Unified variation discovery and genotyping from high throughput sequencing data: supplementary material

This document describes in detail the algorithms and models implemented in the variant calling software Octopus.

Contents

1 Introduction

2 Methods

2.1 Read QC

2.2 Candidate allele generation

2.2.1 Raw CIGAR alignment

2.2.2 Local re-assembly

2.3 Haplotype generation

2.4 Haplotype likelihood calculation

2.4.1 Kmer based re-mapping

2.4.2 Read error modelling

2.4.3 Pair Hidden Markov Model evaluation

2.5 Haplotype filtering

2.6 Calling models

Although each genotype model is in a sense independent, they do share common attributes which we define here for brevity.

Constants	Description
P_s	The ploidy of organism s
h	A haplotype
$\mathbb{M}$	The sequencing error model

Latent variables	Description
r	A sequencing read

Then the conditional probability of a read given a haplotype is:

$$p(r|h, \mathbb{M}) \tag{1}$$

Which has already been described. For brevity, we will from here on omit the $\mathbb{M}$ term and just write $p(r|h)$.

2.6.1 Coalescent mutation model

$$p(\boldsymbol{g}|\mathcal{M}_{coal}) = \binom{\theta_1}{\theta_1+\theta_2}^{k_1} \binom{\theta_2}{\theta_1+\theta_2}^{k_2} \binom{k_1+k_2}{k_1} p_{\theta=\theta_1+\theta_2}(k_1+k_2)$$

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

2.6.2 Somatic mutation prior model

2.6.3 De novo mutation prior model

2.6.4 Individual model

This model is designed to model reads coming from a single individual with a known fixed ploidy. It is the simplest genotype model that Octopus offers, and is also the fastest to compute. It is used by multiple Octopus variant callers.

2.6.5 Definitions

Constants	Description
P	The individuals ploidy
G	The number of genotypes
$\boldsymbol{\phi}$	Vector of G probabilities
h_{ij}	The j^{th} haplotype of the i^{th} genotype
$\mathcal{R}$	The set of sequencing reads for the individual
N	$\mathcal{R}$

Note that h_{ij} is technically a deterministic surjective function $h_{ij} = f(\boldsymbol{g}, i, j)$ which maps genotypes to haplotypes, but for brevity we just write h_{ij}.

Observed latent variables	Description
$\mathcal{R}$	The set of sequencing reads for the individual
N	$\mathcal{R}$

Hidden latent variables	Description
$\boldsymbol{g}$	Binary (1-of-G)

2.6.6 Marginal distributions

The marginal distribution for $\boldsymbol{g}$ is categorical:

$$p(\boldsymbol{g}|\boldsymbol{\phi}) = \prod_{i=1}^{G} \phi_i^{g_i} \tag{2}$$

The marginal distribution for a single read r is a mixture distribution, which follows from the assumption that the haplotypes that make up a known ploidy genotype are exchangeable, and thus we must assign them equal probabilities of being sequenced:

$$p(r|g_i = 1) = \frac{1}{P} \sum_{k=1}^{P} p(r|h_{ik}) \tag{3}$$

which can also be written as:

$$p(r|\boldsymbol{g}) = \prod_{i=1}^{G} \frac{1}{P} \sum_{k=1}^{P} p(r|h_{ik})^{g_i} \tag{4}$$

2.6.7 Joint distribution

$$p(\mathcal{R}, \boldsymbol{g}) = \prod_{i=1}^{G} \phi_i^{g_i} \prod_{n=1}^{N} \frac{1}{P} \sum_{k=1}^{P} p(r_n|h_{ik})^{g_i} \tag{5}$$

2.6.8 Posterior distribution

$$p(\boldsymbol{g}|\mathcal{R}) = \frac{\prod_{i=1}^{G} \phi_i \prod_{n=1}^{N} \frac{1}{P} \sum_{k=1}^{P} p(r_n|h_{ik})^{g_i}}{\sum_{\boldsymbol{g}'} \prod_{i=1}^{G} \phi_i \prod_{n=1}^{N} \frac{1}{P} \sum_{k=1}^{P} p(r_n|h_{ik})^{g_i}} \tag{6}$$

