I Need A Title

Contents

1 TUF

Thermodynamically ultra-fastened (TUF) regions are stretches of the DNA which fail to denature even after the application of extreme melting conditions [9]. This behavior effectively reduces the amplification efficiency in these regions. It has also been reported that TUF regions contain a core sequence which exhibits an increased GC concentartion relative to the surrounding DNA. It is in fact these locally concentrated spikes of GC content which is believed to remain duplexed despite the application of denaturation processes [9].

G-quadruplexes (G4) are four-stranded secondary structures fromed by particular G-rich nucleic acid sequences. Notably, G4s can adopt intramolecular folds whenarising from a single G-rich DNA or RNA strand, or in-termolecular folds, through dimerization or tetramerizationof two or more strands [**?**]. Extensive evidence implicates G4 sequences in various essential biological functions, including telomere maintenance,DNA replication, genome rearrangements,DNA damage response, chromatin structure, RNA processing and transcriptional ortranslational regulation, see [**?**] and references therein.

Our aim is to develop a mathematical model in order to investigate this assumption i.e. that TUF cores exhibit high GC concentration when compared to the surrounding DNA. In this regard, a model should be able to capture the existence of TUF regions and subsequently allow for their investigation. Analysis of genome sequences can be a time consuming and error prone process if done manually. In whole genome amplified (WGA) read-depth (RD) sequencing samples, TUF regions are often represented as regions of low coverage, resembling deletions. Thus, borrowing from studies related to copy number variation (CNV), we develop a hidden Markov model that classifies segregated chromosome regions into states depending on their average read-depth. Using a RD characterization we can distinguish commonly found behavior in sequences such single copy or full copy deletion. In this regard, a TUF region can be assumed to represent one extra state. Although the exact characteristics in terms of RD of such a state are not known, the assumption is that a low RD observed in a WGA sample in combination of a normal RD observed in the same region for a non-WGA sample is indicative of TUF.

2 Hidden Markov model

A hidden Markov model (HMM) is a probabilistic framework that uses two interrelated probabilistic mechanisms; a Markov chain of a finite number of states, N, and a set of random functions each associated with a respective state [8]. The set of state is denoted by $S = \{S_0, S_1, \cdots, S_{N-1}\}$. At a given time instant, the system is assumed to be in some state and an observation is generated by the random function corresponding to this state [8]. State transitioning occurs according to transition probability matrix $\mathbf{A}$. Within the HMM framework, an observer only sees the random output generated by the random functions cooresponding to the states and not the states themselves. Thus, the state at which the system is in can only be probabilistically inferred.

We use more or less standard notation and denote with q_n the state of the system under consideration at the discrete time instance n. Hence,

$$q_n \in S \tag{1}$$

where S is a set of discrete states. A sequence of states, each of which belongs in S, is denoted with Q; $Q = \{q_1 q_2, \cdots q_T\}$. A sequence of observations is denoted with O; $O = \{o_1 o_2, \cdots o_T\}$. Overall, an HMM is characterized by the following parameters, see [6] and [8]

- $\mathbf{A}$ a probability transition matrix
- $\mathbf{B}$ a probability emission matrix
- $\boldsymbol{\pi}$ an initialization vector

Each a_{ij} of $\mathbf{A}$ expresses the probability of transitioning to state j given that the previous state was i:

$$a_{ij} = P(q_n = j | q_{n-1} = i), \forall i, j \in S \tag{2}$$

Equation 2 expresses the assumption that the system states form a Markov chain. In other words, the current system state depends only on the previous state. Since the a_{ij}s represent probabilities, the following conditions should be respected [8]

$$a_{ij} \geq 0, \sum_j a_{i,j} = 1 \tag{3}$$

Similarly, each element b_{jk} of the emission matrix $\mathbf{B}$ specifies the probability that at time instant n and state j, the observation is o_k:

$$b_{jk} = P(O_n = o_k | q_n = j) \tag{4}$$

We have the following constraints for the $\mathbf{B}$ matrix

$$b_{jk} \geq 0, \sum_k b_{jk} = 1 \tag{5}$$

A HMM does not require the number of states is the same as the number of observation symbols. Finally, the vector $\boldsymbol{\pi}$ provides the probability distributions at time $n = 0$ meaning

$$\pi_j(0) = P(q_0 = j) \tag{6}$$

Collectively, we denote an HMM using the letter λ:

$$\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) \tag{7}$$

Finally, we assume that we are dealing with a time invariant system. In other words, the transition probability matrix remains constant.

