Unified variation discovery and genotyping from high-throughput sequencing data

Daniel Cooke[1]& Gerton Lunter[1]

Haplotype-based variant callers, which consider physical linkage between variation sites, have become the *de facto* choice for germline variation discovery and genotyping. However, despite being intrinsically more powerful and informative than naïve approaches, there is a notable absence of these tools beyond those aimed at detecting common germline variation in diploid individuals. In particular, there are currently no published methods for targeted *de novo* mutation discovery in parent-offspring trios, nor for somatic mutation detection in tumour samples. One possible reason for this is the technical difficulty in implementing these algorithms. Here we show a flexible haplotype-based variant detection algorithm, Octopus, that incorporates a polymorphic Bayesian genotyping model capable of accurately characterising multiple types of variation sources. Octopus outperforms all existing germline, somatic, and *de novo* mutation detection tools, offers the first general approach to probabilistic phasing directly from raw sequence data, and is readily extendible for new and atypical types of samples.

Detection of genetic variation and genotyping from High-Throughput Sequencing (HTS) data is now commonplace in clinical diagnosis pipelines, and is rapidly replacing array based assays for phenotype association studies. The high resolution of sequencing based assays empowers detection of rare variants, resolution of structural variation (indels), and the potential to resolve haplotype structure with *physical phasing* which provides important information for disease risk analysis and phylogenetic inference. This has lead to the development of a number of variant detection algorithms [34].

Other than a handful of *de novo* based assembly approaches [3], the majority of these algorithms are mapping based [4, 6, 5], which require raw sequencing reads first be mapped to a reference sequence using a read mapper. Naïve variant callers use unaltered read alignments provided from a read mapper, and make calls on a positional basis [1]. These methods make the simplifying - and mistaken - assumption that the read alignments provided by the mapper against the reference are always true. In contrast, haplotype-based methods only assume read mapping location is approximately correct and compare evidence for one or more read alignments under multiple hypothetical haplotypes. This is an improvement for two reasons. First, alignment-based errors are avoidable reducing dependence on the read mapper. Second, the genotype posterior signal-to-noise ratio improves with haplotype length because the number of erroneous haplotypes increases exponentially while the number of true haplotypes remains constant. This makes haplotype based methods more powerful than equivalent positional methods [33].

While there are now several haplotype-based methods available, they have all focused on germline variation in pooled sequencing studies; employing similar genotyping models derived from population genetics to explain observed read data [4, 6, 5]. Such models offer a poor approximation to data typically generated from other types of sequencing experiment, such as tumour or bacterial sequencing. In other cases, valuable prior information, such as pedigree structure, is not fully utilised. This can lead to a significant reduction in power to detect and classify real variation. Although naïve methods that implement more appropriate models are often available, they often perform worse than poorly fitted haplotype-based tools. Researchers often resort to spending valuable time implementing bespoke pipelines that involve post-hoc intersections of multiple caller outputs to obtain required inferences. This is especially prevalent for studies requiring somatic and *de novo* mutation detection, which are often severely limited in ability to characterise all types of variation, potentially leading to misleading conclusions.

There are other important shortcomings to existing tools: (i) sequencing errors are not adequately modelled; most methods [6, 4] assume uniform sequencing errors, but real sequencing errors are often systematic [7], which inflates false positive calls; (ii) read mapping qualities [2] are not properly modelled which can severely reduce specificity; (iii) haplotype generation lacks sensitivity due to approaches that are either too simplistic [6], or are too dependent on parametrisation [4]; (iv) some difficult regions are ignored due to 'windowing' problems - where there are too many alternative alleles in a region to jointly consider.

Finally, variant phase information is becoming increasingly valuable to resolve population haplotype structure. Obtaining phased variants directly from raw sequence data is challenging because there is uncertainty in variant and genotype calls in addition to uncertainty in any particular haplotype structure. Although some existing methods implement non-probabilistic physical phasing algorithms [6], these approaches cannot resolve regions with weak read support or capture biologically meaningful information available in the genotype model, and

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

are usually only effective over regions smaller than one read length with unambiguous high-coverage read support.

We sought to address these issues by designing a method that treats haplotype generation, read error modelling, genotype inference, and phasing independently. This allows us to apply the most appropriate statistical model at each stage, whilst taking advantage of common haplotype generation machinery. We designed a genotype posterior based phasing algorithm that is able to achieve far greater phase lengths than existing tools, sometimes well beyond one read length. Our algorithm is implemented in a C++14 application, Octopus.

