GenCHORD: A Pipeline for Genomic Rearrangement Analysis

JACK FRASER-GOVIL and RONNIE CRAWFORD, Wellcome Sanger Institute, United Kingdom
MAX STAMMNITZ and ELIZABETH MURCHISON, University of Cambridge, United Kingdom
ZEMIN NING, Wellcome Sanger Institute, United Kingdom

We present *GenCHORD*, a specialised step-detection algorithm suitable for detecting large-scale copy number variations from the coverage data of a sequenced, aligned genome. This algorithm is specifically tuned for detecting Structural Variations which arise during chromothripsis-induced cancer, and so permits a high degree of data reduction, whilst preserving pertinent features for cancer classification. We describe the underlying statistical and algorithmic model, demonstrate the ability of the to preserve, compress and encode information such that a simple Machine Learning model can classify and identify subtypes of Devil Facial Tumour Disease, a cancer affecting *Sarcophilus harrisii*, the Tasmanian Devil, and discuss how this tool may be used in future.

ACM Reference Format:
Jack Fraser-Govil, Ronnie Crawford, Max Stammnitz, Elizabeth Murchison, and Zemin Ning. 2024. *GenCHORD*: A Pipeline for Genomic Rearrangement Analysis. In *2024 8th International Conference on Computational Biology and Bioinformatics (ICCBB 2024), November 28–30, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3715020.3715051

1 INTRODUCTION

The term Genomic Rearrangements refer to large-scale Structural Variations (SVs) amongst individual genomes; including large-scale deletion, insertion, duplication and translocation of megabase-length sequences of DNA, and often lead to genetic disorders (when present in the germ cells), or cancer (when arising from somatic mutations). When a Genomically Rearranged sample is sequenced and aligned against a reference the true sequences novel adjacencies can be obscured, especially when the size of the rearranged regions are significantly larger than the read length. This is because the majority of reads will still align to the reference despite the fact that their spatial location within the genome has been altered: only those reads which bridge the unusual adjacencies contain the information that a rearrangement has occurred.

Reconstructing these rearrangements requires complex analysis, such as a haplotype phased-assembly , which is not only computationally costly, but requires the DNA be sequenced to a very high coverage, as well techniques such as Hi-C[1]. We should like to be able to identify, analyse and classify the presence of genomic rearrangements in a simpler fashion.

To do this, we leverage the knowledge that chromothripsis-induced rearrangements are often associated with highly variable copy-number variations[1, 2]. Where portions of the genome have been duplicated or deleted (generically termed 'gains' and 'losses'), the erroneous alignment leads to variations in the base-coverage reported by the alignment tool, as demonstrated in Figure 1. The locations of regions which have undergone gains and losses corresponds strongly to the edges of the regions which have been rearranged.

Authors' addresses: Jack Fraser-Govil; Ronnie Crawford, Wellcome Sanger Institute, Hinxton, United Kingdom; Max Stammnitz; Elizabeth Murchison, University of Cambridge, Cambridge, United Kingdom; Zemin Ning, Wellcome Sanger Institute, Hinxton, United Kingdom, CB10 1RQ.

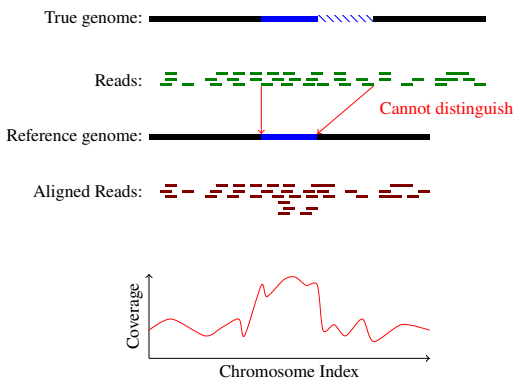

Fig. 1. A depiction of duplication leading to a higher coverage. The blue hatched region is duplicated with respect to the reference, but since the reads are much shorter than the duplicated region, they manifest as a higher coverage rate.

Many tools to analyse the copy-number variations exist, however in this work we develop **Gen**ome **C**overage **H**armonic **O**ptimiser, **R**educer and **D**e-noiser (*GenCHORD*), a novel tool tailored for this problem which leverages the power of a Bayesian hypothesis testing to infer a statistically robust denoised copy-number analysis, which is suitable for encoding into a Neural Network, and hence permit accurate detection, analysis, and classification of the genomic rearrangement.

