araCNA: Somatic copy number profiling using long-range sequence models

Ellen Visscher[1] and Christopher Yau[1,2]

[1]Nuffield Department for Women's & Reproductive Health, University of Oxford, Oxford, UK
[2]Health Data Research UK, London, UK

Abstract

Somatic copy number alterations (CNAs) are hallmarks of cancer. Current algorithms that call CNAs from whole genome sequenced (WGS) data have not exploited deep learning methods owing to computational scaling limitations. Here, we present a novel deep-learning approach, araCNA, trained only on simulated data that can accurately predict CNAs in real WGS cancer genomes. araCNA uses novel transformer alternatives (e.g Mamba) to handle genomic-scale sequence lengths ($\sim$1M) and learn long-range interactions. Results are extremely accurate on simulated data, and this zero-shot approach is on par with existing methods when applied to 50 WGS samples from the cancer genome atlas. Notably, our approach requires only a tumour sample and not a matched normal sample, has fewer markers of overfitting, and performs inference in only a few minutes. araCNA demonstrates how domain knowledge can be used to simulate training sets that harness the power of modern machine learning in biological applications.

Introduction

Somatic copy number alterations (CNAs) are genomic regions that are amplified or deleted when somatic cells replicate. They are a hallmark of many cancers and a driving factor of tumorigenesis (Zhang and Pellman, 2022), driving the amplification of oncogenes (Kim et al., 2020; Rosswog et al., 2021). CNA abundance is known to be associated with disease stage, prognosis and response to treatment, and particular cancer types can be characterised by specific classes of structural variation, often in specific chromosomal regions (Hadi et al., 2020). Accurate CNA profiles of cancer samples are important for downstream analysis such as association studies to identify underlying cancer signatures or as a prognostic biomarker (Steele et al., 2022).

The copy number landscape of a tumour can be profiled using array- and sequencing-based technologies where increases and decreases in signal intensity or sequence read depth correspond to gains and loss of genomic segments (Nakagawa et al., 2015; Pinkel and Albertson, 2005). Allelic information from single nucleotide polymorphisms (SNPs) can also be exploited to determine copy number (see **Figure 1**, **Figure 2**). Data can then be processed using CNA calling algorithms to derive copy number states.

CNA calling algorithms generally consist of an approach to segment the genome into regions of constant copy number and a copy number calling or classification step (into deletion, duplication, etc). Methods adopt a variety of approaches. This includes calling total copy numbers only (Chiang et al., 2009; Ivakhno et al., 2010; Ha et al., 2012; Xi et al., 2016) while others provide allele-specific major and minor copy number calling capabilities (Van Loo et al., 2010; Nik-Zainal et al., 2012). Others adopt a hybrid approach, first calling total copy numbers and

then assigning these to major/minor copy numbers (Boeva et al., 2012; Talevich et al., 2016). Approaches that account for tumour heterogeneity can also provide sub-clonal copy numbers (Carter et al., 2012; Oesper et al., 2013; Fischer et al., 2014; Ha et al., 2014), again sometimes calling allele-specific copy numbers (Shen and Seshan, 2016; McPherson et al., 2017; Cun et al., 2018; Zaccaria and Raphael, 2020). When multiple tumour samples are available from the same individual, information from across samples can be used to posit evolutionary trees upon which somatic point mutations and CNAs can be placed (Satas et al., 2021; Myers et al., 2024).

Standard CNA callers adopt the use of segmentation algorithms or hidden Markov models which are effective at processing one cancer sequencing sample at a time but the low information capacity of these methods or models mean they cannot *learn* (Janiesch et al., 2021; Mohamed et al.). CNA callers process every sample as though it were the very first sample they have seen. Therefore, despite large-scale mapping exercises such as The Pan-Cancer Genome Atlas studies (Steele et al., 2022), few new CNA calling approaches have been developed.

While many areas of 'omics analysis have been heavily influenced by deep learning in recent years, the technology has had relatively little impact on CNA calling algorithms. For example, the use of transformer-based models has led to the creation of massive foundation models for single-cell and integrative 'omics (Szałata et al., 2024). However, the quadratic scaling limits of transformers have meant that these same principles are often not applicable to genome-based applications, like CNA calling, where distances between genomic regions of interest might be vast (Consens et al., 2023). One of the few examples is ECOLE, (Mandiracioglu et al., 2024), which is designed for whole exome sequencing (WES), which uses a transformer architecture to classify exome regions into neutral, deletion, and duplication, etc. Since WES has lower data resolution as it captures only the coding portion of the genome, ECOLE tackles the simpler task of classifying exome regions into these broad categories and does call the exact copy number, or allele-specific copy numbers.

In this paper, we propose a novel deep learning-based approach to the CNA calling problem, which we call `araCNA`, and is based on the use of recently developed transformer alternatives that provide genome-scale modeling capabilities. While existing copy number callers can be considered to be complex signal segmentation algorithms, `araCNA` *learns* to call copy number alterations and other features of sequencing data obtained from cancer samples. We describe how to train such a model and provide empirical evidence of its performance. The approach we suggest therefore opens up new opportunities for creating CNA callers that learn and can be fine-tuned for specific sub-tasks of interest, or integrated directly with deep learning models for other data modalities.

Materials & Methods

Model overview

The input data for `araCNA` is assumed to be a sequence of allele-specific read counts at genome-wide loci (see **Figure 1A**) which can be converted into total read depth (R) and B allele frequency (B) values as is commonplace for copy number callers. These are fed into a long-range sequence model which converts the input sequence into an output sequence consisting of the probabilities of major and minor copy number values at corresponding loci. Global parameters of interest such as tumour purity or ploidy can also be provided (**Figure 1B**).

Mathematical Preliminaries

We first define the mathematical construction, similar to that first shown in Van Loo et al. (2010). We define $C_{T,i}$ as the total copy number at locus i in the tumour, $C_{P,i}$ as the copy

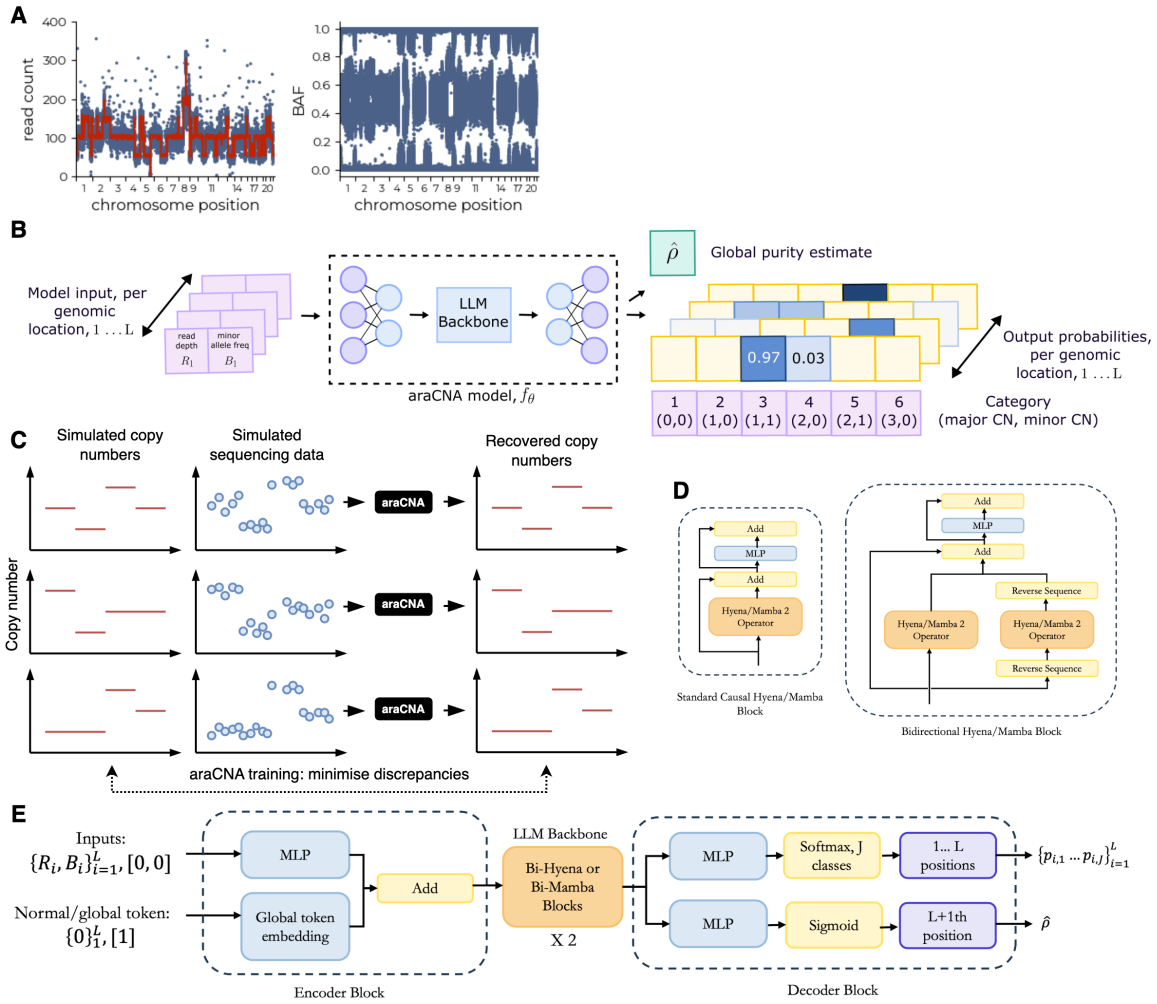

Figure 1: Overview of `araCNA` model. The (A) input data (read count and B allele frequency) is (B) converted by `araCNA` into a sequence of probabilistic copy number calls and global parameter estimates. (C) `araCNA` is trained using simulated data. (D) Bi-directional variants of the causal Hyena or Mamba blocks are used in the (E) `araCNA` model architecture.

