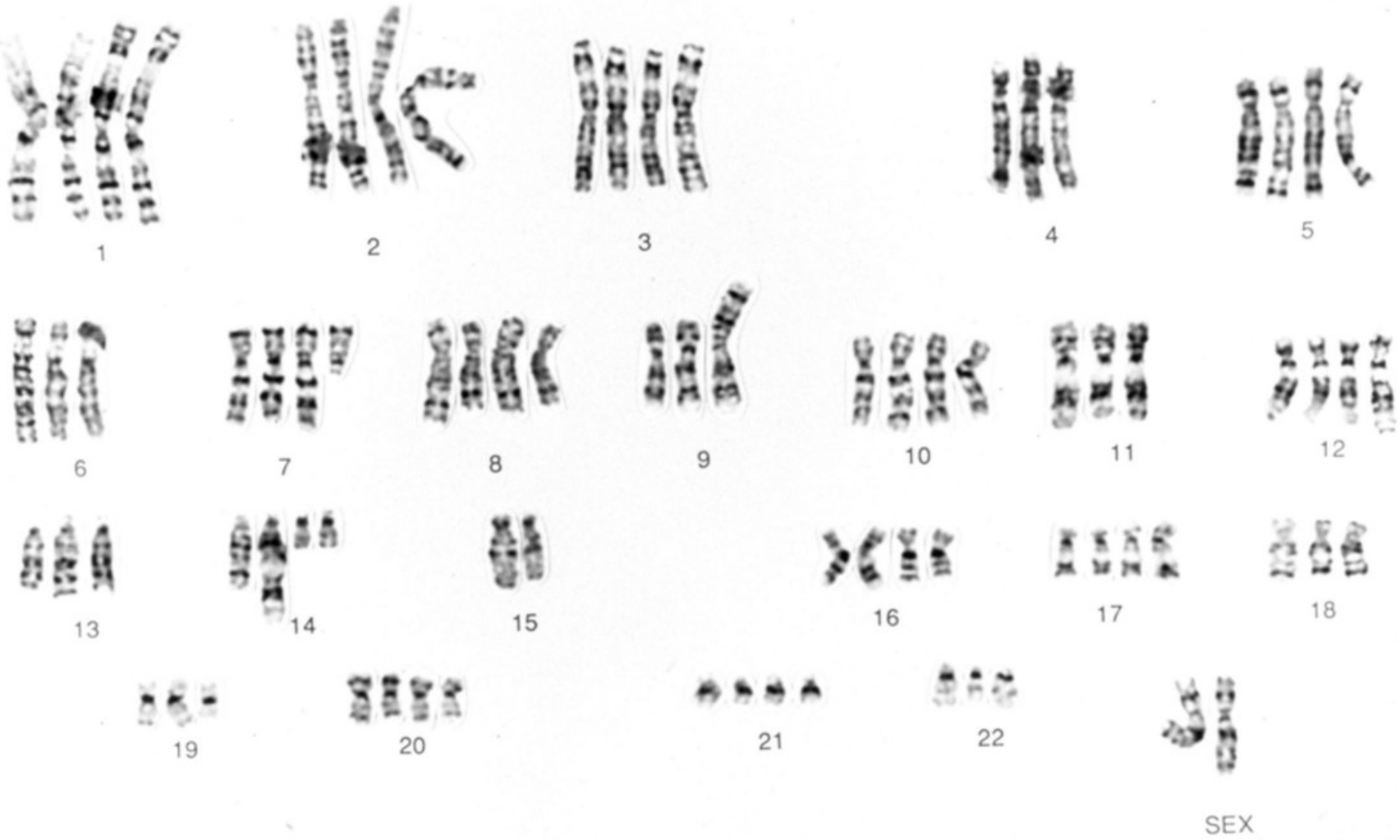


- Unknown population haplotype frequencies π
- Unknown sample genotypes $\boldsymbol{g}_s$
- Known sample ploidy

