Genetics and Population Analysis

**geck**: trio-based comparative benchmarking of variant calls

Péter Kómár* and Deniz Kural

Seven Bridges Inc., 1 Main St, Cambridge MA 02142, USA
Totient Inc., 1 Main St, Cambridge MA 02142, USA

* To whom correspondence should be addressed.

Abstract

**Motivation:** Classical methods of comparing the accuracies of variant calling pipelines are based on truth sets of variants whose genotypes are previously determined with high confidence. An alternative way of performing benchmarking is based on Mendelian constraints between related individuals. Statistical analysis of Mendelian violations can provide truth set-independent benchmarking information, and enable benchmarking less-studied variants and diverse populations.

**Results:** We introduce a statistical mixture model for comparing two variant calling pipelines from genotype data they produce after running on individual members of a trio. We determine the accuracy of our model by comparing the precision and recall of GATK Unified Genotyper and Haplotype Caller on the high-confidence SNPs of the NIST Ashkenazim trio and the two independent Platinum Genome trios. We show that our method is able to estimate differential precision and recall between the two pipelines with $10^{-3}$ uncertainty.

**Availability:** The Python library geck, and usage examples are available at the following URL: https://github.com/sbg/geck, under the GNU General Public License v3.

Contact: peter.komar@totient.bio

Supplementary information: Supplementary materials are available at Bioinformatics online.

1 Introduction

Next-generation sequencing (NGS) has enabled entire human genomes to be sequenced in a single experiment and has, since its development, played a crucial role in many areas including population genetics (Auton et al., 2015; Mallick et al., 2016), disease genetics (Bamshad et al., 2011), and the study of human mutations (Veltman and Brunner, 2012). Many variant callers have been developed to call genetic variants from NGS data (Sandmann et al., 2017), employing multiple steps of data processing and statistical inference (Li, 2011). Due to their wide impact, benchmarking variant calling pipelines is of utmost importance (Li, 2014; Shringarpure et al., 2015).

A widely-used method of measuring the accuracy of variant calling pipelines is tallying mis-called variants on samples that have been characterized previously (Olson et al., 2015; Hwang et al., 2015; Cornish and Guda, 2015). High-confidence variant sets, or “truth sets”, are compiled by teams of experts, such as the Genome in a Bottle (GIAB) Consortium, by integrating variant calls from several sequencing and data processing technologies. Due to careful curation, these data sets reach very high fidelity, and provide a benchmarking reference for SNPs, indels (Zook et al., 2014; Talwalkar et al., 2014; Boutros et al., 2014), and structural variants (Parikh et al., 2016; Fang et al., 2016; Nutsua et al., 2015), Human Genome Structural Variant Consortium (2017)). In all cases, high-coverage sequencing data from a handful of (often related) individuals are required to produce a truth set (Zook et al., 2016; Ebetole et al., 2017). Such truth sets are expensive to create, specific to the population, and biased toward the methods that produced them.

Sequencing data from related individuals provide an alternative way to detect errors made by a variant calling pipeline. Since de-novo mutation rates ($\sim 10^{-8}$ per locus (Veltman and Brunner, 2012)) are several orders of magnitude smaller than calling and genotyping error rates (often $10^{-2} - 10^{-5}$), one can use the variant calls inconsistent with the rules
of Mendelian inheritance to assess genotyping errors. While not all errors result in a Mendelian violation, which limits the sensitivity of this strategy, as noted by Li (2014), additional assumptions about the error probabilities permit the estimation of genotyping error rates from parent-offspring dyads (Haaland and Skag, 2013), siblings (Wang, 2004; Johnson and Haydon, 2007; Korostochlevsky et al., 2009), trios (Sobel et al., 2002; Douglas et al., 2002; Saunders et al., 2007; Jostins, 2011) and larger pedigrees (Eberle et al., 2017). If sequencing reads from many related individuals are available, recombination events can be inferred and used to estimate locus-specific genotyping errors (Markus et al., 2011; Chen et al., 2013; Peng et al., 2013; Kojima et al., 2013). All these approaches have the advantage of not relying on a truth set, thereby eliminating the bias towards established variant callers.

When modeling genotyping calling errors in related individuals, one is tempted to parameterize the genotyping error profile with a couple of parameters and estimate their values independently for different callers. This method has two significant shortcomings by neglecting the following correlations: First, errors made by different variant calling pipelines are often correlated due to identical assumptions (e.g. well-mapped reads) which, if violated in the input data, make mistakes of the pipelines coincide at particular loci. Second, mapping and calling errors depend on sequence context, which is similar among closely related individuals and induces correlations between genotyping errors in samples of different family members. To the extent of our knowledge, there exists no published method that addresses both correlations.