We can factor out the constant P term, and write more compactly as:

$$p(g_i = 1|\mathcal{R}) = \frac{\phi_i \prod_{n=1}^{N} \sum_{k=1}^{P} p(r_n|h_{ik})}{\sum_{j=1}^{G} \phi_i \prod_{n=1}^{N} \sum_{k=1}^{P} p(r_n|h_{ik})} \tag{7}$$

2.6.9 Evidence

The evidence for the model is simply the denominator of (6):

$$p(\mathcal{R}) = \sum_{i=1}^{G} \phi_i \prod_{n=1}^{N} \frac{1}{P} \sum_{k=1}^{P} p(r_n|h_{ik}) \tag{8}$$

2.6.10 Population genotype model

2.6.11 Trio genotype model

2.6.12 CNV genotype model

The copy-number-variant genotype model attempts to model copy number changes in a single individual from which multiple samples have been taken. This model can be contrasted with the individual model which assumes the haplotypes in the individuals germline genotype are present in a $1:1$ ratio (for diploid cases).

2.6.13 Definitions

Constants	Description
P	The individuals germline ploidy
S	The number of samples for the individual
G	The number of genotypes
$\boldsymbol{\phi}$	Vector of G probabilities
h_{ij}	The j^{th} haplotype of the i^{th} genotype
$\boldsymbol{\alpha_s}$	A vector of P Dirichlet counts

Observed latent variables	Description
$\mathcal{R}_s$	The set of sequencing reads for sample s of the individual
N_s	$\mathcal{R}_s$

Hidden latent variables	Description
$\boldsymbol{g}$	Binary (1-of-G)
$\boldsymbol{\pi}_s$	A vector of P probabilities of sequencing each haplotype in the individuals genotype
$\boldsymbol{z}_{sn}$	Binary (1-of-P), specifying which haplotype was sequenced in read n of sample s

2.6.14 Marginal distributions

The marginal distribution for $\boldsymbol{g}$ is categorical:

$$p(\boldsymbol{g}|\boldsymbol{\phi}) = \prod_{i=1}^{G} \phi_i^{g_i} \tag{9}$$

The marginal distribution of ϕ is assumed to be Dirichlet, this is mostly because it simplifies the mathematics.

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{P} \pi_k^{\alpha_k - 1} \tag{10}$$

where $\mathrm{B}(\boldsymbol{\alpha})$ is the multivariate Beta function:

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \tag{11}$$

The marginal distribution for each $\boldsymbol{z}$ is categorical:

$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{k=1}^{P} \pi_k^{z_k} \tag{12}$$

The marginal distribution for a single read r is a mixture distribution:

$$p(r|\boldsymbol{\pi}, g_i = 1) = \sum_{\boldsymbol{z}} \sum_{k=1}^{P} p(\boldsymbol{z}|\boldsymbol{\pi}) p(r|\boldsymbol{z}, h_{ik}) \tag{13}$$

$$= \sum_{k=1}^{P} \pi_k p(r|h_{ik}) \tag{14}$$

2.6.15 Joint distribution

$$p(\mathcal{R}, \boldsymbol{g}, \boldsymbol{Z}, \boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\phi}) = p(\boldsymbol{g}|\boldsymbol{\phi}) \prod_{s=1}^{S} p(\boldsymbol{\pi}_s|\boldsymbol{\alpha}_s) p(\boldsymbol{Z}_s|\boldsymbol{\pi}_s) p(\mathcal{R}_s|\boldsymbol{Z}_s, \boldsymbol{g}) \tag{15}$$

$$= \prod_{i=1}^{G} \phi_i^{g_i} \prod_{s=1}^{S} \frac{1}{\mathrm{B}(\boldsymbol{\alpha}_s)} \prod_{k=1}^{P} \pi_{sk}^{\alpha_{sk}-1} \prod_{n=1}^{N_s} \prod_{k=1}^{P} \pi_{sk}^{z_{snk}} \prod_{i=1}^{G} \prod_{n=1}^{N_s} \prod_{k=1}^{P} p(r_{sn}|h_{ik})^{g_i z_{snk}} \tag{16}$$

2.6.16 Posterior distribution

We are interested in the posterior distributions of $\boldsymbol{g}$ and $\boldsymbol{\pi}$, which due to the hidden latent variables $\boldsymbol{z}$ cannot be evaluated in closed form, we therefore resort to approximations. We choose a Variational Bayes approximation because it is deterministic and faster to compute than Monte Carlo based methods. A description of Variational Bayes is given in the appendix.