2 STATISTICAL THEORY & ALGORITHM

The Data

Throughout this work we shall assume that our data is the coverage (per-base sequencing depth) extracted from an aligned BAM file[1]. Our data therefore takes the form of an ordered sequence of integers, $S = \{k_1, k_2, k_3, ...\}$, where k_i is a nonnegative integer corresponding to the coverage reported by the i^{th} base in the sequence.

We shall have C such sequences, corresponding to the C distinct chromosomes present in the sample. Each sequence is of length N_c (the length of the c^{th} chromosome). We shall assume that, aside from a prior that the global parameters should be similar[2], these sequences are to be analysed independently.

The values of k are distributed around a central value by a degree of noise arising from biological variation and experimental error in the pipeline. Where this central value undergoes a drastic, discontinuous change is indicative of the beginning of a gain or a loss, and hence (in the context of chromothripsis) a region which has undergone rearrangement.

2.1 Smoothing & Binning

The most obvious solution to denoise the Coverage Data is to simply pass a smoothing kernel over the data, potentially in combination with a binning algorithm, and use that to infer the underlying mean function that the data is oscillating around. Smoothing in this fashion serves as a generic way to extract a smooth curve from the extremely noisy data present, however it fails us on a number of grounds. Most importantly, this method can act to bias and manipulate the data in unforeseen and undesirable ways. Anecdotally, we found many occasions where the severity of an inferred deletion or duplication could be manipulated by altering the analysis lengthscale.

[1] such as with the samtools *depth* command
[2] But not identical, since the coverage depth can vary per-chromosome in even a healthy sample

2.2 Harmonic Fitting

We should therefore attempt to identify the coverage-discontinuities directly from the dataset. This, in essence is a form of *step detection*, a well known problem in signal processing. However, whilst there exist several out-of-the-box algorithms which might provide us with robust detections, we note that knowledge about the form of the data can be leveraged to provide a significantly more powerful and biologically meaningful inference.

2.2.1 Model Assumptions. We use the following knowledge and assertions as the underpinnings of our model.

Firstly, as argued in Appendix A, we assume that the coverage, k, is distributed according to a (slightly modified) Negative Binomial probability mass function, which is characterised by the mean μ and variation $\sigma^2_{\text{model}} = \mu + \sigma^2$, where $\sigma^2 > 0$ is the variation associated with the biological variability.

We assume that the mean μ is constant across a chromosome, except during gains or losses, where it changes discontinuously, variations in sampling frequency (such as those induced by GC bias) are accounted for by the (generous) error rate of the probability model. If we assume that the gains and losses we are identifying occur in all cells within the sample (i.e. no contamination or subclonality), then μ can only ever be integer multiples of the single-homolog coverage depth, ν, such that $\mu = q\nu$, where q is a non-negative integer equal to the multiplicity of the domain. We have a strong prior that most of the data should lie at $q = d_c$, the normal ploidy of the creature (d can vary per chromosome, for example, the sex chromosomes, or in cases where the sample is known *a priori* to be aneuploid).

We know that discontinuities must be separated by at least a distance L on the linear genome, else they would have been resolved by the alignment tool. However, we do not know where these lengths L begin or end, so a strict binning is not sufficient.

The task of detecting the 'steps', therefore, requires that we assign each base a value of q, the multiplicity (or 'harmonic', in signal processing terms) of the base: gains and losses are trivially detectable wherever this integer value changes, in doing so we should ensure that our prior knowledge and model restrictions are obeyed. In Appendix A.2 we produce a statistical model, Eq. (6) which permits us to assign a statistical score to a given set of proposed $\{q\}$ values for a genome.

2.2.2 Effects of Subclonality and Contamination. Should the biological sample contain some cells which do not contain the same mutation - either as a function of subclonality in the mutation, or contamination of the sample with healthy cells[3] - we shall see a violation of one of our assumptions: the value of μ will not lie at integer multiples of ν. When this effect is small, it does not meaningfully affect our analysis. At high degrees of contamination, however, it can blur the distinction between different values of q, and the classifier will struggle to reliably assign a harmonic since it will interpret the sample as lying halfway between two q values – something it is forbidden to include.

We model this behaviour (whilst retaining the power of the integer-assumption) by making a simplifying assumption: that the degree of contamination is small and constant across the chromosome, and always tends to bias the coverage back towards the expected ploidy of the sample. If the degree of contamination is η, then we find that the expected 'contamination-biased frequency' is

$$\begin{aligned} \mu'_q &= [(1-\eta)q + \eta d_c]\,\nu \\ &= q'\nu \qquad \Longleftrightarrow \qquad q' = (1-\eta)q + \eta d_c \end{aligned} \tag{1}$$

The effects of subclonality can therefore modelled by everywhere replacing q with $q'(q, \eta)$.

