Supplementary Materials for "*geck*: trio-based comparative benchmarking of variant calls"

Péter Kómár and Deniz Kural

Contents

A Derivations

A.1 Number of independent model parameters

We count the number of algebraically independent model parameters for both with and without parameter sharing.

A.1.1 Without parameter sharing

Our model employs three sets of parameters:

- $f = \{f_g \ : \ g \in t\}$ - the frequencies of each of the 15 correct genotype trios g,
- $\theta = \{\theta_{g,m} \ : \ g \in t, \ m \in M\}$ - the fractions of variants, with a certain correct genotype trio g, that is subject to, m, one of the 5 error categories,
- $E = \{E^{(g,m)}_{i,j,k} \ : \ g \in t, \ m \in M, \ i,j,k \in I\}$ - the probabilities that correct genotype i of a variant (that is subject to an error category m and has a certain correct trio genotype g) is called one of the 3×3 possible joint genotypes (j,k) by two callers,

where the sizes of the sets are, $|t| = 15$, $|M| = 5$, $|I| = 3$.

Since there is only one normalization constraint for f, namely $\sum_g f_g = 1$, the total number of algebraically independent parameters of f is $d_f = |t| - 1 = 15 - 1 = 14$.

Without parameter sharing, only the normalization constraints $\sum_m \theta_{g,m} = 1, \quad \forall g \in t$ limit the independent values of θ. They amount to $d_\theta = |t| \cdot (|M| - 1) = 15 \times (5 - 1) = 60$.

Similarly, we could evaluate the number of independent parameters of E: Number of entries is $|t| \cdot |M| \cdot |I|^3$, while the number of normalization constraints, $\sum_{j,k} E^{(g,m)_{i,j,k}} \forall g \in t, \ m \in M, \ i \in I$ is $|t| \cdot |M| \cdot |I|$, which would amount to $|t| \cdot |M| \cdot (|I|^3 - |I|)$. This calculation, however, does not take into account that some of entries of E has no effect. E.g. the value of $E^{((00,00,00),e)}_{11,00,00}$ has no consequence on the result, because it corresponds to an impossible event where the correct trio genotype is $(00, 00, 00)$, but one particular correct (individual) genotype is 11.

First, let's consider the possible g, i combinations. In Table 1, we summarize how many different values i can actually take for a given choice of g. The total is 29, instead of $|t| \cdot |I|$.

Table 1: Number of unique individual genotypes i in each trio genotype trio g.

correct genotype trio g	possible i values	number
(00,00,00)	00	1
(00,01,00)	00, 01	2
(00,01,01)	00, 01	2
(00,11,01)	00, 01, 11	3
(01,00,00)	00, 01	2
(01,00,01)	00, 01	2
(01,01,00)	00, 01	2
(01,01,01)	01	1
(01,01,11)	01, 11	2
(01,11,01)	01, 11	2
(01,11,11)	01, 11	2
(11,00,01)	00, 01, 11	3
(11,01,01)	01, 11	2
(11,01,11)	01, 11	2
(11,11,11)	11	1
		29

At the same time, for every g, i combination, the error category $m \in M = \{a, b, c, d, e\}$ requires different values of E to be set to zero:

- $m = a$ makes $E^{g,a}_{i,j,k} = 0$, wherever $j \neq i$ and $k \neq i$, which amounts to an additional $(|I|^2 - 1) = 8$ constraints.
- $m = b$ makes $E^{g,b}_{i,j,k} = 0$, wherever $j \neq i$, which amounts to and additional $(|I| - 1) \cdot |I| = 6$ constraints.
- $m = c$ makes $E^{g,c}_{i,j,k} = 0$, wherever $k \neq i$, which amounts to and additional $|I| \cdot (|I| - 1) = 6$ constraints.
- $m = d$ makes $E^{g,d}_{i,j,k} = 0$, wherever $j \neq k$, which amounts to and additional $\cdot(|I| - 1) \cdot |I| = 6$ constraints.
- $m = e$ makes no additional constraints.

The normalization requirements $\sum_{j,k} E^{(g,m)}_{i,j,k} = 1 \quad \forall g, m, i$ amount to one additional constraint for each group. This means that the total possible m, j, k combinations (for every g and i) is $|M| \cdot |I|^2 - (8 + 1) - (6 + 1) - (6 + 1) - (6 + 1) - 1 = 14$. The total number of independent (and relevant) values of E is therefore $d_E = 29 \times 14 = 406$.

Adding up the number of independent values for f, θ and E yields $d_f + d_\theta + d_E = 14 + 60 + 406 = \boxed{480}$.

A.1.2 With parameter sharing

With parameter sharing, the parameters of our model can be written as:

- $f = \{f_g \ : \ g \in t\}$ - the frequencies of each of the 15 correct genotype trios g,
- $\theta = \{\theta_{s,m} \ : \ s \in S, \ m \in M\}$ - the fractions of variants, with a true trio genotype in a certain subset t_s, that is subject to one of the 5 error categories m,
- $E = \{E^{(s,m)}_{i,j,k} \ : \ s \in S, \ m \in M, \ i, j, k \in I\}$ - the probabilities that the i correct genotype of a variant (that is subject to an error category m and has a certain correct trio genotype in the subset t_s) is called one of the 3×3 possible joint genotypes (j, k) by two callers,

where $S = \{0, 1, 2\}$.

Since the structure of f did not change, we can use the previous result, $d_f = |t| - 1 = 15 - 1 = 14$.

With parameter sharing, θ is indexed by $s \in S$ (instead of $g \in t$). Considering the normalization constraints $\sum_m \theta_{s,m} = 1 \ \forall s \in S$, we can express the number of independent parameters as $d_\theta = |S| \cdot (|M| - 1) = 3 \times (5 - 1) = 12$.

First, let's consider the number of possible (s, i) combinations for E. Table 2 shows the number of possible i values for each s (describing the subsets t_s of t). There are a total of 5 different s, i combinations such that $E^{(s,m)}_{i,j,k}$ actually affects the model.

Table 2: Number of unique individual genotypes i in each subset t_s.

subset index s	subset t_s	possible i values	number
0	$\{(00, 00, 00)\}$	00	1
1	{trios with at least one 01}	00, 01, 11	3
2	$\{(11, 11, 11)\}$	11	1
			5

The constraints associated with different m error categories does not change due to parameter sharing. This means that the total number of independent m, j, k combinations is still $|M| \cdot |I|^2 - (8 + 1) - (6 + 1) - (6 + 1) - (6 + 1) - 1 = 14$. Therefore the total number of independent E values is $d_E = 5 \times 14 = 70$.

Adding up the number of independent values yields $d_f + d_\theta + d_E = 14 + 12 + 70 = \boxed{96}$.

A.2 Formula for $P(\mathcal{N} \mid N, f, \theta, E)$

In our paper, we show that the generative probability of a variant having correct genotype trio g, error category m, and called genotype trios G^1 and G^2 can be written as

$$P(G^1, G^2, g, m \mid f, \theta, E) = f_g \theta_{g,m} \prod_p E^{(g,m)}_{g_p, G^1_p, G^2_p} =: P_{G^1,G^2,g,m}, \tag{S.1}$$

where $p \in \{1, 2, 3\}$ denotes the index of a family member in the trio. Assuming that the variants are independently generated by this model, their complete distribution $\mathcal{N} = \{\mathcal{N}_{G^1,G^2,g,m} \ : \ G^1, G^2 \in T, \ g \in t, \ m \in M\}$ is multinomially distributed according to the probabilities $P = \{P_{G^1,G^2,g,m} \ : \ G^1, G^2 \in T, \ g \in t, \ m \in M\}$:

$$P(\mathcal{N} \mid f, \theta, E) = \text{Mult}(\mathcal{N} \mid \mathcal{N}_{\text{tot}}, P), \tag{S.2}$$

where $\mathcal{N}_{\text{tot}}$ is the total number of variants, and Mult is the multinomial distribution, i.e. $\text{Mult}(n \mid n_{\text{tot}}, p) = n_{\text{tot}}! \prod_\xi (p_\xi)^{n_\xi} / n_\xi!$, where ξ runs through all possible combinations of indices of n and p, which are assumed to have the same shape.

From Equation S.1, we can use the definition of conditional probability to express

$$P(g, m \mid G^1, G^2, f, \theta, E) = \frac{P(G^1, G^2, g, m \mid f, \theta, E)}{P(G^1, G^2 \mid f, \theta, E)} = \frac{P_{G^1,G^2,g,m}}{\sum_{g',m'} P_{G^1,G^2,g,m}} =: R_{G^1,G^2,g,m}. \tag{S.3}$$

Again, since we assume that the variants are independently generated by our model, the slice of the distribution $\mathcal{N}_{G^1,G^2,:,:} = \{\mathcal{N}_{G^1,G^2,g,m} \ : \ g \in t, \ m \in M\}$ is multinomially distributed according to the probabilities $R_{G^1,G^2,:,:} = \{R_{G^1,G^2,g,m} \ : \ g \in t, \ m \in M\}$:

$$P(\mathcal{N}_{G^1,G^2,:,:} \mid N, f, \theta, E) = \text{Mult}(\mathcal{N}_{G^1,G^2,:,:} \mid N_{G^1,G^2}, R_{G^1,G^2,:,:}) \quad \forall G^1, G^2, \tag{S.4}$$

where N_{G^1,G^2} is the number of variants whose genotypes are called G^1 by pipeline 1 and G^2 by pipeline 2. From this result, we can write the distribution of the full distribution $\mathcal{N}$ as a product of the distributions of its slices:

$$P(\mathcal{N} \mid N, f, \theta, E) = \boxed{\prod_{G^1,G^2} \text{Mult}(\mathcal{N}_{G^1,G^2,:,:} \mid N_{G^1,G^2}, R_{G^1,G^2,:,:})}, \tag{S.5}$$

where R is a function of P (via Equation S.3), which is a function of f, θ and E (via Equation S.1).

A.3 Formula for $P(f, \theta, E \mid \mathcal{N})$

Equation S.2 expresses the probability of generating a complete (observed + hidden) distribution $\mathcal{N}$, provided that the model parameters f, θ, E are known. Now, we use Bayes theorem to express the posterior probability of f, θ, E when $\mathcal{N}$ is known.

$$P(f, \theta, E \mid \mathcal{N}) \propto P(\mathcal{N} \mid f, \theta, E)\ P(f, \theta, E), \tag{S.6}$$

where $P(f, \theta, E)$ is the *a priori* distribution of the model parameters.

With parameter sharing, the expression of the generative probability of a variant (from Equation S.1), can be written as

$$\begin{aligned} P(G^1, G^2, g, m \mid f, \theta, E) &= f_g \theta_{g,m} \prod_p E^{(g,m)}_{i,j,k} && \text{(S.7)} \\ &= f_g \times \left[\prod_s (\theta_{s,m})^{[g \in t_s]}\right] \times \left[\prod_s \prod_{i,j,k} \prod_p (E^{(s,m)}_{i,j,k})^{[g \in t_s][i=g_p][j=G^1_p][k=G^2_p]}\right] && \text{(S.8)} \\ &= f_g \times \left[\prod_s (\theta_{s,m})^{Q_{s,g}}\right] \times \left[\prod_s \prod_{i,j,k} (E^{(s,m)}_{i,j,k})^{Q_{s,g} K_{i,j,k,g,G^1,G^2}}\right], && \text{(S.9)} \end{aligned}$$

where $[g \in t_s]$ evaluates to 1, if g is in the t_s subset of t (and 0 otherwise), and $[x = y]$ evaluates to 1 if $x = y$ (and 0 otherwise). Above, we defined the following two arrays:

$$\begin{aligned} Q_{s,g} &:= [g \in t_s] \quad \in \{0, 1\}, && \text{(S.10)} \\ K_{i,j,k,g,G^1,G^2} &:= \sum_{p=1,2,3} [i = g_p] \times [j = G^1_p] \times [k = G^2_p] \quad \in \{0, 1, 2, 3\}. && \text{(S.11)} \end{aligned}$$

The above expression of $P(G^1, G^2, g, m \mid f, \theta, E)$ can be directly substituted in Equation S.2 to give the generative distribution

$$\begin{aligned} P(\mathcal{N} \mid f, \theta, E) = \mathcal{N}_{\text{tot}}! \prod_{G^1,G^2,g,m} \frac{1}{\mathcal{N}_{G^1,G^2,g,m}!} \Bigg\{ & (f_g)^{\mathcal{N}_{G^1,G^2,g,m}} \times \\ & \left[\prod_s (\theta_{s,m})^{Q_{s,g} \mathcal{N}_{G^1,G^2,g,m}}\right] \times \\ & \left[\prod_s \prod_{i,j,k} (E^{(s,m)}_{i,j,k})^{Q_{s,g} K_{i,j,k,g,G^1,G^2} \mathcal{N}_{G^1,G^2,g,m}}\right] . \Bigg\} \end{aligned} \tag{S.12}$$