Marginal distribution: diploid case

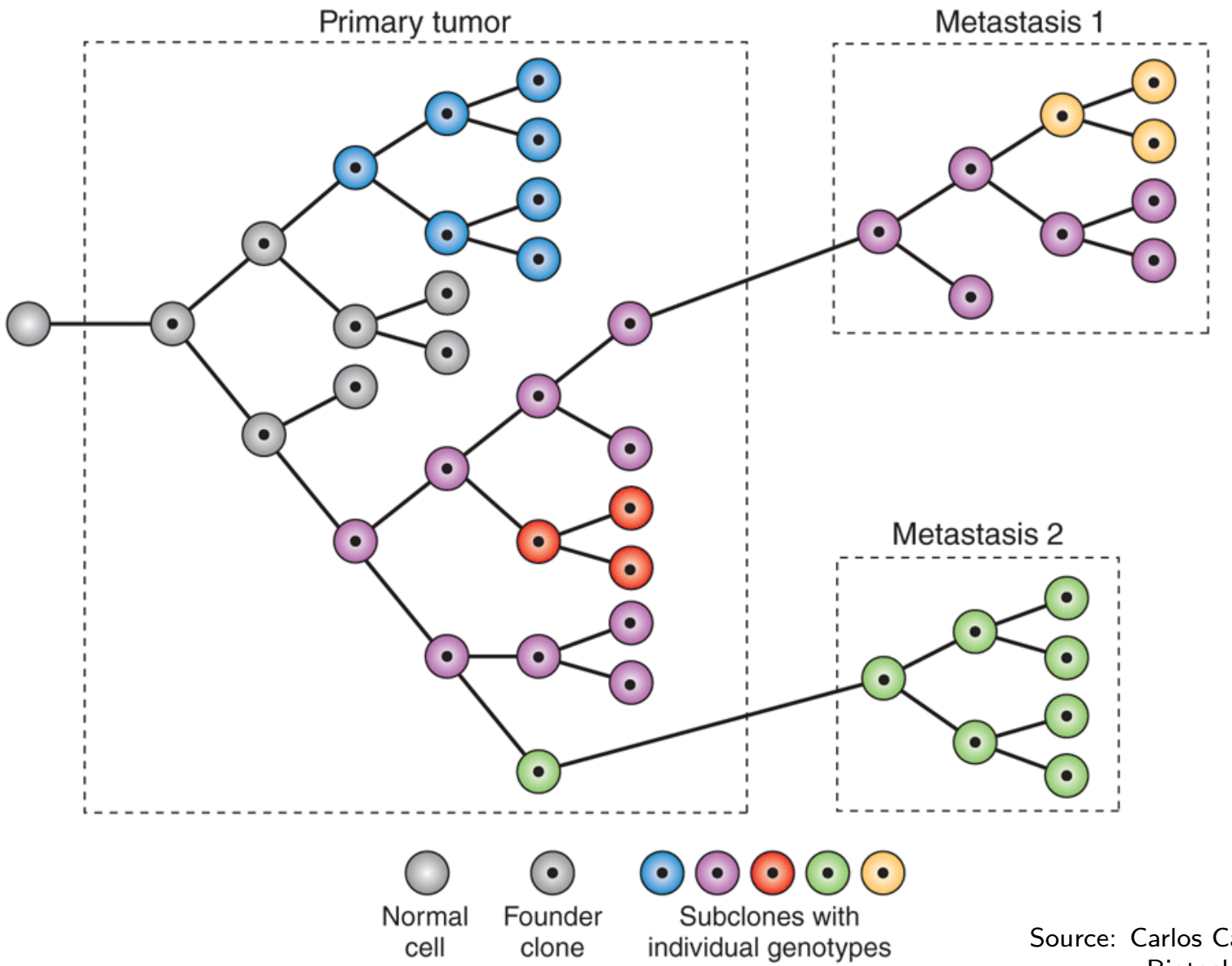
$$p(\boldsymbol{R}, \pi) = p(\pi|\alpha) \prod_{s=1}^{S} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\pi) \prod_{r \in R_s} \left\{ \frac{1}{2} p(r|\boldsymbol{g}_1) + \frac{1}{2} p(r|\boldsymbol{g}_2) \right\}$$

