GATK gCNV: accurate germline copy-number variant discovery from sequencing read-depth data

Mehrtash Babadi
mehrtash@broadinstitute.org
Data Sciences Platform
Broad Institute
Cambridge, MA, USA

Samuel K. Lee
slee@broadinstitute.org
Data Sciences Platform
Broad Institute
Cambridge, MA, USA

Andrey N. Smirnov
asmirnov@broadinstitute.org
Data Sciences Platform
Broad Institute
Cambridge, MA, USA

Abstract

We introduce GATK gCNV, a novel algorithm and tool for the discovery of rare and common copy-number variants (CNVs) from next-generation sequencing (NGS) read-depth data. In GATK gCNV, sequencing biases are modeled via negative-binomial factor analysis, and copy-number states and genomic regions of low and high CNV activity are modeled using a hierarchical hidden Markov model (HHMM). We use automatic differentiation variational inference (ADVI) and variational message passing to infer continuous and discrete latent variables in a principled framework. We further use a deterministic annealing protocol to deal with the non-convexity of the variational objective function. Inference is implemented using the PyMC3 probabilistic programming language (PPL) and Theano. We demonstrate that GATK gCNV outperforms existing tools for CNV detection.

1 Introduction

Copy-number variation (CNV) is a class of genomic *structural variation* where the *integer copy-number state* of a large genomic region (typically >1000 base pairs, spanning several consecutive exons or genes) is altered with respect to a reference genome. Inferring these copy-number states is an important problem in computational genomics, both for research applications and clinical practice [11]. Inferring integer CNV states from NGS read-depth data begins with aligning short reads (sequences of ∼100 base pairs from a sample genome) to the reference genome and counting the number of reads that align to each region. Calling CNV events from these counts is a challenging problem due to strong systematic biases, which can arise from batch effects in sample preparation and sequencing library preparation protocols, variation in sequencing efficiency across genomic regions, and numerous other hidden processes.

Many previous methods for CNV detection from read-depth data attempt to remove systematic biases via PCA denoising [1] or regression [3, 4], or try to obviate the issue by pre-clustering samples and genomic regions [2, 8]. CNVs are subsequently detected using hidden Markov models (HMM) or non-parametric change-point detection algorithms [7]. Crucially, these methods suffer from a lack of self-consistency between data normalization and event detection, which results in inadvertent removal of signal in the former and decreased sensitivity in the latter (as discussed in, e.g., [3]).

Here, we present GATK gCNV, a principled Bayesian approach for learning global and sample-specific biases of read-depth data from large cohorts while *simultaneously* inferring copy-number states. Our model combines a negative-binomial factor analysis module for learning batch effects with a hierarchical HMM (HHMM) for detecting both sample-specific CNV events and global regions of high and low CNV activity. Self-consistency between bias modeling and CNV calling greatly improves the performance of our algorithm with respect to existing methods.

2 Summary of Methods

Modeling read-depth — We seek to model the data n_{st}, the integer count of aligned reads for sample $s = 1, 2, \ldots, S$ in genomic region $t = 1, 2, \ldots, T$. We model n_{st} with a negative-binomial distribution; taking λ_{st} as the Poisson parameter and α_{st} as the Gamma distribution parameter, we may write

$$\begin{aligned} n_{st} &\sim \mathsf{NegativeBinomal}(\lambda_{st}, \alpha_{st}), \\ \lambda_{st} &= d_s \left(c_{st}\, \mu_{st} + \varepsilon_M\right), \\ \log(\mu_{st}) &= m_t + \sum_{\nu=1}^{D} W_{t\nu} z_{\nu s} + \sum_{\nu=1}^{K} \overline{W}_{t\nu} \overline{z}_{\nu s}, \\ \log[\alpha_{st}/(1+\alpha_{st})] &= \Psi_s + \Psi_t, \end{aligned} \tag{1}$$

where $d_s \sim \mathsf{LogNormal}(\mu_d, \sigma_d)$ is the mean read-depth per copy, ε_M is a small alignment error rate, $c_{st} \in \mathbb{N}^0$ is the integer copy-number matrix, and $\mu_{st} > 0$ is the multiplicative bias matrix. We model μ_{st} in the logarithmic space using a generalized linear model:

A slightly edited version of this extended abtract was accepted for poster presentation at *PROBPROG 2018, October 04–06, 2018, Cambridge, MA, USA*. (Last edited March 22, 2019.)