Now, we show that by assuming a prior which is a product of Dirichlet distributions results in a posterior of the same form. This *conjugate prior* has the following form:

$$\begin{aligned} P(f, \theta, E) &= P(f)\ P(\theta)\ P(E), && \text{(S.13)} \\ &\text{where} \\ P(f) &= \text{Dir}(f \mid \alpha^{(0)}) \propto \prod_g (f_g)^{(\alpha^{(0)}_g - 1)}, && \text{(S.14)} \\ P(\theta) &= \prod_s \text{Dir}(\theta_{s,:} \mid \beta^{(0)}_{s,:}) \propto \prod_s \prod_m (\theta_{s,m})^{(\beta^{(0)}_{s,m} - 1)}, && \text{(S.15)} \\ P(E) &= \prod_{s,m,i} \text{Dir}(E^{(s,m)}_{i,:,:} \mid \gamma^{(0)}_{s,m,i,:,:}) \propto \prod_{s,m,i} \prod_{j,k} (E^{(s,m)}_{i,j,k})^{(\gamma^{(0)}_{s,m,i,j,k} - 1)}, && \text{(S.16)} \end{aligned}$$

where $\theta_{s,:} = \{\theta_{s,m} : m \in M\}$, $E^{(s,m)}_{i,:,:} = \{E^{(s,m)}_{i,j,k} : j, k \in I\}$, and $\alpha^{(0)} = \{\alpha^{(0)}_g \in \mathbb{R}^+ : g \in t\}$, $\beta^{(0)} = \{\beta^{(0)}_{s,m} \in \mathbb{R}^+ : s \in \{0, 1, 2\},\ m \in M\}$, $\gamma^{(0)} = \{\gamma^{(0)}_{s,m,i,j,k} \in \mathbb{R}^+ : s \in \{0, 1, 2\},\ m \in M,\ i, j, k \in I\}$. Dir

stands for the Dirichlet distribution, $\mathrm{Dir}(x \mid \alpha) = \frac{1}{Z(\alpha)} \prod_\xi (x_\xi)^{(\alpha_\xi - 1)}$, where $Z(\alpha) = \left[\prod_\xi \Gamma(\alpha_\xi)\right] / \Gamma\left(\sum_\xi \alpha_\xi\right)$ and ξ is assumed to run through all allowed combinations of the indices of x and α (which are assumed to have the same shape). We note that the products over s, m, i and j, k in Equation S.16 run through the allowed set of values, which we describe in Section A.1.2.

Multiplying the likelihood (from Equation S.12) with the prior (from Equation S.13) according to Equation S.6 yields the posterior

$$P(f, \theta, E \mid \mathcal{N}) \propto (f_g)^{(\alpha_g - 1)} \times \left[\prod_s \prod_m (\theta_{s,m})^{(\beta_{s,m} - 1)}\right] \times \left[\prod_{s,m,i} \prod_{j,k} (E^{(s,m)}_{i,j,k})^{(\gamma_{s,m,i,j,k} - 1)}\right], \tag{S.17}$$

where the new exponents are

$$\alpha_g = \alpha^{(0)}_g + \sum_{G^1, G^2, m} \mathcal{N}_{G^1, G^2, g, m}, \tag{S.18}$$

$$\beta_{s,m} = \beta^{(0)}_{s,m} + \sum_{G^1, G^2, g} Q_{s,g}\, \mathcal{N}_{G^1, G^2, g, m}, \tag{S.19}$$

$$\gamma_{s,m,i,j,k} = \gamma^{(0)}_{s,m,i,j,k} + \sum_{G^1, G^2, g} Q_{s,g}\, K_{i,j,k,g,G^1,G^2}\, \mathcal{N}_{G^1, G^2, g, m}, \tag{S.20}$$

where Q and K are defined in Equation S.10 and S.11. Considering the normalization constraints (which are also enforced by the Dirichlet priors), we note that, since the posterior is a product of the model parameters, it can be written as a product of Dirichlet distributions.

$$P(f, \theta, E \mid \mathcal{N}) = P(f|\alpha)\, P(\theta|\beta)\, P(E|\gamma) \tag{S.21}$$

$$= \boxed{\mathrm{Dir}(f \mid \alpha) \times \left[\prod_s \mathrm{Dir}(\theta_{s,:} \mid \beta_{s,:})\right] \times \left[\prod_{s,m,i} \mathrm{Dir}(E^{(s,m)}_{i,:,:} \mid \gamma_{s,m,i,:,:})\right]}, \tag{S.22}$$

where α, β and γ are functions of $\mathcal{N}$, given by Equation S.18, S.19 and S.20.

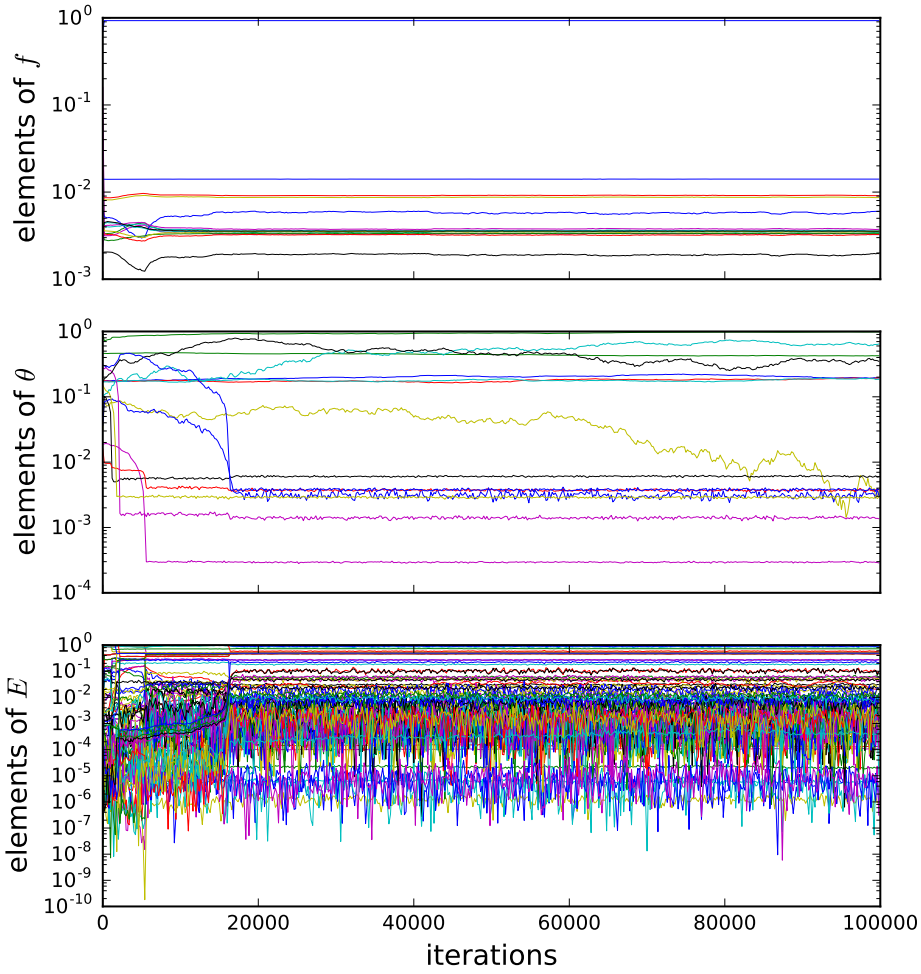
B Gibbs sampling

Two operational parameters of the Gibbs sampler, burn-in (τ_0) and thinning ($\Delta\tau$), have to be set so that the generated samples approximate the joint posterior $P(\mathcal{N}, f, \theta, E \mid N)$ well.

B.1 Burn-in

Figure S.1 shows the components of model parameters (f, θ, E) as a function of iterations. After $\sim$ 20,000 iterations almost all components of the parameters stabilize. To be on the safe side, we set the burn-in iterations to $\tau_0 = 50{,}000$.

Figure S.1: Model parameters drawn from the Gibbs sampler as a function of iterations, run for the Platinum-78 trio. Traces for all components of f, θ and E are shown.



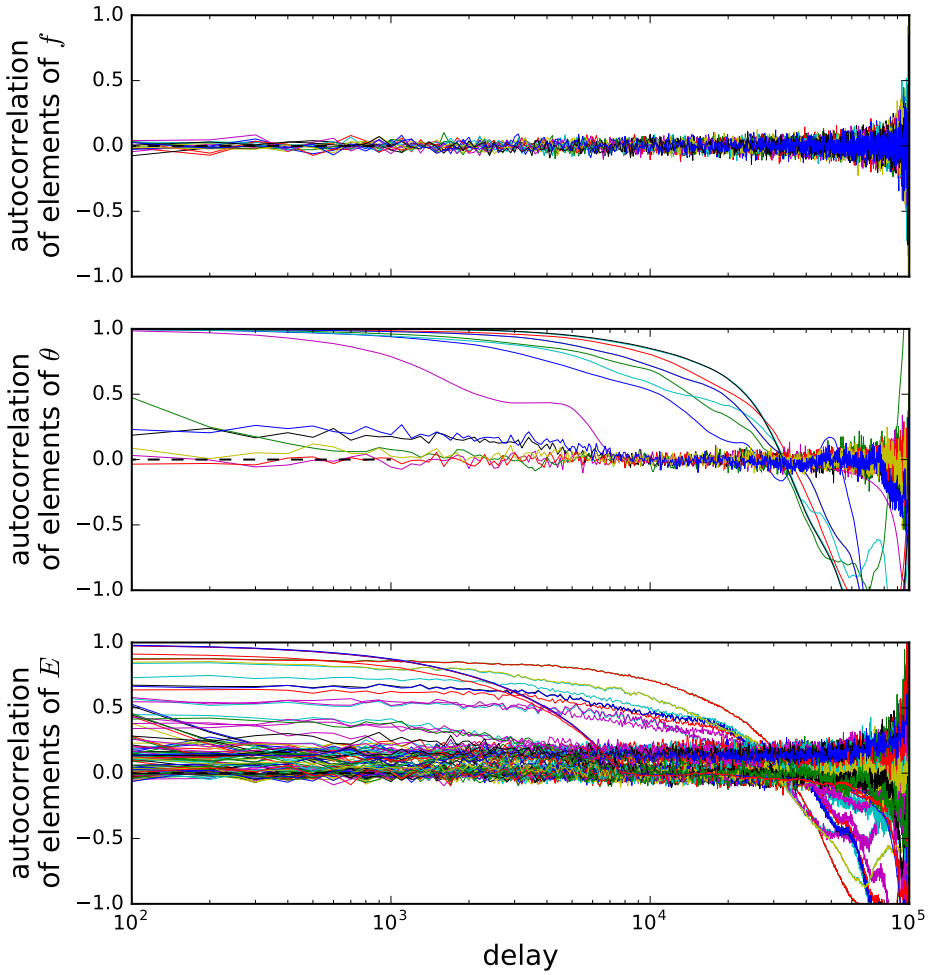
B.2 Thinning

To determine the optimal thinning, we calculate the autocorrelation function $A(t)$ of the series $\{X_\tau : \tau = 0, 1, 2, \ldots T\}$ of model parameter X (e.g. $X = \theta_{1,e}$) as a function of the delay t,

$$A_X(t) := \mathrm{Corr}(X_\tau, X_{\tau+t}) = \left[\frac{1}{T-t}\sum_{\tau=0}^{T-t-1}(X_\tau - \bar{X})(X_{\tau+t} - \bar{X})\right] \Big/ \left[\frac{1}{T}\sum_{\tau=0}^{T-1}(X_\tau - \bar{X})^2\right], \tag{S.23}$$

where $\bar{X} = \frac{1}{T}\sum_\tau X_\tau$. Figure S.2 shows this autocorrelation function for all model parameters as a function of delay. We chose a thinning of $\Delta\tau = 10^3$, as a good balance between low correlation between consecutive samples and high number of samples.

Figure S.2: Autocorrelation function $A_X(t)$ of model parameters $X = f, \theta, E$ as a function of the delay t.



C Additional validation results

Here, we present additional data from the results of our validation experiments on three trios (father, mother, child):

- GIAB-AJ: (HG003, HG004, HG002)
- Platinum-77: (NA12889, NA12890, NA12877)
- Platinum-78: (NA12891, NA12892, NA12878)

C.1 Observed and complete data

The joined counts of observed genotype trios N, and the our model's estimate of the complete (observed + hidden) distribution $\mathcal{N}$ are shown on the Figures S.3, S.4 and S.5.