RESULTS

Haplotype generation

The performance of haplotype-based methods is bounded by the ability to propose true haplotype sequences. The probability a true haplotype is observed in the read data is conditional on the underlying sample genotypes; some haplotypes may inherently have low observation probability (e.g. in tumour samples). It is therefore crucial to be able to achieve high sensitivity at this stage whilst being conscious of the increased computational burden of proposing many wrong haplotypes.

Haplotype construction begins by extracting putative alleles directly from the read alignments, although this can be augmented by external sources. Our algorithm distinguishes alleles from haplotypes: alleles are atomic mutation events whilst haplotypes are ordered sets of non-overlapping alleles. By default two methods are used for read-based allele generation: (i) variants supported directly from read alignments; (ii) variants identified via local re-assembly. The former method is usually more robust as it is only dependent on the alignment routine used by the mapper. Whilst the latter is able to resolve complex variation which may not be present in the read alignments due to inadequate read length. The main drawback of the latter approach is that it is highly sensitive to kmer size used and the genomic region selected for assembly, which can result in high quality variants being missed. By integrating both approaches, our algorithm is able to achieve higher sensitivity than other approaches which use one or the other [4, 6].

Haplotypes are exhaustively constructed from all viable allele combinations. This differs from other approaches which construct haplotypes directly. The advantage of our approach is that haplotypes not directly supported by any particular read alignment can be proposed which enables arbitrary extension of existing haplotypes, enabling posterior based phasing beyond a single read length. The main limitation is that the number of haplotypes is exponential in the number of alleles, which restricts the length of candidate haplotypes (windowing). We have solved this issue by developing a graph-based data-structure, dubbed a 'haplotype-tree'. Nodes in the tree are alleles, and branches are unique haplotypes. Branches can be readily extended and pruned, accelerated by a hash-table lookup. This enables haplotypes with low posterior support to be removed from the tree before proliferation. This procedure can be repeated to obtain arbitrary long haplotypes whilst keeping the number of haplotypes under consideration roughly linear.

Genotype inference and variant calling

The algorithm proceeds by calculating likelihoods for each read conditional on each haplotype using a Hidden Markov Model (HMM). Base mismatch and indel emission distributions are dependent on surrounding sequence context and sequencing technology, the former is derived from base quality scores. Error penalties occurring outside the current active region are deducted by following the Viterbi path. It may then be necessary to eliminate some haplotypes using likelihood based statistics to produce a final set of haplotypes.

The next step of the algorithm is to calculate posterior probability distributions for all Bayesian genotype model latent variables (**Figure 1**), which are used by *calling models* to make variant, genotype calls, and any other sample-type specific inferences.

The population model is applicable to a set of samples of known ploidy and unknown pedigree. Variations of this model are implemented by most existing methods [6, 4, 5, 1]. For the case of a single sample the model simplifies to:

$$p(\boldsymbol{g}|\boldsymbol{R}, \mathcal{M}_g) \propto p(\boldsymbol{g}|\mathcal{M}_g) \prod_{r \in \boldsymbol{R}} \frac{1}{|\boldsymbol{g}|} \sum_{i=1}^{|\boldsymbol{g}|} p(r|\boldsymbol{g}_i)$$

where $p(\boldsymbol{g}|\mathcal{M}_g)$ is a germline prior model. For more than one sample exact inference is usually intractable and approximations are required. Genotype likelihoods are first computed under a simpler model using Expectation Maximisation [5], and then high likelihood genotype combinations are used to compute posterior probabilities under the Bayesian model.

The trio model is appropriate for parent-offspring trios as the genetic inheritance of genetic information from parent to child can be explicitly modelled. The parent-child relationship is stochastic as there is uncertainty in the inherited haplotype and because of *de novo* mutations during meiosis. Recombination is not considered. Parental genotypes are assumed to share information as in the population model. This model can be computed exactly in most cases.