Here we present a statistical model that addresses both shortcomings for the case of a family trio. First, we aggregate the number of different genotype trios for the two pipelines jointly, and model this data. This way, we take correlations between two variant callers explicitly into account. Second, our model consists of a statistical mixture of different genotyping error profiles. Although each error profile assumes independence between individual family members, enabling an efficient mathematical solution, their mixture captures the correlations.

We validate our model in a series of experiments on real trios. First, we run two variant caller pipelines on publicly available NGS data from trios, and let our model estimate the precision and recall of the two pipelines. Then we compare these estimated performance metrics with their true values. The truth is directly calculated by comparing the calls on the children with their correct genotypes from the truth data sets. We demonstrate that our model is able to detect differential performance (absolute difference of precision or recall) between two pipelines as small as \(10^{-3}\). With this level of accuracy, our model enables informed development of pipelines targeting less-studied variants and populations for which no curated truth set is available.

The rest of the paper is divided into three parts: In section 2, we introduce our notation, describe the structure and assumptions of our model, and show how one can use it to infer the number of genotyping errors and stratify them by type. In section 3, we describe how we validate our method on publicly available trio data. Finally, in section 4, we present the results of the validation experiment.

## 2 Model

### 2.1 Definitions and notation

**Trio genotypes.** Let \(I = \{00, 01, 11\}\) denote the set of individual, bi-allelic, unphased genotypes: homozygous reference (00), heterozygous (01) and homozygous alternate (11). Let \(g = \{g_i \in I : p = 1, 2, 3\}\) be an ordered trio of individual genotypes, representing the genotypes of variants in the genomes of father (\(g_1\)), mother (\(g_2\)) and child (\(g_3\)). Let \(T\) denote the set of mathematically possible genotype trios, i.e. \(T = I^{3	imes}\), and let \(T^c\) denote its subset that conforms with Mendelian inheritance, i.e. \(t = (g \in T : (g_{11} \in g_1 \text{ and } g_{12} \in g_2) \text{ or } (g_{11} \in g_2 \text{ and } g_{12} \in g_1))\), where \(g_{11}\) and \(g_{12}\) are the first and second symbols (0 or 1) of \(g_1\). One can show that \(|T^c| = 3^3 = 27\) and \(|t| = 15\) by enumeration.

### Observed genotypes

Two variant calling pipelines run independently on samples from the three family members produce six separate sets of variant calls. We require that variant representations are harmonized, after which we can aggregate the counts of different genotype trio pairs of one class of variants (e.g. SNPs, short indels, structural variants) called by the two pipelines across all loci. This yields the joint counts of genotype trios \(N = \left\{N_{G_1G_2} : G_1, G_2 \in T\right\}\), where \(N_{G_1G_2}\) is the number of variants whose trio genotypes are called \(G_1\) by pipeline 1, and \(G_2\) by pipeline 2. An examples is shown in Figure 1.

**Correct genotypes.** Explicit information about the correct trio genotypes \(g \in t\) is missing from the observed counts \(N\). Let \(N' = \left\{N'_{G_1G_2} : G_1, G_2 \in T, g \in t\right\}\) denote the (hidden) complete distribution of called and correct genotype trios, where \(N'_{G_1G_2}\) is the number of variants with called genotype trios \(G_1, G_2\) and correct genotype trio \(g\), such that \(\sum N'_{G_1G_2,g} = N_{G_1G_2}, G_1, G_2\). In the next section, we explicitly model the genotyping errors in order to infer the hidden components in \(N'\) examples of which are shown on Supplementary Figures S.3, S.4, and S.S. Uncovering this data, \(N'\), from the observed mixture, \(N\) is our main goal.

### Benchmarking metrics

If obtained, the complete distribution \(N'\) contains all necessary information to directly evaluate any benchmarking metric. We will use the notation \(\mu(\cdot) : N' \rightarrow \mu\) to denote any benchmarking metric, derivable from \(N'\). One important metric is the genotype confusion matrix, \(\nu^{(p)} = \{\nu^{(p)}_{ij,k} \in N : i, j, k \in I\}, for
individual $p$, where
\[
n_{t,j,k}^{(p)} = \sum_{G_{t}^{2}, G_{j}^{2}, G_{k}^{2}} [i = g_{t}] [j = G_{j}^{2}] [k = G_{k}^{2}] N_{G_{t}^{2}, G_{j}^{2}, G_{k}^{2}, g_{t}}
\]
where $[x = y]$ is 1, if $x = y$ and 0 otherwise. $n_{t,j,k}^{(p)}$ counts how many variants of person $p$, with correct genotype trio $i$, are called $j \in I$ by pipeline 1, and $k \in I$ by pipeline 2.