number of the paternal chromosome at locus i, $C_{M,i}$ as the copy number of the maternal chromosome at locus i and $C_{B,i}$ is the B allele copy number. We further define ρ as the purity of the tumour sample (the proportion of tumour vs non-tumour) and r_d as the expected number of sequencing reads per copy number. In practice, samples are sequenced up to a given average read depth, which assumes a uniform coverage of the short reads across the whole genome, (Sims et al., 2014). When many duplicated regions exist, the actual average read depth per copy number r_d is unknown. Finally, we assign R_i to mean the total number of reads at a locus and B_i as the B allele frequency (BAF). The sequence data is therefore a collection $\{R_i, B_i\}_{i=1}^L$ for L loci.

For a pure tumour sample, we have the total copy number as

$$C_{T,i} = C_{P,i} + C_{M,i}.$$

Considering sample impurity, we define the sample copy number C_T^s as

$$C_{T,i}^s = \rho C_{T,i} + 2(1 - \rho),$$

where we assume the contaminating normal cells have copy number 2 at all loci. While normal cells may possess some copy number variants, the size of these regions are typically negligible

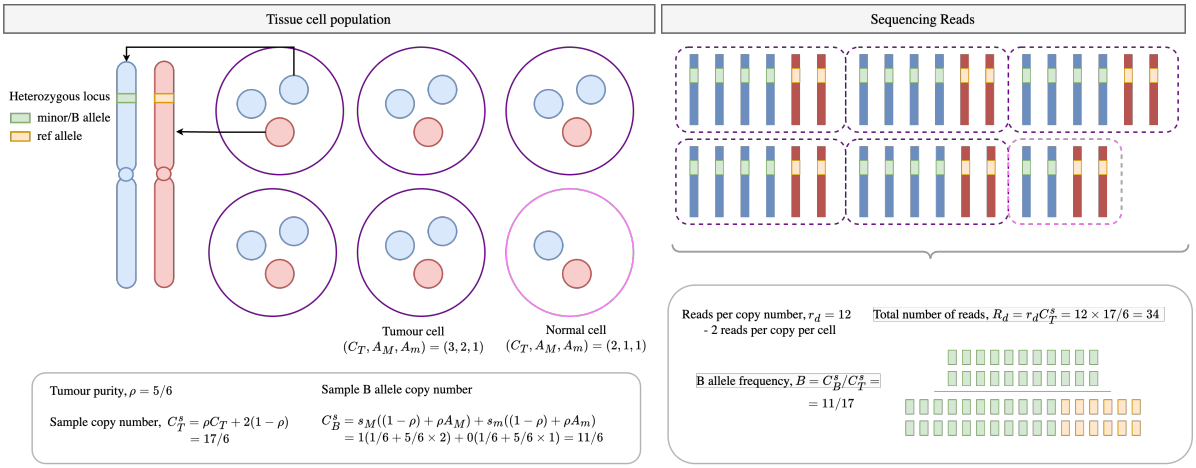

Figure 2: Mathematical construction of copy number calling. Illustration of how purity, copy number, read depth per-copy number and heterozygous loci result in measured read depth and B allele frequency.

compared to the cancer-associated alterations we aim to detect and so we ignore these for simplicity.

The B allele copy number is defined as:

$$C_{B,i} = s_{p,i} C_{P,i} + s_{m,i} C_{M,i},$$

where $(s_{p,i}, s_{m,i}) \in \{(0,0), (0,1), (1,0), (1,1)\}$ denotes whether the paternal and maternal chromosomes respectively have the specified SNP B allele at locus i.

Adding sample impurity we have:

$$C_{B,i}^s = s_{p,i}((1-\rho) + \rho C_{P,i}) + s_{m,i}((1-\rho) + \rho C_{M,i})$$

The total number of reads at a locus R_i are then:

$$R_i = r_d C_{T,i}^s$$

While the B allele frequency (BAF) B_i at locus i is given by:

$$B_i = \frac{C_{B,i}^s}{C_{T,i}^s}$$

Both R_i and B_i are measured data, obtained from sequencing a tumour sample. Figure 2 illustrates the relationship from $(\rho, r_d, \{C_{P,i}, C_{M,i}\}) \rightarrow \{R_i, B_i\}$. Our aim is to do the reverse- to infer $(\rho, r_d, \{C_{P,i}, C_{M,i}\})$ from $\{R_i, B_i\}$. However, we can see in the above formulation, that both R_i and B_i remain the same if we were to swap the values for $C_{P,i}$ and $C_{M,i}$, meaning that we are unable to infer the parental copy numbers - they are non-identifiable. We can instead introduce, $A_{M,i}, A_{m,i} = \max(C_{P,i}, C_{M,i}), \min(C_{P,i}, C_{M,i})$ as the major/minor allele-specific copy numbers, which are identifiable. Hence, we amend our aim to infer $(\rho, r_d, \{A_{M,i}, A_{m,i}\})$ from $\{R_i, B_i\}$.

The araCNA model

The function of `araCNA` can be summarised as $f_\theta(\{R_i, B_i\}) \rightarrow (\{p_{i,k}\}, \hat{\rho})$ where f_θ is the long-range sequence model parameterised by network weights θ. $\hat{\rho}$ is `araCNA`'s global purity

estimate. While $p_{k,i}$ is the probability that `araCNA` assigns the locus as belonging to copy number profile K_j. The profile categories $K_1, \ldots, K_J$ correspond to major/minor parental copy number combinations, with $K_1 := (A_M = 0, A_m = 0)$, $K_2 := (A_M = 1, A_m = 0)$ etc. Here, $J = \sum_{i=0}^{T_{\max}} \min(i+1, T_{\max} - i + 1)$, where $T_{\max}$ is a hyper-parameter corresponding to the maximum modelled total parental copy number. From these, we can also estimate $\hat{r}_d = \mu_{\text{robust}}(R)/(\mathbb{E}[\hat{C}^s])$, where $\mu_{\text{robust}}(R)$ is the robust or trimmed mean of the read depth vector and $\mathbb{E}[\hat{C}^s]$ is the expected value of the overall sample copy number (ploidy).

We trained our model using simulated datasets where the ground truth copy numbers and purity are known. The details of this simulation are discussed later in this section. The loss function consists of a supervised sequence loss and a supervised global loss. The supervised sequence loss is the cross entropy:

$$\mathcal{L}_{ss} = -\frac{1}{L} \sum_{i=1}^{L} \sum_{j=1}^{J} I\{c_i = K_j\} \log(p_{k,i})$$

where L is the sequence length, $c_i \in K_1 \ldots K_J$ is the known target profile of a genomic locus. The supervised global parameter losses are:

$$\mathcal{L}_{sr} = |r_d - \hat{r}_d|,$$
$$\mathcal{L}_{s\rho} = |\rho - \hat{\rho}|.$$

The total loss is then given by:

$$\mathcal{L} = \mathcal{L}_{ss} + \lambda_r \mathcal{L}_{sr} + \lambda_\rho \mathcal{L}_{s\rho}.$$

We found $\lambda_r = \lambda_\rho = 1$ to work well. The model was trained by iteratively increasing the complexity of the simulated data.

Model Architecture

Sequence inputs are entered into `araCNA` together with global placeholders, and projects these through an encoder block, backbone and decoder block to give the copy number and purity outputs, **Figure 1C**. Elements used as global estimates (i.e. purity) in the output are appended to the end of the normal sequence, with zero values for read depth and minor allele frequency.

To differentiate these elements from normal elements, we included a global token feature, encoded as 0 for normal elements and 1 for global prediction sequence elements. The normal observational data (read depth and minor allele frequency), is projected into a dimension d using a multi-layer perception (MLP) with ReLU non-linearities, while the token feature is projected into dimension d using a simple embedding. The two inputs are combined in this projected space, to serve as the encoder block for the main backbone of the model. The output from the main backbone diverges into two streams in the decoder- one for prediction of the copy numbers, and one for prediction of the global parameters. Both streams use MLPs with ReLU activation, with a softmax to project the copy number stream into a probability over the J possible classes, and a sigmoid to project the purity estimate between 0 and 1.