To evaluate approximate posterior distributions in the Variational Bayes framework we need to factorise (15) into independent factors, there is only one possibility here:

$$q(\boldsymbol{g}, \boldsymbol{Z}, \boldsymbol{\pi}) = q(\boldsymbol{g}) \prod_{s=1}^{S} q(\boldsymbol{Z}_s) q(\boldsymbol{\pi}_s) \tag{17}$$

where non-latent variables have been omitted for brevity. We evaluate the optimal factors in tern.

$$\ln q^*(\boldsymbol{Z}_s) = \mathbb{E}_{\boldsymbol{g}, \boldsymbol{\pi}, \boldsymbol{Z}_{s' \neq s}}[\ln p(\mathcal{R}, \boldsymbol{g}, \boldsymbol{Z}, \boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\phi})] + \text{const} \tag{18}$$

$$= \mathbb{E}[\ln p(\boldsymbol{Z}_s|\boldsymbol{\pi}_s)] + \mathbb{E}[\ln p(\mathcal{R}_s|\boldsymbol{Z}_s, \boldsymbol{g})] + \text{const} \tag{19}$$

$$= \sum_{n=1}^{N_s} \sum_{k=1}^{P} z_{snk} \mathbb{E}[\ln \pi_{sk}] + \sum_{n=1}^{N_s} \sum_{k=1}^{P} \sum_{i=1}^{G} \mathbb{E}[g_i] \ln p(r_{sn}|h_{ik})^{g_i z_{snk}} + \text{const} \tag{20}$$

$$= \sum_{n=1}^{N_s} \sum_{k=1}^{P} z_{snk} \ln \rho_{snk} + \text{const} \tag{21}$$

where we have defined

$$\ln \rho_{snk} = \ln \tilde{\pi}_{sk} + \sum_{i=1}^{G} \phi_i \ln p(r_{sn}|h_{ik}) \tag{22}$$

and

$$\ln \tilde{\pi}_{sk} = \psi(\alpha_{sk}) - \psi(\hat{\alpha}_s) \tag{23}$$

where ψ is the digamma function and $\hat{\alpha}_s = \sum_{k=1}^{P} \alpha_{sk}$. Exponentiating both sides of (18) and normalising we obtain:

$$q^*(\boldsymbol{Z}_s) = \prod_{n=1}^{N_s} \prod_{k=1}^{P} \tau_{snk}^{z_{snk}} \tag{24}$$

where

$$\tau_{snk} = \frac{\rho_{snk}}{\sum_{j=1}^{P} \rho_{snj}} \tag{25}$$

Next we look at each $q^*(\boldsymbol{\pi}_s)$:

$$\ln q^*(\boldsymbol{\pi}_s) = \mathbb{E}_{\boldsymbol{g},\boldsymbol{\pi}'_s \neq s,\boldsymbol{Z}}[\ln p(\mathcal{R},\boldsymbol{g},\boldsymbol{Z},\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{\phi})] + \text{const} \tag{26}$$

$$= \mathbb{E}[\ln p(\boldsymbol{\pi}_s|\boldsymbol{\alpha}_s)] + \mathbb{E}[\ln p(\boldsymbol{Z}_s|\boldsymbol{\pi}_s)] + \text{const} \tag{27}$$

$$= \sum_{k=1}^{P}(\alpha_k - 1)\ln \pi_{sk} + \sum_{k=1}^{P}\sum_{n=1}^{N_s}\tau_{snk}\ln \pi_{sk} + \text{const} \tag{28}$$

and therefore

$$q^*(\boldsymbol{\pi}_s) = \prod_{k=1}^{P}\pi_{sk}^{\alpha_{sk}-1}\prod_{k=1}^{P}\pi_{sk}^{\sum_{n=1}^{N_s}\tau_{snk}} \tag{29}$$

$$= \prod_{k=1}^{P}\pi_{sk}^{\alpha_{sk}+\sum_{n=1}^{N_s}\tau_{snk}-1} \tag{30}$$