Genotyping can be understood as a classification task with one negative (00) and two positive classes (01 and 11). For person $p$, we define the true positive, false negative, false positive and “mixed-up positive” counts for pipeline 1 as $TP^{(1)}_{p} = n_{p,0,1,0} + n_{p,1,1,1}$, $FN^{(1)}_{p} = n_{p,1,0,0} + n_{p,1,1,0}$, $FP^{(1)}_{p} = n_{p,0,0,1} + n_{p,0,1,1}$ and $MP^{(1)}_{p} = n_{p,1,0,1} + n_{p,1,1,0}$, respectively, where $\cdot$ denotes summation over the third index. For pipeline 2, similar formulas apply where the summations are prescribed for the second index. We define precision and recall with the usual formulas: prec = $TP/TP + FP + MP$, and recall = $TP/(TP + FN + MP)$, for each person and each pipeline.

### 2.2 Generative model

We introduce a generative model in a form of a mixture of categorical distributions over $(G_{t}^{2}, G_{j}^{2}) \in T \times T$ pairs, where each mixture component is labeled by a correct trio genotype $g_{t} \in t$, and an error category $m$ (defined in this section). The graphical representation of this model is shown in Fig. 2. We define the parameters $a, f, E$ below.

**Correct frequencies.** Let $F = \{f_{g} : g \in t\}$ denote the frequencies of each correct genotype trio $g$ among the variants under investigation, i.e.
\[
P(g) = f_{g}.
\]
where $F$ is subject to normalization, i.e. $\sum_{g} f_{g} = 1$. By treating all 15 components of $F$ as independent parameters, our model is able to fit aggregate data of variants not showing Hardy-Weinberg equilibrium (a limitation of the models used by Haaland and Skang (2013) and Douglas et al. (2002)). This way of parameterizing the correct frequencies makes our model more flexible than previous approaches that assume global allele frequencies (Haaland and Skang, 2013; Wang, 2004; Johnson and Haydon, 2007; Browning and Browning, 2013; Douglas et al., 2002; Saunders et al., 2007; Korostishevsky et al., 2009). Supplementary E contains more discussion.

### Error categories.** We assume that each variant is affected by one of the error categories $M = \{a, b, c, d, e\}$: (a) Both pipelines are correct, (b) Pipeline 1 is correct, (c) Pipeline 2 is correct, (d) Pipelines agree (but not necessarily correctly), and (e) Unrestricted joint distribution. Let $\theta = \{\theta_{g,m} : g \in t, m \in M\}$ denote the fractions of variants, with correct genotype trio $g$, that are affected by error category $m$, i.e.
\[
P(m | g) = \theta_{g,m}.
\]
where $\theta$ is subject to normalization, i.e. $\sum_{m} \theta_{g,m} = 1$, $\forall g$. The error category $m$ is assumed to be the same for all three genomes at a given locus. Now, we extend the definition of $\mathcal{N}$ with this additional index, $\mathcal{N}' = \{N_{G_{t}^{2}, G_{j}^{2}, G_{k}^{2}, m} : G_{t}^{2}, G_{j}^{2}, G_{k}^{2} \in T, g \in t, m \in M\}$, where $\sum_{m} N_{G_{t}^{2}, G_{j}^{2}, G_{k}^{2}, m} = N_{G_{t}^{2}, G_{j}^{2}, G_{k}^{2}}$. The idea of alternative error categories is used in Korostishevsky et al. (2009) and Heid et al. (2008), but none considered a mixture over error categories.

### Error rates.** Let $E = \{E_{i,j,k}^{(g,m)} : g \in t, i,j,k \in I\}$ denote the probabilities of calling the genotypes $(j,k)$, when the correct genotype is $i$ for a variant whose correct genotype trio is $g$ and is subject to error category $m$. Under a given $g$ and $m$, we assume that errors happen independently, and with the same rate, for all members of the trio. In practice, this requires that the samples are prepared, sequenced and analyzed in the same way, separately from each other. This allows us to write the probability of each type of genotyping event $g_{m}^{(g,m)} (G_{t}^{2}, G_{j}^{2})$ as the product
\[
P(G_{t}^{2}, G_{j}^{2} | g, m) = \prod_{p} P(G_{t}^{2}, G_{j}^{2} | g, m) = \prod_{p} E_{i,j,k}^{(g,m)}
\]
where $E$ is subject to the normalization $\sum_{p} E_{i,j,k}^{(g,m)} = 1$, $\forall g, m, i$. Different error categories $m \in M$ allow only certain components of $E$ to be non-zero:

- $E_{i,j,k}^{(g,m)} \neq 0$, only if
  - $m = a$ and $j = i$ and $k = i$ (both are correct)
  - $m = b$ and $j = i$ (pipeline 1 is correct)
  - $m = c$ and $k = i$ (pipeline 2 is correct)
  - $m = d$ and $j = k$ (pipelines agree)
  - $m = e$ (unrestricted)

Previsouly published models either assume an error rate independent of the correct genotype (Sobel et al., 2002; Saunders et al., 2007; Jostins, 2011; Browning and Browning, 2013) or model the $E$ matrix with a few parameters (Wang, 2004; Johnson and Haydon, 2007; Korostishevsky et al., 2009; Douglas et al., 2002; Heid et al., 2008). It is worth emphasizing that we only assume conditional independence of the errors in different family members (conditioned on fixed $g$ and $m$). The marginal probability $P(G_{t}^{2}, G_{j}^{2}) = \sum_{g,m} P(G_{t}^{2}, G_{j}^{2}, g, m)$ does not factorize under our assumptions, enabling our model to capture correlated errors across family members.