We created two implementations of `araCNA` which use two recently developed long-range sequence models, Hyena (Poli et al., 2023) and Mamba (Gu and Dao, 2024; Dao and Gu, 2024), as part of its architecture. These long-range sequence models known as state space like models (SSMs), were developed to achieve similar results to transformers but avoid using the computationally expensive self-attention mechanism, where the number of operations scales quadratically with sequence length and therefore is prohibitive for genome-scale data (Consens et al., 2023). In contrast, Mamba and Hyena-based models offer training times that scale linearly

with sequence length. We call the `araCNA` models with Mamba/Hyena blocks `araCNA`-mamba and `araCNA`-hyena respectively. Mamba-2 is the more popular of the SSM models, but being constrained to GPUs with Nvidia A100 architecture (for training and inference). We adapt the blocks to be bidirectional rather than causal (unidirectional) as used in natural language processing (**Figure 1E**).

The hyperparameters of `araCNA`-mamba and `araCNA`-hyena are outlined in the **Supplementary Information**, resulting in parameter counts of 70K and 12.09M respectively. Each hyena block contains two positional encodings of length $L_{\max}$ corresponding to a hyperparameter of the maximum input sequence length. For our models, $L_{\max}$ was set to 1M, hence the majority of the parameters, 12M, can be attributed to the positional encodings. We note that 70K parameters for `araCNA`-mamba is relatively small for a modern neural network, especially given the task at hand is not trivial (Habib and Qureshi, 2022).

Smoothing

Given the noise in real data, many copy number calling methods employ smoothing or thresholding in some way to reduce the number of CNA segments and reduce false positives. For example, ASCAT has a 'penalty' parameter that controls smoothing (Van Loo et al., 2010), Battenberg has several 'gamma' hyperparameters (Nik-Zainal et al., 2012), HMMCopy has 'e' and 'strength' parameters (Ha et al., 2012) and CNV Kit uses thresholding to segment and call different copy numbers (Talevich et al., 2016). We also employ a smoothing technique, based on the predicted output probabilities.

Here, we wish to find the 'best' joint sequence $\mathbf{S} \in \{1, \dots J\}^L$ given the model output probabilities, $p_{k,i}$, that a locus, i, is in copy number state k. We can find the optimal sequence by balancing the probabilistic evidence of being in a given state, with the penalty λ_t of transitioning to a different state as:

$$\hat{\mathbf{S}} = \arg\min_{\mathbf{S}} \left[-\sum_{i=1}^{L} \sum_{k=1}^{J} I\{S_i = k\} \log p_{k,i} + \lambda_t \sum_{i=1}^{L-1} I\{S_i \neq S_{i+1}\} \right]$$

For a given λ_t, we solve this using a dynamic programming approach similar to the Viterbi algorithm used in HMMs for finding the most likely sequence of hidden states. We found $\lambda_t = 500$ to be suitable.

Synthetic data simulation

We generated synthetic copy number profiles using the following procedures:

1) Sampling the number of segments. We sample the approximate number of copy number segments, $\hat{N}_s$ using a mixture approach; first, we sample a uniform variable, u such that under a user-defined swap probability, q_s, the number of segments is sampled uniformly between 1 and N. When $u > q_s$, a Poisson distribution is used to skew sampling towards smaller total segments. This is to oversample harder cases with fewer segments where it is harder to estimate global parameters like read depth per copy number and purity.

2) Sampling the segment breakpoints. This is done by randomly sampling $b_1, \dots, b_{\hat{N}_s}$ breakpoints from $1 \dots L$, the unique set of these breakpoints defines the segments, and $N_s = |b_1, \dots, b_{\hat{N}_s}|$. We only keep the segments that have a minimum segment length of $L_{\min}$.

3) Sampling the segment profiles. We sample A_M, A_m of each segment from the possible copy number profiles. We inject logic here to preferentially sample profiles closer in copy number to 1-1 when there are fewer segments. This is due to the identifiability issue. When there are

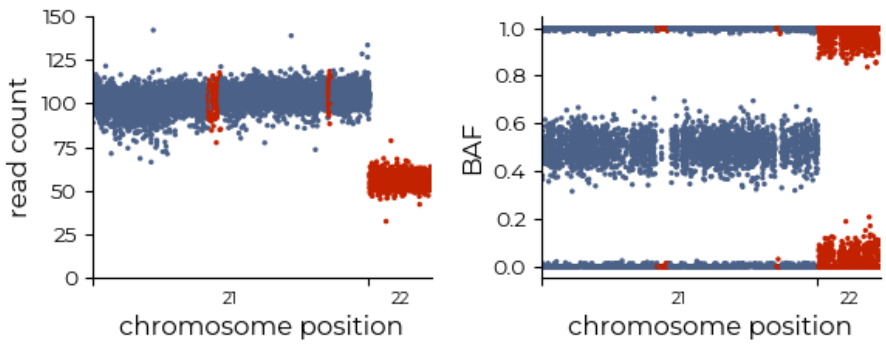

Figure 3: Real sample example illustrating prolonged homozygous regions on chromosome 21, where the BAF noise is very low. In contrast, the region on chromosome 22 indicates a minor copy number of 0, and hence although the BAF profile is close to 1 and 0, the noise profile is larger, where normal contamination at heterozygous loci brings the BAF closer to 0.5. Red colouring indicates regions of interest.

more segments, profiles are sampled more uniformly but still with a preference for lower copy numbers, to inject an implicit bias towards lower ploidy solutions when the model is unsure.

From a sampled profile, we simulate the sequencing read depth and B allele frequency data. Each of the L loci is considered a commonly varying single nucleotide polymorphism, SNP. For both parental alleles, A_M, A_m, we sample each SNP as binomial with a probability of 0.5, we also sample the purity, ρ, uniformly from a range between 0.5 and 1. This gives the sample minor allele copy number, $C^s_{B,i}$ and the sample total copy number, $C^s_{T,i}$. The read depth per copy number, r_d is sampled uniformly between 5 and 70, and together with $C^s_{T,i}$ the overall read depth, R_i is sampled from this mean with additional noise. The BAF, B_i is sampled using total reads sampled based on R_i and the subset of B-allele reads using a binomial probability of $C^s_{B,i}/C^s_{T,i}$, with added noise.

In real data, there exist regions of prolonged homozygosity that can be attributed to identity-by-descent (IBD) regions (i.e identical regions inherited from both parents due to a common ancestor), Figure 3. To emulate this, we also randomly inject regions of prolonged homozygosity into the model. In these IBD regions, the BAF cannot be used to infer copy-number, and the model must use context from before/after the homozygous region for correct prediction.

From this sampling procedure, we therefore have a set of targets $(\{A_{M,i}, A_{m,i}\}, \rho, r_d)$ that generate inputs $\{R_i, B_i\}$, which together are used in the training of `araCNA`. Hence, `araCNA` can be interpreted as performing inference on the above statistical approach, when $(\{A_{M,i}, A_{m,i}\}, \rho, r_d)$ are treated as unknowns. Further details of this procedure are included in **Supplementary Information**.

Training Procedure

For this work, we trained `araCNA` using simulated copy number profiles. We elected to follow this approach since (i) there exists no ground-truth, high-resolution copy number profiles upon which `araCNA` could be trained and (ii) we avoid using copy number profiles produced by other methods in order to perform fair comparisons later on. For the simulations, we set the maximum considered total copy number $T_{\max}$ as a hyperparameter, we then do the following:

1. Sample the approximate number of segments, $\hat{N}_s$ in the total sequence.
2. Sample the segment breakpoints, that gives the actual number of segments N_s in the total sequence.
3. Sample the copy number profiles for each segment, $A_{M,i}, A_{m,i}$ (model targets).

4. Sample the sequence data, R_i, B_i (model input) across all segments, based on sampled purity, ρ read depth per copy number, r_d and copy numbers, $A_{M,i}, A_{m,i}$.

Noise and tumour purity parameters are sampled randomly during training allowing `araCNA` to see different sequence data distributions and learn how to process these into copy number calls. As a consequence, when used at test-time for zero-shot inference, it is not necessary to retrain `araCNA` to process any tumour sample unlike conventional CNA callers which need to be adapted to each sample every time.

To train `araCNA` we adopted an iterative warmup procedure, gradually increasing the complexity of the problem. We found this was necessary for the model to learn, and a similar approach was taken in Nguyen et al. (2023) with gradually increasing the sequence length.

The training procedure was:

1. Begin the synthetic data generating procedure with $\rho = 1$, and without sampling the noise parameters. Use only up to a maximum total copy number of 2, that is profiles, $(A_M, A_m) \in \{(0,0),(1,0),(1,1),(2,0)\}$. Sample r_d, and start with sequence length 10000. Train until convergence.
2. Using the previously trained model weights as initialisation, add in purity and noise parameter sampling. Train until convergence.
3. Using the previously trained model weights as initialisation, slowly increase the maximum total copy number to 8. Train until convergence.
4. Using the previously trained model weights as initialisation, slowly increase the maximum sequence length to 650,000. Train until convergence.