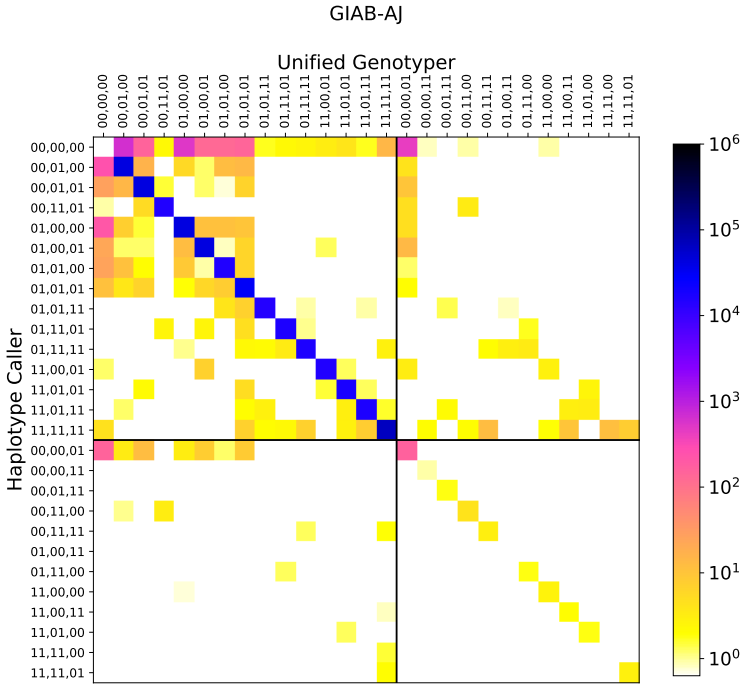
C.2 Confusion matrix of child's genotype

Since the correct genotypes are known for the child (from their truth set), we can compare the estimated genotype confusion matrix $\{n^{(\text{child})}_{i,j,k} \ : \ i,j,k \in I\}$ with its correct value. We plotted $\{n^{(\text{child})}_{i,j,k}\}$ as bars for the three trios on Figure S.6.

C.3 Model's uncertainty of estimating precision and recall

We compare the samples drawn from the posterior of precision and recall (produced by the Gibbs sampler) with the true values of precision and recall (calculated by comparing the calls with the correct calls from the truth set) for the children of each trio. Figure S.7 shows the true values of precision and recall with red crosses, and the samples from the model's Gibbs sampler with blue dots.

Figure S.3: Observed joined genotype trio counts $N = \{N_{G^1,G^2}\}$ (top), and estimated contributions of each correct genotype trio $\mathcal{N} = \{\mathcal{N}_{G^1,G^2,g}\}$ (bottom panels) for GIAB-AJ trio. The correct genotype trios g are printed on top of each panel. The matrix on top is the sum of the 15 matrices on the bottom.



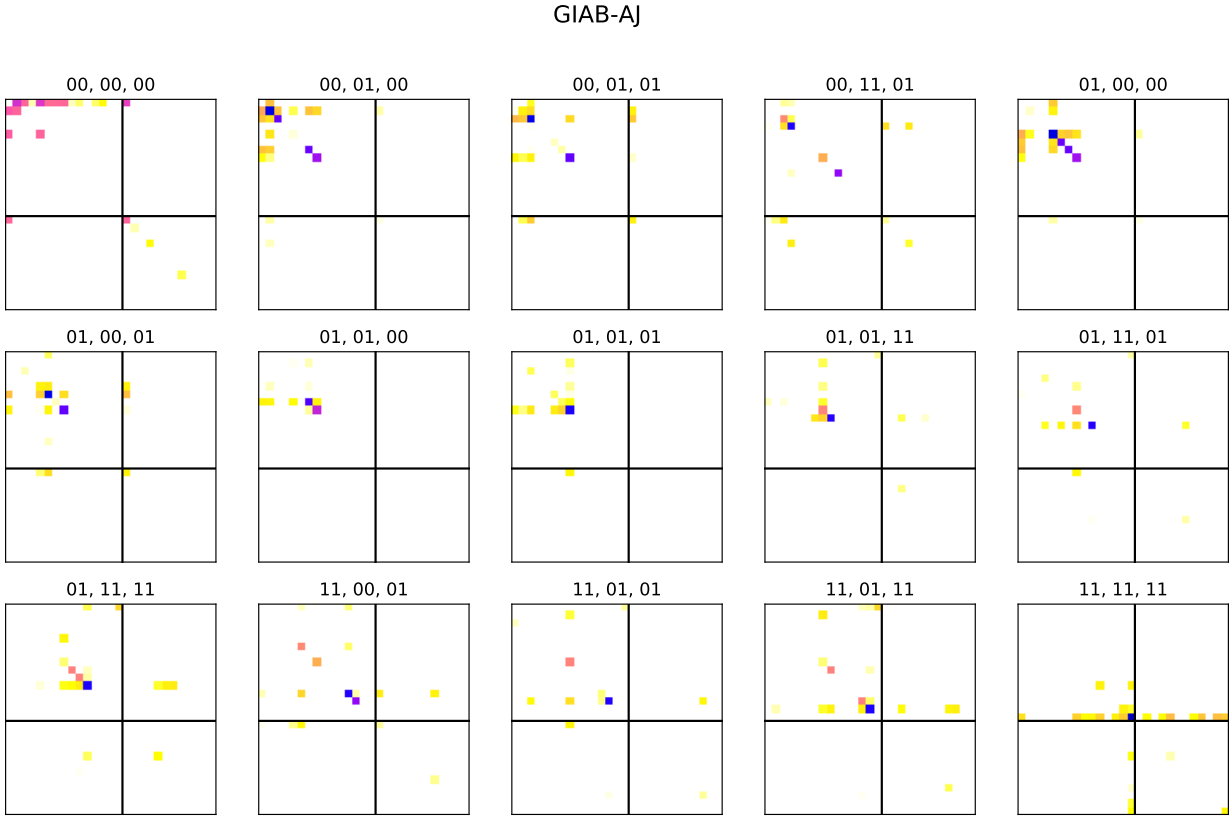
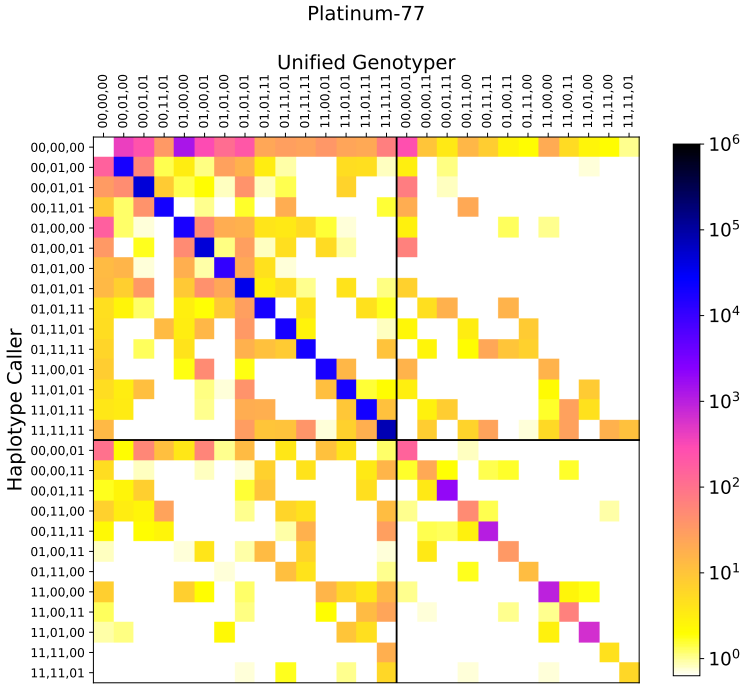


Figure S.4: Observed joined genotype trio counts $N = \{N_{G^1,G^2}\}$ (top), and estimated contributions of each correct genotype trio $\mathcal{N} = \{\mathcal{N}_{G^1,G^2,g}\}$ (bottom panels) for Platinum-77 trio. The correct genotype trios g are printed on top of each panel. The matrix on top is the sum of the 15 matrices on the bottom.



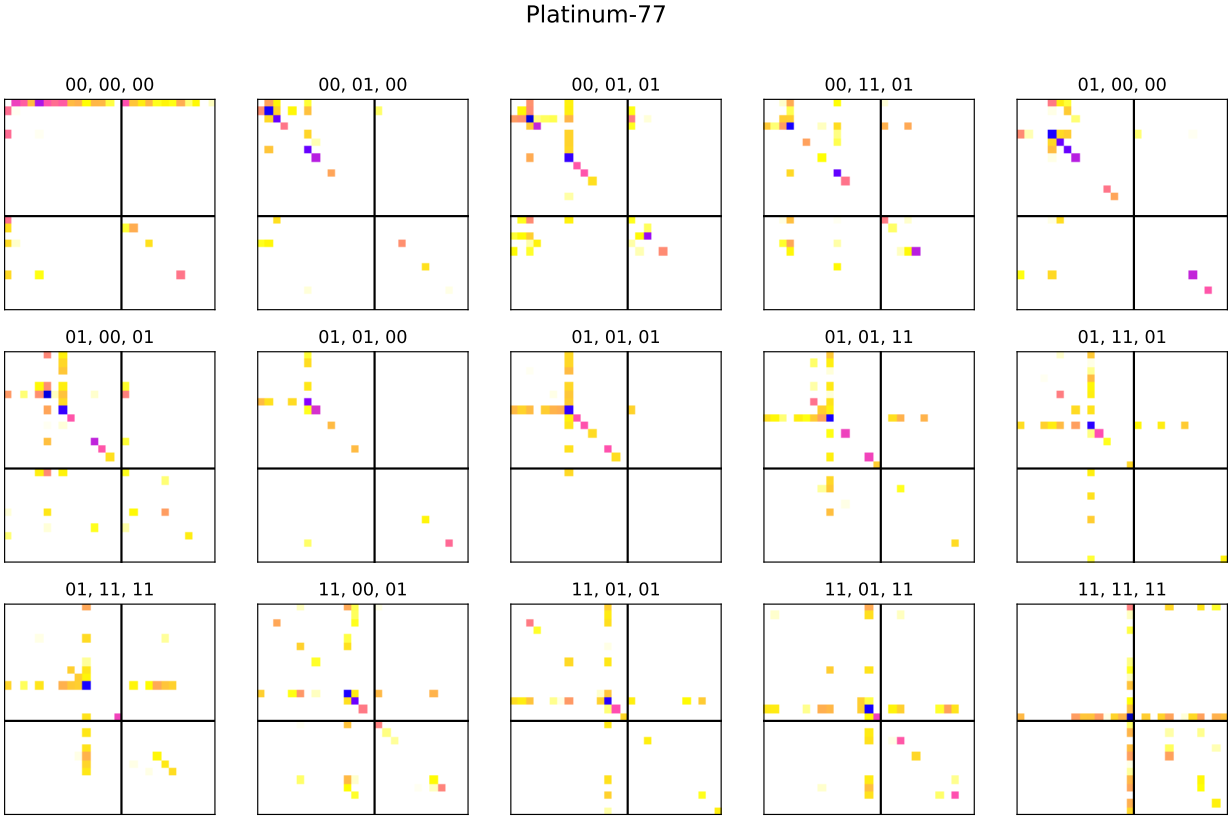
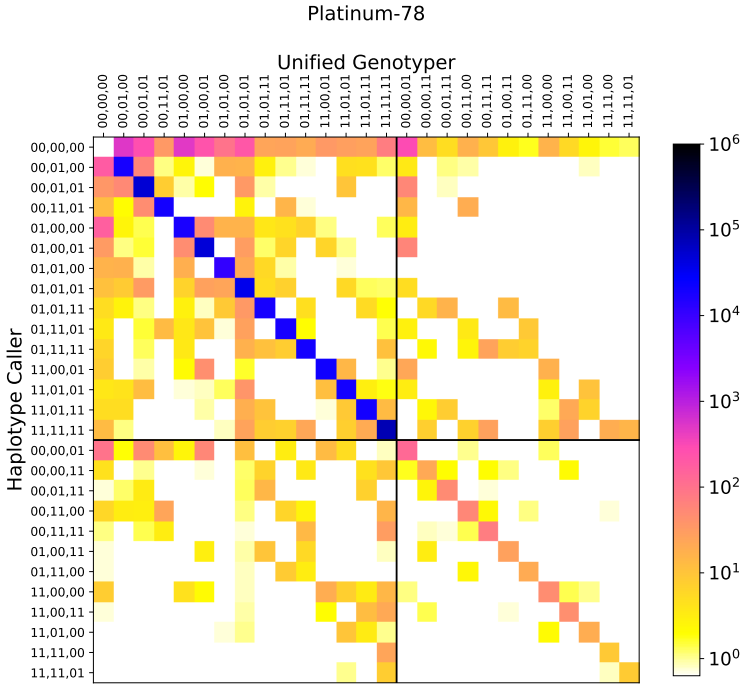


Figure S.5: Observed joined genotype trio counts $N = \{N_{G^1,G^2}\}$ (top), and estimated contributions of each correct genotype trio $\mathcal{N} = \{\mathcal{N}_{G^1,G^2,g}\}$ (bottom panels) for Platinum-78 trio. The correct genotype trios g are printed on top of each panel. The matrix on top is the sum of the 15 matrices on the bottom.



