A unified haplotype-based method for accurate and comprehensive variant calling

Paper Presentation by Chuanyi Zhang

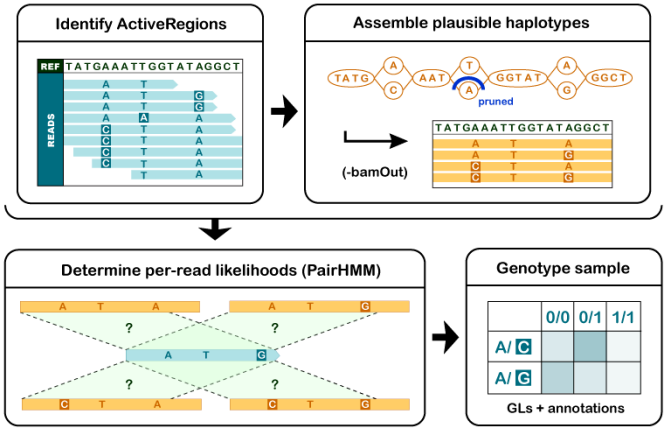
Daniel P Cooke[1,2], David C Wedge[2], and Gerton Lunter[1]

[1]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK
[2]Big Data Institute, University of Oxford, Oxford, UK

March 7, 2019

Table of Contents

Variant Calling



- Ideal scenario: enough read depth
- (1) read processing,
  (2) mapping,
  (3) calling
- haplotype analysis: `HaplotypeCaller` in Genome Analysis Toolkit (GATK)

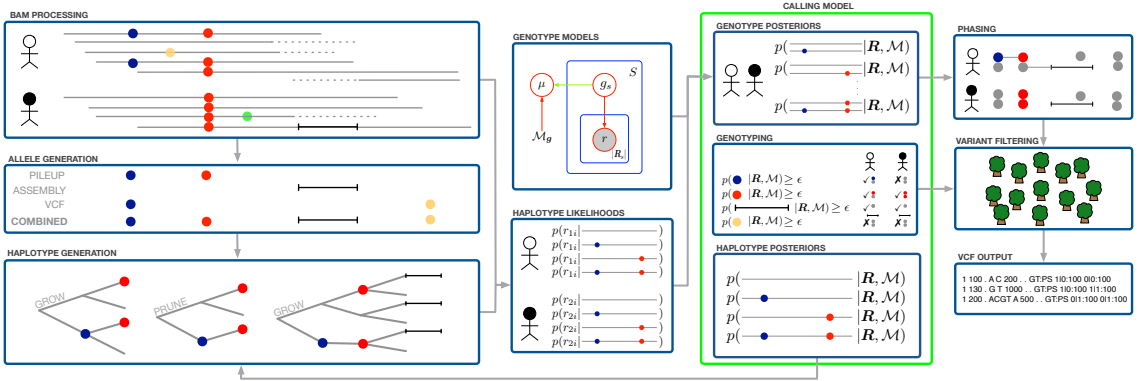
Somatic Variant Calling

Differences from germline calling

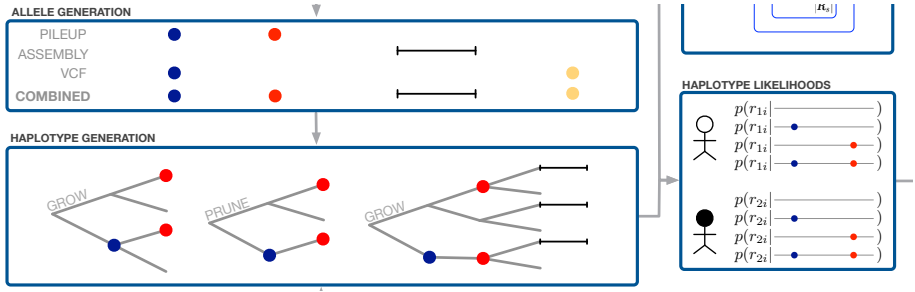
- Allele frequency assumption: purity, multiple subclones, CNA
- Low VAF vs. Artifacts
- Somatic vs. Germline: matched tumor-normal sample