Challenges of cancer calling: messy karyotypes



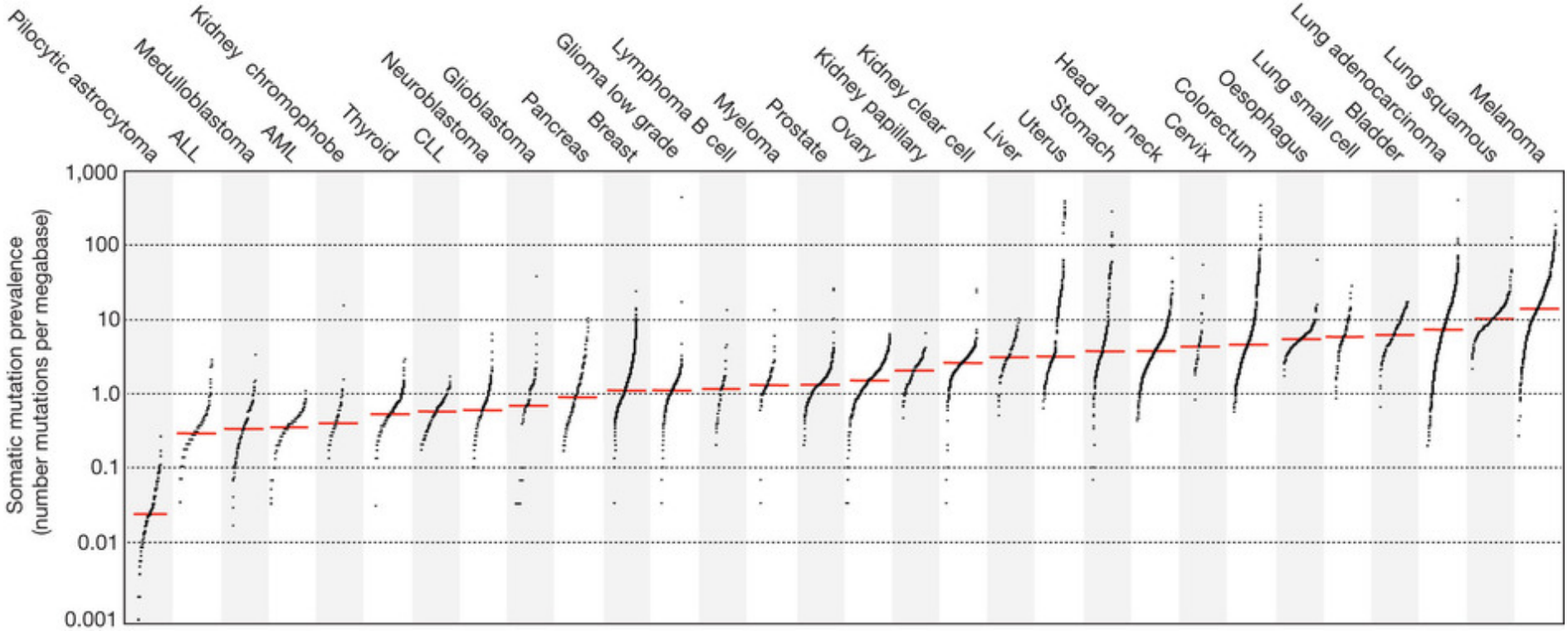
Source: Hillman et al. BMC Cancer 2007

Challenges of cancer calling: tumor heterogeneity