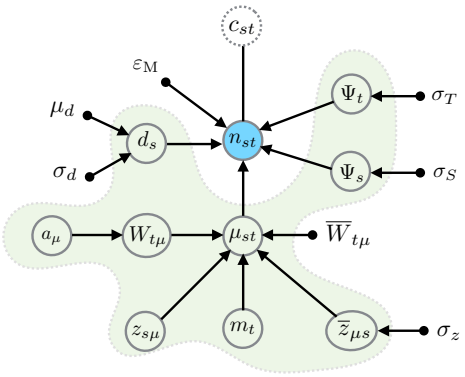

Figure 1. The read-depth model of GATK gCNV.

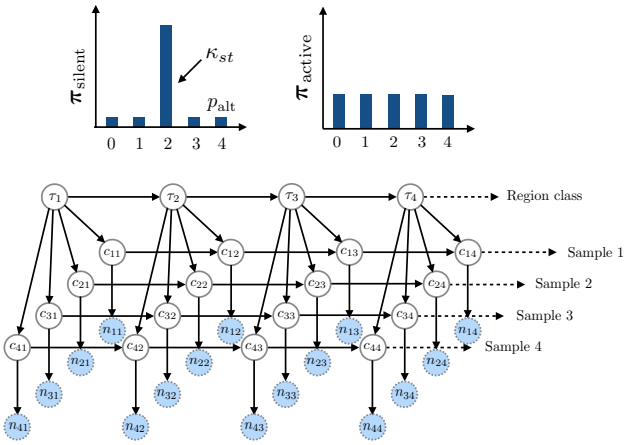

Figure 2. (top) Copy-number priors for silent and active region classes. (bottom) The hierarchical HMM of region class states and sample-specific copy-number states.

a global region-specific mean bias $m_t \sim \mathsf{N}(0, \sigma_m)$, a low-rank matrix factorization term with stochastic global bias factors $W_{t\nu} \sim \mathsf{N}(0, a_\nu^{-1})$ and sample-specific loadings $z_{\nu s} \sim \mathsf{N}(0, 1)$, and a contribution from deterministic bias factors $\overline{W}_{t\nu}$ (arising from known biological features of each genomic region) and their respective loadings $\overline{z}_{\nu s} \sim \mathsf{N}(0, \sigma_z)$. Here, a_ν denotes the automatic relevance determination (ARD) coefficients which are optimized over to select the number of bias factors in a data-driven manner. Finally, we account for the unmodeled overdispersion $\log[\alpha_{st}/(1+\alpha_{st})]$ as a sum of sample-specific $\Psi_s \sim \mathsf{Exp}(\sigma_S)$ and region-specific $\Psi_t \sim \mathsf{Exp}(\sigma_T)$ components. The graphical model for this module is shown in Fig. 1.

Modeling genomic regions and copy-numbers — Certain genomic regions are biologically more prone to CNV events. We introduce a per-region two-state categorical random variable $\tau_t \in \{\text{silent}, \text{active}\}$ to model such global regional differences. The correlation between genomically close region classes is modeled with an HMM with exponentially decaying transition probabilities:

$$p(\tau_t \to \tau_{t+1}) = \exp(-\Delta_{t,t+1}/d_\tau)\,\delta(\tau_t, \tau_{t+1}) + [1 - \exp(-\Delta_{t,t+1}/d_\tau)]\,p(\tau_{t+1}), \quad (2)$$

where $p(\tau_t)$ the prior region class probability, $\Delta_{t,t+1}$ is the genomic distance between the midpoints of region t and $t+1$, and d_τ is the typical size of regions with similar CNV activity rates.

Sample-specific copy-numbers are modeled similarly, with one Markov chain per sample; however, priors and transition probabilities are conditioned on the region class τ_t. We set a *baseline* copy-number state matrix for each sample κ_{st} in a preliminary modeling step that estimates chromosome-level copy numbers. That is, κ_{st} denotes the copy-number state in the absence of any CNV event. We use a prior copy-number probability that strongly prefers the baseline state κ_{st} in regions where $\tau_t = \text{silent}$ and we use a flat prior where $\tau_t = \text{active}$. The copy-number transition matrix is an exponentially decaying process as before, but with a different decay length d_{CNV} and per-region priors. The dependency relations of the resulting hierarchical HMM are shown graphically in Fig. 2.

