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, are currently the *de facto* choice for germline variation discovery and genotyping. However, there is a notable lack of these tools beyond those aimed at detecting common germline variation in diploid human individuals. Here we show a flexible haplotype-based variant detection algorithm, Octopus, that incorporates a polymorphic Bayesian genotyping model capable of accurately characterising multiple 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 to atypical samples and new sequencing technology.

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 offered by sequencing empowers detection of rare variants not covered by genotype arrays; enables identification of small indels and larger structural variation; and in some instances can resolve haplotype structure with *physical phasing* which provides important information for disease risk analysis and phylogenetic inference [1].

Over the past ten years, a number of variant detection algorithms which use HTS as input have been developed. Other than a handful of *de novo* based assembly approaches [2], the majority of these methods are mapping based [3–5], which require raw sequencing reads first be aligned to a reference sequence using a read mapper. Naïve variant callers directly use read alignments provided from a read mapper by forming a pileup of read bases at each reference base, calls are then made at each position by comparing the proportion of reference and non-reference observations [6]. These methods make the simplifying - but mistaken - assumption that the read alignments provided by the mapper against the reference are always correct. 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, certain alignment-based errors, particularly evident around indel variants, are avoidable which reduces 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 [7].

All haplotype-based methods published to date all focus on germline variation in pooled sequencing studies, and employ similar genotyping models derived from population genetics to explain observed read data [3–5]. Such models offer a poor approximation to data typically generated from other types of sequencing experiment, including tumour and bacterial sequencing. In other cases, valuable prior information, such as pedigree structure, is not fully utilised. This can result in a significant reduction in power to detect and classify real variation. Although naïve methods that implement more appropriate statistical models are often available, they often perform worse than poorly fitted haplotype-based tools. Researchers often resort implementing custom pipelines that involve post-hoc intersections of multiple caller outputs to obtain required inferences [8–15]. 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 in existing tools: First, sequencing errors are not adequately modelled; most methods [3, 4] assume uniform sequencing errors, but actual sequencing errors are often context-dependent [16], leading to falls calls; Second, read mapping qualities [17] are often inadequately modelled, which can reduce specificity; Third, haplotype lengths are usually short as the number of possible haplotypes increases exponentially with the number of candidate variants, this is especially problematic in highly polymorphic regions, which can induce boundary artefacts resulting in false calls, or the region being skipped entirely.

Finally, phase information is becoming increasingly valuable to resolve population haplotype structure and infer functional effects of co-mutations [1]. Obtaining variant phase directly from raw sequence data is challenging because phase and genotype inference is confounded as 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 [4], these approaches cannot resolve regions with weak read support or

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

capture biologically meaningful information available in the genotype model, such as inheritance patterns. Therefore, existing methods are usually only able to offer phased calls over regions smaller than one read length with unambiguous strong read support [4].

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 statistical phasing algorithm based off genotype posterior distributions 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*, which is freely available to all.

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 remaining 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 local reassembly 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 [3, 4].

Haplotypes are exhaustively constructed from all viable allele combinations. This differs from other approaches which construct haplotypes directly from read observations. 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 specific haplotypes to be removed from the tree before proliferation. This procedure can be repeated to obtain arbitrary long haplotypes whilst keeping the upper bound on the number of haplotypes under consideration constant, a process we have termed *lagging*. The main drawbacks from this approach is the increased computational work required to recompute alignments on haplotype extension, and the decrease in posterior resolution once haplotypes are pruned. We have addressed these issues by only pruning haplotypes with low posterior support, and allowing the tree to be rebased (i.e. the root node is changed to another node already in the tree) once the tree depth reaches a certain theshold. The amount of lagging can also be user controlled, or switched of entirely, depending on preference of calling and phase accuracy or runtime.

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. If there are many haplotypes, it may be necessary to filter some using likelihood based statistics at this stage to produce a final set of haplotypes for genotyping.

The next step of the algorithm is to calculate posterior probability distributions for all genotype model latent variables (**Figure 1**), which are used by *calling models* to make variant & genotype calls, and other data-type specific inferences (e.g. *de novo* status in trios). There are currently four calling models implemented, which we discuss briefly.

The individual model is the simplest model possible; it models a single individual of known ploidy. The genotype posterior distribution, which can be computed exactly, can be written as

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

where $\boldsymbol{g}$ is the genotype, $\boldsymbol{R}$ is the read set, $p(\boldsymbol{g}|\mathcal{M}_g)$ is the genotype prior, and $|\boldsymbol{g}|$ is the sample ploidy.