Metrics

We define the metrics used for reconstruction accuracy as follows. The BAF root mean square error (RMSE) is calculated as:

$$\sqrt{\frac{1}{L}\sum_{i=1}^{L}(B_i - \hat{B}_i)^2}$$

$$\begin{aligned}\hat{B}_i &= \min_{s_1,s_2} \frac{\hat{C}^s_{B,i}(s_1,s_2)}{\hat{C}^s_{T,i}} \\ &= \min_{s_1,s_2} \frac{s_1((1-\hat{\rho}) + \hat{\rho}\hat{A}_{M,i}) + s_2((1-\hat{\rho}) + \hat{\rho}\hat{A}_{m,i})}{2(1-\hat{\rho}) + \hat{\rho}(\hat{A}_{M,i} + \hat{A}_{m,i})}\end{aligned}$$

where the reconstructed BAF, $\hat{B}_i$ is calculated from the most probable haplotype $(s_1, s_2) \in \{(0,0),(0,1),(1,0),(1,1)\}$ at each loci.

The read depth mean absolute error (MAE) is calculated as:

$$\begin{aligned}&\frac{1}{L}\sum_{i=1}^{L}|R_i - \hat{R}_i|, \\ \hat{R}_i &= \hat{r}_d \hat{C}^s_{T,i}, \\ \hat{r}_d &= \frac{\mu_{\text{robust}}(R)}{\frac{1}{L}\sum_{i=1}^{L}\hat{C}^s_{T,i}}.\end{aligned}$$

We opted to use absolute error here, due to mapping errors that can lead to read depth and hence RMSE inflation. Here, $\mu_{\text{robust}}(R)$ is the robust or trimmed mean of the data, excluding data at the highest/lowest 5% of values.

Battenberg may assign a fraction of each segment to a subclone with a different set of copy numbers (Nik-Zainal et al., 2012). Hence, for the Battenberg multiclonal reconstruction, we account for the fraction attributed to some sub-clone, for that segment. This gives an extra set of parameters for each loci/segment; τ_i, the fraction of the sample at that segment attributed to clone 1, $A^1_{M,i}, A^1_{m,i}$, as well as $(1-\tau_i)$, $A^2_{M,i}, A^2_{m,i}$ for clone B. We therefore adapt the reconstruction to follow the Battenberg formulation (Nik-Zainal et al., 2012), and in their implemented code (`https://github.com/Wedge-lab/battenberg/`), where the sample copy numbers for each locus is given by:

$$C^s_{B,i} = s_1((1-\rho) + \rho(\tau_i A^1_{M,i} + (1-\tau_i)A^2_{M,i}) + s_2((1-\rho) + (\tau_i A^1_{m,i} + (1-\tau_i)A^2_{m,i}),$$
$$C^s_{T,i} = 2(1-\rho) + \rho(\tau_i(A^1_{M,i} + A^1_{m,i}) + (1-\tau_i)(A^2_{M,i} + A^2_{m,i})),$$

The analysis for the non-multiclonal Battenberg reconstruction uses the copy number predictions of the clone estimated as having the highest proportion (i.e $\max(\tau_i, 1-\tau_i)$ at each locus, with the normal formulation.

The concordance, CC between two methods, 1 and 2, is calculated according to:

$$CC_{\text{major}} = \frac{1}{L}\sum_i^L I(A^1_{M,i} = A^2_{M,i})$$
$$CC_{\text{minor}} = \frac{1}{L}\sum_i^L I(A^1_{m,i} = A^2_{m,i})$$
$$CC_{\text{both}} = \frac{1}{L}\sum_i^L I(A^1_{M,i} = A^2_{M,i} \text{ and } A^1_{m,i} = A^2_{m,i})$$
$$CC_{\text{total}} = \frac{1}{L}\sum_i^L I(A^1_{M,i} + A^1_{m,i} = A^2_{M,i} + A^2_{m,i})$$
$$= \frac{1}{L}\sum_i^L I(C^1_{T,i} = C^2_{T,i})$$

For methods like HMMCopy, only C_T is output, so only CC_{total} can be measured. When method 1 is the ground truth, like in simulations, this can be interpreted as an accuracy.

Results

Simulation Study

We first compared results from our two `araCNA` variants (`araCNA`-mamba and `araCNA`-hyena) using simulated data, sampled from the same generating procedure from which the training data was also sampled. We show results from 100 simulated test genomes, with a maximum sampled sequence length of 650k, to emulate the data size of a real SNP dataset. **Figure 4A** shows both models achieve high copy number classification accuracy for the task though `araCNA`-mamba slightly outperforms `araCNA`-hyena. Since in this case, we know the ground truth, we can directly measure accuracy. However, in real data, there is no known ground truth, so we also include the reconstruction error for the two models. **Figure 4C** shows unsupervised metrics for performance, where better methods are likely to have a lower BAF and read depth reconstruction. When examining tumour ploidy and purity estimates (**Figure 4B**) and reconstruction

performance (**Figure 4C**), the differences are small though `araCNA`-mamba does achieve slightly better metrics than `araCNA`-hyena. **Figure 4D,E** show the reconstructed read depth and BAF and predicted copy numbers from each model for one of the example simulated genomes.

The Hyena and Mamba models were trained using the same underlying simulation and training procedure. Differences in results can therefore be traced to differences in the model architecture or sensitivity to hyperparameters. We found that Hyena often took longer to converge during training, becoming stuck in local minima, likely due to the size of the model (12M parameters). In contrast, the Mamba model (70K parameters) achieved better performance despite its significantly smaller size and need for an A100 GPU.

The Cancer Genome Atlas

Next, we compared `araCNA` to a number of existing CNA calling tools with whole genome sequencing data from a selection of 50 tumour samples chosen from the colorectal (CRC), breast (BRCA) and ovarian (OV) cancer cohorts of The Cancer Genome Atlas (TCGA). We chose these three cancer types due to the extensive and well-characterised aneuploidy present in tumours of these types. We utilised two of the most well-used algorithms that can call allele-specific copy numbers: ASCAT, (Van Loo et al., 2010), and an adaption of ASCAT called Battenberg, (Nik-Zainal et al., 2012), which can also model sub-clonal populations, by optionally assigning a fraction of each CNA segment to another subclone. We compared results to Battenberg without and with its default sub-clonality modeling.

We also compared to CNVkit, (Talevich et al., 2016), which works with WGS although being intended primarily for WES and gives allele-specific copy numbers after first calling total copy numbers. Finally, we report results for HMMCopy, which only calls total copy number (Ha et al., 2012). Both ASCAT and CNVkit performed well in a recent benchmark (Masood et al., 2024), while Battenberg and HMMCopy are popular methods included in other method benchmarks (Zaccaria and Raphael, 2020; Alkodsi et al., 2015). Details of their implementation can be found in the **Supplementary Information**.

Since there are no ground truth copy number profiles for these tumours, we used proxy measures such as reconstruction error as in the simulation study to provide an unsupervised metric for performance, where better methods will be expected to have low reconstruction error. However, it is also important to consider both the number of segments and the range of modeled copy numbers produced by each CNA calling approach. A low reconstruction error accompanied by a large number of highly variable copy number segments may suggest overfitting, while high copy number calls at high read depth regions will have a lower reconstruction error but could be less plausible. **Figure 5** illustrates the analysis of a TCGA ovarian cancer sample using all methods.

Figure 6A shows the distribution of the root mean-squared error (RMSE) and mean absolute deviation (MAE) for the reconstruction of the B allele frequency and read depth over the fifty samples. Using both `araCNA` variants for zero-shot inference of the copy number states gives comparable reconstruction performance to the existing CNA calling methods while using similar numbers of segments. While ASCAT and Battenberg are able to achieve slightly improved reconstruction error performance, when we examined the top five percent of total copy number calls produced, we found that ASCAT and Battenberg sometimes assigned as high as 100 copies to localised genomic regions. It is difficult to verify the correctness of such calls given the lack of ground-truth, however, we found most of these regions contained few or no known cancer or other functional genes (**Supplementary Figure SI-1**).