Hybrid ADVI framework — The GATK gCNV model contains both continuous and discrete latent random variables (RVs). In order to leverage PPLs and automatic variational inference, we assume a factorized posterior of the form $p(\mathsf{C}, \mathsf{D}|n_{st}) \approx q(\mathsf{C})\,q(\mathsf{D})$, where $\mathsf{C} = \{m_t, W_{t\mu}, \ldots\}$ and $\mathsf{D} = \{c_{st}, \tau_t\}$ denote the set of continuous and discrete latent RVs, respectively. Subsequently, the full evidence lower bound (ELBO) admits a natural partitioning $\mathsf{ELBO} = \mathbb{E}_{\mathsf{C}\sim q(\mathsf{C})}[\mathsf{ELBO}|\mathsf{C}] + \mathbb{E}_{\mathsf{D}\sim q(\mathsf{D})}[\mathsf{ELBO}|\mathsf{D}]$. Given $q(\mathsf{D})$ and a variational ansatz for $q(\mathsf{C})$, the ELBO can be increased by performing gradient descent steps on the parameters of $q(\mathsf{C})$. Likewise, given $q(\mathsf{C})$, one can gather sufficient statistics via posterior sampling to apply Bayesian updates on $q(\mathsf{D})$; see below. We refer to this scheme as *Hybrid ADVI*.

Continuous sector: *ADVI* — We implement inference for the continuous RVs using the PyMC3 PPL [9]. We assume a fully factorized Gaussian variational posterior for $q(\mathsf{C})$, which is updated using ADVI [5]. Once partial convergence is achieved, we move on to the discrete sector. It can be shown that a sufficient statistic for a Bayesian update of $q(\mathsf{D})$ is the $q(\mathsf{C})$-averaged *log emission probability*, $\mathbb{E}_{\mathsf{C}\sim q(\mathsf{C})}[\log p(n_{st}|c_{st}, \mathsf{C})]$, which we obtain via sampling.

Discrete sector: *variational message passing* — Performing an exact Bayesian update of $q(\mathsf{D})$ is feasible via message passing; however, this has an exponential complexity in the number of per-sample Markov chains (S). Assuming $\mathsf{S} \gg 1$, we expect the posterior $q(\tau_t)$ to become sharp, resulting in an effective decoupling of τ_t (parent chain) and c_{st} (child chains). Therefore, we assume an ansatz $q(\mathsf{D}) \approx q(\tau_t) \prod_{s=1}^{S} q_s(c_{st})$, neglecting correlations between τ_t and c_{st} *and* between different child chains, but retaining genomic correlations along each chain. This allows us to update $q(\tau_t)$ and $q_s(c_{st})$ using the standard forward-backward algorithm, although we must use mean-field effective prior and transition probabilities that need to be self-consistently determined. For brevity, we omit the derivation of the iterative scheme, but note that self-consistency is achieved quickly after 2-3 rounds of forward-backward updates with relaxation. The complexity of our variational treatment of this HHMM is $\mathcal{O}(STC^2)$, where C is the number of allowed integer copy-number states; crucially, this is linear in S.

Marginalized warm-up and deterministic annealing — The complete inference scheme involves interleaving updates of $q(\mathsf{C})$ and $q(\mathsf{D})$. In practice, we found that the non-convexity of ELBO yielded spurious local minima with poor initialization. To alleviate this issue, we utilize two techniques: (1) We initialize $q(\mathsf{C})$ and $q(\mathsf{D})$ by *approximately* marginalizing all discrete RVs, and obtain the first estimate of $q(\mathsf{C})$ from the marginalized model. (2) In the spirit of the deterministic annealing expectation-maximization (DA-EM) algorithm [10], we apply entropic regularization to both the continuous and the discrete RVs. In brief, we initially encourage high-entropy variational posteriors by replacing the standard ELBO with $\mathsf{ELBO}(\beta) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \beta q(\mathbf{z})]$, where $\beta \geq 1$ is the inverse temperature and is slowly annealed during learning. A similar recipe applies to the variational message passing, where all prior, transition, and emission probabilities are scaled by β^{-1} in log space and renormalized. Our PyMC3 implementation of DA-ADVI is provided with GATK [6].