The tumour model is intended to model populations of somatically mutated heterogeneous tumour cell samples from the same individual. In contrast with the other models, the genotype latent variable in this model is augmented with an additional haplotype to capture somatically mutated tumour genotypes. Although more than one unique *somatic haplotypes* are possible, such observations are rare over typically considered region lengths (up to a few kb) [35]. Also unlike the other models, the tumour model does not assume equal genotype mixture weights to reflect the possibility of low-frequency alleles observations due to local copy-number changes and tumour cell impurity caused by heterogeneous sub-clonal sampling and normal contamination. To explicitly model sub-clonal cell populations, each sample has independent mixture priors, which also allows implicit modelling of a normal sample. Approximate

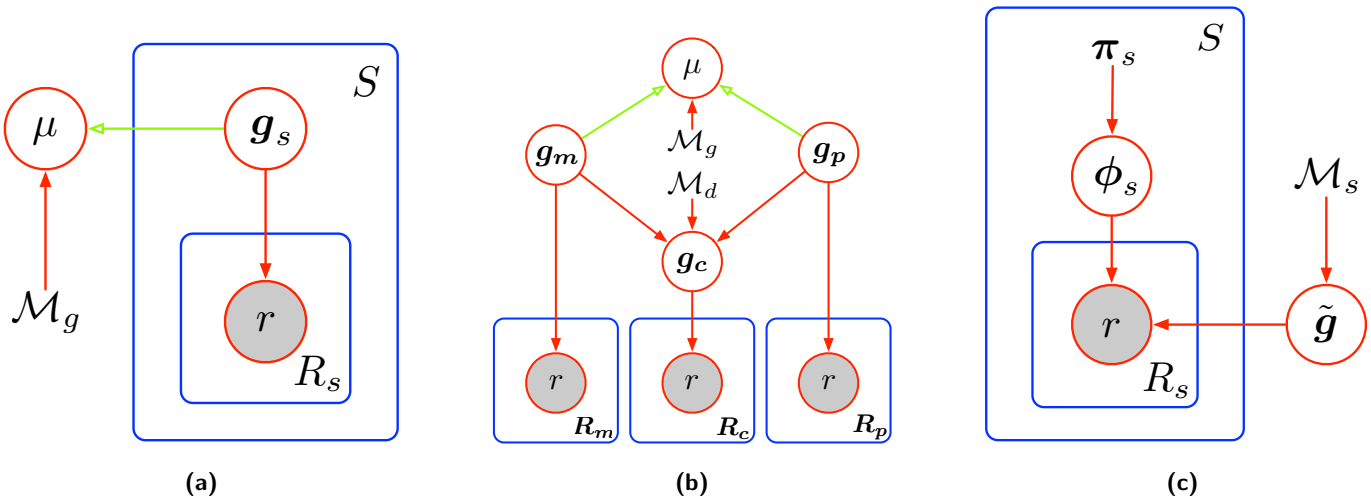

Figure 1 Genotype models shown in graph notation: **(a)** population, **(b)** trio, and **(c)** tumour. Symbols insides circles are latent variables, observed variables are shaded. Symbols inside boxes are repeated. Symbols not inside a circle are parameters or models. Arrows define conditional relationships, red for stochastic and green for deterministic.

posterior probability distributions are assigned to genotypes and sample mixtures using a Variational Bayes algorithm.

Variant calls are made by marginalising over the relevant posterior probability distributions assigned by the calling model. Posterior inferences may also include a Bayesian model comparison to reduce false positive calls due to poor model fit. In particular, somatic mutation calls are only made if the *evidence* for the tumour model is significantly higher than the evidence for the population model (with a single sample). For germline calling, evidence comparison with a incremented ploidy model may be used to detect false positive variant calls caused by mapping errors. For each allele, unphased genotype probabilities (in allele space) are calculated by marginalising over full genotype posterior distributions (in haplotype space).

Statistical phasing

The genotype posterior probability distributions computed in the last step are used to compute physical phasing of called variant sites. No read data is required as all information available from the read data as well as any prior information is contained in the posterior distributions.

Samples are phased independently. First, all possible genotypes are partitioned into 'phase complement' sets (genotypes that share the same alleles at every site), there is only one such partitioning (**Figure 2**). A phase score is defined as the weighted entropies of each normalised phase complement set, with respect to the genotypes present in each phase complement set. Low phase scores indicate less ambiguous phasing. The phasing algorithm finds the smallest set of genomic subregions such that each subregion has a phase score below a given value.

Germline variants from WGS

Although germline variant calling is a mature field, recent benchmark studies have shown there is still surprisingly low concordance between callers, and performance is highly dependent on the mapper used [28, 29]. To assess the performance of our algorithm on whole-genome data we used variants from the well studied 1000G sample NA12878, using the gold-standard high confidence calls provided by the Genome in a Bottle (GIAB) consortium [15] as a benchmark. We made nine call-sets for each caller, one for each coverage-mapper combination (mappers: BWA-mem [30], Bowtie2 [31], and Novoalign). Calls were compared using RTG-tools [26] which is more accurate than naïve intersection methods as calls are compared at a haplotype level. We compared our algorithm to four widely used tools, GATK HaplotypeCaller [4], Freebayes [6], Platypus [5], and Samtools [1].