Table of Contents

Overview


BAM PROCESSING
ALLELE GENERATION
PILEUP
ASSEMBLY
VCF
COMBINED
HAPLOTYPE GENERATION
GROW
PRUNE
GROW
GENOTYPE MODELS
HAPLOTYPE LIKELIHOODS
CALLING MODEL
GENOTYPE POSTERIORS
GENOTYPING
HAPLOTYPE POSTERIORS
PHASING
VARIANT FILTERING
VCF OUTPUT
1 100 . A C 200 . . GT:PS 1|0:100 0|0:100
1 130 . G T 1000 . . GT:PS 1|0:100 1|1:100
1 200 . ACGT A 500 . . GT:PS 0|1:100 0|1:100

Haplotype generating



- build from candidate alleles
- haplotype tree
- prune, stages: (1) pre: haplotype likelihood, (2) post: haplotype posterior

Genotype Prior Models

1. Uniform
2. Hardy-Weinberg-Equilibrium (HWE)
3. Coalescent-HWE
4. Trio
5. ★ Cancer

For ploidy m, genotypes: $g = (h_1, \ldots, h_m)$;
for n populations inside a tumor, joint genotypes: $\mathbf{g} = (g_1, \cdots, g_n)$.

Cancer

$g_{cancer} = (g_{germ}, g_{som})$
$p(g_{cancer} \mid \mathcal{M}_{cancer}) = p(g_{germ} \mid \mathcal{M}_{germ}) p(g_{som} \mid g_{germ}, \mathcal{M}_{som})$, $\mathcal{M}_{germ}$ can be Coalescent-HWE prior model,
if there's only 1 somatic haplotype

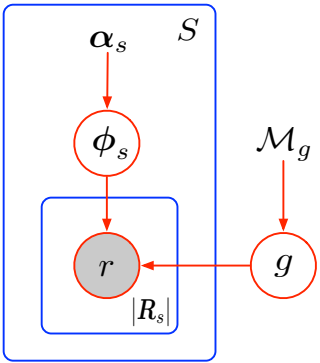
$$p(g_{som} \mid g_{germ}, \mathcal{M}_{som}) = \frac{1}{|g_{germ}|} \sum_{i=1}^{|g_{germ}|} p(g_{som} \mid g_{germ,i}, \mathcal{M}_{som})$$

if multi somatic haplotypes: (assume all haplotypes originate from germline, independently)

$$p(g_{som} \mid g_{germ}, \mathcal{M}_{som}) = \prod_{j=1}^{|g_{som}|} p(g_{som,j} \mid g_{germ}) = \prod_{j=1}^{|g_{som}|} \frac{1}{|g_{germ,j}|} \sum_{i=1}^{|g_{germ,j}|} p(g_{som} \mid g_{germ,j,i}, \mathcal{M}_{som})$$

Graphical model & joint posterior

We want to know joint posterior distribution



$$p(g, \boldsymbol{\pi} \mid \mathcal{R}, \boldsymbol{\alpha}, \mathcal{M}_g) = \frac{p(\boldsymbol{\pi}, g, \boldsymbol{\alpha}, \mathcal{M}_g, \mathcal{R})}{p(\boldsymbol{\alpha}, \mathcal{M}_g, \mathcal{R})}$$

$$= \frac{p(\mathcal{R} \mid \boldsymbol{\pi}, g) p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) p(g \mid \mathcal{M}_g)}{p(\boldsymbol{\alpha}, \mathcal{M}_g, \mathcal{R})}$$

$$\propto p(g \mid \mathcal{M}_g) \prod_{s=1}^{S} p(\mathcal{R}_s \mid \pi_s, g) p(\pi_s \mid \boldsymbol{\alpha}_s)$$

$$= p(g \mid \mathcal{M}_g) \prod_{s=1}^{S} \int p(\mathcal{R}_s \mid \phi_s, g) p(\phi_s \mid \boldsymbol{\alpha}_s) \, \mathrm{d}\phi_s$$