Which we can see from inspection is another Dirichlet distribution with pseudo-counts:

$$\alpha_{sk}^{post} = \alpha_{sk}^{prior} + \hat{N}_{sk} \tag{31}$$

where $\hat{N}_{sk} = \sum_{n=1}^{N_s}\tau_{snk}$. Finally we evaluate $q^*(\boldsymbol{g})$:

$$\ln q^*(\boldsymbol{g}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}[\ln p(\mathcal{R},\boldsymbol{g},\boldsymbol{Z},\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{\phi})] + \text{const} \tag{32}$$

$$= \mathbb{E}[\ln p(\boldsymbol{g}|\boldsymbol{\phi})] + \mathbb{E}[\ln p(\mathcal{R}_s|\boldsymbol{Z}_s,\boldsymbol{g})] + \text{const} \tag{33}$$

$$= \mathbb{E}\left[\sum_{i=1}^{G} g_i \ln \phi_i\right] + \mathbb{E}\left[\sum_{n=1}^{N_s}\sum_{k=1}^{P}\sum_{i=1}^{G} g_i z_{snk}\ln p(r_{sn}|h_{ik})\right] + \text{const} \tag{34}$$

$$= \sum_{i=1}^{G}\mathbb{E}[g_i]\ln\phi_i + \sum_{i=1}^{G}\mathbb{E}[g_i]\sum_{n=1}^{N_s}\sum_{k=1}^{P}\mathbb{E}[z_{snk}]\ln p(r_{sn}|h_{ik}) + \text{const} \tag{35}$$

$$= \sum_{i=1}^{G}\mathbb{E}[g_i]\left\{\ln\phi_i + \sum_{n=1}^{N_s}\sum_{k=1}^{P}\tau_{snk}\ln p(r_{sn}|h_{ik})\right\} + \text{const} \tag{36}$$

So we immediately see that

$$q^*(\boldsymbol{g}) \propto \prod_{i=1}^{G}\theta^{g_i} \tag{37}$$

where

$$\ln\theta_i = \ln\phi_i + \sum_{n=1}^{N_s}\sum_{k=1}^{P}\tau_{snk}\ln p(r_{sn}|h_{ik}) \tag{38}$$

2.6.17 Evidence

The Variational Bayes framework also gives a lower-bound on the evidence for the full posterior distribution. The lower bound is given by:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}[\ln p(\mathcal{R}, \boldsymbol{g}, \boldsymbol{Z}, \boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\phi})] - \mathbb{E}[\ln q^*(\boldsymbol{g}, \boldsymbol{Z}, \boldsymbol{\pi})] \\ &= \mathbb{E}[\ln p(\boldsymbol{g}|\boldsymbol{\phi})] + \sum_{s=1}^{S} \{\mathbb{E}[\ln p(\boldsymbol{\pi}_s|\boldsymbol{\alpha}_s)] + \mathbb{E}[\ln p(\boldsymbol{Z}_s|\boldsymbol{\pi}_s)] + \mathbb{E}[\ln p(\mathcal{R}_s|\boldsymbol{Z}_s, \boldsymbol{g})]\} \\ &\quad - \mathbb{E}[\ln q^*(\boldsymbol{g})] - \sum_{s=1}^{S} \{\mathbb{E}[\ln q^*(\boldsymbol{\pi}_s) + \mathbb{E}[\ln q^*(\boldsymbol{Z}_s)]\}\end{aligned}$$