Germline variants from exome-capture

Many clinical pipelines use exome-capture sequencing to detect disease risk variants. While in principal exome variant calling is easier than whole genome calling, due to differences in library preparation and sequencing depth, there are subtle artefacts present in exome data not in whole genome sequence data. We sought to evaluate the performance of our algorithm on clinical exome-capture sequence data. We compared our method to other tools by callings variants in the ICR 142 validation series data [16] which includes positive and negative Sanger validation sites for a number of exome samples.

Somatic mutations in paired tumour-normal samples

Cancer is a genetic disease caused by mutations that lead to uncontrollable cell division. Different cancers are driven by different sets of mutations to different gene pathways. A crucial step to understanding cancer; how to provide new diagnos-

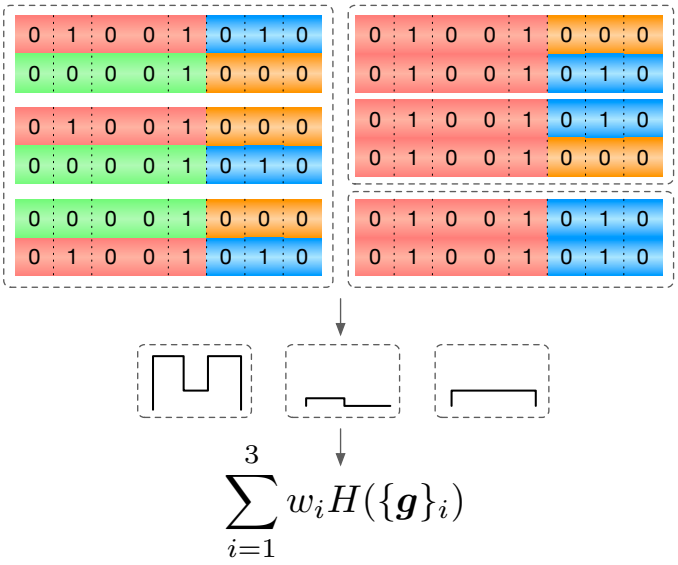

Figure 2 Posterior based variant phasing with phase complement sets: Genotypes are partitioned into phase complement sets which are used to calculate a phase score, the weighted entropy of each phase complement set. The painted haplotypes show one possible local allele boundary.

Table 1 ICR142 series summary

	ICR142 sites			
	TP	FP	TN	FN
Octopus	0	0	0	0
GATK HC	0	0	0	0
Freebayes	0	0	0	0
Platypus	0	0	0	0
Samtools	0	0	0	0

tics and therapies, is to accurately be able to genotype somatic mutations in tumours. Cancer genomes are often highly mutated, containing structural variation that alters local copy number. Cells from the same tumour form pockets of genetically heterogenous clones, and there is little guarantee of clone representation in the final sequencing library. All this makes the detection of somatic mutations challenging, particularly if the inaccurate assumptions are made when modelling the observed data. In particular, germline ploidy assumptions are likely to lead to low sensitivity to low allelic frequency somatic mutations. A number of bespoke somatic mutation detection tools have been developed in an attempt to address these issues [8, 9, 11, 10, 12]. While these algorithms provide more appropriate models than advanced germline variant callers, none, other than a recent update to MuTect (Mutect2) are haplotype-based.

To demonstrate the accuracy of our method at calling somatic mutations in tumour-normal paired cancer samples, we compared the sensitivity and specificity of our method to popular somatic callers MuTect, Mutect2, SomaticSniper, Radia, Strelka, and VarScan2.