[3]We shall term both mechanisms 'contamination', as they manifest similarly in the data

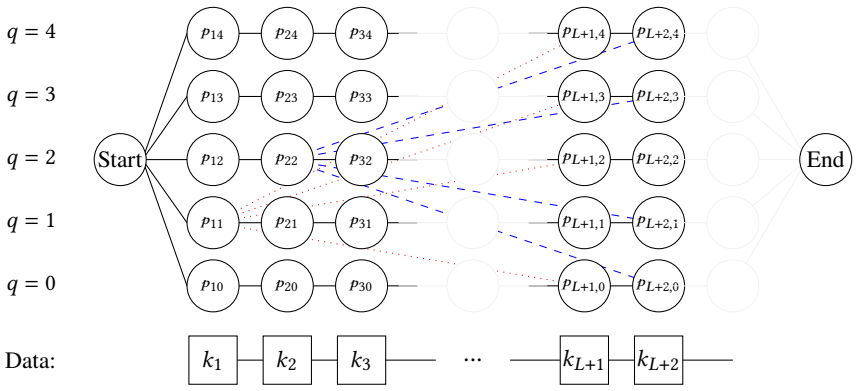

Fig. 2. An example of a harmonic network – only two sets of 'jump edges' are shown (in red and blue) for clarity. In the full network, every node p_{iq} is connected to $p_{i+1,q}$ and $p_{i+L,k\neq q}$. The graph is directional - vertices can only be traversed from left to right.

2.3 The *GenCHORD* Algorithm

After determining the value of the free parameters (discussed in Appendix B), Equations (5) and (7) provide everything we need to compute the score Eq. (6) for a proposed set of $\{q\}$. The task of detecting the most likely (subject to our priors) set of copy-numbers is therefore equivalent to identifying the set of $\{q\}$ which maximise Eq. (6). This is an integer optimisation problem, with the limiting factor that the dimensionality of the solution is equal to the chromosome size.

Provided that we are able to make the assumption that there is some finite Q such that $0 \leq q < Q$ across the entire genome, it is most convenient to think in terms of a *network*, such that an assignment of $\{q\}$ corresponds to a unique path through the network: the optimal path is the one with the highest score.

The network is arranged in a series of layers (not dissimilar in appearance to the network diagrams of feedforward neural networks), each layer corresponding to a given base in the genome. The vertices (or 'nodes') in the layer then correspond to different values of q that might be assigned to that base. In short, a path which passes a node n_{iq} corresponds to the i^{th} base being assigned a harmonic of q.

The edges between vertices are directional, and obey the following rules, which are displayed graphically in Figure 2:

(1) The start node is connected to all nodes in the first layer, $n_{0,0\leq q<Q}$
(2) The node $n_{i,q}$ is connected to nodes $n_{j>i}$ through *step edges* and *jump edges*
   - Step Edges connect $n_{i,q} \rightarrow n_{i+1,q}$, i.e. the same harmonic in the next layer. Step edges have a cost equal to

$$\text{Cost}(n_{iq} \rightarrow n_{i+1,q}) = p_{\vec{\theta}}(k_{i+1}|q) + \begin{cases} 0 & q = d_c \\ \mathfrak{p}_{\text{ploidy}} & \text{else} \end{cases} \tag{2}$$

     That is, they incur a cost equal to the posterior probability that k_{i+1} arises from the same q as k_i, the $\mathfrak{p}_{\text{ploidy}}$ is the prior probability that the q is equal to the normal ploidy.
   - Jump Edges connect $n_{i,q} \rightarrow n_{i+L,j\neq q}$, i.e. a different harmonic in a layer L further along. This enforces the prior restriction that changes in q must be above a certain length (L) in order to be considered valid. Jump Edges incur a cost equal to the posterior probabilty of assigning *all* k_j for $i < j \leq i + L$ to the new harmonic, and so the