Where each expectation is taken with respect to q^*. Evaluating each term separately

$$\mathbb{E}[\ln p(\boldsymbol{g}|\boldsymbol{\phi})] = \mathbb{E}\left[\sum_{i=1}^{P} g_i \ln \phi_i\right] = \sum_{i=1}^{P} \mathbb{E}[g_i] \ln \phi_i = \sum_{i=1}^{P} \phi_i^{post} \ln \phi_i^{prior} \tag{39}$$

$$\begin{aligned}\mathbb{E}[\ln p(\boldsymbol{\pi}_s|\boldsymbol{\alpha}_s)] &= \mathbb{E}\left[\sum_{k=1}^{P} (\alpha_k^{prior} - 1) \ln \pi_{sk} - \ln \mathrm{B}(\boldsymbol{\alpha}^{prior})\right] \\ &= \sum_{k=1}^{P} (\alpha_k^{prior} - 1)\mathbb{E}[\ln \pi_{sk}] - \ln \mathrm{B}(\boldsymbol{\alpha}^{prior}) \\ &= \sum_{k=1}^{P} (\alpha_k^{prior} - 1) \ln \tilde{\pi}_{sk} - \ln \mathrm{B}(\boldsymbol{\alpha}^{prior})\end{aligned} \tag{40}$$

$$\mathbb{E}[\ln p(\boldsymbol{Z}_s|\boldsymbol{\pi}_s)] = \mathbb{E}\left[\sum_{n=1}^{N_s}\sum_{k=1}^{P} z_{snk} \ln \pi_{sk}\right] = \sum_{n=1}^{N_s}\sum_{k=1}^{P} \mathbb{E}[z_{snk} \ln \pi_{sk}] = \sum_{n=1}^{N_s}\sum_{k=1}^{P} \tau_{snk} \ln \tilde{\pi}_{sk} \tag{41}$$

as τ and z are independent under q.

$$\mathbb{E}[\ln p(\mathcal{R}_s|\boldsymbol{Z}_s, \boldsymbol{g})] = \mathbb{E}\left[\sum_{i=1}^{G}\sum_{n=1}^{N_s}\sum_{k=1}^{P} g_i z_{snk} \ln p(r_{sn}|h_{ik})\right] = \sum_{i=1}^{G} \phi_i^{post} \sum_{n=1}^{N_s}\sum_{k=1}^{P} \tau_{snk} \ln p(r_{sn}|h_{ik}) \tag{42}$$

$$\mathbb{E}[\ln q^*(\boldsymbol{g})] = \mathbb{E}\left[\sum_{i=1}^{P} g_i \ln \phi_i\right] = \sum_{i=1}^{P} \mathbb{E}[g_i] \ln \phi_i = \sum_{i=1}^{P} \phi_i^{post} \ln \phi_i^{post} \tag{43}$$

$$\begin{aligned}\mathbb{E}[\ln q^*(\boldsymbol{\pi}_s)] =& \mathbb{E}\left[\sum_{k=1}^{P} (\alpha_k^{post} - 1) \ln \pi_{sk} - \ln \mathrm{B}(\boldsymbol{\alpha}^{post})\right] \\ &= \sum_{k=1}^{P} (\alpha_k^{post} - 1)\mathbb{E}[\ln \pi_{sk}] - \ln \mathrm{B}(\boldsymbol{\alpha}^{post}) \\ &= \sum_{k=1}^{P} (\alpha_k^{post} - 1) \ln \tilde{\pi}_{sk} - \ln \mathrm{B}(\boldsymbol{\alpha}^{post})\end{aligned} \tag{44}$$

$$\mathbb{E}[\ln q^*(\boldsymbol{Z}_s)] = \mathbb{E}\left[\sum_{n=1}^{N_s}\sum_{k=1}^{P} z_{snk} \ln \tau_{snk}\right] = \sum_{n=1}^{N_s}\sum_{k=1}^{P} \tau_{snk} \ln \tau_{snk} \tag{45}$$

These equations can also be used as a test of correctness of the implementation, as $\mathcal{L}$ should increase on each iteration.

2.6.18 Somatic genotype model

The somatic genotype model is identical to the CNV model, with a slight modification to the sample space. Specifically the genotype variable $\boldsymbol{g}$ from the CNV model is replaced with a 'somatic' genotype $\tilde{\boldsymbol{g}}$ which is simply a 1-of-$P+1$ binary random variable. A further restriction of the sample space $\tilde{\boldsymbol{g}}$ is that $h_{i(P+1)} \neq h_{ij} \forall j \leq P$, which simply states that the last (somatic) component always contains a novel haplotype (i.e. not present in the germline).