### 2.3 Limiting complexity

The model described so far has 480 algebraically independent parameters (See Supplementary A.1.1), which is comparable to $27 \times 27 = 729$, the number of entries in the input data. To improve the stability of our estimates, we lower the complexity of the model in two ways: limiting the range of error rates, and using parameter sharing.

**Limiting error rates.** We limit the error rates by requiring that $\max_{i,j,k} E_{i,j,k}^{(g,m)} = E_{i,j,k}^{(g,m)} \forall i,j,k$, $\forall g, m$. This guarantees that the largest fraction of the calls $(j,k)$ are indeed correct ($= i$), separately for each error category.

**Parameter sharing.** We partition $t$ into three disjoint subsets $t = t_{0} \cup t_{1} \cup t_{2}$, where $t_{0} = \{00,00,00\}$, $t_{2} = \{11,11,11\}$ and
$t_1 = t \setminus (t_0 \cup t_2)$, and require, whenever both $g, g'$ are in $t_1$, that $\theta_{g, m} = \theta_{g', m}$, and $E_{g_1, g_2}^{(j)}(m) = E_{g_1', g_2'}^{(j)}(m)$, $i, j, k$. We choose this partitioning expecting that variants that are present in none ($t_0$) or all ($t_2$) of the genomes will be subject to different error rates than variants for which at least one family member is heterozygous ($t_1$). These constraints allow us to introduce the notation $\theta_{g, m} := \theta_{g', m}$ and $E_{g_1, g_2}^{(j, k)} := E_{g_1', g_2'}^{(j, k)}$ for all $g \in t_1$, where $x = 0, 1, 2$, and reduce the number of algebraically independent parameters to 96. (See Supplementary A.2.)

2.4 Inference

Hidden Components. The product of Equations (2), (3) and (4) yields the joint distribution of observed ($G^1, G^2$) and hidden ($g, m$) variables,

$$P(G^1, G^2, g, m) = f_0 \theta_{g, m} \prod_{p} E_{g_1, g_2}^{(j)}(m),$$

(5)

Under the assumption that the attributes ($G^1, G^2, g, m$) of different variants are independent and identically distributed (given $f, \theta, E$), one can write the full and conditional distributions of the complete distribution $N$ (as (See Supplementary A.2))

$$P(N | f, \theta, E) = \text{Mult}(N | N_{\text{obs}}, P),$$

(6)

$$P(N | f, \theta, E) = \prod_{G^1, G^2} \text{Mult}(N_{G^1, G^2}, s_{G^1, G^2}, R_{G^1, G^2}),$$

(7)

where $N_{\text{obs}}$ is the total number of variants (both in $N$ and $N_i$), $N_{i, G^1, G^2}$ is a shorthand for $N_{G^1, G^2, g, m}$, $P = \{P_{G^1, G^2, g, m}, R_{G^1, G^2, g, m} : g \in t, m \in M\}$, $P = \{P_{G^1, G^2, g, m}, R_{G^1, G^2, g, m} : g \in t, m \in M\}$ where $R_{G^1, G^2, g, m} = P_{G^1, G^2, g, m} \sum_{i_2} \alpha_{i_1} \theta_{g, m}$ and $\text{Mult}(N_{G^1, G^2}) \propto \prod_i (\alpha_{i_{1234}} N_{G^1, G^2})$ is the multinomial distribution. The distribution in Equation (6) describes an unconstrained sampling process where the variants can be freely distributed between bins with different ($G^1, G^2, g, m$) labels. The distribution of Equation (7) represents a constrained sampling process where the number of variants in each observable bin ($G^1, G^2$) are fixed to the actually observed value $N_{G^1, G^2}$, and only their hidden attributes ($g, m$) are subject to random sampling.