A general review of HMM in relation to bioinformatics is given in [8]. Hidden Markov models have been used for copy number variation detection research e.g. [2], [10] and [1], the analysis of of array CGH data [4], and the analysis of profile series [7].

3 HMM for TUF

Our intention is to investigate the applicability of hidden Markov models in terms of identifying TUF regions in the genome. In this regard, we develop an HMM model using the `pomegranate` [1] Python library. This section discusses the current state of the approach we use. It further attempts to justify certain modeling choices that have been made.

The develped HMM uses two sequences; a sequence that it underwent WGA treatment before sequencing (sample m605), and one that was not treated (sample m585). This is necessary in order to amplify the existence of TUF regions. Figure 1 shows a snapshot of the applification in the WGA sample compared to the non-treated one as these are viewed in the IGV browser.

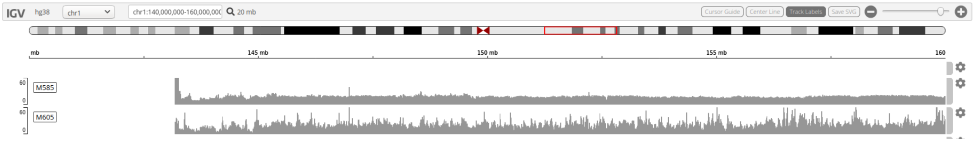

Fig. 1: m605 and m585 samples.

Furthermore, we assume the following set of discrete states:

- Deletion
- TUF
- Normal copy
- Duplication
- TUFDUP
- Gap

The Gap state corresponds to the case where there is not base present in either of the sequences used. The TUF state assumes that low WGA sample mean when compared to normal sample mean for the non-WGA sample. However, this does not fully capture the spectrum of the data, see section 3.1 and figure 9. Thus, we introduce the TUFDUP state in order to represent data where the WGA

[1] `https://github.com/jmschrei/pomegranate`

sample mean is rather small whilst the non-WGA sample mean is large enough to assume that this is

As mentioned previously, an HMM model assumes that the system in hand can be in a state from a specified set S. This set of states can be assumed a priori implying some knowledge of the data. Examples of this methodology are given in [2] and [10] where six states are used. Another approach is to use clustering techiniques in order to determine an optimal number e.g. [4] and [5]. In the latter approach, each cluster is assumed to represent a state. Clusters being very similar under some assumed similarity metric can be merged together.

One advantage of the clustering approach is that it allows for an educated guess about the parameters of the distributions that the HMM framework requires. Frequently used states in CNV studies are full and single copy deletion, normal and duplication. TUF as well as gap regions can be represented naturally as an extra state.

The second point of major concern, is the appropriate probability distribution that best models each state in terms of emission probabilities. This, in general, seems to be more important than how one is modeling the transition probability matrix $\mathbf{A}$, [6]. There is a variety of methods to achieve this. The simplest being to assume a priori a given probability mass function with given parameters. Another approach is to use an estimation technique such as histograms, kernel estimation or clustering.

In this work, we assume that the states follow a two dimensional Gaussian distribution. The exception to this is the Gap state (see below). The empircal distributions that we compute when investigate the data, suggest that he assumption of Gaussian distributions is not unreasonable see figures 10, 11 and 11.