2.6.19 Bacterial genotype model

2.7 Variant callers

2.7.1 Individual

2.7.2 Population

2.7.3 Cancer

The cancer caller is designed to detect somatic mutations in a set of tumour samples from a single individual. The set of samples may also include a single normal sample, which is assumed to be tumour free.

2.7.4 Pre-processing

A set of candidate germline and 'cancer' genotypes is generated from the set of candidate haplotypes. If the number of haplotypes is strictly greater than the organism ploidy then each germline genotype will be present in the set of cancer genotypes at-least once.

Each set of genotypes may then be filtered to reduce the complexity of the model fitting step.

2.7.5 Model comparison

We then fit the individual $\mathcal{M}_g$, CNV $\mathcal{M}_c$, and somatic $\mathcal{M}_s$ models to the data. For the individual case all reads are pooled together, which as can be seen from 5 results in the same likelihood function as treating each sample separately. We then calculate the posterior probability of each model by evaluating Bayes theorem:

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{M})p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D})} \tag{46}$$

In particular the term $p(\mathcal{D}|\mathcal{M})$ is the evidence the model $\mathcal{M}$, and $p(\mathcal{M})$ is the prior probability.

2.7.6 Germline genotype calling

The probability of each germline genotype is then:

$$p(\boldsymbol{g}|\mathcal{R}) = p(\mathcal{M}_g|\mathcal{R})p(\boldsymbol{g}|\mathcal{M}_g) + p(\mathcal{M}_c|\mathcal{R})p(\boldsymbol{g}|\mathcal{M}_c) + p(\mathcal{M}_s|\mathcal{R})p(\boldsymbol{g}|\mathcal{M}_s) \tag{47}$$

where $p(\boldsymbol{g}|\mathcal{M}_s) = \sum_{\tilde{\boldsymbol{g}} s.t. \boldsymbol{g} \in \tilde{\boldsymbol{g}}} p(\tilde{\boldsymbol{g}}|\mathcal{M}_s)$.

The called genotype is then the MAP genotype according to this posterior distribution.

2.7.7 Germline variant calling

The posterior probability of each candidate variant allele a is then:

$$p(a|\mathcal{R}) = \sum_{\boldsymbol{g} s.t. a \in \boldsymbol{g}} p(\boldsymbol{g}|\mathcal{R}) \tag{48}$$

Germline candidates are called if the posterior is above a user defined threshold, and if the candidate is present in the called germline genotype. If the candidate is not above the user-defined threshold, it is added to a list of candidate somatic candidates.

2.7.8 Somatic allele calling

Noting that $\mathcal{M}_s$ defines a probability distribution of probabilities over each component of each samples 'cancer genotype', we define the probability that a particular sample contains a somatic haplotype, $\tilde{h}$, as:

$$p(\tilde{h}_s|\mathcal{M}_s) = p(\tilde{h}_s > c_s|\mathcal{M}_s) = \int_c^1 \text{Beta}(\alpha_{P+1}, \sum_{i=0}^{P} \alpha_i) \tag{49}$$

where c_s is some user-defined threshold. The idea being somatic haplotypes with most probability mass less than c_s are considered noise.

We can then calculate the probability that a somatic haplotype is present in any of the samples as:

$$p(\tilde{h}|\mathcal{M}_s) = 1 - \prod_s 1 - p(\tilde{h}_s|\mathcal{M}_s) \tag{50}$$

If this probability is greater than some user-defined threshold, we proceed to call somatic mutations.

2.8 Local phasing of called sites

2.9 Variant filtering

2.10 VCF output

3 Results

3.1 Germline variants from WGS

3.1.1 Data

Whole genome germline calls were made using sequencing data derived from the well studied individual NA12878. We found four sequencing libraries for NA12878, three high coverage and one low coverage, for evaluation:

1. High coverage Illumina reads from the 1000G phase 3 project, Aligned reads were downloaded from `ftp.1000genomes.ebi.ac.uk:/vol1/ftp/phase3/data/NA12878/high_coverage_alignment/NA12878.mapped.ILLUMINA.bwa.CEU.high_coverage_pcr_free.20130906.bam`.