Further, Battenberg's multi-clonal formulation uses the same input data with more degrees of freedom to model clonal fractions (see **Methods**). In the presence of clones, we would expect

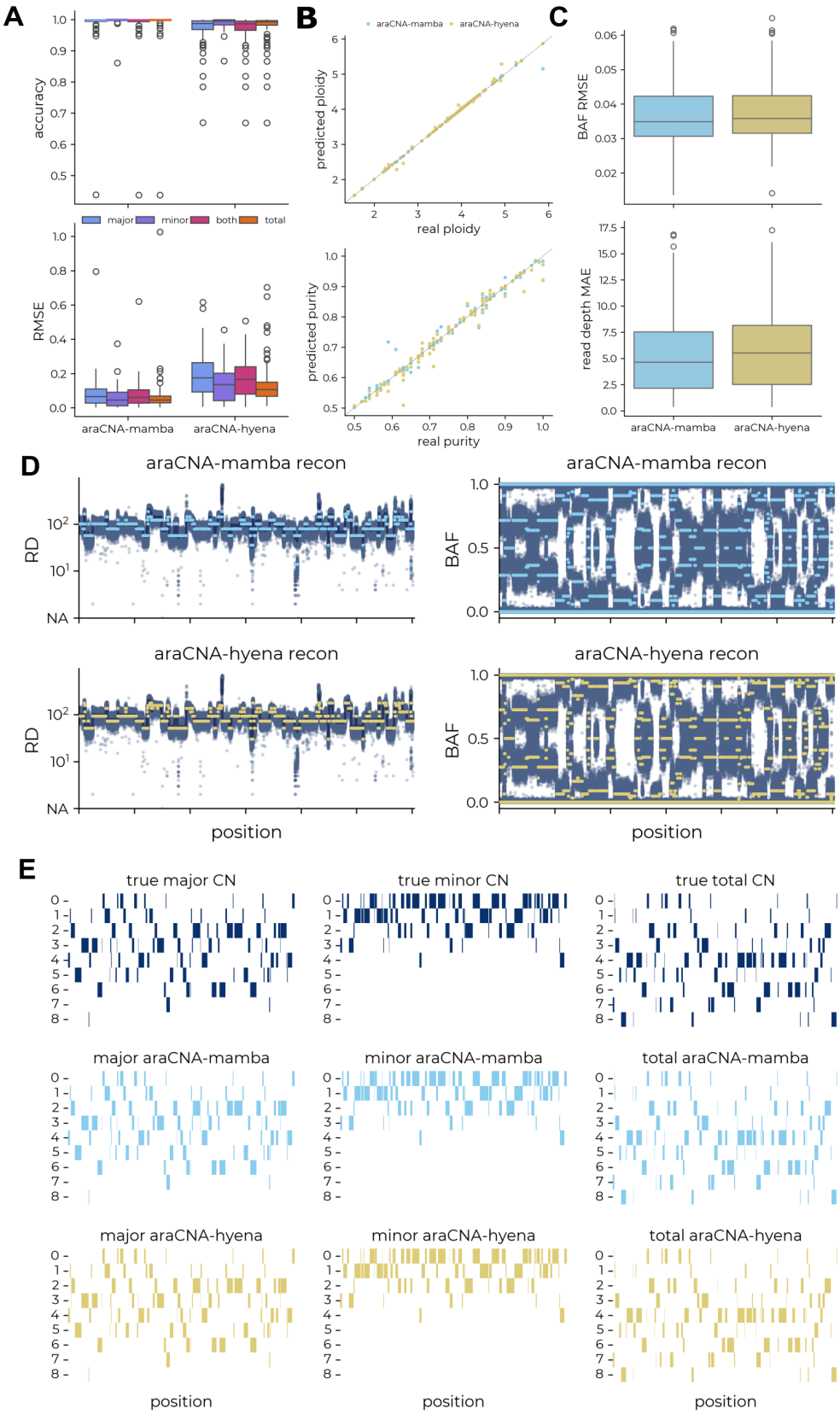

Figure 4: Results for `araCNA` model variants on simulated data. A-C show aggregated results across 100 simulated test genomes. (A) The distribution of concordance and RMSE between the predicted and true copy numbers. (B) The predicted ploidy and purity against the true purity and ploidy. (C) The distribution of mean reconstruction error: BAF root mean squared error (RMSE) and read-depth mean absolute error (MAE). (D) The reconstructed read depth and BAF and E) the underlying predicted copy number segments for `araCNA`-mamba and `araCNA`-hyena on an example simulated test set.

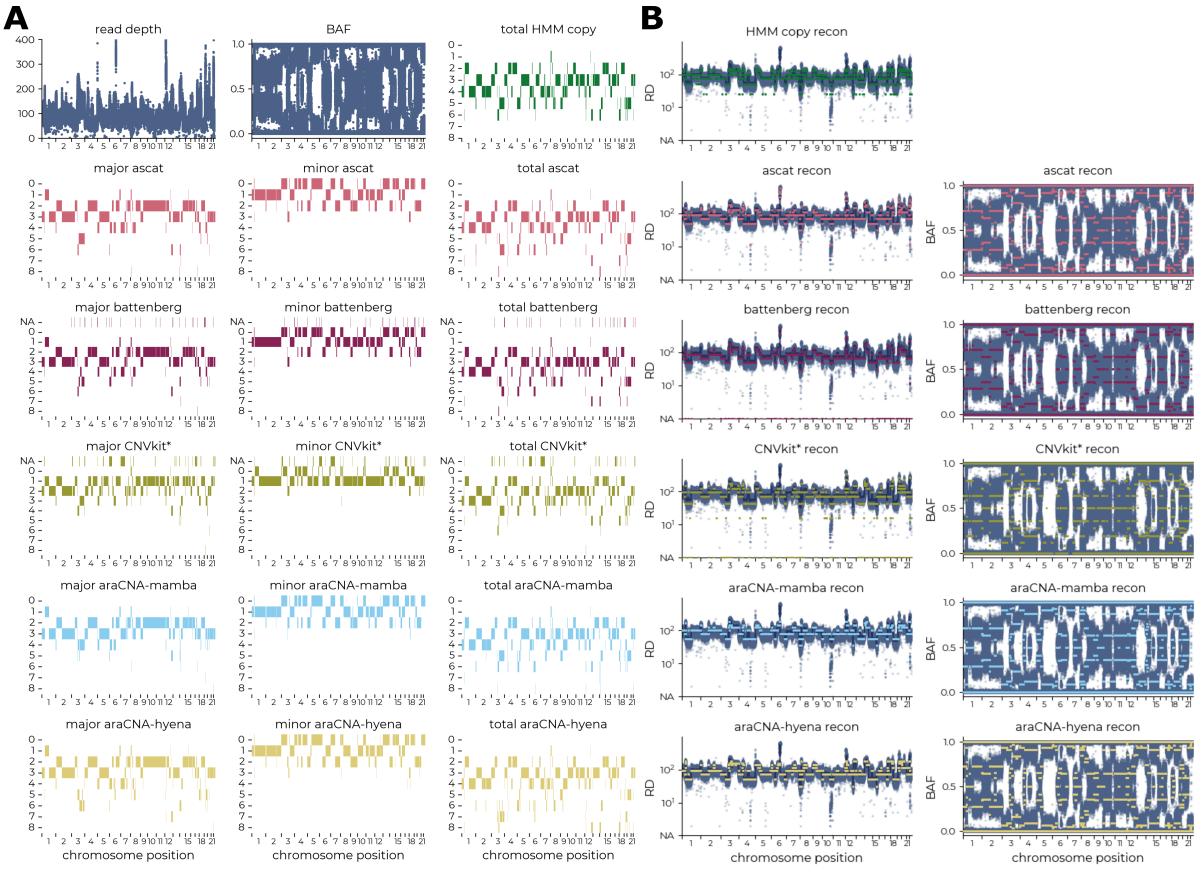

Figure 5: Representative TCGA ovarian cancer sample. Each row shows A) the called copy numbers and B) respective reconstruction of each caller, where `araCNA` has approx. 80% concordance with ASCAT, Battenberg and HMM Copy.

to see a concentration of fraction values around discrete modes, corresponding to the fraction of the sample originating from a subclone with a diverging copy number pattern across the genome.

We found that in most tumours there were no clear modes in the distribution of clonal fraction estimates, **Supplementary Figure SI-2**, indicating overfitting through this additional degree of freedom, to give the observed lower reconstruction error, **Figure 6A**.

We next looked at pairwise copy number classification concordance and the root mean squared difference between copy number calls from different methods (**Figure 6B**). `araCNA`-mamba achieves a median of 80% concordance with ASCAT for all copy number calls across the 50 TCGA cancer samples and a median of 70% concordance with Battenberg while `araCNA`-hyena was substantially less concordant with ASCAT and Battenberg. However, when the discrepancies are measured using root mean squared difference, the average differences were much less than one and we found that most discrepancies were due to small numerical differences (e.g. $1 \leftrightarrow 2, 2 \leftrightarrow 3$, etc). When we further examined the differences between ASCAT and `araCNA`, we found that these appeared to be primarily driven by differences in major and total copy number calls in some samples.