Source: Carlos Caldas Nature Biotechnology 2012

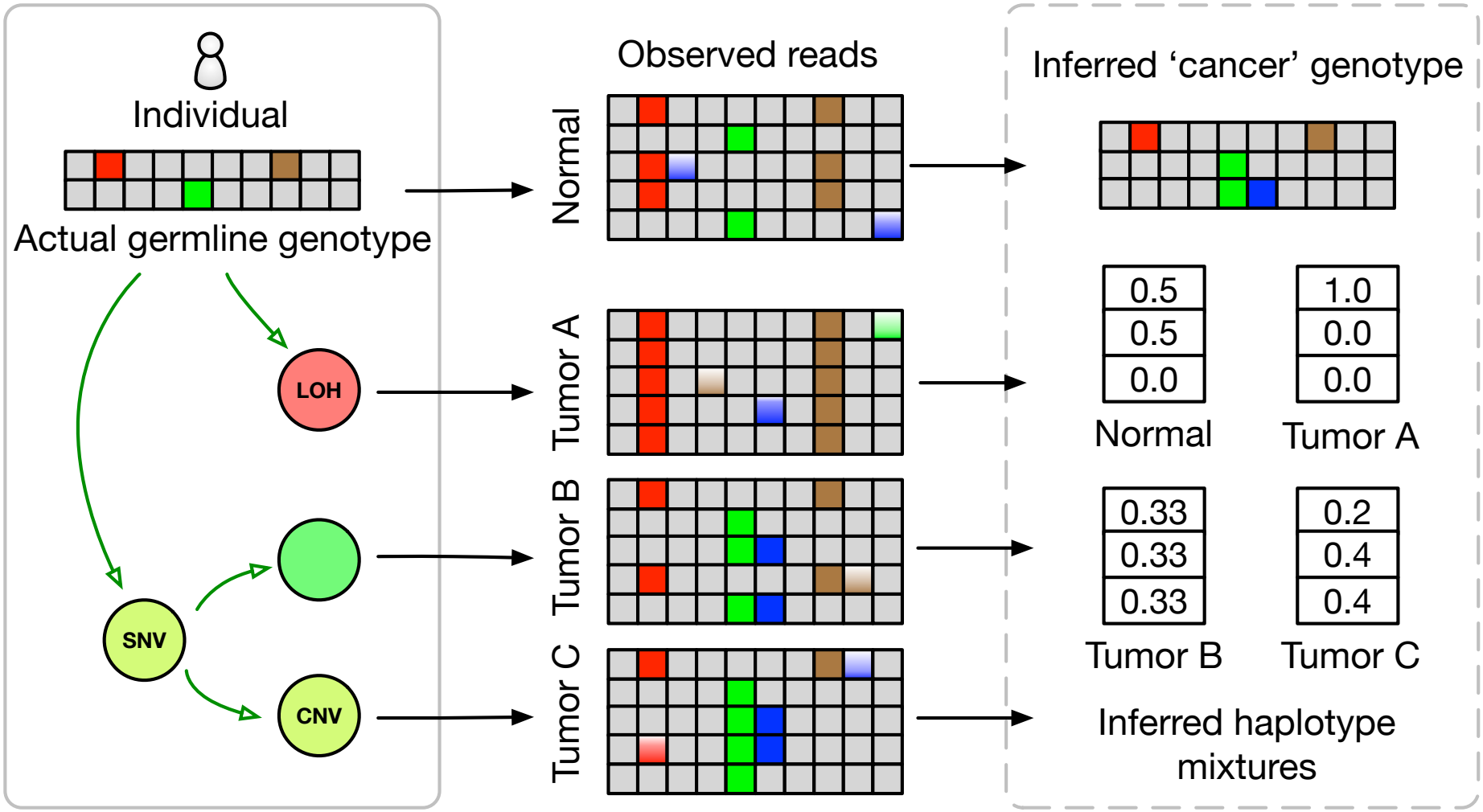
How many haplotypes do we need?



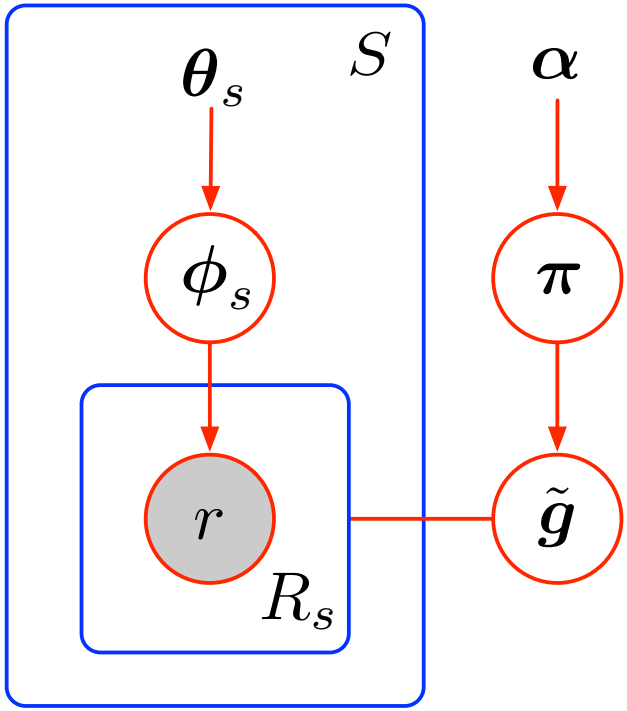
More than three local haplotypes are rare in most cancer types