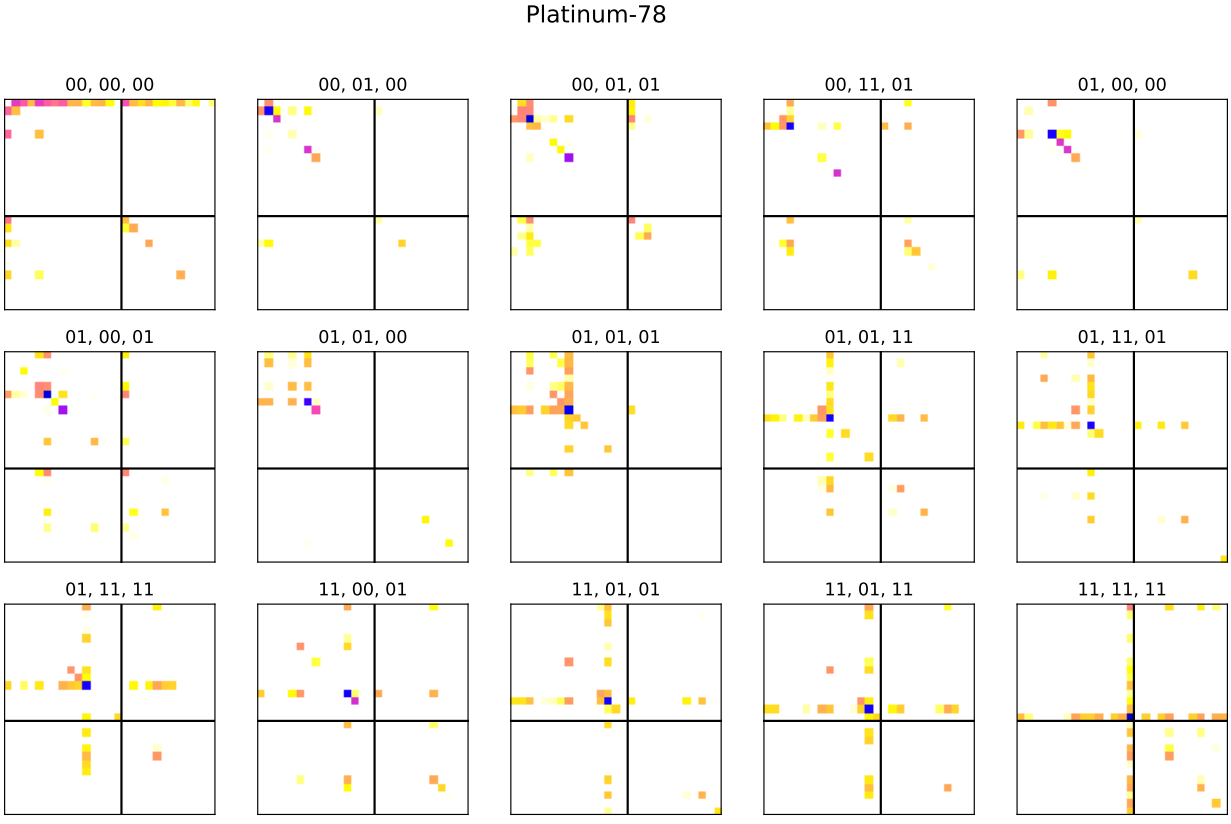


Figure S.6: Genotype confusion matrix entries for the children of all three trios. Colored bars correspond to the true counts calculated by comparing the called genotypes of each variant with the correct genotype (from the truth set): Green bars show cases where both callers made the correct calls, yellow bars show cases where one of them made a mistake, and red bars where both made mistakes. Light blue bars show the estimates of the same counts from our model. The green and yellow bars are estimated with higher accuracy than the red ones, which underlines the fact that our model can estimate the performance difference between pipelines more accurately than the actual values of the performance metrics.

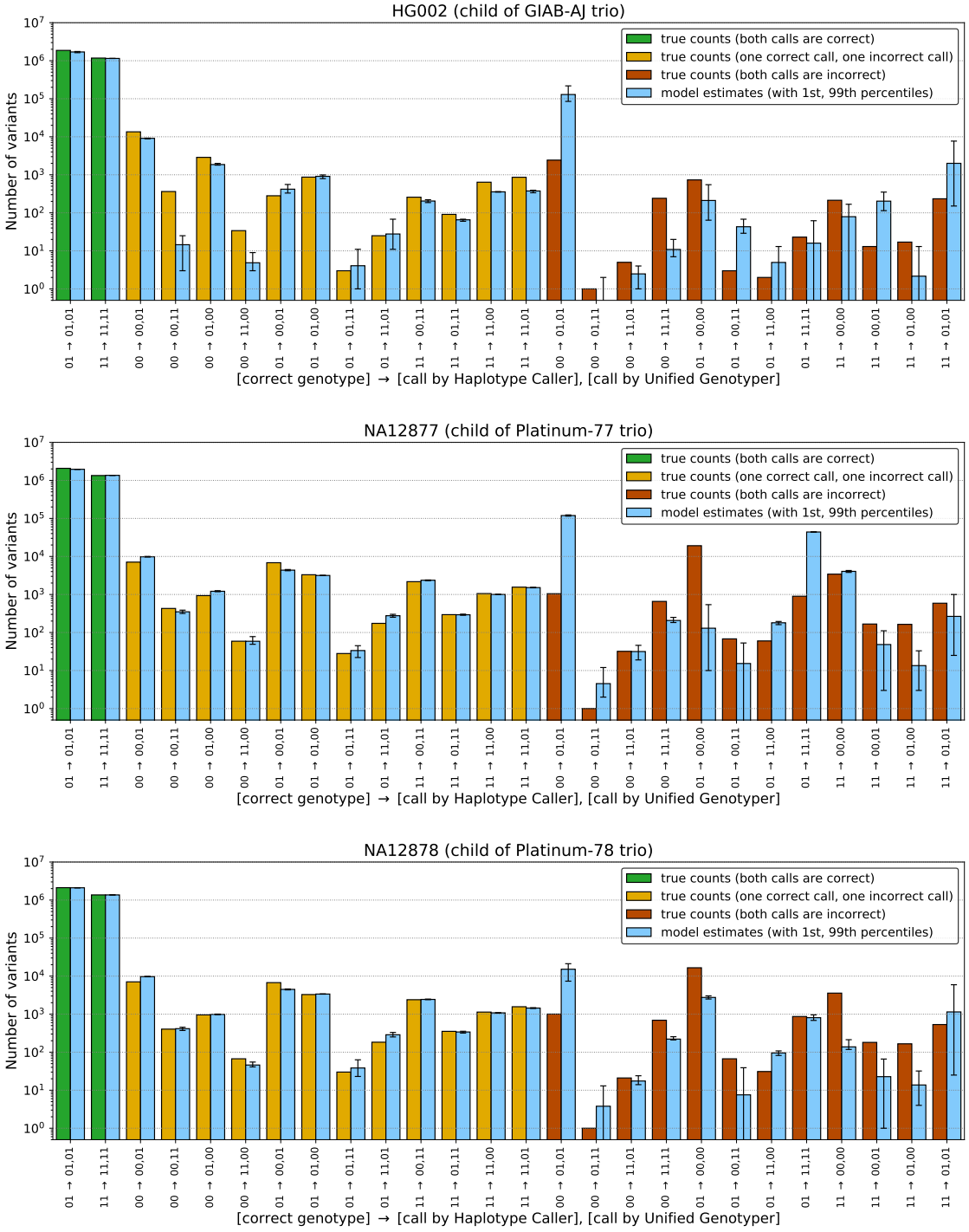
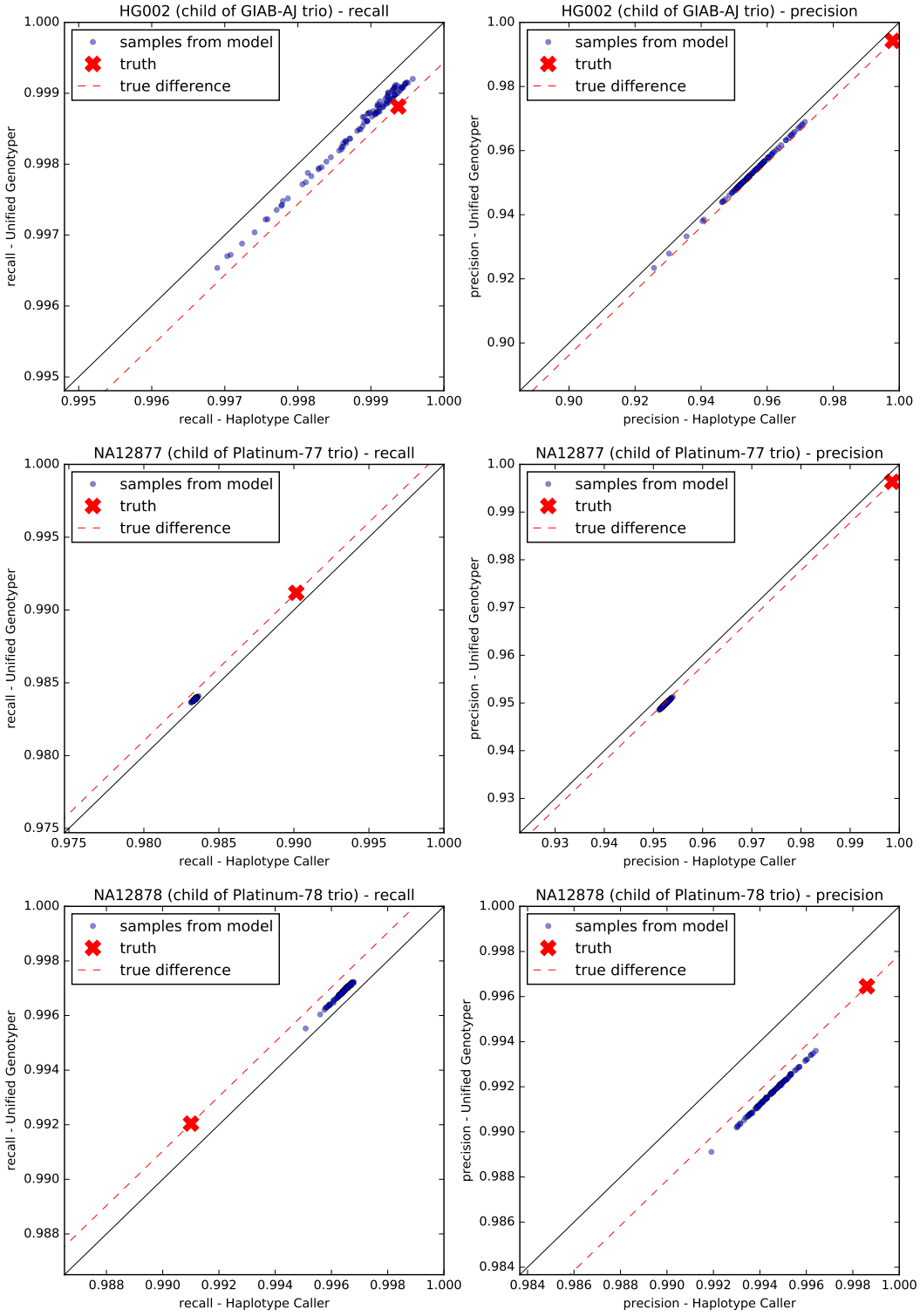


Figure S.7: Joint estimates of precision and recall for the two pipelines (Haplotype Caller, Unified Genotyper). Red cross marks the true values of precision and recall (calculated from the truth set), red dashed line indicates the points where the performance difference between the two pipelines are equal to the true value, and blue dots are samples drawn from the model's estimate about the posteriors. While the model usually makes a significant mistake in estimating the values of precision and recall for the two pipelines, it estimates the differential performance much more accurately.

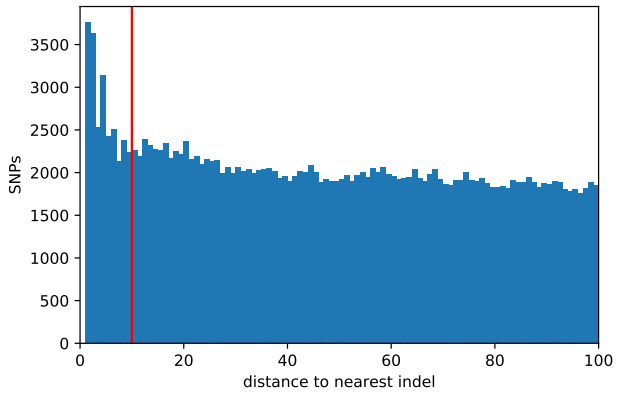


C.4 More careful filtering of SNPs

We restricted our validation experiments to SNPs because we wanted to avoid harmonizing variant representation between members of the trio. Indels are often represented differently by different tools or even by the same tool if the context (nearby variants) is different. The latter problem can also affect SNPs which are close to indels.