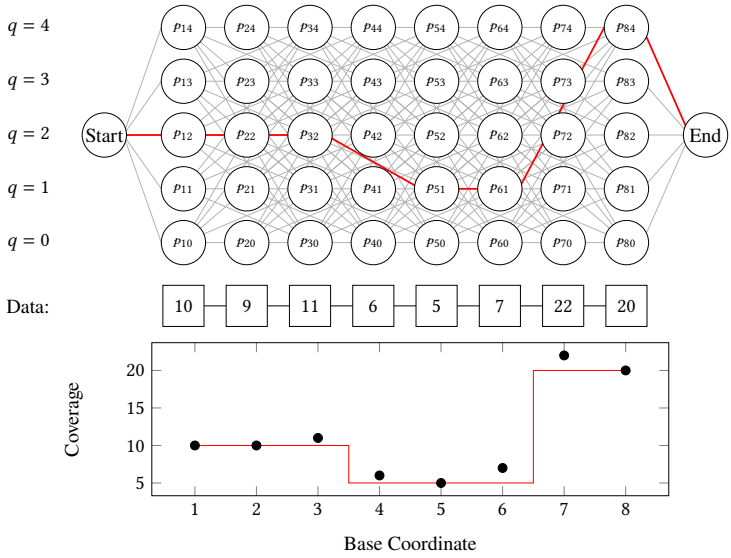

Fig. 3. (Top panel) A demonstration of an optimal path through a network with $L = 2$ and $\nu = 5$, given some example coverage data. (Bottom panel) a projection of this path back onto the coverage distribution. For aesthetic reasons we have placed the transitions at half-integers – in practice non-integer values of base index are meaningless.

statistics must favour all of these k-values being assigned the same q in order to consider a change in q likely.

$$\text{Cost}(n_{iq} \to n_{i+L,j\neq q}) = \sum_{a=i+1}^{i+L} \left(p_{\vec{\theta}}(k_i|j)\right) + \mathfrak{p}_{\text{jump}} + L \times \begin{cases} 0 & j = d_c \\ \mathfrak{p}_{\text{ploidy}} & \text{else} \end{cases} \tag{3}$$

We emphasise again that a path $n_{i,q} \to n_{i+L,j\neq q}$ means that $q_{i+1}, ..., q_{i+L}$ are all assigned the harmonic j, and that this is the only means by which a jump $q \to (j \neq q)$ is possible.

(3) The nodes $n_{G-1,q}$ are connected to the End node

Here we have used the shorthand $p_{\vec{\theta}}(k|q) = p(k|q'\nu, \eta, \sigma, \gamma)$, our underlying distribution function given in Eq. (5). From these rules it is a simple matter of using a basic shortest-path algorithm[4], using the associated cost of each node as the distance: step through each layer and compute the shortest path to each node in that layer, and then iterate until the end node is reached – the longest path through this network corresponds to a unique $\{q\}$, which is then necessarily the one which maximses $\mathcal{L}$. An example of such a minimal path is shown in Fig. 3.

The runtime of this algorithm scales as $O(Q^2G)$, where Q is the maximum harmonic and G is the number of bases being evaluated. However, if we naively construct this network in-memory it occupies a space mQG, where m is the memory footprint of an individual node. A conservative estimate of the naive implementation gives the memory requirements (for a human genome) of ≈650GB, well beyond what most consumer devices are capable of. We note, however, that the network is only semi-local – the network does not need any information more recent than the previous $L + 1$ nodes, and so the network can be constructed on-the-fly. With $L = 10^5$, this reduces the memory requirements to a mere ≈ 35MB.

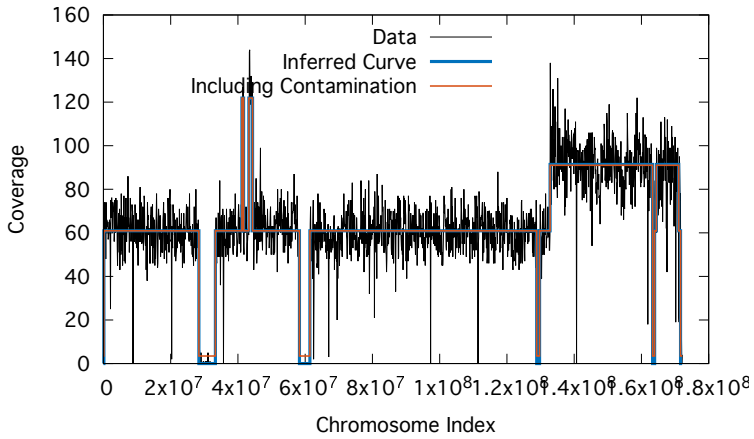

Fig. 4. The *GenCHORD* predictions qv, (blue) for $L = 2 \times 10^5$bp overlaid onto a human oesophageal cancer sample (OES148, Chromosome 6, a known chromothripsis site). This chromosome has minimal contamination, and so the contamination-corrected curve $q'v$ (red) shows only minor deviation. Recall that the blue curve is the most likely (large scale) copy number, and so is equal to an integer multiple, the red curve is the internal model used to derive the value of the blue curve.