The population model is applicable to a set of samples of known ploidy but unknown pedigree. Variations of this model are implemented by most existing methods [3–6]. In most cases exact inference is intractable and approximations are required. Genotype likelihoods are first computed under a simpler model using Expectation Maximisation (EM) [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 inheritance of haplotypes from parent to child can be explicitly modelled. The parent-child relationship is stochastic as there is uncertainty in the inherited haplotype in addition to *de novo* mutations which may occur during meiosis. Recombination

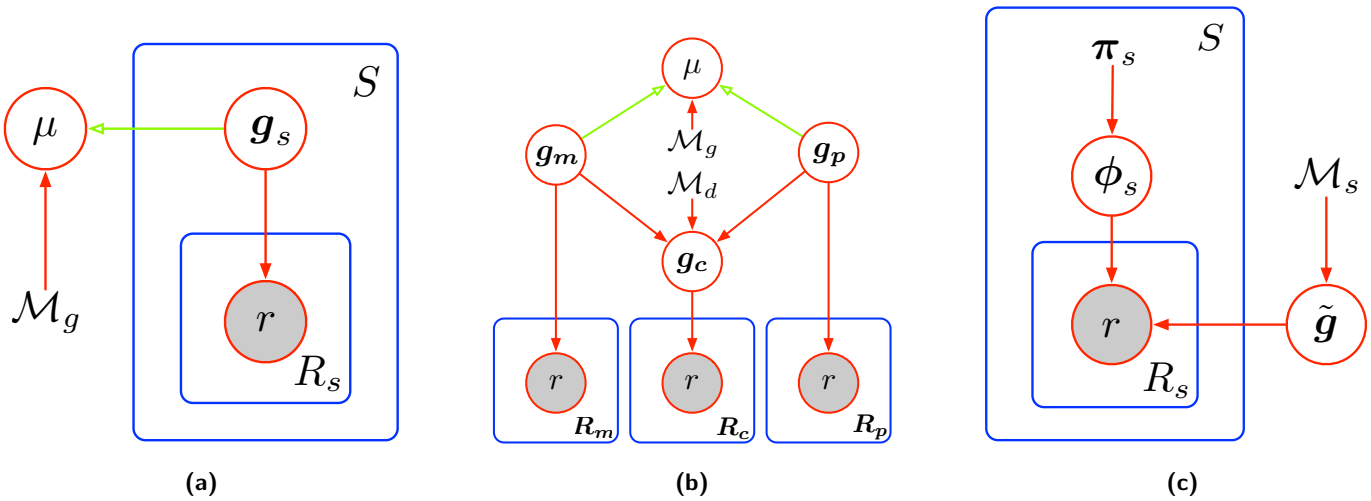


Figure 1 Genotype models shown in graph notation: **(a)** population, **(b)** trio, and **(c)** tumour. Symbols insides circles are latent variables, observed variables are shaded. Symbols inside boxes are repeated. Symbols not inside a circle are parameters or models. Arrows define conditional relationships,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

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 a single 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 tumour samples typically contain many unique *somatic haplotypes*, such observations are rare over the region lengths considered by the algorithm (up to a few kb) [18]. 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 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 individual model. For germline and *de novo* calling, evidence comparison with a incremented ploidy model may be used to detect false positive variant calls caused by mapping errors. For each allele, local genotype probabilities (i.e. at allele resolution) are calculated by marginalising over full genotype posterior distributions.

Statistical phasing

The genotype posterior probability distributions computed in the previous step are used to compute physical phasing of called variant sites. No read data is required as all information available from the reads, as well as any prior information, is already contained in the posterior distribution. The advantage of this method is exemplified when calling trios as the genotype prior can be strongly informative about phase due to identity by decent restrictions.