To get a sense of how big of an effect such SNPs are contributing to our validation experiments, we first calculated the empirical distribution of distances of SNPs to the nearest indel. We did this for truth set of the child of the GIAB-AJ trio, HG002. The result is shown in Figure S.8. The resulting distribution is fairly flat for small values, except for distances shorter than 5-10 bps. We took this as an indication that some of these SNPs may be artefacts due to discordant variant representation. To be on the safe side we decided to exclude all SNPs that fall within 10 bps from an indel. This amounted to 0.87% (27,001) of all SNPs in the truth set (containing a total of 3,097,996 SNPs).

Figure S.8: Histogram of distances of SNPs to nearest indel. The vertical red line indicates our choice of a threshold for dropping SNPs that are within 10 bps to the nearest indel.



Using the 10-bp-radius exclusion windows around true indels from HG002's truth set, we ran geck on SNPs from the resulting "restricted" high-confidence regions. The results, compared to the results obtained for SNPs in non-restricted high-confidence regions are shown in Table 3, where we also show the true values of precision and recall for the two pipelines which we obtained by benchmarking the SNPs in the same "restricted"high-confidence regions. Comparing the runs on all high-confidence SNPs and the run on the restricted set shows no striking differences. While the Δ of precision is estimated more accurately in the latter case, absolute precision is estimated with bigger error. Other estimated values and their errors are not changed significantly after excluding SNPs near indels.

Table 3: Precision and recall of the two pipeline (HC: Haplotype Caller, UG: Unified Genotyper) on HG002 for all SNPs in the high-confidence regions ("HG002") and for those SNPs that are more than 10 bp away from any indel ("HG002 - restricted"). We compare the true values (which we obtained by merging the true variants with calls using bcftools), and the estimated value obtained with trio-based benchmarking. Δ denotes the difference between HC and UG values. σ_{trio} is self-reported uncertainty of the model, i.e. standard deviation of the Gibbs samples.

		precision			recall		
		HC	UG	$\Delta\,(10^{-3})$	HC	UG	$\Delta\,(10^{-3})$
HG002	truth	0.9980	0.9942	3.81	0.9994	0.9988	0.56
	trio	0.9554	0.9529	2.50	0.9988	0.9985	0.33
	σ_{trio}	$\pm$ 0.0076	$\pm$ 0.0076	$\pm$ 0.05	$\pm$ 0.0006	$\pm$ 0.0006	$\pm$ 0.05
HG002 - restricted	truth	0.9986	0.9966	1.97	0.9997	0.9995	0.22
	trio	0.9212	0.9192	2.00	0.9981	0.9977	0.41
	σ_{trio}	$\pm$ 0.0090	$\pm$ 0.0090	$\pm$ 0.02	$\pm$ 0.0013	$\pm$ 0.0013	$\pm$ 0.01

C.5 Outside of high-confidence region

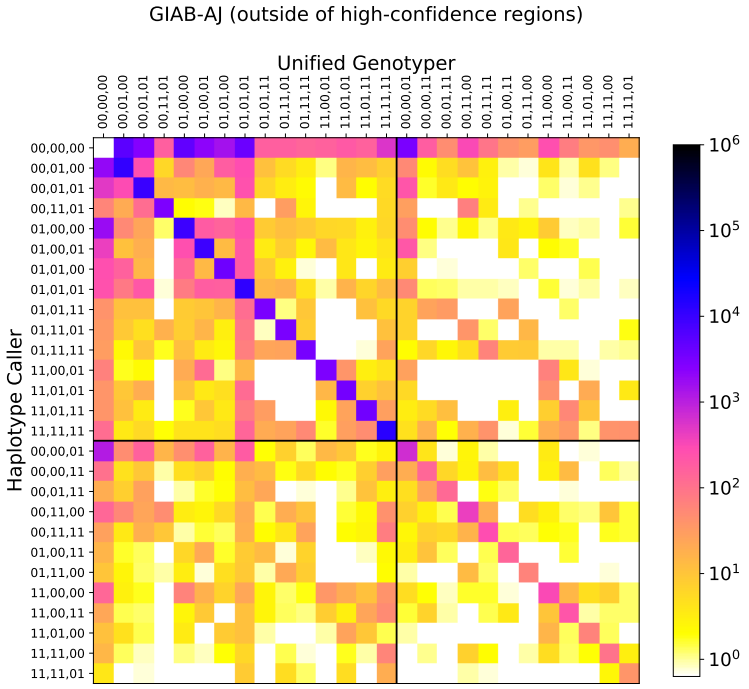
Although we can only accurately validate our model on regions that are reported to be high-confidence by the truth sets, we can run geck on variants from outside of these regions, and check if the estimated benchmarking metrics make sense qualitatively. To do this, we created the complement of the GIAB-AJ high-confidence BED file, and re-run the preprocessing and geck. We found 1,381,544 variants that are called by either tools (HC or UG) in at least one family member (father, mother or child) outside of the high-confidence regions. Their observed genotype trio pair distribution is shown in the upper panel of Figure S.9.

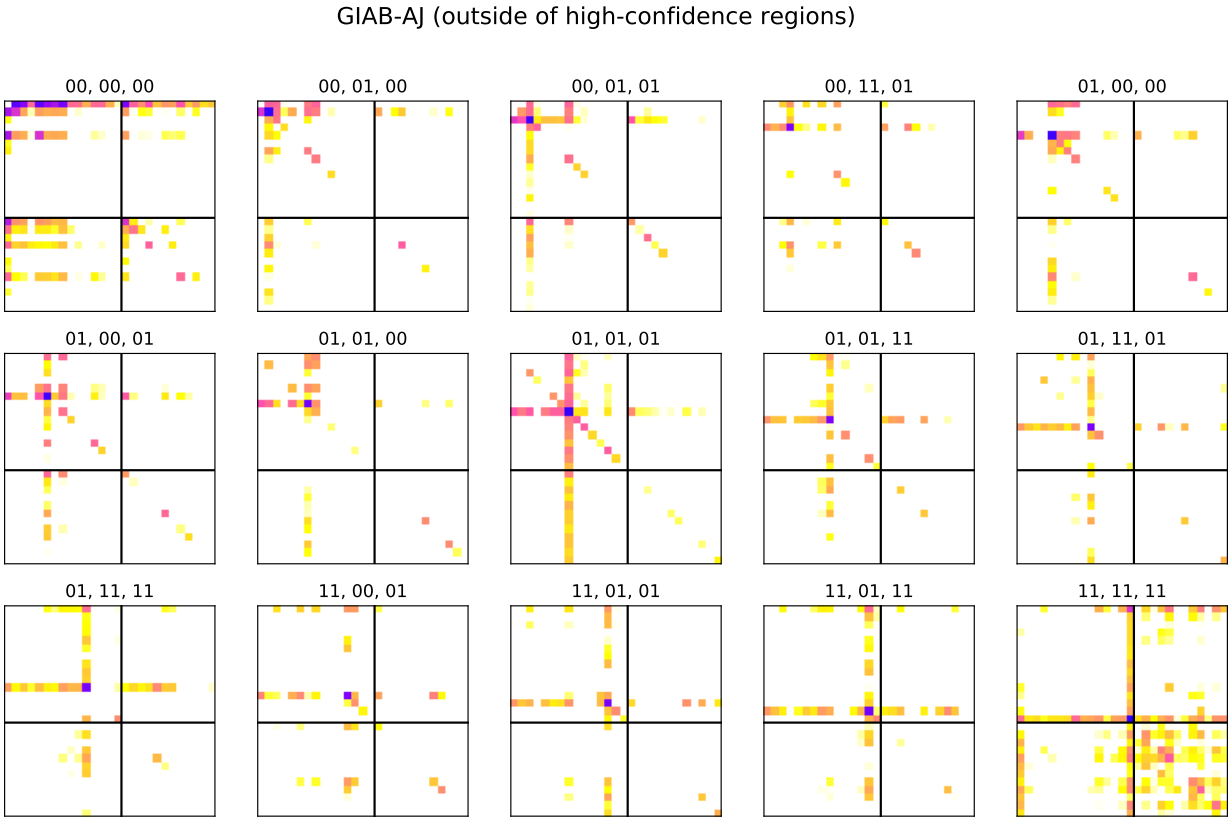
The lower panel of the same figure shows the estimated $\mathcal{N}_{G^1,G^2,g}$ counts, produced by our method. This result is qualitatively similar to the estimates for the high-confidence regions (shown on the lower panels of Figure S.3, S.4 and S.5). The estimated precision and recall and delta between the variant callers are shown in Table 4. In agreement with our expectations, Haplotype Caller outperforms Unified Genotyper in precision, and performs on the same level in terms of recall within margin of posterior uncertainty.

Table 4: Precision and recall of the two pipeline (HC: Haplotype Caller, UG: Unified Genotyper) on HG002 for all SNPs in **outside** of high-confidence regions. σ_{trio} is self-reported uncertainty of the model, i.e. standard deviation of the Gibbs samples.

		precision			recall		
		HC	UG	$\Delta\,(10^{-3})$	HC	UG	$\Delta\,(10^{-3})$
HG002 ("low-confidence" regions)	trio	0.9504	0.8273	123	0.9356	0.9374	-1.84
	σ_{trio}	$\pm$ 0.0192	$\pm$ 0.0061	$\pm$ 13.1	$\pm$ 0.0152	$\pm$ 0.0033	$\pm$ 12.3

Figure S.9: Observed joined genotype trio counts $N = \{N_{G^1,G^2}\}$ (top), and estimated contributions of each correct genotype trio $\mathcal{N} = \{\mathcal{N}_{G^1,G^2,g}\}$ (bottom panels) for variants **outside** of the high-confidence regions in the GIAB-AJ trio. The correct genotype trios g are printed on top of each panel. The matrix on top is the sum of the 15 matrices on the bottom.





C.6 Comparing different aligners

Additional to the above validation experiments, we also run a validation experiment with datasets where different aligners (BWA and Novoalign) were used. This data is available at the ftp site of GIAB, under

- HG003 (father): ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG003_NA24149_father/NISTv3.3.2/GRCh38/supplementaryFiles/inputvcfsandbeds/
- HG004 (mother): ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG004_NA24143_mother/NISTv3.3.2/GRCh38/supplementaryFiles/inputvcfsandbeds/
- HG002 (son): ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/NISTv3.3.2/GRCh38/supplementaryFiles/inputvcfsandbeds/

We selected the VCF files produced by using BWA (with suffix "bwa_mem_IlmnMatePair_sentieonHC") and Novoalign (with suffix "novoalign_Ilmn150bp300X_sentieonHC") in the alignment step, and compared them using geck, using SNPs in the intersection of the three high-confidence BED files of the main release. The input aggregated trio genotype pair counts are shown in the upper panel of Figure S.10. The difference between the two pipelines is bigger then when we were comparing Haplotype Caller and Unified Genotyper: The Novoalign pipeline called significantly less inconsistent trios than the BWA pipeline. Running our estimation method produced $\mathcal{N}_G^1, G^2, g$, shown on the bottom panel of Figure S.10.

To validate the estimated benchmarking metrics, we also performed truth-based benchmarking on the child's sample, using the truth set at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/NISTv3.3.2/GRCh38/. First, we calculated the number of different $g \to G^1, G^2$ mis-genotyping events in the child's sample using the truth set, and compared this to the values estimated by geck. This is shown in Figure S.11. The same trend of Figure S.6 can be seen here: Mis-genotyping events where only one pipeline makes a mistake are accurately estimated (the only exception here is $01 \to 01, 00$, i.e. when Novoalign misses a heterozygous variant) within the margins of uncertainty reported by geck.

Then, we calculated the true precision and recall of the two pipelines and plotted it against the estimates of their joint performance produced by geck. This is shown in Figure S.12. The same information is show in tabular format in Table 5. Geck can correctly capture the sign and order of magnitude of the deltas. In this particular experiment, geck also estimated the absolute recall of both pipelines correctly. Our trio-benchmarking method is useful even for accurately comparing pipelines with significant performance difference.

Table 5: Precision and recall of the two pipeline (BWA, Novoalign) on HG002 for all SNPs in the high-confidence regions. We compare the true values (which we obtained by merging the true variants with calls using bcftools, and estimated with trio-based benchmarking. Δ denotes the difference between HC and UG values. σ_{trio} is self-reported uncertainty of the model, i.e. standard deviation of the Gibbs samples.

		precision			recall		
		BWA	Novoalign	$\Delta\,(10^{-3})$	BWA	Novoalign	$\Delta\,(10^{-3})$
HG002	truth	0.9950	0.9958	−0.89	0.9739	0.9998	−25.9
	trio	0.9881	0.9883	−0.13	0.9743	0.9994	−25.1
	σ_{trio}	± 0.0013	± 0.0013	± 0.28	± 0.0003	± 0.0001	± 0.27

Figure S.10: Observed joined genotype trio counts $N = \{N_{G^1,G^2}\}$ (top), and estimated contributions of each correct genotype trio $\mathcal{N} = \{\mathcal{N}_{G^1,G^2,g}\}$ (bottom panels) for variants in the high-confidence regions in the GIAB-AJ trio called by pipelines using **different aligners**: BWA and Novoalign. The correct genotype trios g are printed on top of each panel. The matrix on top is the sum of the 15 matrices on the bottom.