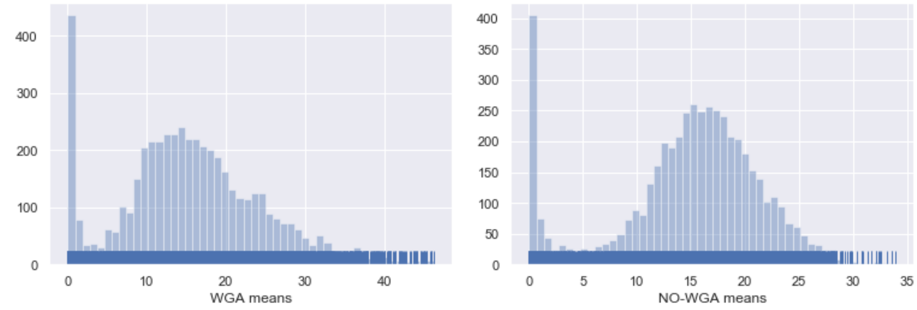

Fig. 2: Full copy deletion histogram for WGA and non-WGA samples.

We estimate the parameters for these distributions as follows. We cluster a dataset which contains manually identified portions of the DNA that match the assumed states by using a Gaussian mixture model (GMM) [3]. The dataset corresponds to small regions from chromosome 1 that contain the states that the model assumes. These regions are then discretized into non-overlapping windows each of which has size 100 bases. We calculate the means for the two samples, i.e. WGA and non-WGA, and then apply a cutoff filter to exclude outliers. The filter is simply a threshold on the means. Hence, a window is assumed as an outlier if either $\mu_{WGA} > 140$ or $\mu_{NWGA} > 120$. The remaining windows, form the input for the GMM [2]. We also investigated more traditional approaches like K-Means and PAM. However, these techniques tend to create eqully sized clusters.

[2] We use the `sklern` implementation

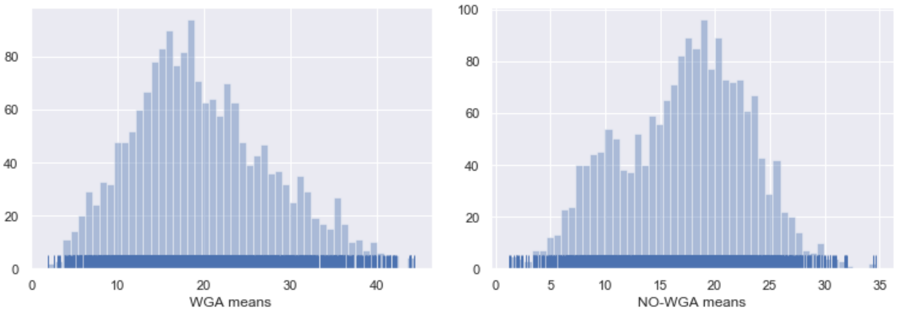

Fig. 3: Single copy deletion histogram for WGA and non-WGA samples.

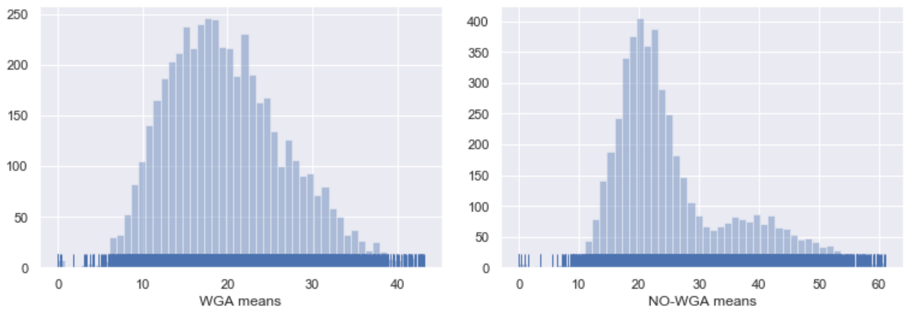

Fig. 4: Duplication histogram for WGA and non-WGA samples.

This is something that we do not anticipate to be the case (for example the Normal state is expected, in general, to dominte the data). A GMM approach allows for more flexibility on the shapes of the clusters whilst we can use the parameters of the ensued Gaussian distributions, i.e. μ_{WGA}, μ_{NWGA} and $\mathbf{\Sigma} = diag(\sigma^2_{WGA}, \sigma^2_{NWGA})$ where μ_i, σ_i are the window mean and standard deviation for the WGA and non-WGA sample, into the HMM model. In GMM clustering the hard cluster assignement of K-means, is changed into a soft one [3]. Note that the windows which have been identified to contain gaps are excluded from the clustering calculations however they are kept in the HMM.