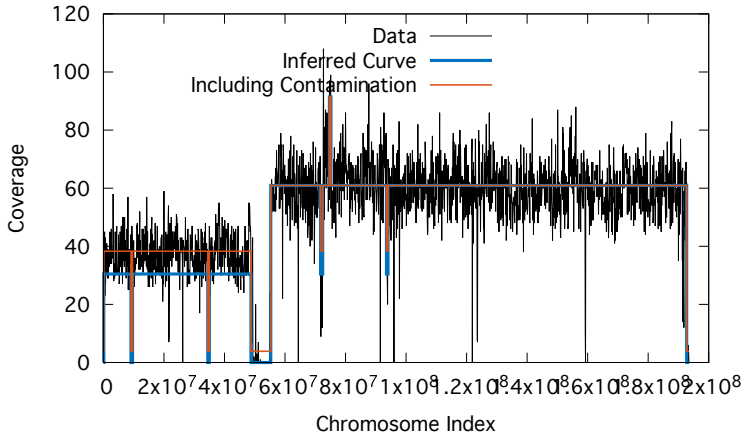

Fig. 5. The *GenCHORD* predictions (blue) overlaid onto a human oesophageal cancer sample (OES148, Chromosome 4), demonstrating how the contamination-modelling shifts the position of the harmonic. Reported values (the blue curve)are always equal to qv, though the model internally uses $q'v$, the contamination is assumed to be a laboratory artefact.

3 RESULTS OF THE *GENCHORD*

3.1 Human Oesophageal Cancer

In Figures 4 and 5 we show the results of the *GenCHORD* algorithm applied to a human oesophageal chromothripsis cancer case, OES148, which was sequenced with PacBio HiFi, a long-read sequencing technology. Fig. 4 is chromosome 6, known to be the main shattering site, and we see that the algorithm is indeed capable of identifying the large-scale

[4]We are actually looking for the longest path, but minimising $|\mathcal{L}|$ is equal to maximising $\mathcal{L}$, so the distinction is trivial

changes present. It is interesting to note which of the deletions the model considered valid - a human might consider the deletions at indices 10^7, $\approx 10^8$ and $\approx 4 \times 10^8$ as important as those at 1.3×10^8 and $\approx 6 \times 10^8$, however this is an artefact of the image compression - closer exmaination reveals this first set to be only a thousand bp in width, whilst the set that the algorithm identified are deletions in excess of 10^5bp. *GenCHORD* has therefore successfully performed as a data-reduction tool, preserving only those features which are most statisticaly meaningful.

Fig. 5 shows chromosome 4 of the same sample. This chromosome is not thought to be a major chromothripsis site, but we have included it in order to demonstrate the contamination-correction, since there is evidence that the p-end of one copy of chromosome 4 is deleted in a majority of cells in the sample, but not all of them – we see that *GenCHORD* has correctly identified that a partial aneuploidy (with a high degree of contamination) is the most likely explanation for the available data.

3.2 Machine Learning & Tasmanian Devils

We turn now to the case of Devil Facial Tumour Disease (DFTD), a transmissabel cancer affecting *Sarcophilus harrisii*, the Tasmanian Devil[3, 4], of which there are two distinct subtypes (DFT1, the original and most prevalent, and DFT2, a more recent, aggressive but rarer form)[5], and there is evidence that DFTD is caused by chromothripsis[6, 7].

The goal of the *GenCHORD* is not merely to serve as a copy-number analyser for its own sake, but as a preprocessing layer with the goal of converting the extremely large, extremely noisy coverage data into a meaningful encoding for use in Machine Learning. Due to the genetic homogeneity of both Tasmanian Devils as a whole, and of DFTD, this provides us with a useful proof-of-concept for *GenCHORD* in coverage-based cancer inference.

GenCHORD is intended to serve as an initial transformation of the data (an input layer) for traditional Machine Learning methods. The Machine Learning models we present, therefore, are intentionally simple and out-of-the-box, in order to prove that the *GenCHORD* has succeeded in preserving the large scale behaviour of the data over the course of its data reduction.

Using a dataset of 177 Tasmanian Devils[7] sequenced using standard Illumina Short Read sequencing, and annotated as being either 'normal' (i.e. non-cancerous), 'DFT1' or 'DFT2', we transformed the data using *GenCHORD*, and encoded the data into a 20-dimensional feature vector, as described in Appendix C.