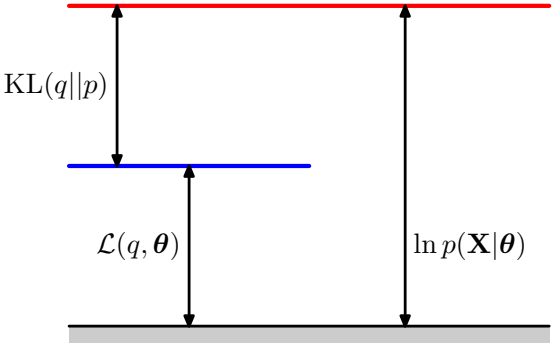
$$= p(g \mid \mathcal{M}_g) \prod_{s=1}^{S} \int \prod_{r \in \mathcal{R}_s} \sum_{i=1}^{|g|} \phi_{si} p(r \mid h_i) p(\phi_s \mid \boldsymbol{\alpha}_s) \, \mathrm{d}\phi_s$$

Problem of computing

- This posterior is intractable Since $\phi_s \sim \mathrm{Dir}(\boldsymbol{\alpha}_s)$ So the integration over ϕ_s is intractable. ϕ is latent variables.
- Using **Variational Bayes** (VB)
  Approximate $p^*(x) \triangleq p(x \mid \mathcal{D})$ (intractable posterior) with $q(x)$. Maximize $L(q) \triangleq -D_{KL}(q\|\tilde{p})$ (not $D_{KL}(p^*\|p)$), where $\tilde{p} = p(x, \mathcal{D}) = p^*(x)p(\mathcal{D})$

$$
\begin{aligned}
L(q) &= -\mathbb{E}_q\left[\log \frac{q}{\tilde{p}}\right] = -\int q(x) \log \frac{q(x)}{p^*(x)p(\mathcal{D})} \,\mathrm{d}\mu(x) \\
&= -\int q(x) \log \frac{q(x)}{p^*(x)} - q(x) \log p(\mathcal{D}) \,\mathrm{d}\mu(x) \\
&= -D_{KL}(q\|p^*) + \log p(\mathcal{D}) \\
&\leq \log p(\mathcal{D})
\end{aligned}
$$

ELBO



$L(q)$ is *evidence* lower bound (ELBO). Maximizer is $q = p^*$.

VB cont'

Bayes:

$$p(x \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid x)p(x)}{p(\mathcal{D})} \quad , \left(\text{post} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \right)$$

And by assuming this factorization

$$q(\mathbf{g}, \boldsymbol{Z}, \phi) = q(\mathbf{g}) \prod_{s=1}^{S} q(\boldsymbol{Z}_s) q(\phi_s)$$

where we introduce the latent binary matrix $\boldsymbol{Z}_s$, $q(Z_{snk})$ are so-called *responsibilities* of assuming haplotype k for read n in sample s. By this factorization (mean field) we can optimize on them alternately. Moreover, if we assume these priors are Dirichlet, then prior and posterior are *conjugated*. Categorical (likelihood) and Dirichlet are conjugate distributions.

Calling Model

Assume 3 possible cases:

1. No somatic mutations, clean germline, the individual model with any germline prior (merge) $\mathcal{M}_{ind}$
2. Copy number changes, but no somatic, the *subclone model* with germline prior (e.g. Coalescent-HWE) $\mathcal{M}_{ind}$
3. Somatic occurs, possible CNA, the *subclone model* with cancer genotype prior.

Calling Model

Germline genotype posterior

$$
\begin{aligned}
p(g \mid \mathcal{R}) &= \sum_x p(g, \mathcal{M}_x \mid \mathcal{R}) \\
&= \sum_x p(g, \mid \mathcal{M}_x, \mathcal{R}) p(\mathcal{M}_x \mid \mathcal{R}) \\
&= p(g, \mid \mathcal{M}_{ind}) p(\mathcal{M}_{ind} \mid \mathcal{R}) \\
&+ p(g, \mid \mathcal{M}_{CNV}) p(\mathcal{M}_{CNV} \mid \mathcal{R}) \\
&+ p(g, \mid \mathcal{M}_{somatic}) p(\mathcal{M}_{somatic} \mid \mathcal{R})
\end{aligned}
$$

where $p(\mathcal{M}_x \mid \mathcal{R}) = p(\mathcal{M}_x) p(\mathcal{R} \mid \mathcal{M}_x)$, and $p(\mathcal{R} \mid \mathcal{M}_x)$ is the "evidence"; and $p(g \mid \mathcal{M}_{somatic}) = \sum_{\tilde{g}: g \in \tilde{g}} p(\tilde{g} \mid \mathcal{M}_{somatic})$, $\tilde{g} = (g_{germ}, g_{som})$, from *cancer prior*
Germline allele posterior $p(a \mid \mathcal{R}) = \sum_{g: a \in g} p(g \mid \mathcal{R})$.