Model parameters. Using Bayes theorem, the conditional posterior of $(f, \theta, E)$ can be written as a product

$$P(f, \theta, E | N) \propto P(N | f, \theta, E) P(f, \theta, E),$$

(8)

where the last term stands for the prior of the model parameters. Since $P(N | f, \theta, E)$ depends only on products of the model parameters (via Equation (5) and (6)), using a product of Dirichlet distributions as prior for $(f, \theta, E)$,

$$P(f, \theta, E) = P(f | \gamma(0))^0) \times P(\theta | \gamma(0))^0 \times P(E | \gamma(0))^0,$$

(9)

yields a posterior with identical structure (See Supplementary A.3),

$$P(f, \theta, E | N) = P(f | \alpha) P(\theta | \beta) P(E | \gamma),$$

(10)

where, in Equations (9) and (10),

$$P(f | \alpha) = \text{Dir}(f | \alpha)$$

(11)

$$P(\theta | \beta) = \prod_{\theta_{g, m}} \text{Dir}(\theta_{g, m} | \beta_{g, m})$$

(12)

$$P(E | \gamma) = \prod_{\gamma_{g_1, g_2}} \text{Dir}(E_{g_1, g_2}^{(j, k)} | \gamma_{g_1, g_2, i, j, k})$$

(13)

Here, Dir stands for the Dirichlet distribution, $\text{Dir}(x | \alpha) = \Gamma \left( \sum_{\alpha} \alpha \right) \prod_{x} x^{\alpha - 1} / \Gamma(\alpha)$, and its parameters are $\alpha$, $g \in t$, $\beta = \{\beta_{g, m} : s \in [0, 1, 2], m \in M\}$, $\gamma = \{\gamma_{g_1, g_2, i, j, k} : s \in [0, 1, 2], m \in M, i, j, k \in I\}$. In Supplementary A.3, we show that the posterior values of $\alpha, \beta, \gamma$ are

$$\alpha_g = \alpha^{(0)}_g + \sum_{G^1, G^2} N_{G^1, G^2, g, m},$$

(14)

$$\beta_{g, m} = \beta^{(0)}_{g, m} + \sum_{g \in t_2} N_{G^1, G^2, g, m},$$

(15)

$$\gamma_{g_1, g_2, i, j, k} = \gamma^{(0)}_{g_1, g_2, i, j, k} + \sum_{G} \left[ \gamma | t_2 \sum_{p} \delta_{p} \gamma[G_{p}^{1} = G_{p}^{1}, G_{p}^{2} = G_{p}^{2}, N_{G_{p}^{1}, G_{p}^{2}, g, m} \right.$$}

(16)

where $[\cdot]$ evaluates to 1 if the statement inside is true, and 0 otherwise.

Gibbs sampling. The conditional posterior of $N$ (see Equation (7)) and $f, \theta, E$ (see Equation (10)) are products of multinomial and Dirichlet distributions, respectively, both of which can be efficiently sampled. This enables Gibbs sampling, where we start with an initial guess $(f, \theta, E)^{(0)}$. In step 1, we sample the independent multinomials of Equation (7) and obtain the $n$th sample of the hidden components, $N^{(n)}$. Then, in step 2, we sample the independent Dirichlet variables of Equation (10) and obtain the next sample of the model parameters, $(f, \theta, E)^{(n+1)}$. We make sure that $\text{max}_{g, m}[E_{g_1, g_2}^{(j, k)}] = E_{g_1, g_2}^{(j, k)}$ by first sampling each slice $E_{g_1, g_2}^{(j, k)}$ from an unconstrained Dirichlet distribution, followed by lowering all components that violate this criteria to the level of $E_{g_1, g_2}^{(j, k)}$, and re-normalizing. Repeating step 1 and 2 yields approximate samples from the joint posterior $P(N, f, \theta, E | N)$. With appropriate “burn-in” ($\tau_0$) and “thinning” ($\Delta \tau$), this procedure provides estimates of the hidden components in the form of a list of samples ($X^{(n+\tau_0)}$, $X^{(n+\tau_0+\Delta \tau)}$,...). Details of choosing $\tau_0$ and $\Delta \tau$ are given in Supplementary B.

Benchmarking Metrics. From the samples of $\{X^{(n)}\}$, the $\text{est}$imate of any metric $\mu$ can be directly calculated, $\mu^{(\text{est})} := \mu(X^{(n)})$, enabling accurate estimation of its posterior.

3 Materials and Methods

To assess the accuracy of our model, we carried out the validation experiment shown on Figure 3. First, we selected trios for which not only validation experiment. We create two sets of trio variants called by a pipeline using GATK Haplotype Caller and another using GATK Unified Genotyper. Running our trio-based benchmarking model on this data produces estimates of various benchmarking metrics for both pipelines on all three data sets. At the same time, we use the correct genotypes for the child’s truth set to calculate the true values of the same benchmarking metrics. Grey-shading marks the components that would be present even in a regular usage of our model, when the truth set is not available.
the NGS reads but also curated truth-sets of the children are published, and used two variant caller pipelines to produce variant calls from the read data. Second, we simultaneously used our model to estimate benchmarking metrics for both pipelines on all three members of the trio, and used the correct genotypes from the child’s truth set to calculate the true values of the same benchmarking metrics. Finally, we compared the estimated metrics with the true metrics to calculate the accuracy of our model.