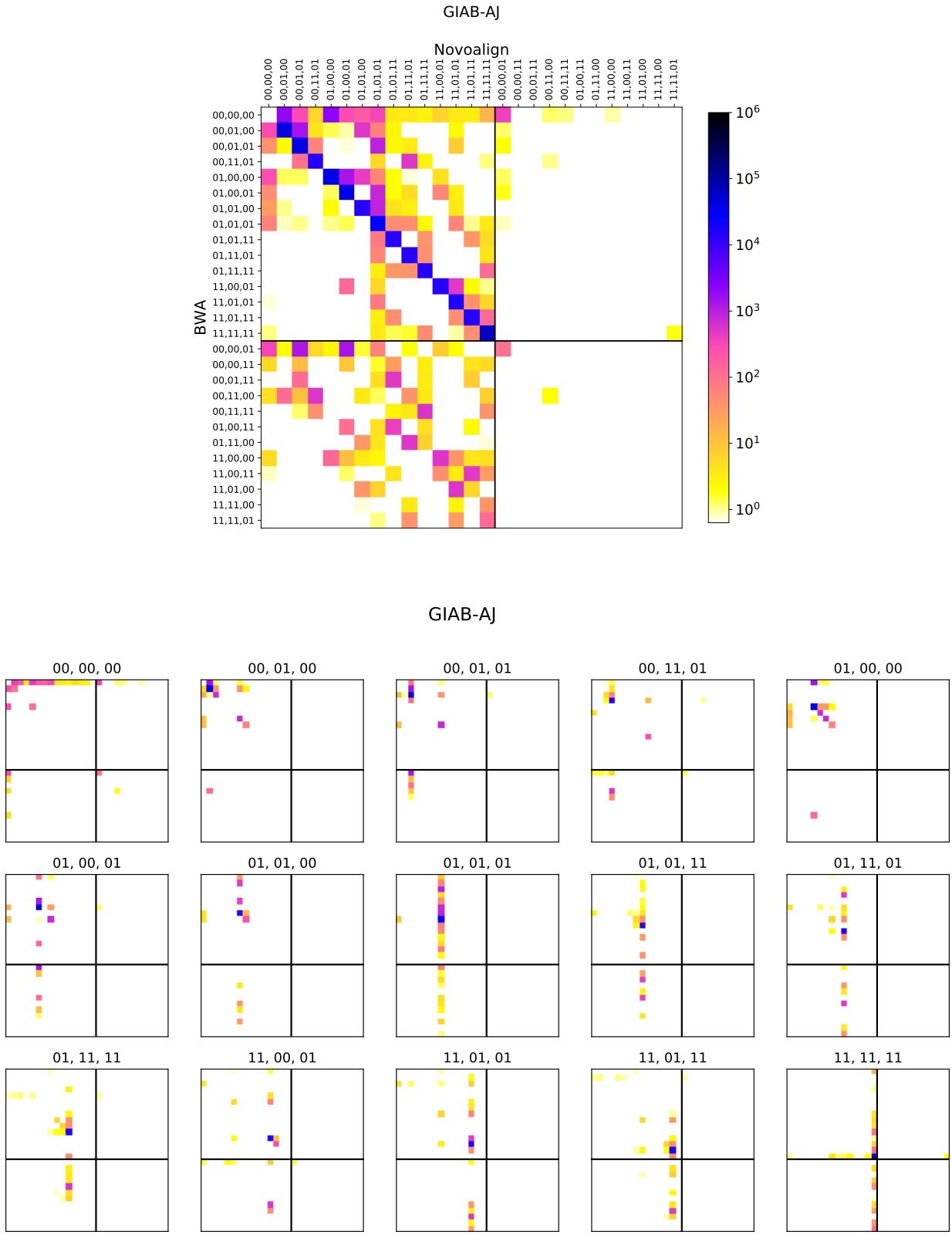


Figure S.11: Genotype confusion matrix entries for HG002 for the experiment comparing a pipeline using BWA with another pipeline using Novoalign. Colored bars correspond to the true counts calculated by comparing the called genotypes of each variant with the correct genotype (from the truth set): Green bars show cases where both callers made the correct calls, yellow bars show cases where one of them made a mistake, and red bars where both made mistakes. Light blue bars show the estimates of the same counts from our model. The green and yellow bars are estimated with higher accuracy than the red ones, which underlines the fact that our model can estimate the performance difference between pipelines more accurately than the actual values of the performance metrics.

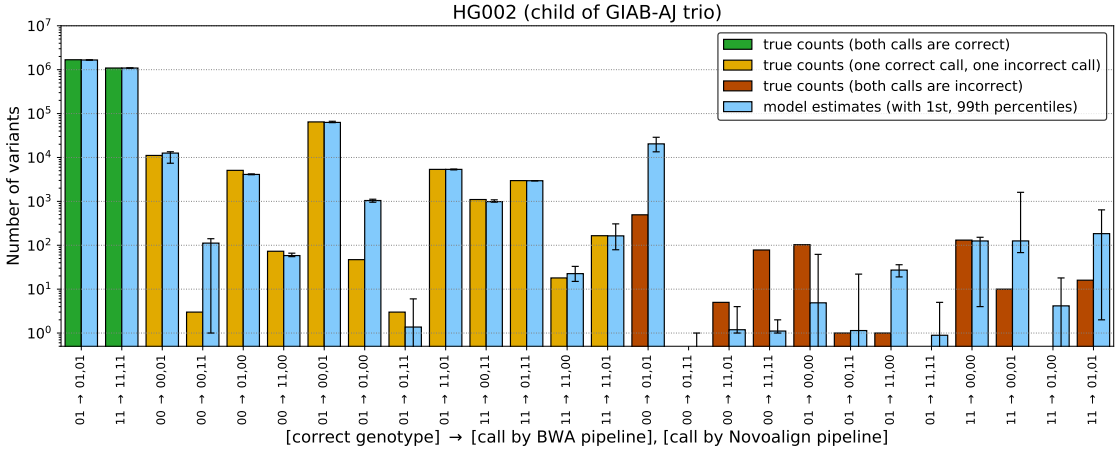
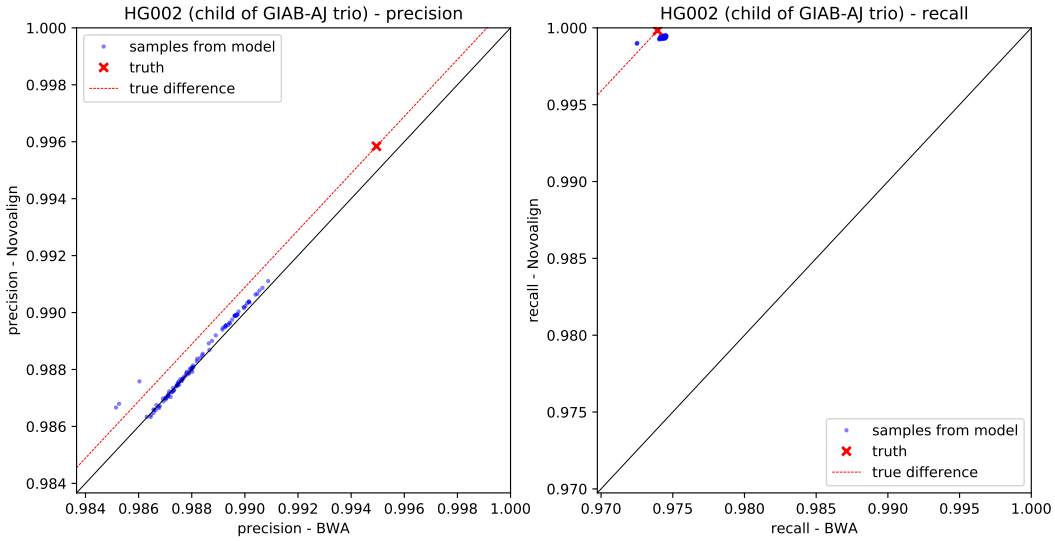


Figure S.12: Joint estimates of precision and recall for the two pipelines (BWA, Novoalign). Red cross marks the true values of precision and recall (calculated from the truth set), red dashed line indicates the points where the performance difference between the two pipelines are equal to the true value, and blue dots are samples drawn from the model's estimate of the posteriors.



D Pipelines

Here, we list the command lines used for preparing read data and in the whole genome analysis pipelines. The alignment step with BWA-MEM, and the preprocessing step with GATK tools were common in the two pipelines, only the variant calling step was different: Pipeline 1 used GATK Haplotype Caller v3.5, and pipeline 2 used GATK Unified Genotyper v2.7. We also list the command lines used in the experiments to validate our model. While indicated the connection between command lines by using the same filename placeholders `<...>`, we also give a detailed overview of their connections in Figure S.13.

D.1 Read preparation

We used `samtools` v.1.4.1 to sort the publicly available alignment files (`<input.bam>`) by read name, and extract reads into two files `<1.fq>` and `<2.fq>`.

```
$ samtools sort
  -n
  <input.bam>
> <qname-sorted.bam>
```

```
$ samtools bam2fq
  -1 <1.fq>
  -2 <2.fq>
  -F 3840
  <qname-sorted.bam>
```

The resulting fq files contained exact duplicates and unpaired reads. We removed them using our Python script `run_fastq_purge.py`.

```
$ python run_fastq_purge.py  <1.fq> <2.fq>  -o <1-purged.fq> <2-purged.fq>
```

D.2 Alignment

We used `BWA-MEM` v0.7.13 to align the paired-end Illumina reads (`<1.fastq.gz>`, `<2.fastq.gz>`), and `samblaster` and `sambamba` to filter out secondary alignments and sort the reads by position, producing `<aligned.bam>` alignment file.

```
$ /bwa-0.7.13/bwa mem
  -M
  -R '@RG\tID:1\tLB:hiseq-X-v1-HLI\tPL:illumina\tPU:reads\tSM:<sample>'
  -t 30 human_g1k_v37_decoy.fasta
  <1.fastq.gz>
  <2.fastq.gz>
| /samblaster/samblaster
  -i /dev/stdin
  -o /dev/stdout
| /sambamba_v0.6.0 view
  -t 30
  --filter 'not secondary_alignment'
  -f bam
  -l 0
  -S /dev/stdin
| /sambamba_v0.6.0 sort
  -t 30
  -m 18GiB
  --tmpdir ./
  -o <aligned.bam>
  -l 5 /dev/stdin
```

D.3 Realignment and recalibration

First, GATK RealinerTargetCreator is used to identify the regions in `<aligned.bam>` where realignment needs to be called, producing `<intervals>`.

```
$ java
  -Xmx2048M
  -jar /GenomeAnalysisTK_3.5-0-g36282e4.jar
  --analysis_type RealignerTargetCreator
  -nt 4
  --out <intervals>
  --reference_sequence human_g1k_v37_decoy.fasta
  --input_file <aligned.bam>
  --phone_home NO_ET
  --known 1000G_phase1.indels.b37.vcf
  --known Mills_and_1000G_gold_standard.indels.b37.sites.vcf
  --gatk_key <GATK-key>
```

Second, GATK IndelRealigner is run to realign the reads from `<aligned.bam>` in the target regions `<intervals>`, producing `<realigned.bam>`.

```
$ java
  -Xmx2048M
  -jar /GenomeAnalysisTK_3.5-0-g36282e4.jar
  --analysis_type IndelRealigner
  --out <realigned.bam>
  --targetIntervals <intervals>
  --reference_sequence human_g1k_v37_decoy.fasta
  --input_file <aligned.bam>
  --phone_home NO_ET
  --knownAlleles 1000G_phase1.indels.b37.vcf
  --knownAlleles Mills_and_1000G_gold_standard.indels.b37.sites.vcf
  --gatk_key <GATK-key>
```

Third, GATK BaseRecalibrator is run to recalibrate the base qualities of the realigned reads `<realigned.bam>`, producing `<realigned_recal_data.grp>`.

```
$ java
  -Xmx50000M
  -jar /GenomeAnalysisTK_3.5-0-g36282e4.jar
  --analysis_type BaseRecalibrator
  --out <realigned_recal_data.grp>
  --disable_indel_quals
  --reference_sequence human_g1k_v37_decoy.fasta
  --input_file <realigned.bam>
  --knownSites dbsnp_137.b37.vcf
  --intervals 20
  --gatk_key <GATK-key>
```

Finally, GATK PrintReads is run to create the recalibrated alignment file, `<realigned.base_recalibrated.bam>`.

```
$ java
  -Xmx2048M
  -jar /GenomeAnalysisTK_3.5-0-g36282e4.jar
  --analysis_type PrintReads
  -nct 4
  --out <realigned.base_recalibrated.bam>
  --reference_sequence human_g1k_v37_decoy.fasta
  --input_file <realigned.bam>
  --phone_home NO_ET
  --gatk_key <GATK-key>
  --BQSR <realigned_recal_data.grp>
```

D.4 Variant calling

In the final step, we use two different variant callers to call variants.