To further investigate these discrepancies, we compared tumour ploidy and purity estimates using ASCAT as a baseline. While tumour purity estimates were well correlated between all methods, tumour ploidy estimates differed between methods for some tumours (**Figure 6C**). Interestingly, although `araCNA` is giving a different copy number profile for some samples compared to ASCAT, it still gives low reconstruction error. We surmised that for some tumours

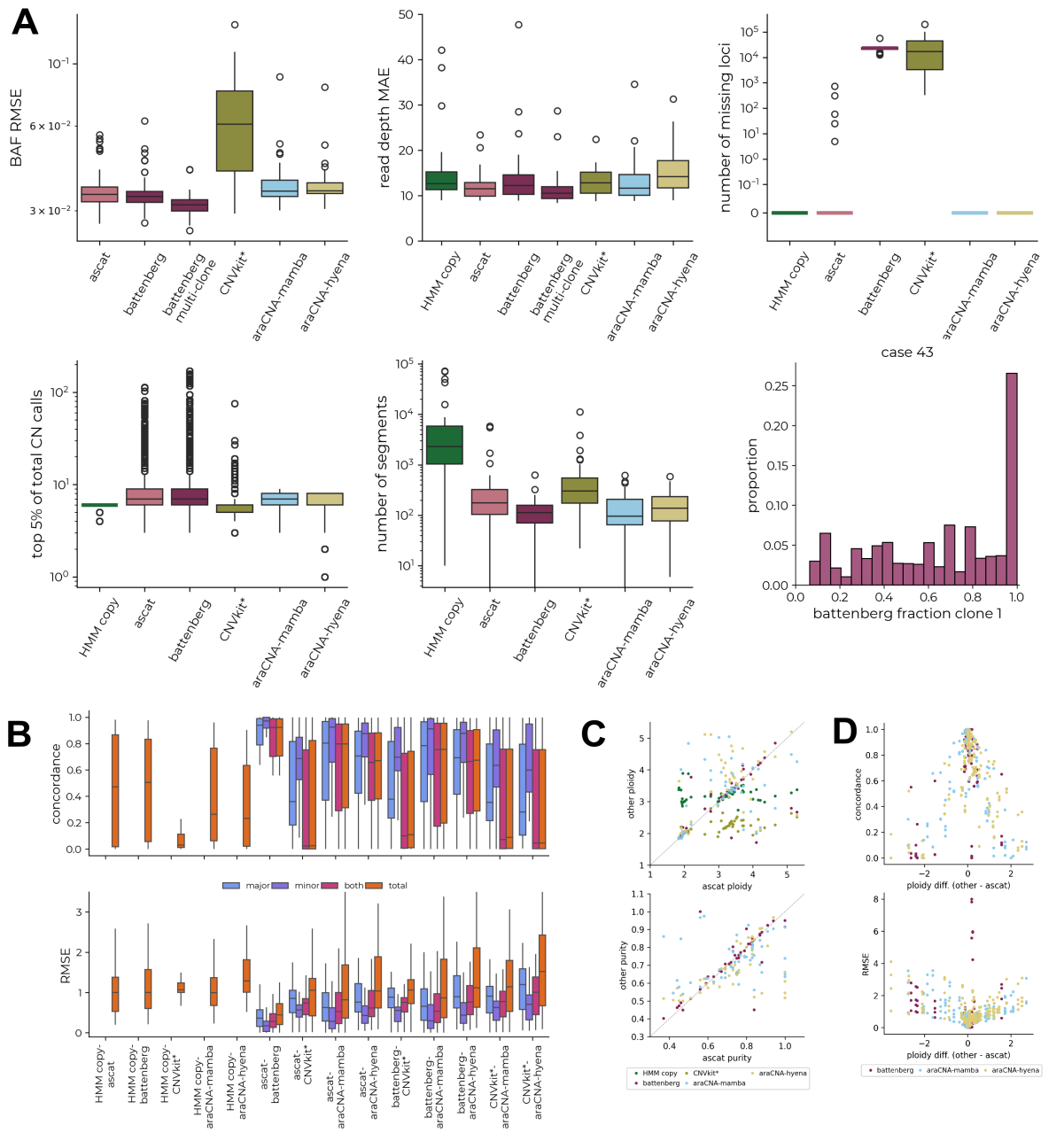

Figure 6: Comparison between CNA callers across 50 TCGA cancer samples (A) Boxplots showing the B-allele-frequency root-mean-squared reconstruction error (RMSE), read-depth mean absolute error (MAE), number of missing loci after calling, copy number distribution of top 5% of copy number calls and the number of different copy number segments identified across the tumour samples. Example distribution of Battenberg clonal fraction for a particular tumour. (B) The distribution of concordance and RMSE between the copy number predictions of each method. (C) The predicted tumour purity and ploidy against the ASCAT-derived purity and ploidy. (D) Concordance and RMSE of predicted copy numbers against the ploidy difference between methods and ASCAT.

there is weak identifiability (see **Figure 7**) and `araCNA` identifies a different ploidy state to ASCAT but with a corresponding copy number profile that is still compatible with the observed data. Indeed when we examined the concordance between `araCNA` and Battenberg with ASCAT, we see a reduction when there is a significant departure from the ploidy state estimated by ASCAT (**Figure 6D**). Interestingly, where there are disagreements, tumours classified as near-triploid by ASCAT will generally be classified in a higher ploidy state by `araCNA` and vice-versa which is a hallmark of the identifiability issue. There also exists a subset of tumours

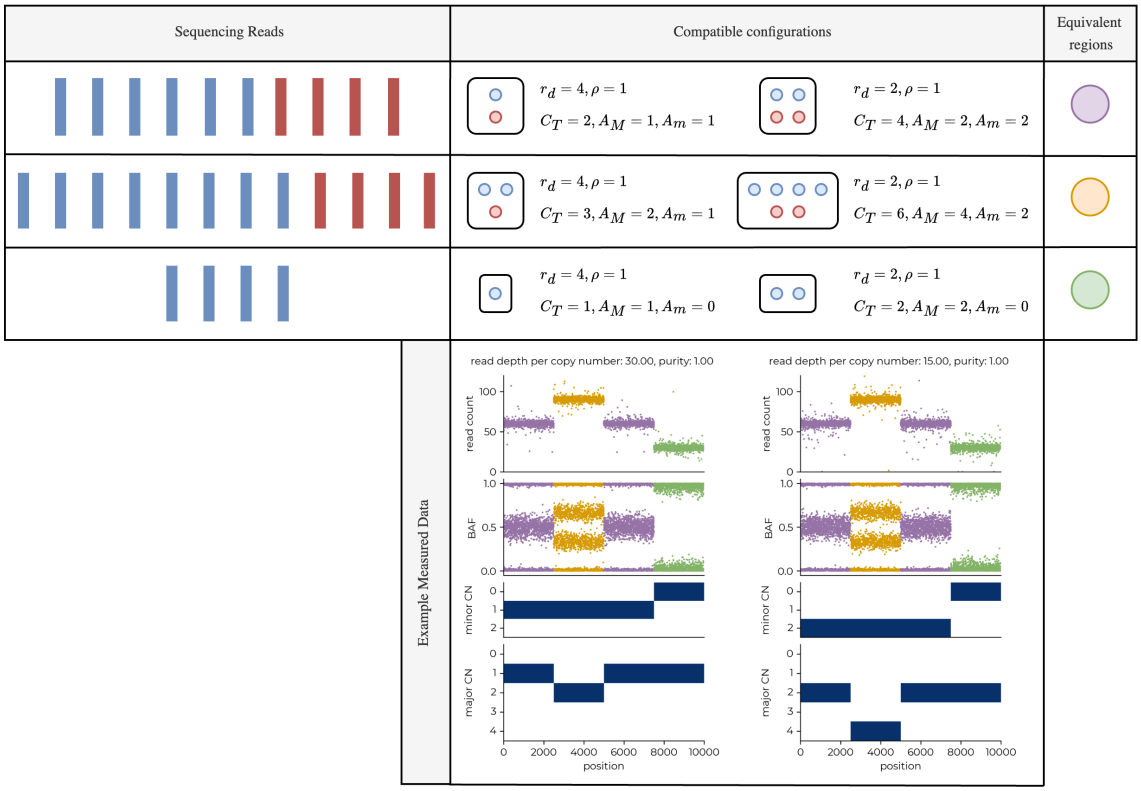

Figure 7: Identifiability of copy number profiles. Example illustrating how different copy number profiles can result in the same observed sequencing read depth and B-allele frequency measures. Here, each copy number profile on the right is exactly double that of the left, while the read depth per copy number value is half the left. Blue and red reads are those from the paternal/maternal chromosome respectively, where the B-allele frequency will reflect their chromosome ratio at heterozygous loci. This combination results in the exact same measured data, as highlighted in the bottom panel, where the colours link to each configuration profile above. Parameters are defined in **Methods**. In general, many combinations of copy number profile, tumour purity and read depth per copy number parameters may give rise to similar observed data. In the presence of noise, segments that may uniquely determine the copy number profile can be missed.

where ASCAT and Battenberg disagreed on tumour ploidy despite their algorithmic similarities which further highlights sensitivity of methods to the identifiability issue.

Overall, while `araCNA` was only trained with simulated data only and applied using zero-shot inference, it appears to perform as well as standard CNA callers on real tumour sequencing data with the caveat that there is a limit to our ability to quantify performance due to a universal lack of ground-truth data. This suggests that this form of

Discussion

Our results have shown that it is possible to develop a novel deep-learning approach to predict copy number alterations from whole genome sequencing. We have built a training simulation framework to generate supervised training data for this biological problem that otherwise would have no known ground truth. We applied our model using zero-shot inference to TCGA cancer samples and found that `araCNA` achieves comparable performance to existing methods despite being trained only on simulated datasets. This highlights the utility of having strong mechanistic models that relate copy numbers to sequencing data. Overall, these results demonstrate how we can use domain knowledge to develop simulation training sets for simulation-based inference in biological applications, where computing the posterior may be intractable or computationally

intensive (Cranmer et al., 2020).