We then used a simple Feedforward Neural Network (5 layers, dimensions 20,30,10,8,3, ReLu activation throughout) as a classifier, including a 50% dropout layer after the second layer to prevent overfitting. This model is a standard implementation – the novelty of our method is in the input layer and the feature representation, not the ML model itself.

In Fig. 6 we show the result of the accuracy of this classifier: it exceeded 95% accuracy everywhere, indicating that the encoding is retaining the information necessary to detect cancer-causing SVs, and to identify which SVs are responsible for which cancer. We note that there were a number of troublesome misclassifications; investigating the dataset revealed a number of low-quality sequences from early in the data collection programme. Future work shall incorporate additional information into the *GenCHORD* output and subsequent encoding to allow the identification of such artefacts, and hence allow the classifier to focus on the cancer signal.

Unsupervised methods yielded similarly positive results. In the right panel of Fig. 6, we demonstrate the result of a UMAP clustering[8] on a $N = 200$ encoding of our dataset. The result is a spectacular dissection of the feature space: aside from two misclassifications, UMAP correctly clustered all of the non-cancerous, DFT1 and DFT2 datasets into isolated clusters.

In addition, we notice that there is additional substructure present: the 'normal' datasets are bifurcated, investigation revealed this to be based solely on the sequencing depth of the sex chromosomes (the normal set contained both male and

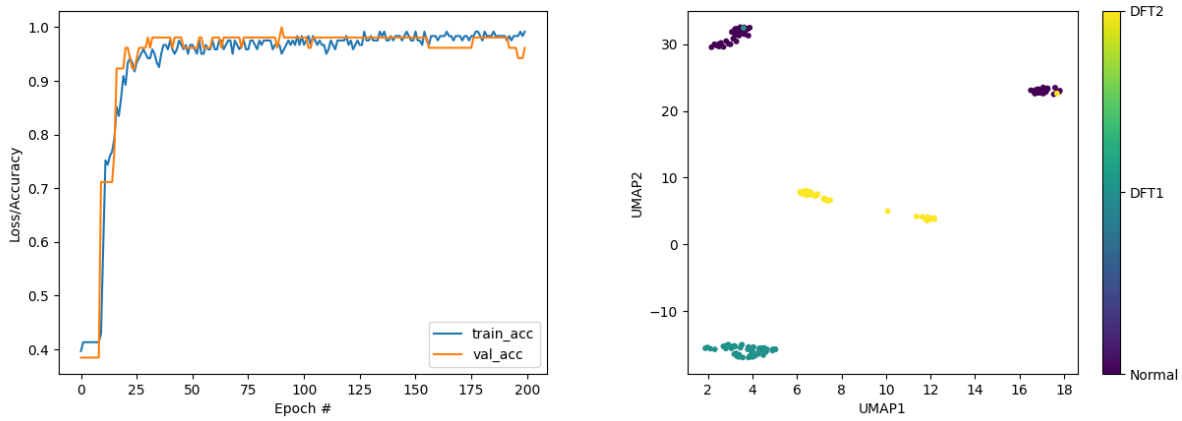

Fig. 6. (*Left*:)The accuracy of a 5-layer FNN with a 50% dropout rate after the second layer. We see a convergence to extremely high prediction accuracy. (*Right*:) a UMAP clustering of the Tasmanian Devil dataset

female Devils); however (since the transmissable cancer is spread primarily through male-male sexual competition), the cancer samples are entirely male. The subtypes within the DFT2 cluster, arises from substructure within DFT2, where two distinct clades have already been identified[7] - our clusterings match up with the assigned clades with other methods. That the DFT1 cluster is does not similarly split up into the 6 clades identified elsewhere is most likely a feature of our (somewhat naive) encoding, the loci which differentiate the clades would fall within the same spatial bin, and so do not permit the separation of the clades. We have begun work on a more nuanced encoding which would permit such spatial resolution within the encoding.

The work in this section is preliminary, and serves only to demonstrate that the *GenCHORD* algorithm has great potential as a data reduction and transformation tool for future Machine Learning work, which we intend to include novel encodings and architectures.