Source: Alexandrov et al. Nature 2013

Cancer genotype model: overview



Cancer genotype model: maths

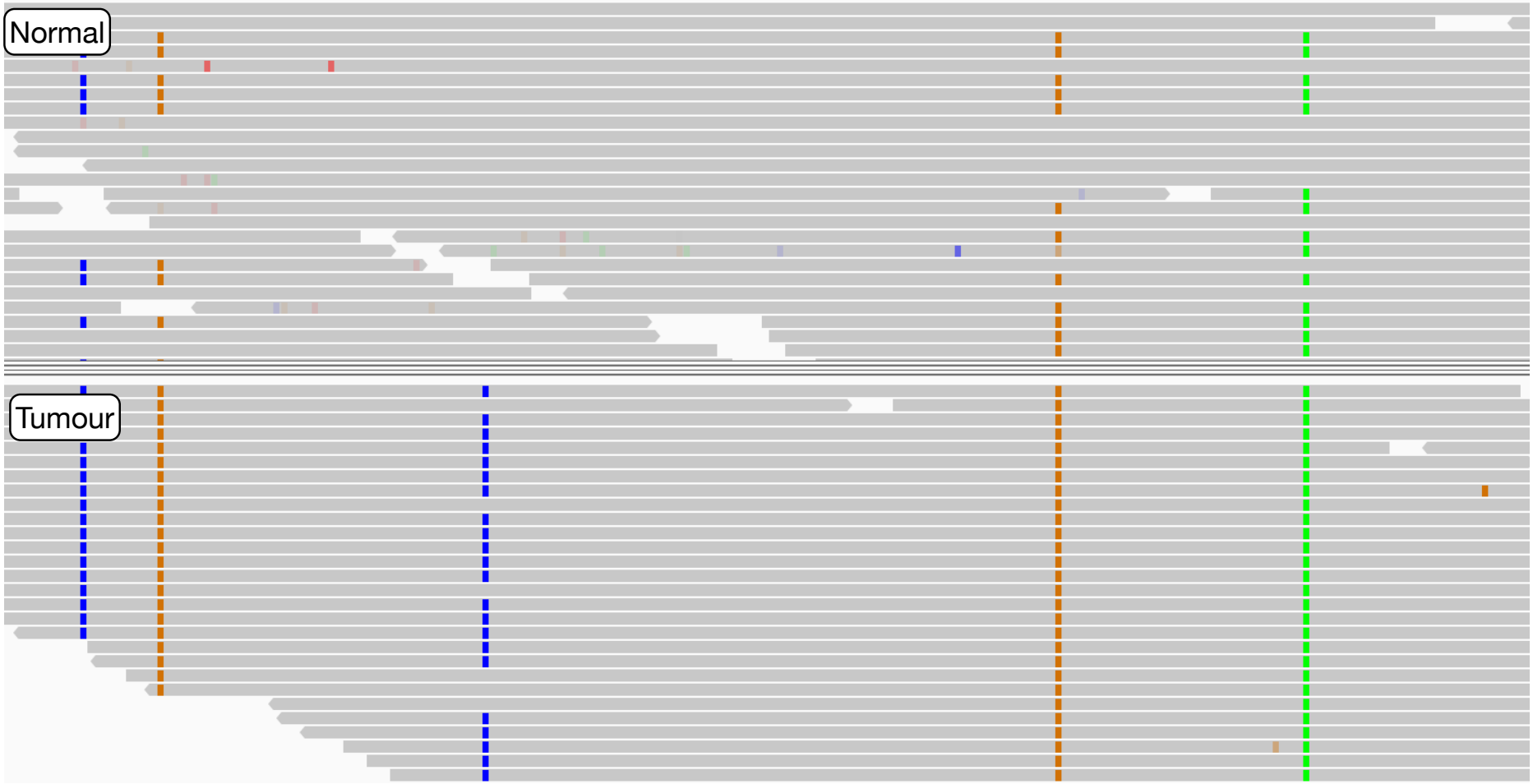


- One unknown 'cancer' genotype $\tilde{\boldsymbol{g}}$
- Unknown haplotype mixtures ϕ_s
- Mixture priors θ_s implicitly model 'normal'