Figure 5 shows the clustered data when using five clusters. Only four clusters are actually visible. The cluster that represented deletion was dropped in favor of the red cluster in the figure.

The yellow cluster is used to extract the parameters for the Duplication state whilst the pink and blue are used to model two different Normal states.

The TUF state is represented as a mixture model with two components each of which is represented as a two dimensional Gaussian distribution. Each component is weighted using a coefficient of 1/2. Figure 6 shows the clustering for identifying the properties of the TUF state. The green componet shown in figure 6 is used in order to initialize the TUF state.

In order to model the Gap state we assume that both components follow a uniform distribution $U(-999.5, -998.5)$. Our intention is to make this state to stand out from the rest so that the model is forced to select this when a gap window is found (see subsection 4).

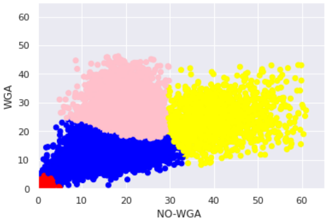

Fig. 5: GMM clustering with five clusters.

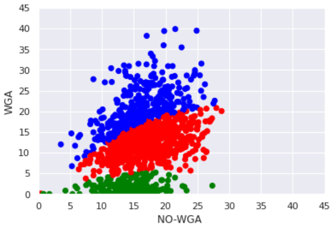

Fig. 6: GMM clustering for TUF state with three clusters.

Finally, the TUGAP state is...

The HMM also requires as input the initialization vector $\boldsymbol{\pi}$ and the transition matrix $\mathbf{A}$, see equation 7. For the former we assume a uniform probability for every state i.e. every state is equally likely to initiate the sequence of hidden states. Hence,

$$\pi_i = \frac{1}{|S|}, \quad \forall i \in S \tag{8}$$

where $|S|$ denotes the number of discrete states. For the latter, we assume that every state can transition to any other state including itself. However, we assign a signfinicantly higher probability to the latter scenario than the former. In other words, we assume that the model is more likely to stay in a given state than transitioning to another. This is summarized by the matrix $\mathbf{A}$ in equation 9

$$\mathbf{A} = \begin{bmatrix} 0.85 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.85 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.85 & 0.85 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.85 & 0.025 & 0.85 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.85 & 0.025 & 0.025 & 0.85 & 0.025 & 0.025 \\ 0.025 & 0.85 & 0.025 & 0.025 & 0.025 & 0.85 & 0.025 \\ 0.025 & 0.85 & 0.025 & 0.025 & 0.025 & 0.025 & 0.85 \end{bmatrix} \tag{9}$$

In summary, the HMM has as follows

- $\pi_i = \frac{1}{|S|}, \quad \forall i \in S$

- Gap state $G \sim U(-999.5, -998.5)$
- Every state $S \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- TUF state $TUF \sim \sum_{i=1}^{2} c_i N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), c_i = 1/2$

Figure 7 shows the HMM model with the transition probabilities used in a graphical form. We remark however, that the framework we use is flexible enough to assume different paramteric models and add or remove states.

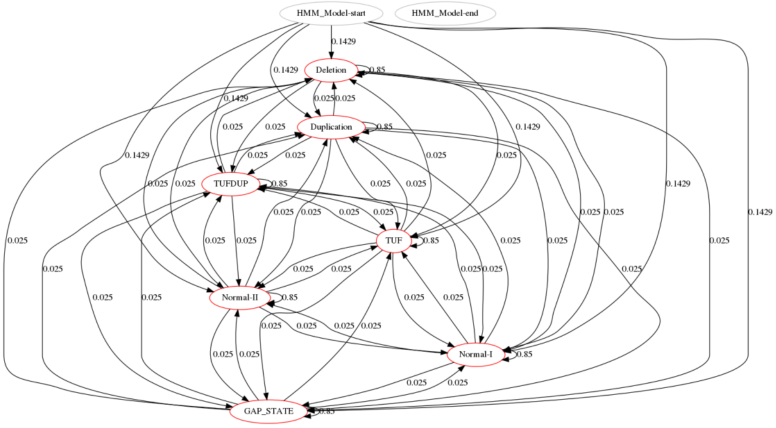

Fig. 7: States and transition probabilities for HMM model.