4 CONCLUSIONS & FUTURE WORK

In this work we have demonstrated the *GenCHORD* algorithm, a powerful data-reduction method based on the principles of step-detection. *GenCHORD* permits us to incorporate the effects of biological variability, contamination and minimum jump-sizes to retain only the most statistically important deviations from the expected ploidy, and works equally well on both long and short read technologies. *GenCHORD* is publicly available: https://github.com/wtsi-hpag/GenCHORD

We have demonstrated that this method permits us to transform the high-dimensional and complex data associated with Tasmanian Devil Facial Tumour Disease into a low-dimensional representation from which a small 5-layer FNN model can learn, with unerring accuracy, to predict the presence and type of cancer. We also showed that the *GenCHORD* preserved information by which an unsupervised method could cluster the cancers into familiar groupings, but also uncover new subtypes, the nature of which we are still investigating.

This proves a useful testbed of *GenCHORD* as a tool for studying cancer, however, the exact properties which make Tasmanian Devils an attractive test case of this method (the genetic homogeneity of the population and near-clonality of the cancer) mean that application to the human case will be far from trivial.

Future work on this topic will focus on the encoding methodology and novel ML architecture: how this reduced data can be represented in the machine in such a way to make feature extraction meaningful, explainable and useful for scientific and diagnostic applications.

REFERENCES

[1] Jannat Ijaz, Edward Harry, Keiran Raine, Andrew Menzies, Kathryn Beal, Michael A Quail, Sonia Zumalave, Hyunchul Jung, Tim HH Coorens, Andrew RJ Lawson, et al. Haplotype-specific assembly of shattered chromosomes in esophageal adenocarcinomas. *Cell Genomics*, 4(2), 2024.

[2] Philip J Stephens, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, King Wai Lau, David Beare, Lucy A Stebbings, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell*, 144(1):27–40, 2011.

[3] Elizabeth P Murchison. Clonally transmissible cancers in dogs and tasmanian devils. *Oncogene*, 27(2):S19–S30, 2008.

[4] Elizabeth P Murchison, Ole B Schulz-Trieglaff, Zemin Ning, Ludmil B Alexandrov, Markus J Bauer, Beiyuan Fu, Matthew Hims, Zhihao Ding, Sergii Ivakhno, Caitlin Stewart, et al. Genome sequencing and analysis of the tasmanian devil and its transmissible cancer. *Cell*, 148(4):780–791, 2012.

[5] Ruth J Pye, David Pemberton, Cesar Tovar, Jose MC Tubio, Karen A Dun, Samantha Fox, Jocelyn Darby, Dane Hayes, Graeme W Knowles, Alexandre Kreiss, et al. A second transmissible cancer in tasmanian devils. *Proceedings of the National Academy of Sciences*, 113(2):374–379, 2016.

[6] Janine E Deakin, Hannah S Bender, Anne-Maree Pearse, Willem Rens, Patricia CM O'Brien, Malcolm A Ferguson-Smith, Yuanyuan Cheng, Katrina Morris, Robyn Taylor, Andrew Stuart, et al. Genomic restructuring in the tasmanian devil facial tumour: chromosome painting and gene mapping provide clues to evolution of a transmissible tumour. *PLoS genetics*, 8(2):e1002483, 2012.

[7] Maximilian R Stammnitz, Kevin Gori, Young Mi Kwon, Edward Harry, Fergal J Martin, Konstantinos Billis, Yuanyuan Cheng, Adrian Baez-Ortega, William Chow, Sebastien Comte, et al. The evolution of two transmissible cancers in tasmanian devils. *Science*, 380(6642):283–293, 2023.

[8] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

A MATHEMATICAL THEORY

A.1 Negative Binomial

The statistical model $\mathcal{P}$ for the coverage k, is derived from the common assumption that the genome is sampled via a Poisson distribution with mean λ, but where there are biological and experimental errors which change λ, such that the model is the convolution of a Poisson distribution, and some 'error distribution', which encodes the degree to which λ varies. Whilst maximal-entropy arguments might imply a lognormal distribution as the optimal choice, for convenience (since the outcome is analytically computable), we use the Gamma distribution with mean $\mu > 0$ and variance $\sigma^2 > 0$, which gives the Negative Binomial Distribution as our model:

$$\mathcal{P} = \int_0^\infty \Gamma(\lambda|\mu,\sigma)\text{Poisson}(k|\lambda)\text{d}\lambda \iff k \sim \text{NB}\left(k;\mu,\mu+\sigma^2\right) \tag{4}$$

We make a modification to this model, since it will often be necessary to truncate the probability distribution to some finite K to avoid contamination with short but highly repetitive regions which fall outside the scope of our model. We therefore normalise the model up to this finite K:

$$p(k;\mu,\mu+\sigma^2) = \frac{\text{NB}\left(k;\mu,\mu+\sigma^2\right)}{\sum_{k'=0}^{K}\text{NB}\left(k';\mu,\mu+\sigma^2\right)} \tag{5}$$