3.1 Data

Read data. We used 50x, 148-bp-long, paired-end, Illumina reads from three trios: the GIAB Ashkenazim trio (HG003, HG004, HG002), published by the Genome in a Bottle Consortium (Zook et al., 2016), and 101-bp-long, paired-end, Illumina reads from the two unrelated trios from Platinum Genome2 (NA12889, NA12890, NA12877) and (NA12891, NA12892, NA12878), published by Illumina, Inc. (Eberle et al., 2017). From these alignment files, we extracted and sorted the reads using samtools5, and dropped duplicates and unpaired reads, using a custom script. Command lines are listed in Supplementary D.

Truth set for child. Curated, high-confidence variants are publicly available for all three children, HG002, NA12877 and NA12878. We used bcftools6 to select bi-allelic SNPs on chromosomes 1–22 (~4 million). Merging and comparing with the variants produced by the variant caller pipelines allows direct, “truth set-based”, benchmarking of the pipelines. (See Supplementary D for command lines.) We restricted the scope of our validation experiment to SNPs to avoid having to reconcile different representations of identical variants, which often arise for indels. Supplementary C.4 and C.5 contain validation results where we further excluded SNPs near indels, and where we included only SNPs outside of high-confidence regions, respectively.

Pipelines. We used two variant caller pipelines, both of which use BWA-MEM 0.7.13 (Li and Durbin, 2009) for alignment, GATK tools (McKenna et al., 2010) for local realignment and base quality recalibration, and take advantage of known SNPs and indels (HapMap, dbsNP, 1000GP, Omni) according to GATK Best Practices recommendations (DePristo et al., 2011, Van der Auwera et al., 2013). In the variant calling step, one pipeline used GATK HaplotypeCaller v3.5, and the other GATK UnifiedGenotyper v2.7. Command lines are listed in Supplementary D. (Additionally, Supplementary C.6 shows results of comparing pipelines with different aligners.)

Aggregating genotype calls. For each trio, we merged the SNPs called by the two pipelines using bcftools, and filtered out the variants that fell outside the high-confidence regions or become multi-allelic. We aggregated the remaining SNPs (~4 million) by counting the occurrences of each (G1, G2) combinations, where the trio genotype G1 is called by Haplotype Caller, and G2 by Unified Genotyper. During aggregation, we recorded missing variants as genotype 00 and disregarded phasing information, i.e. regarded both heterozygous genotypes, 01 and 10, as 01. This yielded the observed counts of joint trio calls \( N = \{N_{G1, G2}\} \).

Subsampling. To test the robustness of our model, we repeated the validation experiments for subsets of SNPs of different sizes. We subsampled a pre-defined number of variants from \( N \) by choosing a random subset of SNPs with uniform probability.

3.2 Running the estimator

Imputing “uncalled variants”. Neither Haplotype Caller nor Unified Genotyper reports homozygous reference genotypes (i.e. 00) when run on a single sample. Due to this limitation, the number of potential variants that do not get called 01 or 11 by either pipeline in any family member, i.e. \( N_{0(00,00,00)} \), is technically zero. With this “bulk” of the 00 genotypes missing, we would not find the correct 00 \( \rightarrow 01 \) and 00 \( \rightarrow 11 \) error rates. To fix this, we used an estimate of the total number of polymorphic loci: the total number of variants reported in dbSNP7 (\( 5.2 \times 10^7 \)).

Details of Gibbs sampling. For each set of input data \( N \), we perform Gibbs sampling as described in Section 2.4. We set the parameters of the Dirichlet priors to \( \alpha = 1 \forall g \), \( \beta_{g, i, j} = 1 \forall g, i, j \) and \( \gamma_{g, i, j, k} = 1000 \), if \( i = j = k \) and 1 otherwise \( \forall g, i, j \). The skewed choice for \( \gamma_{g, i, j, k} \) represents our expectation that the error rates are small \( \ll 1 \times 10^{-2} \). We draw the initial values of \( f, \theta, E \) from their prior, and let the Gibbs sampler run for \( T_0 = 50,000 \) iterations, allowing it to reach equilibrium. We proceeded by running 100,000 iterations, while recording samples with \( \Delta T = 1000 \) period, to decrease the correlation between consecutive samples. This procedure resulted in 100 samples, \( \{\mathbb{N}_{\text{true}}\} \), drawn from its posterior \( P_N \mid \mathbb{N} \). With our scipy-based Python implementation, this procedure takes 4.5 hours, using 1 CPU on the AWS instance m1.small.

3.3 Metrics

For each trio, we run the trio-based benchmarking, which produces samples \( \{\mathbb{N}_{\text{true}}\} \). From this, we directly calculate the samples of the confusion matrix for the child \( \{n_{\text{(child)}}(i,j)\} \) (from Equation (1)). Using the formulas for precision and recall, we obtain their samples \( \{\text{prec}(\text{child})\} \) and \( \{\text{rec}(\text{child})\} \) for both pipelines. We estimate each metric with the average of its samples.