We have taken advantage of recent advancements in deep learning, such as SSM-like models, that allow for extremely long sequence lengths, on the order of genomic data, and to learn long-range interactions for fast CNA inference. This approach is not limited to CNA calling, and could be repurposed for inference in many biological applications where simulations are already used. Our experiments with Mamba and Hyena highlight that the choice of architecture can be important and performance can vary between models for a given application.

Importantly, once constructed it is possible to further refine and retrain a model like with additional data. Unlike classic CNA callers, `araCNA` can continue to learn and improve using standard deep learning techniques such as fine-tuning or transfer learning. For instance, these approaches maybe used to adapt the model due to some distributional shift between the simulated training data used in training and the actual sequencing data. This might occur if, for example, there was additional sequencing noise and artefacts from sequencing of FFPE tumour samples.

While we limited ourselves in this work to training on simulated data, it is possible to use CNA profiles from existing copy number callers for training. However, since these would not be ground-truth data, this would result in a model which is an *emulator* of the existing copy number caller. While emulators could not normally exceed the original CNA caller in terms of performance, they can offer usability advantages since, by pre-training the emulator, it can be applied directly at run-time without retraining on each specific sample. There may also be possibilities for producing ensemble callers by using outputs from multiple existing CNA callers for training.

Our approach could be further improved by explicitly modelling patterns of sequencing variation, due to mappability, amplification or platform sequencing issues. It could also be fine-tuned on real genomes, where profiles have been curated and validated using external data sources (e.g flow cytometry for ploidy estimation, or single cell whole genome sequencing). Further, our modelling approach could likely be extended to integrate sub-clonal calling and holistically account for clone-specific mutations and clone-specific copy numbers. It could likely also be adapted to include multiple longitudinal samples from the same patient, or for low-quality data or different data modalities (e.g SNP array).

A challenge we have highlighted is the issue of identifiability when multiple copy number profiles could explain the same measured sequencing data and the correct copy number profile might be non-identifiable as a result. Multiple tumour sampling in space and/or time, such as used in TracerX (Al Bakir et al., 2023), can alleviate the problem by increasing the probability of unique solutions but may also introduce issues since different evolutionary trajectories could also produce similar patterns of CNAs. In `araCNA` we have addressed this by biasing the training simulations toward lower complexity (ploidy) explanations as is often done with standard CNA callers. We have found that without such bias training data in which there exists a one-to-many relationship between input and output causes `araCNA` to output the average CNA over compatible solutions. Further work is required to enable multiple distinct CNA profiles to be produced.

While deep learning approaches in genomics have previously been limited by a lack of labeled ground truth data and sequence length (Consens et al., 2023), our model `araCNA` demonstrates how simulation-based learning in tandem with emerging deep-learning architectures can address both these issues. We show how `araCNA` performs on par with traditional genomic models despite requiring only a tumour sample and not a matched normal sample, fewer markers of overfitting, and performing inference in only a few minutes. On top of this, `araCNA` has the ability to learn and adapt using fine-tuning or transfer learning.

Acknowledgements

CY is supported by an UKRI Turing AI Acceleration Fellowship (Ref: EP/V023233/1) and EPSRC grant (Ref: EP/Y018192/1). EV is supported by the Oxford EPSRC Centre for Doctoral Training in Health Data Science (Ref: EP/S02428X/1).

Data Availability

The latest source codes and tutorials are available on Github (urlhttps://github.com/e-vissch/araCNA). A permanent archive of the software is also available at Figshare (10.6084/m9.figshare.29063492).

References

Cheng-Zhong Zhang and David Pellman. Cancer Genomic Rearrangements and Copy Number Alterations from Errors in Cell Division. *Annual Review of Cancer Biology*, 6(Volume 6, 2022):245–268, April 2022. ISSN 2472-3428. doi: 10.1146/annurev-cancerbio-070620-094029. URL https://www.annualreviews.org/content/journals/10.1146/annurev-cancerbio-070620-094029. Publisher: Annual Reviews.

Hoon Kim, Nam-Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D. Gujar, Jens Luebeck, Jihe Liu, Viraj Deshpande, Utkrisht Rajkumar, Sandeep Namburi, Samirkumar B. Amin, Eunhee Yi, Francesca Menghi, Johannes H. Schulte, Anton G. Henssen, Howard Y. Chang, Christine R. Beck, Paul S. Mischel, Vineet Bafna, and Roel G. W. Verhaak. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nature Genetics*, 52(9):891–897, September 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-0678-2. URL https://www.nature.com/articles/s41588-020-0678-2. Publisher: Nature Publishing Group.

Carolina Rosswog, Christoph Bartenhagen, Anne Welte, Yvonne Kahlert, Nadine Hemstedt, Witali Lorenz, Maria Cartolano, Sandra Ackermann, Sven Perner, Wenzel Vogel, Janine Altmüller, Peter Nürnberg, Falk Hertwig, Gudrun Göhring, Esther Lilienweiss, Adrian M. Stütz, Jan O. Korbel, Roman K. Thomas, Martin Peifer, and Matthias Fischer. Chromothripsis followed by circular recombination drives oncogene amplification in human cancer. *Nature Genetics*, 53(12):1673–1685, December 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00951-7. URL https://www.nature.com/articles/s41588-021-00951-7. Publisher: Nature Publishing Group.

Kevin Hadi, Xiaotong Yao, Julie M. Behr, Aditya Deshpande, Charalampos Xanthopoulakis, Huasong Tian, Sarah Kudman, Joel Rosiene, Madison Darmofal, Joseph DeRose, Rick Mortensen, Emily M. Adney, Alon Shaiber, Zoran Gajic, Michael Sigouros, Kenneth Eng, Jeremiah A. Wala, Kazimierz O. Wrzeszczyński, Kanika Arora, Minita Shah, Anne-Katrin Emde, Vanessa Felice, Mayu O. Frank, Robert B. Darnell, Mahmoud Ghandi, Franklin Huang, Sally Dewhurst, John Maciejowski, Titia de Lange, Jeremy Setton, Nadeem Riaz, Jorge S. Reis-Filho, Simon Powell, David A. Knowles, Ed Reznik, Bud Mishra, Rameen Beroukhim, Michael C. Zody, Nicolas Robine, Kenji M. Oman, Carissa A. Sanchez, Mary K. Kuhner, Lucian P. Smith, Patricia C. Galipeau, Thomas G. Paulson, Brian J. Reid, Xiaohong Li, David Wilkes, Andrea Sboner, Juan Miguel Mosquera, Olivier Elemento, and Marcin Imielinski. Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. *Cell*, 183(1):197–210.e32, October 2020. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2020.08.006. URL https://www.cell.com/cell/abstract/S0092-8674(20)30997-1. Publisher: Elsevier.

Christopher D. Steele, Ammal Abbasi, S. M. Ashiqul Islam, Amy L. Bowes, Azhar Khandekar, Kerstin Haase, Shadi Hames-Fathi, Dolapo Ajayi, Annelien Verfaillie, Pawan Dhami, Alex McLatchie, Matt Lechner, Nicholas Light, Adam Shlien, David Malkin, Andrew Feber, Paula Proszek, Tom Lesluyes, Fredrik Mertens, Adrienne M. Flanagan, Maxime Tarabichi, Peter Van Loo, Ludmil B. Alexandrov, and Nischalan Pillay. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, June 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04738-6. URL https://www.nature.com/articles/s41586-022-04738-6. Publisher: Nature Publishing Group.

H. Nakagawa, C. P. Wardell, M. Furuta, H. Taniguchi, and A. Fujimoto. Cancer whole-genome sequencing: present and future. *Oncogene*, 34(49):5943–5950, December 2015. ISSN 1476-5594. doi: 10.1038/onc.2015.90. URL https://www.nature.com/articles/onc201590. Publisher: Nature Publishing Group.

Daniel Pinkel and Donna G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37(6):S11–S17, June 2005. ISSN 1546-1718. doi: 10.1038/ng1569. URL https://www.nature.com/articles/ng1569. Publisher: Nature Publishing Group.

Derek Y. Chiang, Gad Getz, David B. Jaffe, Michael J. T. O'Kelly, Xiaojun Zhao, Scott L. Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S. Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*, 6(1):99–103, January 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1276. URL https://www.nature.com/articles/nmeth.1276. Publisher: Nature Publishing Group.

Sergii Ivakhno, Tom Royce, Anthony J. Cox, Dirk J. Evers, R. Keira Cheetham, and Simon Tavaré. CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, 26(24):3051–3058, December 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq587. URL https://doi.org/10.1093/bioinformatics/btq587.