Credible somatic mass

$p_{somatic}(a \mid \mathcal{R}) \leftarrow \sum_{?} p(\tilde{g} \mid \mathcal{M}_{somatic}, \text{credible})$
There are K somatic haplotypes, then the *credible somatic frequencies* satisfy

$$p(\phi_{sk} > \tau \mid \mathcal{M}_{somatic}) = \int_{\tau}^{1} \text{Beta}\left(\theta\,; \alpha_{P+1}, \sum_{i=1}^{P} \alpha_i\right) d\theta$$

where $\phi_{sk} \sim \text{Beta}(\alpha_k, \sum \alpha - \alpha_k)$ since $\boldsymbol{\phi}_s \sim \text{Dir}(\boldsymbol{\alpha}_s)$ i.e. $p(\boldsymbol{\phi}_s) = \frac{1}{\text{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} \phi_{sk}^{\alpha_k - 1}$.
The *credible somatic mass* is

$$\lambda_s = 1 - \prod_{k} 1 - p(\phi_{sk} > \tau \mid \mathcal{M}_{somatic})$$

means the probability mass of $\exists$ at least 1 credible in K somatic haplotypes.

Calling allele

Then $\lambda = 1 - \prod_s \lambda_s$. $\quad (\overline{\exists_1 \wedge \cdots \wedge \exists_S} = \nexists_1 \vee \cdots \vee \nexists_S)$

$$p_{somatic}(a \mid \mathcal{R}) = \lambda \left(1 - \prod_s \sum_a \mathbb{1}_{\{a \notin \tilde{g}.germ \wedge a \in \tilde{g}.som\}} p(\tilde{g} \mid \mathcal{R}, \mathcal{M}_{somatic}) \right)$$

Might be a typo?

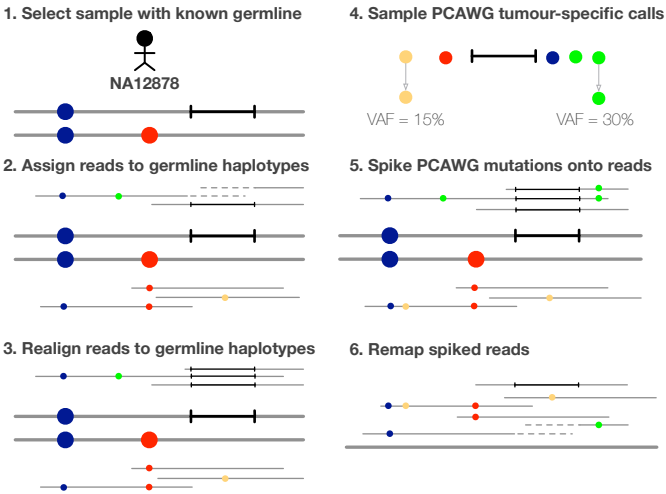
Table of Contents

Synthetic Tumors

Evaluation of somatic mutation calling is challenging

- Real tumor with manual inspected mutations
- Mix reads from unrelated individuals
- ★ Spike mutations directly into raw sequenceing reads from healthy tissue