2. Low coverage Illumina reads from the 1000G phase 3 project. Aligned reads were downloaded from `ftp.1000genomes.ebi.ac.uk:/vol1/ftp/phase3/data/NA12878/alignment/NA12878.mapped.ILLUMINA.bwa.CEU.low_coverage.20121211.bam`.

3. High coverage Illumina reads from the Illumina Platinum Genomes project. Raw FASTQ files were downloaded from `https://storage.cloud.google.com/genomics-public-data/platinum-genomes/fastq/ERR194147_1.fastq.gz?_ga=2.121231755.-2019438195.1508521573` and `https://storage.cloud.google.com/genomics-public-data/platinum-genomes/fastq/ERR194147_2.fastq.gz?_ga=2.134461938.-2019438195.1508521573`.

4. High coverage Illumina X Ten reads. Raw FASTQ files were downloaded from `https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R1.fastq.gz` and `https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R2.fastq.gz`.

3.1.2 Calling pipelines

3.1.3 Evaluation

3.2 Germline variants from exome-capture

3.3 De novo mutations in parent-offspring trios

3.4 Somatic mutations in paired tumour-normal samples

3.5 Somatic mutations in tumour only samples

3.6 Bacterial calling using HTS data

4 Appendix

4.1 Variational Bayes

Given a probability model $p(\boldsymbol{X}, \boldsymbol{Z})$ where $\boldsymbol{X}$ is observed and $\boldsymbol{Z}$ are latent, the true posterior density $p(\boldsymbol{Z}|\boldsymbol{X})$ can be approximated with another distribution $q(\boldsymbol{Z})$ subject to some measure of similarity. A natural choice of similarity is the KullbackLeibler divergence

$$\mathrm{KL}(q \,||\, p) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{51}$$

This measures the additional amount of information (in nats) required to generate codes from q rather than p, it satisfies $\mathrm{KL}(q \,||\, p) \geq 0$, with equality when $p = q$. So we actually try to minimise this quantity.

We now partition the latent variables $\boldsymbol{Z}$ into a set of disjoint groups denoted by $\boldsymbol{Z_i}$ where $i = 1, \ldots, M$ and assume that the q distribution factorises into a product of these groups, i.e.

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z_i}) \tag{52}$$

This is the only assumption made, in particular the functional form of each q_i is not constrained. The idea is then to optimise each group in tern, which can formally be solved using calculus of variations, but we can to see that by substituting (52) into (51) and separating one group $\boldsymbol{Z_j}$ that

$$\begin{aligned}
\mathrm{KL}(q \,||\, p) &= -\int \prod_{i=1}^{M} q_i \left\{ \ln p(\boldsymbol{Z}|\boldsymbol{X}) - \sum_i \ln q_i) \right\} d\boldsymbol{Z} \\
&= -\int q_j \left\{ \int \ln p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z_i} \right\} d\boldsymbol{Z_j} + \int q_j \ln q_j d\boldsymbol{Z_j} + \text{const} \\
&= -\int q_j \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{Z}|\boldsymbol{X})] d\boldsymbol{Z_j} + \int q_j \ln q_j d\boldsymbol{Z_j} + \text{const} \\
&= \mathrm{KL}(q_j \,||\, \exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{Z}|\boldsymbol{X})])) + \text{const}
\end{aligned} \tag{53}$$

Clearly the q_j which minimises this quantity is when $q_j = \exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{Z}|\boldsymbol{X})])$ and therefore we find the optimal q_j

$$q_j^*(\boldsymbol{Z_j}) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{Z}|\boldsymbol{X})] + \text{const} \tag{54}$$

or equivalently if $p(\boldsymbol{X})$ is absorbed into the constant then

$$q_j^*(\boldsymbol{Z_j}) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const} \tag{55}$$

Note these equations do not represent an explicit solution because they are interdependent. The variational Bayes algorithm therefore proceeds similar to EM by cycling through each group, updating q^*, and repeating until convergence. It can be shown that $\mathrm{KL}(q \,||\, p)$ decreases at each step.