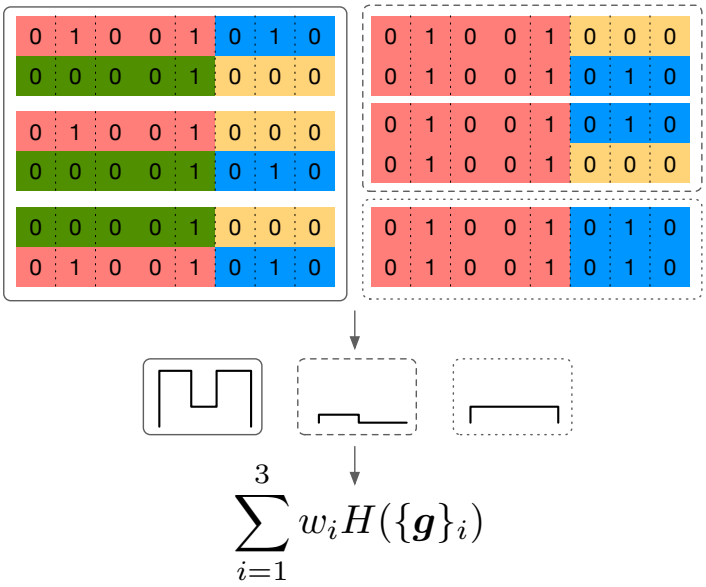


Figure 2 Posterior based variant phasing with phase complement sets: Genotypes are partitioned into phase complement sets which are used to calculate a phase score, the weighted entropy of each phase complement set. The painted haplotypes show one possible local allele boundary.

Samples are phased independently. All possible genotypes are partitioned uniquely into *phase complement* sets (genotypes that share the same allele set 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. Lower phase scores indicate less ambiguous phasing. If the phase score for a given set of genotypes is lower than a given threshold the region is considered phased, otherwise, each genotype in the set can be broken into two parts (all at the same position) which results in two unphased sets of partial genotypes. The phase scores of these two partial sets is never greater than the phase score of the original genotype set. The phasing algorithm therefore iteratively finds the smallest set of breakpoints such that the partial genotype sets defined by those breakpoints are all phased.

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, especially for indels, and performance is highly dependent on the mapper used [19, 20]. To assess the performance of our method on whole-genome data, we called variants in the well studied 1000G sample NA12878, and used the gold-standard high confidence calls provided by the Genome in a Bottle (GIAB) consortium [21] for comparison. We compared our algorithm to four widely used tools, GATK HaplotypeCaller [3], Freebayes [4], Platypus [5], and Samtools [6], following best practise protocols for each where given. Calls were made using four publicly available NA12878 whole genomes: 1000 Genomes high and low coverage reads, Illumina Platinum Genomes, and reads from the Illumina X Ten platform. To account for mapper bias, we remapped all reads using three well known read mappers (BWA-mem [22], Bowtie2 [23], and Novoalign). This resulted in 48 separate call sets. Calls were evaluated using RTG-tools [24] which is more accurate than naïve intersection methods as calls are compared at a haplotype level.

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 [25] which includes positive and negative Sanger validation sites for a number of exome samples.

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 [26–28]. Large scale characterisation of *de novo* mutations in population studies reveals important insights into population structure and demographic history [29]. 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 [30].

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

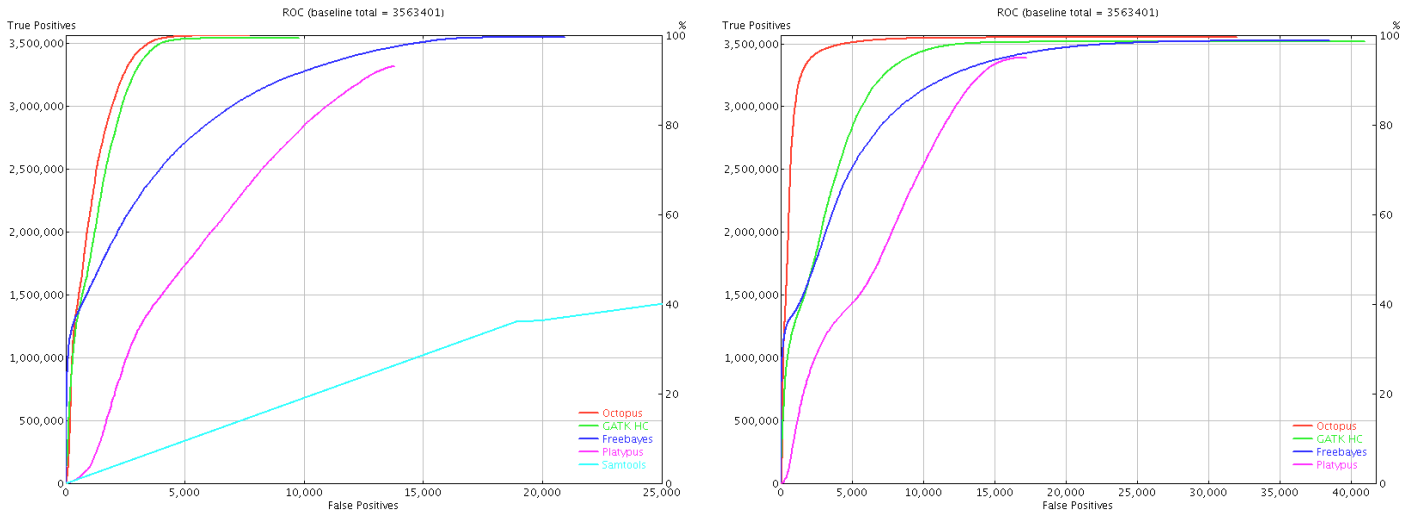
Somatic mutations in paired tumour-normal samples

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 diagnostics 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 [31–35]. 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.