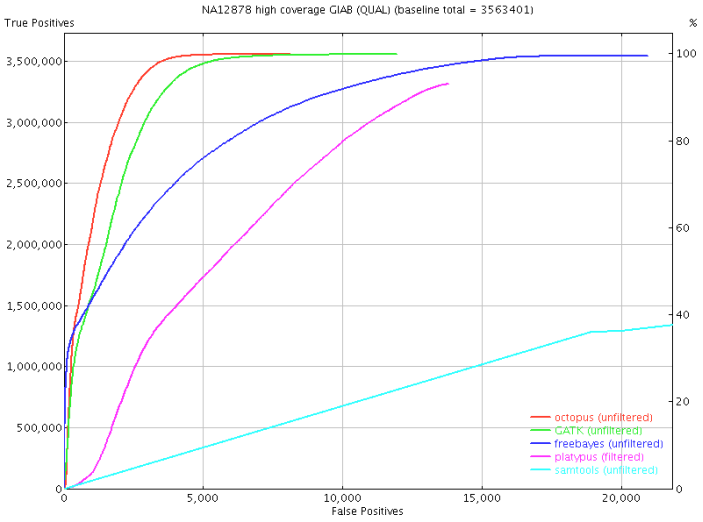

Figure 3 ROC curves comparing variant callers on GIAB high confidence calls.

Somatic mutations in tumour only samples

Reliable paired normal tissue samples are not always available when studying tumours from cancer patients, either because of restrictions present in a clinical setting, or because samples are archival. Most somatic mutation callers require a paired normal sample [8, 9, 11, 10, 12], although a recent tool was made available to specifically to deal analyse tumour only samples [27]. We looked to test the robustness of our algorithm by running the previous samples without the paired normal samples, and compared these calls to our previous calls, and to those made by SomVarIUS.

De novo mutations in parent-offspring trios

Random germline *de novo* mutations resulting from imperfections in the DNA replication process during meiosis provide the necessary genetic variation for evolution. They are also known to be causative of several Mendelian and polygenic diseases [21, 24, 23]. Large scale characterisation of *de novo* mutations in population studies reveals important insights into population structure and demographic history [22]. The fidelity of the DNA replication process means the number of *de novo* mutations expected per genome duplication event is small, usually less than 100 mutations per duplication in humans [25].

Bacterial calling using HTS data

HTS is rapidly replacing culture based assays to diagnose infectious diseases in the clinic. Its potential to detect and identify strains means that can be used to establish drug-susceptibility profiles that potentially offers patients faster, more accurate diagnosis, and hence targeted treatment. This is especially important when patients present with multiple drug-resistant bacterium such a Mycobacterium Tuberculosis; a specific early diagnosis can result in life-saving treatment.

Some bacteria have high mutation rates resulting in high

degrees of heterogeneity in a sample; with small fraction of cells containing a mutation. In addition, patients may be infected with multiple strains of the same species resulting in pseudo-poly-ploidy genotypes. These conditions render many germline variant calling tools inappropriate and underpowered [32].

Genotyping HLA loci

Detection of structural variation

DISCUSSION

As the use of high-throughput sequencing technology expands to new types of organisms, cells, and library preparation techniques, we will need new algorithms to accurately re-assemble the underlying genetic material of these assays. Current approaches are either too rigid or too simple, which limits the potential of the data we generate. We introduce a method that is able to achieve class-leading performance for multiple data types using a powerful core Bayesian framework, and is readily extendible to new types of sample. We hope researchers with niche samples that are unsatisfied with existing tools will be interested in our method.

Single-cell methods will soon enable us to accurately sequence the genomes of individual cells, potentially revolutionising the way we characterise germline and tumour genomes. As our understanding of molecular biology improves, we will be able to devise better statistical models to explain single cell data. Octopus offers an exciting opportunity to implement these models to uncover new insights into genetic variation within individuals.

References

[1] Li H. *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data*. Bioinformatics. 2011.

[2] Li, H., Ruan, J. & Durbin, R. *Mapping short DNA sequencing reads and calling variants using mapping quality scores*. Genome Res. 2008.

[3] Iqbal et al. *De novo assembly and genotyping of variants using colored de Bruijn graphs*. Nature Genetics. 2012.

[4] DePristo et al. *A framework for variation discovery and genotyping using next-generation DNA sequencing data*. Nature Genetics. 2011.

[5] Rimmer et al. *Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications*. Nature Genetics. 2014.

[6] Garrison E, Marth G. *Haplotype-based variant detection from short-read sequencing*. arXiv preprint.

[7] Schirmer et al. *Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform*. Nucl. Acids Res. 2015.

[8] Cibulskis et al. *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples*. Nature Genetics. 2013.

[9] Larson et al. *SomaticSniper: identification of somatic point mutations in whole genome sequencing data*. Bioinformatics. 2012.

[10] Saunders, C.T. et al. *Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs*. Bioinformatics. 2012.

[11] Radenbaugh, A.J. et al *RADIA: RNA and DNA integrated analysis for somatic mutation detection*. PLoS ONE. 2014.