To investigate how sensitive our model is to the total number of variants, we repeatedly sub-sample the observed counts \( N \), and run the trio-based benchmarking all over again. We do this for 8 different number of variants \( (10^3 - 1 \times 10^5) \), repeating each sub-sampling and estimation 10 times for every trio.

4 Results

The simplest method of analyzing trio results from the two pipelines is counting Mendelian violations. This, alongside with the true genotyping errors in the child, is shown in Table 1. Comparing the two columns shows that the difference in the number of Mendelian violations correctly indicate the difference in the actual number of genotyping error between the two pipelines.

Table 1. Number of genotyping errors in the children’s genomes, calculated from comparisons with the truth sets, and the number of Mendelian violations in their respective trios. Counts are shown for Haplotype Caller and Unified Genotyper.

<table>
<thead>
<tr>
<th></th>
<th>Genotyping errors in child</th>
<th>Mendelian violations in trio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HC</td>
<td>UG</td>
</tr>
<tr>
<td>HG002</td>
<td>7,496</td>
<td>20,142</td>
</tr>
<tr>
<td>NA12877</td>
<td>36,759</td>
<td>39,780</td>
</tr>
<tr>
<td>NA12878</td>
<td>31,436</td>
<td>37,224</td>
</tr>
</tbody>
</table>

With *GIAB* (HG002), and * Platinum Genomes truth VCF and BED files. 

To see how accurately our trio-based benchmarking method can compare the two pipelines, we compare estimates of precision and recall with their true values for the children in Table 2. Although the estimated values of the metrics significantly differ from the truth, both the signs and magnitudes of differential metrics are accurately estimated. This is also visible from the smaller estimation uncertainty $\sigma_{est}$ reported by the model.

Table 2. Precision and recall of the two pipeline (HC: Haplotype Caller, UG: Unified Genotyper) on the children’s samples calculated by truth set-based benchmarking, and estimated with trio-based benchmarking, using all 4 million SNPs. $\Delta$ denotes the difference between HC and UG values. $\sigma_{est}$ is self-reported uncertainty of the model, i.e. standard deviation of the Gibbs samples.

<table>
<thead>
<tr>
<th></th>
<th>HC trio</th>
<th>HC truth</th>
<th>HC trio HC truth $\Delta$ ($10^{-3}$)</th>
<th>UG trio</th>
<th>UG truth</th>
<th>UG trio UG truth $\Delta$ ($10^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG002</td>
<td>0.9980</td>
<td>0.9942</td>
<td>3.81</td>
<td>0.9994</td>
<td>0.9988</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{est}$ $\Delta$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
</tr>
<tr>
<td>trio</td>
<td>0.0004</td>
<td>0.0014</td>
<td>0.0020</td>
<td>0.0014</td>
<td>0.0040</td>
<td>0.0012</td>
</tr>
<tr>
<td>NA12877</td>
<td>0.9986</td>
<td>0.9963</td>
<td>2.23</td>
<td>0.9902</td>
<td>0.9912</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{est}$ $\Delta$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
</tr>
<tr>
<td>trio</td>
<td>0.0004</td>
<td>0.0014</td>
<td>0.0020</td>
<td>0.0014</td>
<td>0.0040</td>
<td>0.0012</td>
</tr>
<tr>
<td>NA12878</td>
<td>0.9986</td>
<td>0.9965</td>
<td>2.15</td>
<td>0.9910</td>
<td>0.9920</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{est}$ $\Delta$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
<td>$\sigma_{est}$ ($10^{-3}$)</td>
</tr>
<tr>
<td>trio</td>
<td>0.0004</td>
<td>0.0014</td>
<td>0.0020</td>
<td>0.0014</td>
<td>0.0040</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

According to GIAB (HG002), and Platinum Genome trio VCF and BED files.

In Figure 4, we plot the deviation of the trio-based estimate from the true value as a function of number of variants. While the estimation error of the metrics themselves hovers around 0.1–0.01 even when large number of variants are considered, the model’s error on the differential metrics is around $10^{-3}$4 for large number of variants, and stay below $5 \times 10^{-5}$ even for a mere 1000 variants. Recall values (both absolute and differential) are estimated more accurately than precision.

To understand the reason why our model is much better at estimating differential metrics than the metrics themselves, we plot the entries of the genotype confusion matrix (vi) for NA12878 in Figure 5. Our model only approximately estimates the number of loci where either one of the pipelines makes an incorrect call, but not both. Since these entries of the confusion matrix contribute the most to differential performance, it is no surprise that the difference in precision and recall are accurately estimated, even if their actual values are not. We can intuitively understand this disparity in the following way: We expect the two pipelines to have similar performance. Only the 27 diagonal elements ($G_i^1 = G_i^2$) of $N$ convey information about this common performance. On the other hand, all off-diagonal elements ($G^1_i 
eq G^2_i$) tell us something valuable about the differential performance. Since there are more of the latter, they paint a more detailed picture, and allow us to estimate the differences more accurately.