3.1 Model calibration

Once the basic model is established, we further calibrate it on chromosome 1. This is necessary as it is difficult to identify representative data for every state. Calibration is done by applying the model on various regions and extracting the Viterbi path. The resulting path is then visually evaluated by loading both the region samples and the path on th IGV browser. Figure 8 shows the predicted states for chromosome 1 and region $[1-20] \times 10^6$ using the non-calibrated HMM model. The red spikes correspond to TUF windows.

Figure 9 shows the classification of the windows [3] achieved by the calibrated model on region $[1-20] \times 10^6$. The non-calibrated model did not include the TUFDUP state. These caused the purple dots to be classified as duplication (shown in green). The black dots are classified as TUF state and the purple dots as TUFDUP. This is the case after introducing the TUFDUP state on region

4 Viterbi path

This section presents the Viterbi paths for chromosomes 1, 2, 3, 4, 5, 6, 7. The Viterbi path simply answers the following question; given an HMM λ and a sequence of observations O we seek to find

[3] The windows are represented by the WGA and NWGA means

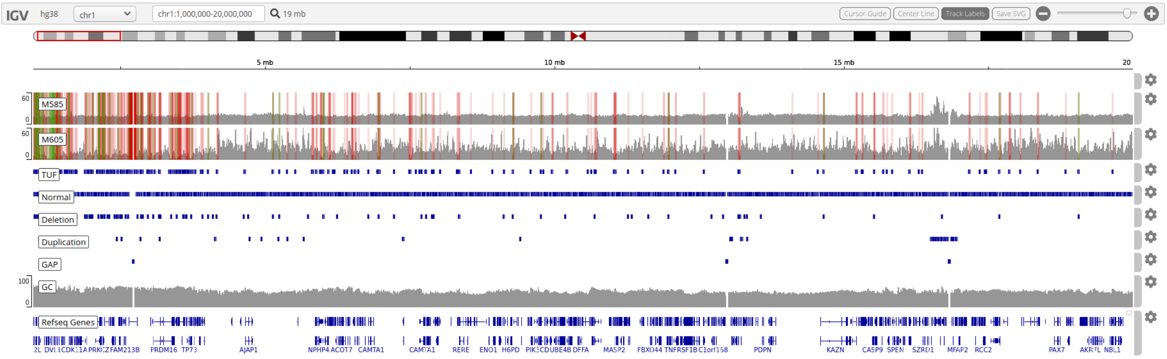
IGV
hg38
chr1
chr1:1,000,000-20,000,000
19 mb
Cursor Guide
Center Line
Track Labels
Save SVG
5 mb
10 mb
15 mb
20
60
0
M585
60
0
M605
TUF
Normal
Deletion
Duplication
GAP
100
0
GC
Refseq Genes
2L DVL1 CDK11A PRKCZ FAM213B PRDM16 TP73 AJAP1 NPHP4 ACOT7 CAMTA1 CAMTA1 RERE ENO1 H6PD PIK3CD UBE4B DFFA MASP2 FBXO44 TNFRSF1B C1orf158 PDPN KAZN CASP9 SPEN SZRD1 MFAP2 RCC2 PAX7 AKR7L NBL1

Fig. 8: Viterbi path classification of windows.