Marginal distribution: diploid case

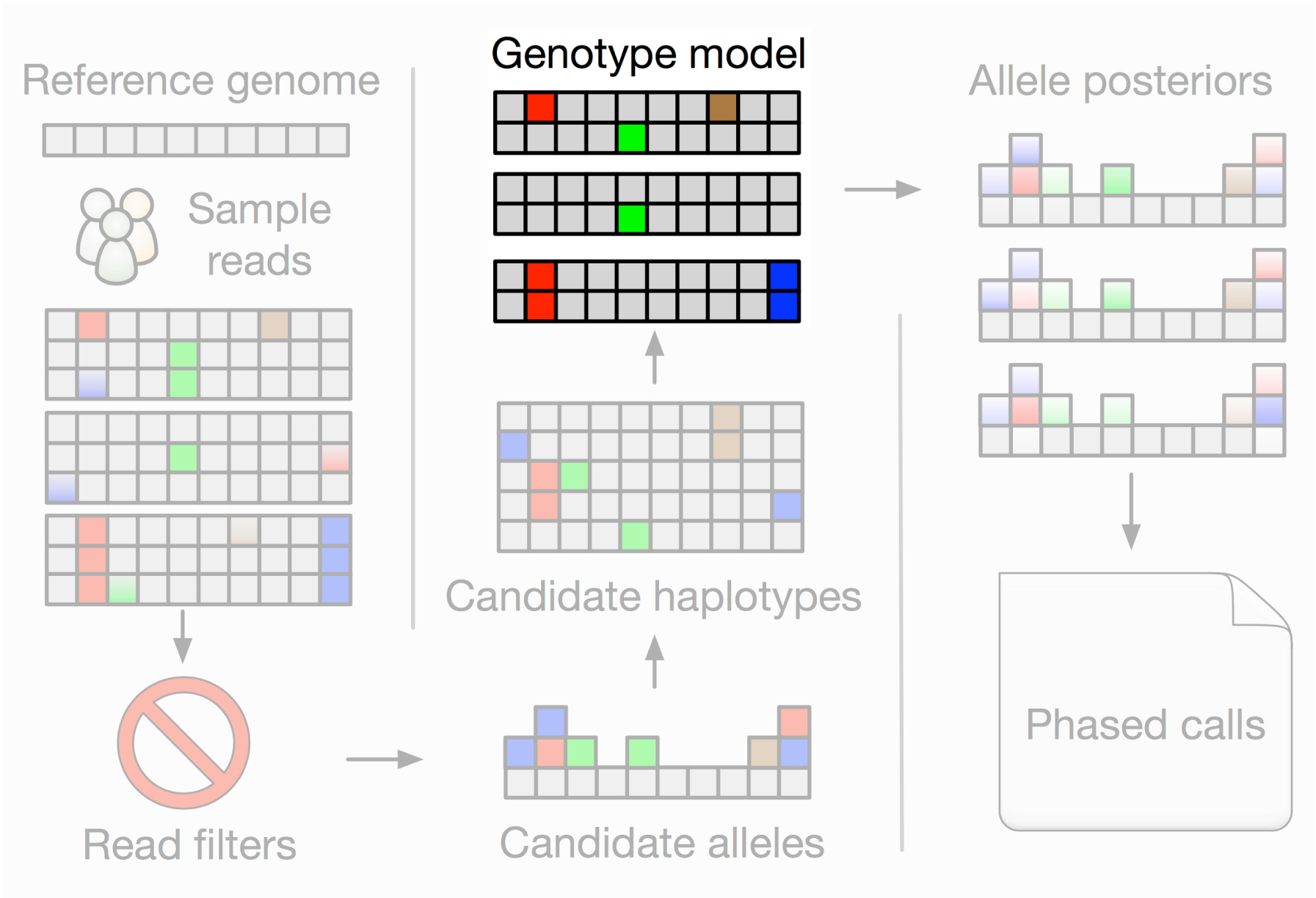
$$p(\boldsymbol{R}, \pi) = p(\pi|\alpha) \sum_{\tilde{\boldsymbol{g}}} p(\tilde{\boldsymbol{g}}|\pi) \prod_{s=1}^{S} \int d\phi_s p(\phi_s|\theta_s) \prod_{r \in R_s} \sum_{i=1}^{3} p(\tilde{\boldsymbol{g}}_i|\phi_{si}) p(r|\tilde{\boldsymbol{g}}_i)$$

Phased somatic mutation calls



Phased somatic mutation calls

Summary & future work



Acknowledgements

Supported by
wellcome trust

UNIVERSITY OF
OXFORD

Lunter group
with special thanks to Gerton Lunter