5 Discussion

Aggregating observed Mendelian violations across many loci in related individuals provides an approximate notion of the accuracy of variant calling pipelines. When comparing two pipelines, we expect the better one to produce fewer Mendelian inconsistencies, and indeed, as shown in Table 1, this is the case: Haplotype Caller produces fewer Mendelian violations in all three trio, and makes fewer genotyping errors for all three children.

In the light of our results in Table 2, two caveats of this naive method are immediately visible: First, the numbers of Mendelian violations do not accurately reflect how much better is the better pipeline. Second, the bar counts of Mendelian violations do not even hint that Unified Genotyper actually has a higher recall on the samples NA12877 and NA12878, something our trio-based benchmarking method estimates correctly (See Table 2). This exemplifies that estimating precise quantitative differences between variant caller pipelines, in the absence of a reliable truth set, is a non-trivial task. As we demonstrated, our method can do this with high accuracy.

Previous trio-based benchmarking approaches range from fitting a calibration curve between observed number of Mendelian violations and true number of errors (Hao et al., 2004), to modeling haplotypes and recombination events during meiosis (Markus et al., 2011; Kojima et al., 2013). All previous approaches focus on a set of variants produced by a single pipeline, disregarding all information about the correlation between pipelines. We showed that by analyzing the joint

---

**Fig. 4.** Estimation errors, Top: Difference between estimated (from trio) and true values (from truth set) of precision and recall. Bottom: Difference of the estimated and true differential precision and recall. Markers show the results obtained from the validation experiments run on sub-sampled variants. Dashed lines highlight the 90th percentile of each group.

**Fig. 5.** Genotype confusion matrix of NA12878. True values are shown alongside with the estimates obtained from trio-based benchmarking. Error bars on the estimates indicate 1st and 99th percentiles of the samples from the Gibbs sampler. (See Supplementary Figure S-6 for HG002 and NA12877.)
counts ($N$) of the called genotype trios from two pipelines one can estimate differential performance with high accuracy (with uncertainty of $10^{-2}$). In fact, exactly because of the information contained in the correlations, the differential metrics can be estimated with more than an order of magnitude lower uncertainty than the metrics themselves. As a consequence, our uncertainty is an order of magnitude lower than it was previously achieved with dyads (Hailand and Skang, 2013), and comparable to the uncertainties obtained from pure simulations for trios (Hao et al., 2004).

Although our current method is limited to using bi-allelic variants from trios, the mathematical framework enables straightforward generalization. Data from multi-allelic variants and larger pedigrees can be incorporated by expanding the sets $I$ and $T$. The demand of computational resources, which increases exponentially with the sizes of $I$ and $T$, can be curbed by truncating the model to limit the number of genotyping errors per locus, which enables implementation with sparse arrays.

One limitation of our method is the need for harmonized variant representation in the input. While this problem is solved for trios by Topaž et al. (2018), in there is no easily-deployable method to faithfully match six samples with different variant representation. Another limitation, stemming from the limitation of the input data, is that some of the Mendelian-consistent trios are in fact results of double (or triple) errors. Our method relies on explicit assumptions about error profiles, its result is accurate only if these assumptions hold.

Our method estimates aggregate metrics. The estimated complete distribution $N$ and the person-specific confusion matrices $n^{(i)}$ contain a wealth of additional information that can be directly used to improve the performance of variant calling. Bayesian model averaging (See e.g. Fragoso and Neto (2015)) can be used to reconcile genotypes at loci where the other is wrong for each called trio combination $(G^1, G^2)$, one can estimate the posterior probability of the correct genotype, $P(G | G^1, G^2)$. E.g. from the confusion matrix $n$ shown in Figure 5, we can calculate the probability of $G$ being correct when $G^1$ and $U$ call 01 and 00, respectively: $P(01 | 01, 00) = \frac{\text{cell}_{01,00}}{\sum_{g=0,1} \text{cell}_{g,00} + \sum_{g=0,1} \text{cell}_{g,01}} \approx 0.72$.

One can imagine extending this scheme to a council of pipelines where, instead of giving equal say to each pipeline, a weighted voting is carried out to determine the genotype of each variant.

Acknowledgements

Discussions with Yilong Li, Lizaoo Li, Maxime Hveret, Amit Jain, Milos Popovic, Aqeel Ahmed, Maria Suciu, Sun-Gou Ji, Ivan Johnson, Kausik Ghose, Eker Gündoğdu were very valuable at multiple stages of this work. We are grateful to them for their insights and suggestions.

Funding

This work was supported in part by UK Department of Health grant SBRI Genomics Competition: Enabling Technologies for Genomic Sequence Data Analysis and Interpretation administered by Genomics England.

References