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 [31–35], although a recent tool was made available to specifically to deal analyse tumour only samples [36]. 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. We benchmarked our method using data



(a) **(b)**

Figure 3 ROC curves comparing variant callers on GIAB high confidence calls generated using RTG Tools using all variants (SNPs and indels). All ROC curves were scored using the VCF QUAL field. **(a)** shows calls from high coverage 1000G data. **(b)** shows calls from X Ten data.

from [37, 38].

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 [39].

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] Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ (2011) The importance of phase information for human genomics. Nat Rev Genet 12: 215-23.

[2] Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de bruijn graphs. Nat Genet 44: 226-32.

[3] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation dna sequencing data. Nat Genet 43: 491-8.

[4] Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. bioRxiv .

[5] Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, et al. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 46: 912-8.

[6] Li H (2011) A statistical framework for snp calling, mutation discovery, association mapping and population genet-

ical parameter estimation from sequencing data. Bioinformatics 27: 2987-93.

[7] Lo Y, Kang HM, Nelson MR, Othman MI, Chissoe SL, et al. (2015) Comparing variant calling algorithms for target-exon sequencing in a large sample. BMC Bioinformatics 16: 75.

[8] Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, et al. (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534: 47-54.

[9] Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, et al. (2017) Whole-genome landscapes of major melanoma subtypes. Nature 545: 175-180.

[10] Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, et al. (2017) The whole-genome landscape of medulloblastoma subtypes. Nature 547: 311-317.

[11] Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, et al. (2015) Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 518: 495-501.

[12] Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, et al. (2016) Multi-nucleotide de novo mutations in humans. PLoS Genet 12: e1006315.

[13] Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, et al. (2017) Parental influence on human germline de novo mutations in 1,548 trios from iceland. Nature 549: 519-522.

[14] Deciphering Developmental Disorders S (2017) Prevalence and architecture of de novo mutations in developmental disorders. Nature 542: 433-438.

[15] Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, et al. (2015) Whole-genome sequencing for prediction of mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis 15: 1193-202.

[16] Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, et al. (2015) Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. Nucleic Acids Res 43: e37.

[17] Li H, Ruan J, Durbin R (2008) Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-8.

[18] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. (2013) Signatures of mutational processes in human cancer. Nature 500: 415-21.

[19] Cornish A, Guda C (2015) A comparison of variant calling pipelines using genome in a bottle as a reference. Biomed Res Int 2015: 456479.

[20] Hwang S, Kim E, Lee I, Marcotte EM (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep 5: 17875.

[21] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, et al. (2014) Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. Nat Biotechnol 32: 246-51.

[22] Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. bioRxiv .

[23] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9: 357-9.

[24] Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, et al. (2015) Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. bioRxiv .

[25] Ruark E, Renwick A, Clarke M, Snape K, Ramsay E, et al. (2016) The icr142 ngs validation series: a resource for orthogonal assessment of ngs analysis. F1000Res 5: 386.

[26] Veltman JA, Brunner HG (2012) De novo mutations in human genetic disease. Nat Rev Genet 13: 565-75.

[27] Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, et al. (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat Genet 44: 1365-9.

[28] Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, et al. (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511: 344-7.

[29] Genome of the Netherlands C (2014) Whole-genome sequence variation, population structure and demographic history of the dutch population. Nat Genet 46: 818-25.

[30] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471-5.

[31] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31: 213-9.

[32] Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, et al. (2012) Somaticsniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28: 311-7.

[33] Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, et al. (2014) Radia: Rna and dna integrated analysis for somatic mutation detection. PLoS One 9: e111516.

[34] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, et al. (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28: 1811-7.

[35] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-76.

[36] Smith KS, Yadav VK, Pei S, Pollyea DA, Jordan CT, et al. (2016) Somvarius: somatic variant identification from unpaired tissue samples. Bioinformatics 32: 808-13.

[37] Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, et al. (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods 12: 623-30.

[38] Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, et al. (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun 6: 10001.

[39] Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, et al. (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6: 235.

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