- In pipeline 1, we used GATK Haplotype Caller.

```
$ java
  -Xmx2048M
  -jar /GenomeAnalysisTK_3.5-0-g36282e4.jar
  --analysis_type HaplotypeCaller
  --out <calls.vcf>
  -nct 4
  --standard_min_confidence_threshold_for_emitting 10
  --reference_sequence human_g1k_v37_decoy.fasta
  --input_file <realigned.base_recalibrated.bam>
  --phone_home NO_ET
  --gatk_key <GATK-key>
  --dbsnp dbsnp_137.b37.vcf
```

- In pipeline 2, we used GATK Unified Genotyper.

```
$ java
  -Xmx2048M
  -jar /GenomeAnalysisTKLite.jar
  --analysis_type UnifiedGenotyper
  -nt 4
  --out <calls.vcf>
  --reference_sequence human_g1k_v37_decoy.fasta
  --input_file <realigned.base_recalibrated.bam>
  --genotype_likelihoods_model BOTH
  --dbsnp dbsnp_137.b37.vcf
```

Both pipeline yield one file `<calls.vcf>` containing the called variants of a single individual.

D.5 Truth-based benchmarking

First, samples in each VCF file are renamed with `bcftools` to avoid name collision in the merging step. The following command lines are used for `<truth.vcf.gz>` and `<calls.vcf.gz>`, producing `<truth-renamed.vcf.gz>` and `<calls-renamed.vcf.gz>`, respectively,

```
$ bcftools reheader
  -s <new-names.txt>
  -o <renamed.vcf.gz>
  <input.vcf.gz>
$ bcftools index --tbi  <renamed.vcf.gz>
```

where `<new-names.txt>` is a plain, tab-delimited text file listing old and new names.
Then we merge the two VCFs, select SNPs from the high-confidence regions, and sort the result.

```
$ bcftools merge
  -o <merged.vcf.gz>
  -O z
  <truth-renamed.vcf.gz>
  <calls-renamed.vcf.gz>
$ bcftools index --tbi  <merged.vcf.gz>
```

```
$ bcftools view
  -v snps
  -R <high-conf-regions.bed>
  -O z
  -o <merged-selected.vcf.gz>
  <merged.vcf.gz>
$ bcftools index --tbi  <merged-selected.vcf.gz>
```

```
$ bcftools sort
  -O z
  -o <merged-selected-sorted.vcf.gz>
  <merged-selected.vcf.gz>
$ bcftools index --tbi  <merged-selected-sorted.vcf.gz>
```

Finally, we aggregate the different genotype combinations, and calculate true benchmarking metrics.

```
$ python aggregate_merged_vcf.py
  <merged-selected-sorted.vcf.gz>
  <merged-counts.txt>
```

```
$ python calculate_truth_metrics.py
  <merged-counts-tool1.txt>
  <merged-counts-tool2.txt>
  <sanitized-confusion-matrix.txt>
  <benchmarking-metrics.json>
```

D.6 Trio-based benchmarking

We first append the name of the pipeline to each sample.

```
$ bcftools view
  -h
  <calls.vcf.gz>
| grep
  "#CHROM"
| awk
  -F'\t'
  '{ printf $10 }'
> sample.tmp
$ echo ":<tool-name>" >> sample.tmp
$ bcftools reheader
  -s sample.tmp
  -o <calls-renamed.vcf.gz>
$ bcftools index --tbi <calls-renamed.vcf.gz>
```

Then, we create a pedigree file for keeping the relationships of the samples.

```
$ python create_ped_file.py
  HC
  UG
  <father1-renamed.vcf.gz>
  <mother1-renamed.vcf.gz>
  <child1-renamed.vcf.gz>
  <father2-renamed.vcf.gz>
  <mother2-renamed.vcf.gz>
  <child2-renamed.vcf.gz>
  <trio.ped>
```

While we also merge the variant files, select the variants in the high-confidence regions and sort the variants.

```
$ bcftools merge
  -o <merged.vcf.gz>
  -O z
  <father1-renamed.vcf.gz>
  <mother1-renamed.vcf.gz>
  <child1-renamed.vcf.gz>
  <father2-renamed.vcf.gz>
  <mother2-renamed.vcf.gz>
  <child2-renamed.vcf.gz>
$ bcftools index --tbi  <merged.vcf.gz>
```

```
$ bcftools view
  -v snps
  -R <high-confidence.bed>
  -O z
  -o <merged-selected.vcf.gz>
  <merged.vcf.gz>
$ bcftools index --tbi  <merged-selected.vcf.gz>
```

```
$ bcftools sort
  -O z
  -o <merged-selected-sorted.vcf.gz>
  <merged-selected.vcf.gz>
$ bcftools index --tbi <merged-selected-sorted.vcf.gz>
```

Then, we aggregate the genotype combinations,

```
$ python aggregate_merged_vcf.py
  <merged-selected.vcf.gz>
  <raw-trio-counts.txt>
```

calculate the observed joint counts N and, optionally, we subsample it.

```
$ python calculate_confusion_matrix.py
  HC
  UG
  <raw-trio-counts.txt>
  <trio.ped>
  <N.txt>
```

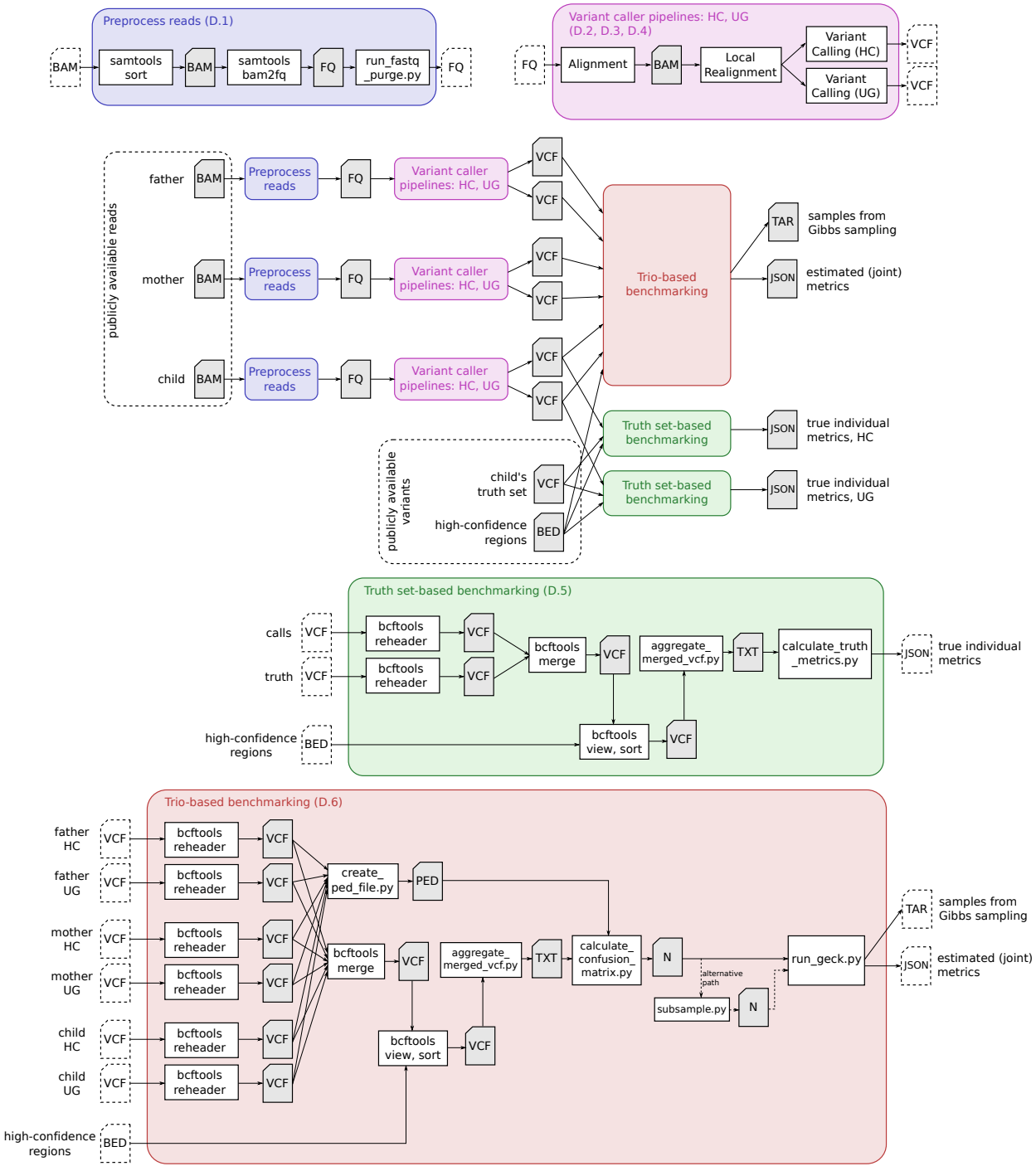
```
$ python subsample.py
  <N.txt>
  <number-of-samples>
  <random-subsampling-seed>
  <N-subsampled.txt>
```

Finally, we run the trio-based benchmarking, producing `<metrics.json>`, and we compress the samples into `<trio-samples.tar.gz>`.

```
$ python run_geck.py
  <N.txt>  # or <N-subsampled.txt>
  52054804
  HC
  UG
  1234567890  # random seed
  50000  # burn-in
  100000  # total iterations after burn-in
  1000  # thinning
  <metrics.json>
  0.01  # list of percentiles to report in json
  0.05
  0.5
  0.95
  0.99

$ tar
  -z
  -c
  -f <trio-samples.tar.gz>
  Ncomplete.txt
  n_family.txt
  n_father.txt
  n_mother.txt
  n_child.txt
  metrics_family.txt
  metrics_father.txt
  metrics_mother.txt
  metrics_child.txt
```

Figure S.13: Pipelines used in the validation experiment. Preprocessing reads, aligning, calling variants, and performing trio-based and truth set-based benchmarking.



E Discussion of possible modifications

Our model assumes that all 15 true genotype trios $g \in t$ have their own frequencies f_g, which are independently estimated. This flexibility would lack motivation, if we were analyzing genotypes of a single variant across a large number of trios, because

1. Random Mendelian segregation is expected to hold, i.e. the ratio of heterozygous and homozygous children is expected to be 1:1 if one parent is homozygous, and 2:1 ratio if both parents are heterozygous.
2. Hardy-Weinberg equilibrium (HWE) is expected to hold, unless one allele exerts significant selection pressure.

HWE is a stronger assumption than random Mendelian segregation, we investigate them separately.

In this section, we show that neither of the above two assumptions hold for truth benchmarking data for the GIAB-AJ trio. This result may seem paradoxical, until we realize that the counts (N) of aggregated genotype trios are obtained by aggregating genotypes of *different variants* in the same three people, instead of one variant across different trios.

We downloaded the true variant files from `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/`, merged them with bcftools, and dropped all non-snp variants and all variants outside of the high-confidence region (the bed file of which we obtained by intersecting the three bed files for the three family members with bedtools). Then, we counted the observed genotype trios, identifying "./." with "00" and not distinguishing between "0|1" and "1|0", counting both under the heterozygous label "01".

E.1 Assuming Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium dictates that the frequencies of 00, 01 and 11 genotypes are determined by a single allele frequency f_0:

$$P(00) = (1-f_0)^2 \tag{S.24}$$
$$P(01) = 2(1-f_0)f_0 \tag{S.25}$$
$$P(11) = (f_0)^2 \tag{S.26}$$

If we assume that the same allele frequency is applicable to the two parents, then we can write the frequency of parental genotype combinations (g_1, g_2) as

$$P(g_1, g_2) = P(g_1) \times P(g_2), \qquad \text{where } (g_1, g_2) \in I^{\times 2}. \tag{S.27}$$

Child genotypes are assumed to be combined from randomly chosen parental alleles, one from each, resulting in the following probabilities for trio genotypes $g = (g_1, g_2, g_3)$:

$$P(g) = P(g_1, g_2, g_3) = P(g_1, g_2) \times P(g_3 \,|\, g_1, g_2) \tag{S.28}$$

$$\text{where } P(g_3 \,|\, g_1, g_2) = \begin{cases} 1 & \text{if } g_1 \neq 01, \text{ and } g_2 \neq 01 \\ 1/2 & \text{if } (g_1 = 01, \text{ and } g_2 \neq 01) \text{ or } (g_1 \neq 01, \text{ and } g_2 = 01) \\ 1/2 & \text{if } g_1 = g_2 = g_3 = 01 \\ 1/4 & \text{if } g_1 = g_2 = 01 \text{ and } g_3 \neq 01 \\ 0 & \text{if not allowed by Mendelian inheritance} \end{cases} \tag{S.29}$$

For a table format, see Table 6.