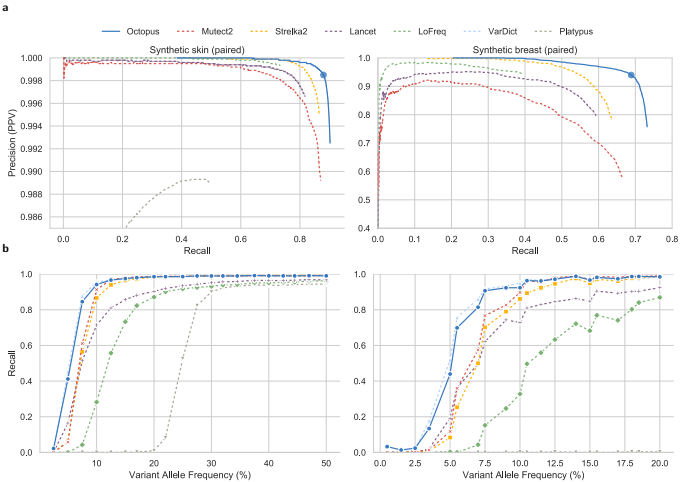
Synthetic Tumors



- Reads from NA12878 $\sim$ $300\times \rightarrow \{30\times, 35\times, 60\times, 65\times\}$
- Assign and realign to make sure spiked mutations fall on same haplotype, and position consistent
- Sample mutations from pan-cancer analysis of whole genomes (PCAWG) uniformly

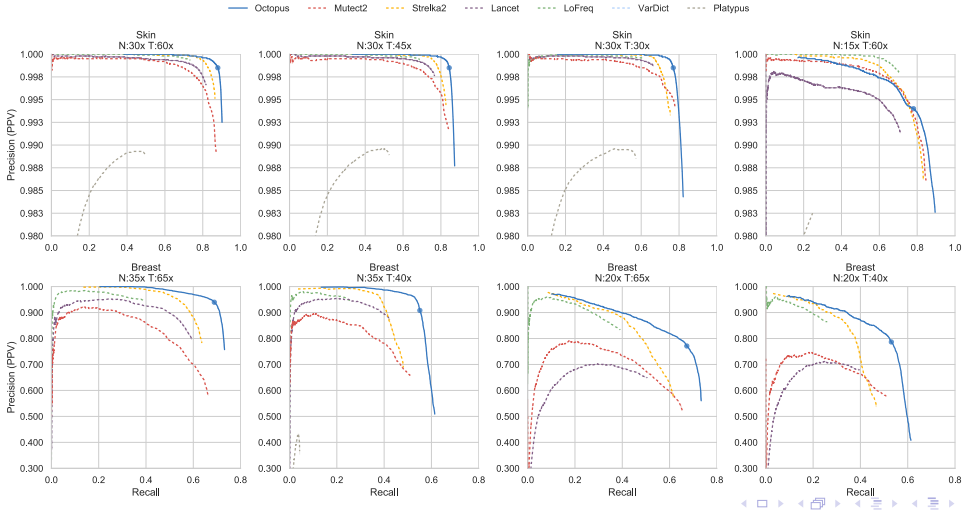
Produce raw unmapped reads (FASTQ) files.

Somatic Mutations Calling Accuracy



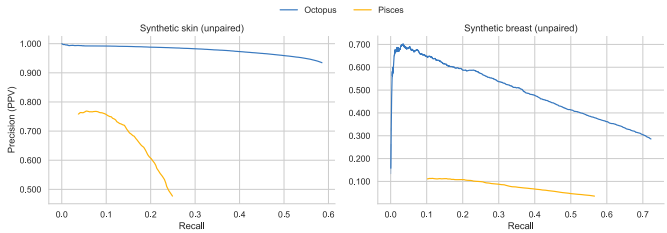
1. Precision-Recall curve: top-right is optimal
2. Recalls for each VAF, using PASS variants
3. Most differences in recall is due to low frequencies

Low coverage

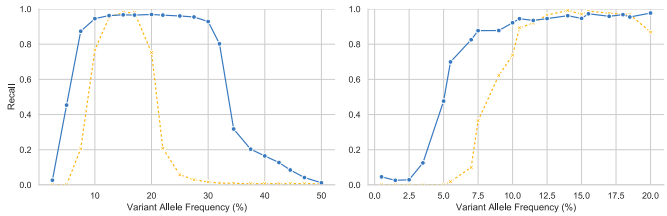


Somatic Mutations Calling Accuracy without paired normal

a



b



1. Octopus is able to discover mutations even without paired normal sample