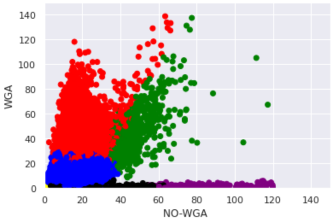

Fig. 9: Viterbi path classification of windows.

the state sequence Q that maximizes the probability

$$P(Q|O, \lambda) \tag{10}$$

We extract regions typically of size 20×10^6 bases. The regions are discretized into non overlapping windows of size 100 bases. Each of the windows has a view of both samples, i.e. m605 and m585. The same cutoff described previously is also applied. Gap windows are included in the formed sequence. In the present context, the observations are pairs of RD means corresponding to the sample view that each window contains. The HMM model discussed in section 3 is used to compute the Viterbi path for the sequence. Figures 10, 11 and 12 present the classification of the windows for the following regions

- $[1 - 20] \times 10^6$
- $[20 - 40] \times 10^6$
- $[40 - 60] \times 10^6$
- $[60 - 80] \times 10^6$

and chromosomes 1,2 and 6 respectively after applying the Viterbi algorithm.

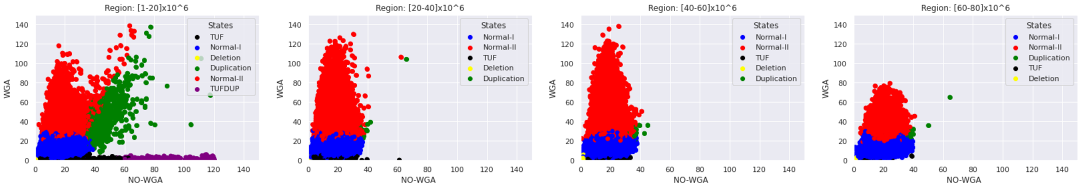

Fig. 10: Regions 1,2,3,4 for chromosome 1 left to right.

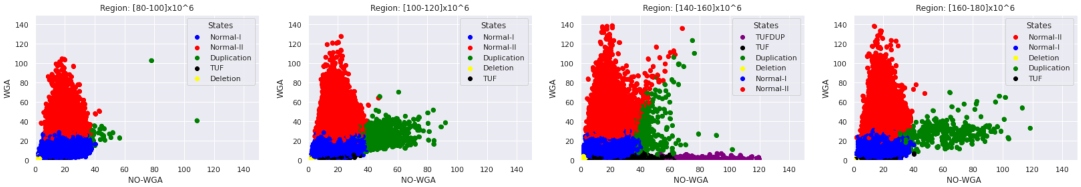

Fig. 11: Regions 5,6,7,8 for chromosome 1 left to right.

4.1 Summary

Currently, we need to

- Establish quantitative mterics for assessing the performance of the HMM
- Compare the performance of the HMM after some training has been performed
- Establish a better clustering approach
- Develop and end-to-end framework for the analysis.

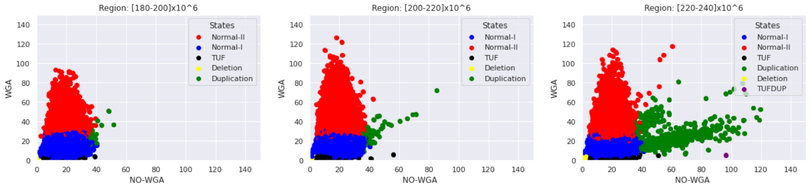

Fig. 12: Regions 9, 10, 11 for chromosome 1 left to right.

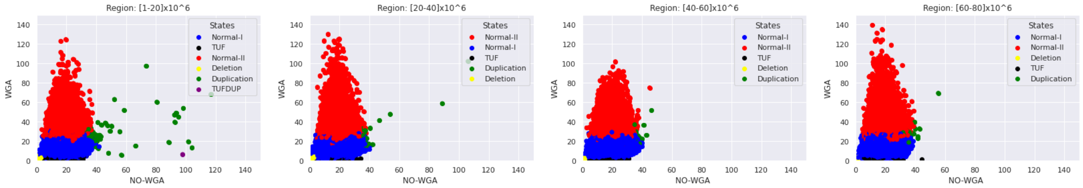

Fig. 13: Regions 1,2,3,4 for chromosome 2.