3 Benchmarking

Finally, we benchmark GATK gCNV using a cohort of whole-exome sequencing (WES) blood-normal samples. We use a manually validated and FDR-controlled callset obtained from matched whole-genome sequencing (WGS) samples using Genome STRiP [2] as the truth callset. To compare, we also include the CNV calls obtained using two popular CNV calling tools, XHMM [1] and CODEX [3], in the benchmark. The results are shown in Fig. 3. We find that GATK gCNV yields nearly 20% higher sensitivity and 50% higher specificity over all CNV events compared to the other methods. Although not presented here, benchmarks that stratify by truth variant frequency also show that GATK gCNV performs favorably on common CNV events. More extensive benchmarking results demonstrate even further improvement, primarily resulting from GATK gCNV hyperparmeter optimization using additional WGS truth callsets, and will be presented elsewhere.

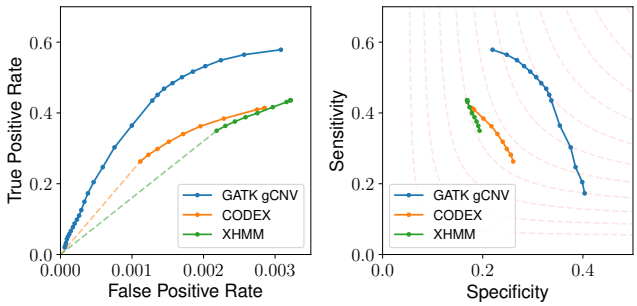

Figure 3. The performance of GATK gCNV, XHMM [1], and CODEX [3] in detecting CNV events from WES data against a matched WGS callset obtained using Genome STRiP [2]. Performance metrics are reported per WES target. All curves are parameterized by call quality and the dashed red isoclines (right) indicate constant F_1 score.

References

[1] Menachem Fromer, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker, Steven A. McCarroll, Michael C. O'Donovan, Michael J. Owen, George Kirov, Patrick F. Sullivan, Christina M. Hultman, Pamela Sklar, and Shaun M. Purcell. 2012. Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. *American Journal of Human Genetics* 91, 4 (Oct. 2012), 597–607. https://doi.org/10.1016/j.ajhg.2012.08.005

[2] Robert E. Handsaker, Vanessa Van Doren, Jennifer R. Berman, Giulio Genovese, Seva Kashin, Linda M. Boettger, and Steven A. McCarroll. 2015. Large multiallelic copy number variations in humans. *Nature Genetics* 47, 3 (March 2015), 296–303. https://doi.org/10.1038/ng.3200

[3] Yuchao Jiang, Derek A. Oldridge, Sharon J. Diskin, and Nancy R. Zhang. 2015. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Research* 43, 6 (March 2015), e39. https://doi.org/10.1093/nar/gku1363

[4] Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arné Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research* 40, 9 (May 2012), e69. https://doi.org/10.1093/nar/gks003

[5] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. 2017. Automatic Differentiation Variational Inference. *J. Mach. Learn. Res.* 18, 1 (Jan. 2017),

430–474. http://dl.acm.org/citation.cfm?id=3122009.3122023

[6] Aaron Henrik McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark Depristo. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* (July 2010), gr.107524.110. https://doi.org/10.1101/gr.107524.110

[7] Adam B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)* 5, 4 (Oct. 2004), 557–572. https://doi.org/10.1093/biostatistics/kxh008

[8] Jonathan S. Packer, Evan K. Maxwell, Colm O'Dushlaine, Alexander E. Lopez, Frederick E. Dewey, Rostislav Chernomorsky, Aris Baras, John D. Overton, Lukas Habegger, and Jeffrey G. Reid. 2016. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics* 32, 1 (Jan. 2016), 133–135. https://doi.org/10.1093/bioinformatics/btv547

[9] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2 (April 2016), e55. https://doi.org/10.7717/peerj-cs.55

[10] Naonori Ueda and Ryohei Nakano. 1998. Deterministic annealing EM algorithm. *Neural Networks* 11, 2 (March 1998), 271–282. https://doi.org/10.1016/S0893-6080(97)00133-0

[11] Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics* 10 (2009), 451–481.