To show how far are the true frequencies in the truth set of the GIAB-AJ trio from what we would expect assuming Hardy-Weinberg equilibrium, we optimize f_0 and the total number of variants n_{tot}, which we need to do because the truth data set does not list variants with (00,00,00) trio genotype. We find the optimal values of f_0 and n_{tot} by minimizing the following cost function

$$\text{cost}(f_0, n_{\text{tot}}) = \sum_{g \in t} \left[\frac{n_g^{\text{true}} - n_g^{\text{exp}}}{\sqrt{n_g^{\text{exp}}}} \right]^2, \qquad \text{where} \quad n_g^{\text{exp}} = n_{\text{tot}} P(g). \tag{S.30}$$

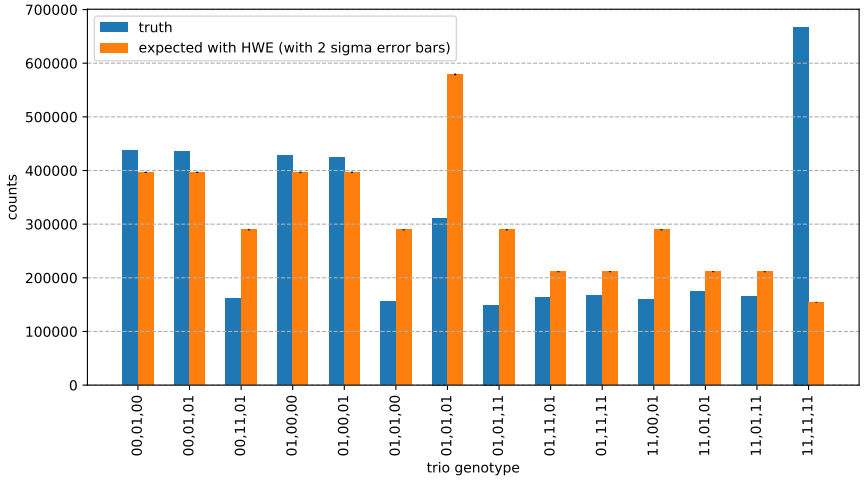
Table 6: Probabilities of child genotypes (g_3), given parental genotype combination (g_1, g_2), i.e. $P(g_3 \mid g_1, g_2)$. (Empty cells stand for zero probability.)

		g_3		
g_1	g_2	00	01	11
00	00	1		
00	11		1	
11	00		1	
11	11			1
00	01	1/2	1/2	
01	00	1/2	1/2	
01	11		1/2	1/2
11	01		1/2	1/2
01	01	1/4	1/2	1/4

We used the $\sqrt{n_g^{\text{exp}}}$ factor in the denominator to account for the Poisson sampling noise around the expected value. Minimization was done with Phython's scipy.optimize.minimize function, and we obtained the optimal values: $n_{\text{tot}} = 4.867 \times 10^6$ and $f_0 = 0.4220$.

Using these, we calculated the expected counts for each trio genotype using Equation S.28, and plotted it alongside with the counts from the truth set in Figure S.14. Differences between true and expected counts are overwhelmingly significant, which we emphasize by showing the tiny 2-sigma error bars representing the expected, Poisson-distributed, sampling noise. This means, the aggregate counts of true data do not obey Hardy-Weinberg equilibrium, and so our model should not be restricted to it either.

Figure S.14: True (blue) and expected (orange) counts of genotype trios. The true numbers are calculated from merging the truth sets for the three members of the GIAB-AJ trio in common high-confidence regions. The expected counts are found by assuming HWE and finding the optimal global allele frequency $f_0 = 0.4220$ and the optimal total number of variants $n_{\text{tot}} = 4.867 \times 10^6$, using the truth data directly. The uncertainty of the expected counts is assumed to be Poisson-distributed around the mean, they are shown (although barely visible, because of their small size) with black bars.



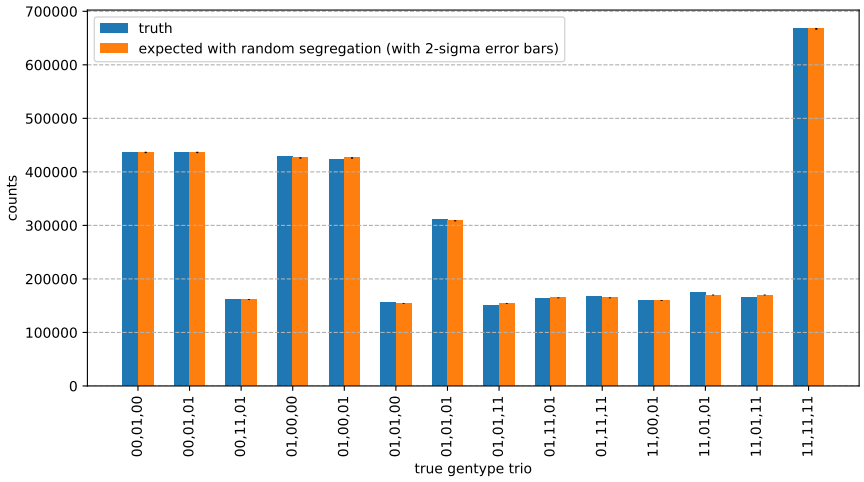
E.2 Assuming random Mendelian segregation

Although it is clear from Figure S.14 that Hardy-Weinberg equilibrium does not hold in this aggregate data, we can ask if a relaxed condition, namely random Mendelian segregation, holds. Under this hypothesis, we assume that each of the 9 parental combinations $(g_1, g_2) \in I^{\times 2}$ has its own independent frequency, and the genotypes of the child's variants are sampled independently from $P(g_3 \mid g_1, g_2)$ (see Equation S.29).

The 9 unknown frequencies can be directly obtained from the truth data by counting how many times we observe each parental combination. After that, we multiply each with $P(g_3 \mid g_1, g_2)$ to obtain the expected counts, and estimate the standard deviation of their expected sampling noise with the square root of the expected counts. Comparison of the true genotype trio counts and the expected ones are shown in Figure S.15. As expected, this flexible model can describe the observed data much better (in fact, exactly, for the parental combinations where the child's genotype is deterministic).

To get a more zoomed in picture of the remaining differences, we plot the difference between expected and true counts in Figure S.16. Apart from the cases where the child's genotype is deterministic, the counts of only two trios agree with the expected values within 2-sigma error margin, and some (namely (01,01,01), (11,01,01) and (11,01,11)) even have differences more than 6 sigma. This indicates that the assumption of random Mendelian segregation does not hold for true trio genotype counts aggregated over different variants. And so our model should not incorporate such a restriction.

Figure S.15: True (blue) and expected (orange) counts of genotype trios. The true numbers are calculated from merging the truth sets for the three members of the GIAB-AJ trio in common high-confidence regions. The expected counts are found by assuming random Mendelian segregation, and fitted directly to the truth data. The uncertainty of the expected counts is assumed to be Poisson distributed around the mean, they are shown (although barely visible, because of their small size) with black bars.

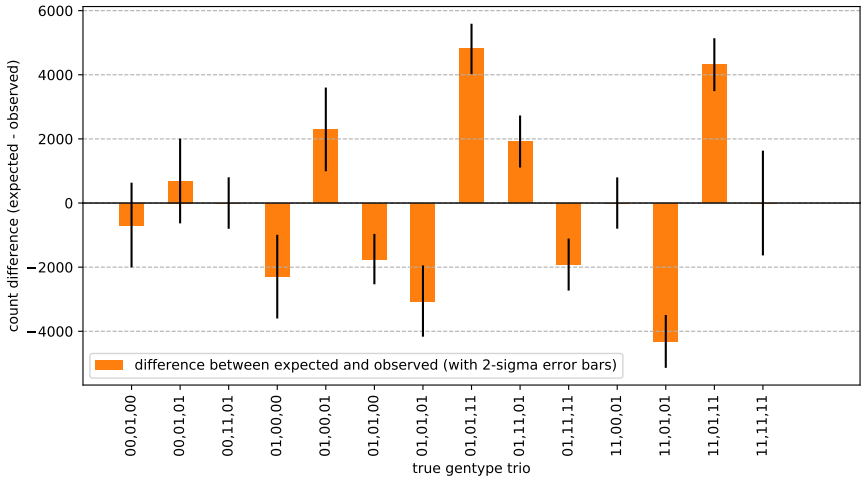


The significant difference between the observed aggregate genotype frequencies and the expected values under the assumption of random Mendelian segregation is surprising. After all, we do not claim to disprove the basis for Mendelian randomization. Although we did not investigate it in details, there is one hypothesis that could explain the observed discrepancy: While investigating the effect of random Mendelian segregation, we assumed that this process acts independently on each SNP. This working hypothesis led to standard deviations equal to the square root of the expected number of variants in for each trio genotype.

In reality, SNPs in strong linkage disequilibrium do not get randomized during gametogenesis, and as a result, they are not inherited independently. While this modification of the process leaves the expected trio genotypes frequencies unchanged, it results in a larger expected sampling noise because the effective sample size is the number of recombination events, which is much smaller than the number of SNPs. We suspect that this process can explain the significant deviation of the observed counts from the expected ones.

Now, one could ask, "So if we used correct error bars on Figure S.16, we would not see significant difference

Figure S.16: Difference between expected (under the assumption of random Mendelian segregation) and true trio genotype counts. Black bars indicate the expected 2-sigma sampling noise.



from random Mendelian segregation. Why should we not incorporate it into our model then?". To answer, let us note that geck assumes that each variant provides an independent data point. The full log likelihood of the model is a sum of independent terms, one for each variant, which translates to assuming that the total number of independent input samples is equal to the number of SNPs. During the parameter estimation procedure, the total number of independent data points determines the posterior variance of each parameter, including the values of f. If we did not allow our model to fit each f value independently, but forced it to assume random Mendelian segregation, it would work hard to get the best (restricted) f values, because they amount to a large log-likelihood penalty due to assuming a large number of independent data points. This would incorrectly prioritize getting slightly better f values above accurately estimating θ and E.

As a quick solution, one could imagine re-weighting or sub-sampling the observed data to match the expected number of independently inherited variants. This, however, would undermine the estimates of the genotyping error rates E, because from their point of view the data points are not correlated since linkage disequilibrium is not likely to cause correlation between genotyping errors. So, now we are faced with this problem: the effective number of independent samples for estimating f (under random Mendelian segregation) is lower than the effective number of independent samples for estimating E.

Instead of introducing a new layer of variables in our graphical model to account for recombination events (which would require additional inputs from the user), we took the shortcut, and granted the f parameters enough flexibility to account for the significant deviations of the true frequencies from what is expected under random Mendelian segregation.

E.3 Origin of the true f values

The results above show that allowing our model to find all 15 f_g values independently is necessary to accurately recover the true frequencies. Beside this evidence, we would like to expose our logic behind this decision.

The parameters f represent the distribution of the true trio genotypes in the data set. This distribution depends not only on the biological mechanisms at play, but also on the choices made by the scientist carrying out the benchmarking analysis. If the scientist restricts their analysis to variants with largely different alternate allele frequencies and which are not expected to be heterozygous (e.g. because they are benchmarking a tool that genotypes 00 and 11 variants with extremely high accuracy), then the aggregate genotype counts will respect neither Hardy-Weinberg equilibrium nor random Mendelian segregation. They will simply describe the true abundance of each trio genotype in the pool of variants under analysis.

Even in this case, the estimated f produced by our method will accurately reflect the distribution of true trios, because we do not restrict the values of f, and allow all 15 of them to move independently (apart from the normalization requirement $\sum_g f_g = 1$) during the Gibbs sampling process.

E.4 Incorporating uncertainty of input data

Throughout this work, we assumed that the two sets of input variants from a trio – most commonly 2×3 VCF files – can be converted to a single merged VCF file which, in turn, can be directly aggregated to give us the genotype trio pair counts, N_{G^1,G^2}. We deliberately chose SNPs to validate our method, because their identity is most certain.

Many complex variants (such as mnps, indels and structural variants) are difficult to call with single-base-pair resolution and, as a result, they are subject to identity ambiguity during merging. In other words, it is hard to tell if two differently represented complex variants in two different VCF files are in fact the same variant. Researchers often tackle this problem by turning to "soft" identity metrics (such as reciprocal overlap for deletion SVs, or local alignment score for crowded series of snps and indels). While the usage of these metrics is commendable, often the following steps involve setting an ad-hoc threshold to obtain a boolean decision. Unfortunately, this leads to additional loss of information, which could be avoided if downstream tools could handle the original soft metrics efficiently.

A straight-forward extension of our model (which is expected to increase the accuracy of the estimated benchmarking metrics if the total number of variants is small) is to admit an ensemble of counts N_{G^1,G^2} as input: Cycling through the variants and drawing the genotype of each from its posterior reported by the variant caller, and aggregating the drawn genotypes in every cycle will result in a stochastically changing N matrix. This is well suited for Gibbs-sampling which, in every iteration, can perform the sampling using the current N matrix. The resulting Markov process will realize a stochastic solution of the inference problem (similar to how stochastic gradient descent approximates the true gradient descent solution). Implementing this efficiently is a non-trivial challenge, but it is worth considering because such a trio benchmarking method will enable accounting for uncertainties originating not only from variant identity ambiguity but also from genotype uncertainties reported by PL scores in VCF files.