[12] Koboldt et al *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing*. Genome Res. 2012.

[13] Wei et al. *A Bayesian framework for de novo mutation calling in parents-offspring trios*. Bioinformatics. 2015.

[14] Deatherage DE, Barrick JE. *Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq*. Methods Mol Biol. 2014.

[15] Zook et al. *Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls*. Nature Biotechnology. 2014.

[16] Ruark et al. *The ICR142 NGS validation series: a resource for orthogonal assessment of NGS analysis*. F1000Research. 2016

[17] Alioto et al. *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing*. Nature Communications. 2015.

[18] Ewing et al. *Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection*. Nature Methods. 2015.

[19] nik-Zainal et al. *Landscape of somatic mutations in 560 breast cancer whole-genome sequences*. Nature. 2016.

[20] Conrad et al. *Variation in genome-wide mutation rates within and between human families*. Nature Genetics. 2011.

[21] Veltman and Brunner. *De novo mutations in human genetic disease*. Nature Reviews Genetics. 2012.

[22] Francioli et al. *Whole-genome sequence variation, population structure and demographic history of the Dutch population*. Nature Genetics. 2014.

[23] Gilissen. et al. *Genome sequencing identifies major causes of severe intellectual disability*. Nature. 2014.

[24] Xu et al. *De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia*. Nature Genetics. 2012.

[25] Kong et al. *Rate of de novo mutations and the importance of fathers age to disease risk*. Nature. 2012.

[26] Cleary et al. *Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines*. arXiv preprint.

[27] Smith et al. *SomVarIUS: somatic variant identification from unpaired tissue samples*. Bioinformatics. 2015.

[28] Cornish, A and Guda, C. *A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference*. Biomed Res Int. 2015.

[29] Hwang et al. *Systematic comparison of variant calling pipelines using gold standard personal exome variants*. Scientific Reports. 2015.

[30] Li H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv preprint.

[31] Ben Langmead & Steven L Salzberg. *Fast gapped-read alignment with Bowtie 2*. Nature Methods. 2012.

[32] Olson et al. *Best practices for evaluating single nucleotide variant calling methods for microbial genomics*. Front Genet. 2015.

[33] Lo et al. *Comparing variant calling algorithms for target-exon sequencing in a large sample*. BMC Bioinformatics. 2015.

[34] Wang and Lin. *Detecting associations of rare variants with common diseases: collapsing or haplotyping?*. Brief Bioinform. 2015.

[35] Alexandrov et al. *Signatures of mutational processes in human cancer*. Nature. 2013.

ONLINE METHODS

Haplotype likelihood filtering

For efficiency reasons, it may sometimes be necessary to reduce the haplotype set before genotype inference. While there is no silver bullet solution, we have found likelihood based approaches are most effective.

Coalescent priors

The default germline prior model, $\mathcal{M}_g$, is a coalescent based model, $\mathcal{M}_{coal}$, is used in all models. The somatic prior model, $\mathcal{M}_s$, is also partially defined in terms of $\mathcal{M}_{coal}$, for the germline component of the augmented genotype. This model has two heterozygosity parameters θ_1 and θ_2 for SNP and indel variants respectively. Indel heterozygosities are known to vary considerably depending on sequence context, we therefore define $\theta_2 = \lambda(s)\theta_2'$ where $\lambda(s)$ is a function of the surrounding sequence context s and θ_2' is the baseline indel heterozygosity. Given a set of haplotypes, H, which could be a genotype, we calculate the union of segregating snp and indel sites, k_1 and k_2, between H and the reference haplotype. Then $p(H|\mathcal{M}_{coal})$ is:

$$p(k_1, k_2|\theta_1, \theta_2) = \binom{\theta_1}{\theta_1+\theta_2}^{k_1}\binom{\theta_2}{\theta_1+\theta_2}^{k_2}\binom{k_1+k_2}{k_1}p_{\theta_1+\theta_2}(k_1+k_2)$$

Where $p_\theta(S=k)\sum_{i=2}^{n}(-1)^i\binom{n-1}{i-1}\binom{i-1}{\theta+i-1}\binom{\theta}{\theta+i-1}^k$.

Population model approximation

Trio model priors

The key component in the trio model is the conditional offspring genotype prior probability $p(\boldsymbol{g_c}|\boldsymbol{g_m}, \boldsymbol{g_p}, \mathcal{M}_{denovo})$.

Estimation of somatic allele frequency

Marginalisation of genotype posteriors