A.2 Statistical Model

The task of idenitfying breakpoints is equivalent to assigning $\{q\}$-values to a given genome. We must therefore form statistical score $\mathcal{L}(\{q\})$, which is maximised at the most likely set $\{q\}$. We make the standard assumption that the coverage k_i of the i^{th} base is drawn independently and identically from an underlying distribution function (*iid*), which is a function of the fundamental frequency, ν, and the harmonic, q_i. Bases are statistically connected through the *prior* function, which encodes many of the points outlined in the previous section. As a result of the *iid* assumption, the global score can be found to be:

$$\mathcal{L}(\{q\}|\{k\},\nu,\sigma,\vec{\theta}) = \sum_i \mathfrak{p}\left(k_i|q_i,\nu,\sigma\right) + \text{Prior}(\{q\},\vec{\theta}) \tag{6}$$

Here $\vec{\theta}$ are the hyperparameters of the model. This is an indirect statement of standard Bayesian Maximum Posterior methodology - the quantity $\mathfrak{p}$ is the log-probability of observing the data k_i given the harmonic q_i and parameter ν; as already discussed, this will take the form of a Negative Binomial distribution. Finally, we must formulate a suitable prior, which enforces the remainder of our assumptions, namely:

(1) Consecutive values of q_i should be similar
(2) Changes in q_i can only happen over a large distance
(3) Most of the chromosome c should be at $q_i = d_c$ (equal to the ploidy)

Therefore, we propose:

$$\begin{aligned} \text{Prior}(\{q\}) =& \sum_i \left[\mathfrak{p}_{\text{ploidy}} u(q_i, d_c) + u(q_i, q_{i-1}) \left(\mathfrak{p}_{\text{jump}} + \mathfrak{p}_\infty \sum_{j=i-L}^{i-1} u(q_j, q_{i-1}) \right) \right] \\ u(a,b) =& (1 - \delta_{ab}) = \begin{cases} 0 & a = b \\ 1 & \text{else} \end{cases} \end{aligned} \tag{7}$$

The terms in the prior therefore act (in order), to penalise every base which is not at the expected diploid level by an amount $\mathfrak{p}_{\text{diploid}}$, penalise every jump between dissimilar qs by an amount $\mathfrak{p}_{\text{jump}}$ and to penalise jumps which occur within a distance L of another jump by an amount $\mathfrak{p}_\infty$, which is a number $\approx -\infty$, but which has the property $0 \times \mathfrak{p}_\infty = 0$ – rather than a penalty, this therefore acts as a *forbiddance*, discarding models with such jumps.

B PARAMETER INFERENCE

The model presented above has a number of free parameters - notably ν (the fundamental sampling frequency of the chromosome), σ (the standard deviation of the sampling-noise function), γ (the degree of alignment noise), η (the contamination amount), as well as the functional form of p_{error}.

We need to determine these values before we may reliably assign the values of $\{q\}$. To do this efficiently, we shall assign these values based on the global coverage distribution - the counts of all observed coverage values binned into a histogram. To this distribution, we then fit a Negative Binomial Mixture Model using the contamination-biased probability:

$$p_{\text{global}}(k|\vec{\theta}) = \gamma p_{\text{error}}(k) + (1-\gamma) \sum_{q=0}^{Q} w_q NB\left(k_i; q'(q,\eta)\,\nu, \sigma\right) \tag{8}$$

Where w_q are the population weightings; equal to the fraction of bases which will be assigned to q when the full assignment is performed. *GenCHORD* uses a standard Maximum Posterior Method to infer the maximal value of these parameters given a set of data.

C FEATURE VECTOR

After the data has been reduced by *GenCHORD*, we encode it into a feature vector $\vec{v}$ of length N. Each element of the vector corresponds to a portion of the genome of length G/N where G is the total length of the genome. Each segment is assigned a score based on the presence, q-value and length of domains that GeneCHORD had assigned, where a domain is a continuous region of length ℓ which has been assigned the same value of q. The score associated with a domain $D(\ell, q)$ on chromosome c (which has ploidy d_c) is:

$$S(\ell, q, d_c) = \exp\left(|q - d_c| + \frac{\ell N}{G} \right) \tag{9}$$

The score assigned to each element of the feature vector $\vec{v}$ is then the sum of scores for all domains in the associated portion of the genome.