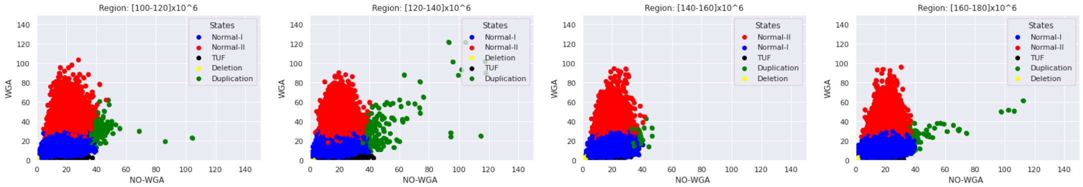

Fig. 14: Regions 5, 6, 7, 8 for chromosome 2.

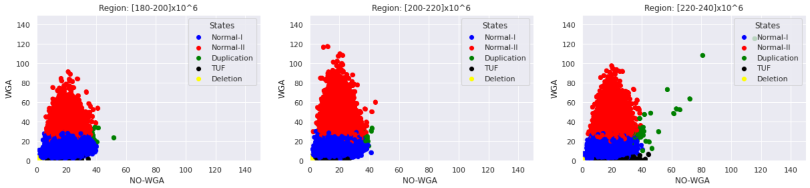

Fig. 15: Regions 9, 10, 11 for chromosome 2.

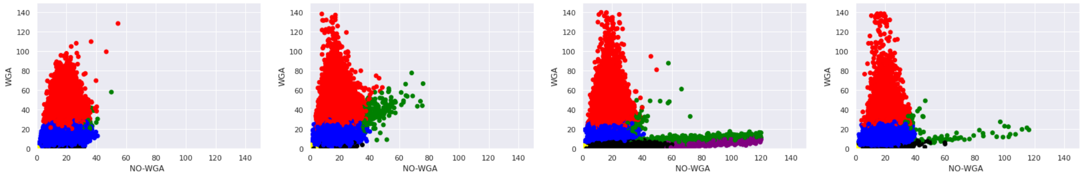

Fig. 16: Regions 1,2,3,4 for chromosome 6.

References

[1] Patric Cahan, Laura E. Godfrey, Peggy S. Eis, Todd A. Richmond, Rebecca R. Selzer, Michael Brent, Howard L. McLeod, Timothy J. Ley, and Timothy A. Graubert. wuhmm: a robust algorithm to detect dna copy number variation using long oligonucleotide microarray data. Nuclic Acids Research, 2008.

[2] Stefano Colella, Christopher Yau, Jennifer M. Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S. Bassett, Anneke Seller, Christopher C. Holmes, and Jiannis Ragoussis. Quantisnp: an objective bayes hidden-markov model to detect and accurately map copy number variation using snp genotyping data. Nuclic Acids Research, 2007.

[3] P. Flach. Machine Learning The art and science of algorithms that make sense of datat. Cambridge University Press, 2012.

[4] Jane Fridlyand, Antoine M. Snijders, Dan Pinkel, Donna G. Albertson, and Ajay N. Jain. Hidden markov models approach to the analysis of array cgh data. Journal of Multivariate Analysis, 2004.

[5] Tingting Liu and Jan Lemeire. Efficient and effective learning of hmms based on identification of hidden states. Mathematical Problems in Engineering, 2017.

[6] Rabiner L. R. A tutorial on hidden markov models and selected applications in speech recognition. Some Jurnal, 2009.

[7] Alexander Schliep, Alexander Schnhuth, and Christine Steinhoff. Using hidden markov models to analyze gene expression time course data. BIOINFORMATICS, 2003.

[8] Koski T. Hidden Markov Models for Bioinformatics. Kluwer Academic Publishers, 2001.

[9] Colin D Veal, Peter J Freeman, Kevin Jacobs, Owen Lancaster, Stphane Jamain, Marion Leboyer, Demetrius Albanes, Reshma R Vaghela, Ivo Gut, Stephen J Chanock, and Anthony J Brookes. A mechanistic basis for amplification differences between samples and between genome regions. BMC Genomics, 2012.

[10] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F.A. Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. Genome Research, 2007.