Gavin Ha, Andrew Roth, Daniel Lai, Ali Bashashati, Jiarui Ding, Rodrigo Goya, Ryan Giuliany, Jamie Rosner, Arusha Oloumi, Karey Shumansky, Suet-Feung Chin, Gulisa Turashvili, Martin Hirst, Carlos Caldas, Marco A. Marra, Samuel Aparicio, and Sohrab P. Shah. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research*, 22 (10):1995–2007, October 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.137570.112. URL https://genome.cshlp.org/content/22/10/1995. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Ruibin Xi, Semin Lee, Yuchao Xia, Tae-Min Kim, and Peter J Park. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Research*, 44(13):6274–6286, July 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw491. URL https://doi.org/10.1093/nar/gkw491.

Peter Van Loo, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M. Perou, Anne-Lise Børresen-Dale, and Vessela N. Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, September 2010. doi: 10.1073/pnas.1009843107. URL https://www.pnas.org/doi/full/10.1073/pnas.1009843107. Publisher: Proceedings of the National Academy of Sciences.

Serena Nik-Zainal, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L. Cooke, Jonathan Hinton, Andrew Menzies, Lucy A. Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J. Mudie, Stephen J. Gamble, Philip J. Stephens, Stuart McLaren, Patrick S. Tarpey, Elli Papaemmanuil, Helen R. Davies, Ignacio Varela, David J. McBride, Graham R. Bignell, Kenric Leung, Adam P. Butler, Jon W. Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerød, Samuel A. J. R. Aparicio, Andrew Tutt, Anieta M. Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L. Richardson, Anne-Lise Børresen-Dale, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell. The Life History of 21 Breast Cancers. *Cell*, 149(5):994–1007, May 2012. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2012.04.023. URL https://www.cell.com/cell/abstract/S0092-8674(12)00527-2. Publisher: Elsevier.

Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleier-

macher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425, February 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr670. URL https://doi.org/10.1093/bioinformatics/btr670.

Eric Talevich, A. Hunter Shain, Thomas Botton, and Boris C. Bastian. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Computational Biology*, 12(4):e1004873, April 2016. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1004873. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4839673/.

Scott L. Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W. Laird, Robert C. Onofrio, Wendy Winckler, Barbara A. Weir, Rameen Beroukhim, David Pellman, Douglas A. Levine, Eric S. Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, May 2012. ISSN 1546-1696. doi: 10.1038/nbt.2203. URL https://www.nature.com/articles/nbt.2203. Publisher: Nature Publishing Group.

Layla Oesper, Ahmad Mahmoody, and Benjamin J. Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, 14(7):R80, July 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-7-r80. URL https://doi.org/10.1186/gb-2013-14-7-r80.

Andrej Fischer, Ignacio Vázquez-García, Christopher J. R. Illingworth, and Ville Mustonen. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, 7(5):1740–1752, June 2014. ISSN 2211-1247. doi: 10.1016/j.celrep.2014.04.055. URL https://www.cell.com/cell-reports/abstract/S2211-1247(14)00373-8. Publisher: Elsevier.

Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M. Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A. Marra, C. Blake Gilks, David G. Huntsman, Jessica N. McAlpine, Samuel Aparicio, and Sohrab P. Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, November 2014. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.180281.114. URL https://genome.cshlp.org/content/24/11/1881. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Ronglai Shen and Venkatraman E. Seshan. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, 44(16):e131, September 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw520. URL https://doi.org/10.1093/nar/gkw520.

Andrew W. McPherson, Andrew Roth, Gavin Ha, Cedric Chauve, Adi Steif, Camila P. E. de Souza, Peter Eirew, Alexandre Bouchard-Côté, Sam Aparicio, S. Cenk Sahinalp, and Sohrab P. Shah. ReMixT: clone-specific genomic structure estimation in cancer. *Genome Biology*, 18(1):140, July 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1267-2. URL https://doi.org/10.1186/s13059-017-1267-2.

Yupeng Cun, Tsun-Po Yang, Viktor Achter, Ulrich Lang, and Martin Peifer. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nature Protocols*, 13(6):1488–1501, June 2018. ISSN 1750-2799. doi: 10.1038/nprot.2018.033. URL https://www.nature.com/articles/nprot.2018.033. Publisher: Nature Publishing Group.

Simone Zaccaria and Benjamin J. Raphael. Accurate quantification of copy-number aberrations

and whole-genome duplications in multi-sample tumor sequencing data. *Nature Communications*, 11(1):4301, September 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17967-y. URL https://www.nature.com/articles/s41467-020-17967-y. Publisher: Nature Publishing Group.

Gryte Satas, Simone Zaccaria, Mohammed El-Kebir, and Benjamin J Raphael. Decifering the elusive cancer cell fraction in tumor heterogeneity and evolution. *Cell systems*, 12(10): 1004–1018, 2021.

Matthew A Myers, Brian J Arnold, Vineet Bansal, Metin Balaban, Katelyn M Mullen, Simone Zaccaria, and Benjamin J Raphael. Hatchet2: clone-and haplotype-specific copy number inference from bulk tumor sequencing data. *Genome biology*, 25(1):130, 2024.

Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, September 2021. ISSN 1422-8890. doi: 10.1007/s12525-021-00475-2. URL https://doi.org/10.1007/s12525-021-00475-2.

Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. Deep Belief Networks for phone recognition.

Artur Szałata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang, and Fabian J. Theis. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02353-z. URL https://www.nature.com/articles/s41592-024-02353-z. Publisher: Nature Publishing Group.

Micaela E. Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. To Transformers and Beyond: Large Language Models for the Genome, November 2023. URL http://arxiv.org/abs/2311.07621. arXiv:2311.07621 [q-bio].

Berk Mandiracioglu, Furkan Ozden, Gun Kaynar, Mehmet Alper Yilmaz, Can Alkan, and A. Ercument Cicek. ECOLE: Learning to call copy number variants on whole exome sequencing data. *Nature Communications*, 15(1):132, January 2024. ISSN 2041-1723. doi: 10.1038/s41467-023-44116-y. URL https://www.nature.com/articles/s41467-023-44116-y. Publisher: Nature Publishing Group.

David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15 (2):121–132, February 2014. ISSN 1471-0064. doi: 10.1038/nrg3642. URL https://www.nature.com/articles/nrg3642. Publisher: Nature Publishing Group.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Re. Hyena Hierarchy: Towards Larger Convolutional Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28043–28078. PMLR, July 2023. URL https://proceedings.mlr.press/v202/poli23a.html. ISSN: 2640-3498.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, May 2024. URL http://arxiv.org/abs/2312.00752. arXiv:2312.00752 [cs].

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality, May 2024. URL http://arxiv.org/abs/2405.21060. arXiv:2405.21060 [cs].

Gousia Habib and Shaima Qureshi. Optimization and acceleration of convolutional neural networks: A survey. *Journal of King Saud University - Computer and Information Sciences*,

34(7):4244–4268, July 2022. ISSN 1319-1578. doi: 10.1016/j.jksuci.2020.10.004. URL https://www.sciencedirect.com/science/article/pii/S1319157820304845.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *Advances in Neural Information Processing Systems*, 36:43177–43201, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html.

Daniall Masood, Luyao Ren, Cu Nguyen, Francesco G. Brundu, Lily Zheng, Yongmei Zhao, Erich Jaeger, Yong Li, Seong Won Cha, Aaron Halpern, Sean Truong, Michael Virata, Chunhua Yan, Qingrong Chen, Andy Pang, Reyes Alberto, Chunlin Xiao, Zhaowei Yang, Wanqiu Chen, Charles Wang, Frank Cross, Severine Catreux, Leming Shi, Julia A. Beaver, Wenming Xiao, and Daoud M. Meerzaman. Evaluation of somatic copy number variation detection by NGS technologies and bioinformatics tools on a hyper-diploid cancer genome. *Genome Biology*, 25(1):163, June 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03294-8. URL https://doi.org/10.1186/s13059-024-03294-8.

Amjad Alkodsi, Riku Louhimo, and Sampsa Hautaniemi. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Briefings in Bioinformatics*, 16(2):242–254, March 2015. ISSN 1467-5463. doi: 10.1093/bib/bbu004. URL https://doi.org/10.1093/bib/bbu004.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, December 2020. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/doi/10.1073/pnas.1912789117. Publisher: Proceedings of the National Academy of Sciences.

Maise Al Bakir, Ariana Huebner, Carlos Martínez-Ruiz, Kristiana Grigoriadis, Thomas BK Watkins, Oriol Pich, David A Moore, Selvaraju Veeriah, Sophia Ward, Joanne Laycock, et al. The evolution of non-small cell lung cancer metastases in tracerx. *Nature*, 616(7957): 534–542, 2023.

E. S. Venkatraman and Adam B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, March 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl646. URL https://doi.org/10.1093/bioinformatics/btl646.

Eric Talevich. CNVkit: Genome-wide copy number from high-throughput sequencing, 2024. URL https://cnvkit.readthedocs.io/en/stable/index.html. Software: Version 0.9.11.