Finding heterogeneously methylated genomic regions using ONT reads

Emanuele Raineri*, Mariona Alberola i Plà, Marc Dabad, Simon Heath†

January 7, 2022

Abstract

Nanopore reads encode information on the methylation status of cytosines in `CpG` dinucleotides. The length of the reads makes it comparatively easy to look at patterns consisting of multiple loci; here we exploit this property to look for regions where one can define subpopulations of cells based on methylation patterns. As a benchmark we run our clustering algorithm on known imprinted genes and show that the clustering based on methylation is consistent with the phasing of the genome; we then scan chromosome 15 looking for windows corresponding to heterogeneous methylation. We can also compute the covariance of methylation across these regions so that it is not confounded by the mixture of different types of reads. The source code for `cvlr` is available from the authors. Tables, figures, and some further explanations of the algorithms are available as supplementary information.

1 Introduction

In this manuscript we exploit the fact that one can measure both DNA sequence and cytosine methylation status in one go via Nanopore sequencing (see eg. [11]) to scan the genome looking for regions where methylation changes between cells. We do this through a software which can cluster Nanopore reads based on the methylation patterns appearing on them. The idea of quantifying the variability in epigenetic patterns on single read was initially discussed in [6]. A probabilistic model to solve this problem has been proposed by [4] on short reads measured with whole genome bisulfite sequencing; [3] discusses the analysis of RRBS libraries and [1] presents a

*emanuele.raineri@cnag.crg.eu
†simon.heath@cnag.crg.eu

GENE	chr:start-end	strand	expression	median abs diff	pval (phasing)
GNAS	chr20:58839718-58911192	+	Both	0.25	$9.22E-7$
PEG10	chr7:94656325-94669695	+	Paternal	0.417	$2.55E-7$
SNRPN	chr15:24953500-24958500	+	Paternal	0.581	3.52E-10
ZDBF2	chr2:206274663-206279663	+	Paternal	0.244	0.0833
ZIM2	chr19:56782266-56787266	-	Paternal	0.130	0.566
PLAGL1	chr6:143940300-144064599	-	Paternal	0.167	2.75E-4
IGF2	chr11:2129112-2141238	-	Paternal	0.385	0.685
H19	chr11:1995163-2001470	-	Maternal	0.351	0.4
MEG3	chr14:100779410-100861031	+	Maternal	0.193	0.136

Table 1: Putatively imprinted genes we looked at.

statistical method for long reads based on an Ising-like model but as far as we know there is no genome wide analysis of methylation heterogeneity based on these techniques out yet. One can first phase the genome and then look for allele specific methylation of course, as done for example in [2] but what we propose here does not rely on a previous phasing step; hence it can capture allele specific differences but also other forms of heterogeneity. We show some results obtained by looking at chomosome 15, which we picked because it contains a number of well studied imprinted regions [9]. We used the sample NA12878 sequenced by the nanopore consortium [5] as our data set.

2 Results

To make sure we can recapitulate known allele specific methylation, we first looked at some genes known to be imprinted in many tissues (table 1). An example of analysis on GNAS is in figure 1. Supplementary material contain more explicit examples. For each gene we clustered in $k = 2$ groups the reads mapping on the corresponding genomic stretch, and also tagged each read with its haplotype as computed by `WhatsHap`[7]. We then check that each cluster is enriched in a specific haplotype through a Fisher's test: the p-values from the test are also listed in table 1. We did not always look at the entire gene region, rather we sometimes selected stretches where the methylation seems to be at intermediate levels or positions surrounding known regulatory sites: the exact loci are under the "sub-region" heading.

We then divided chr15 in windows of 5Kb and clustered (with $k = 2$) the

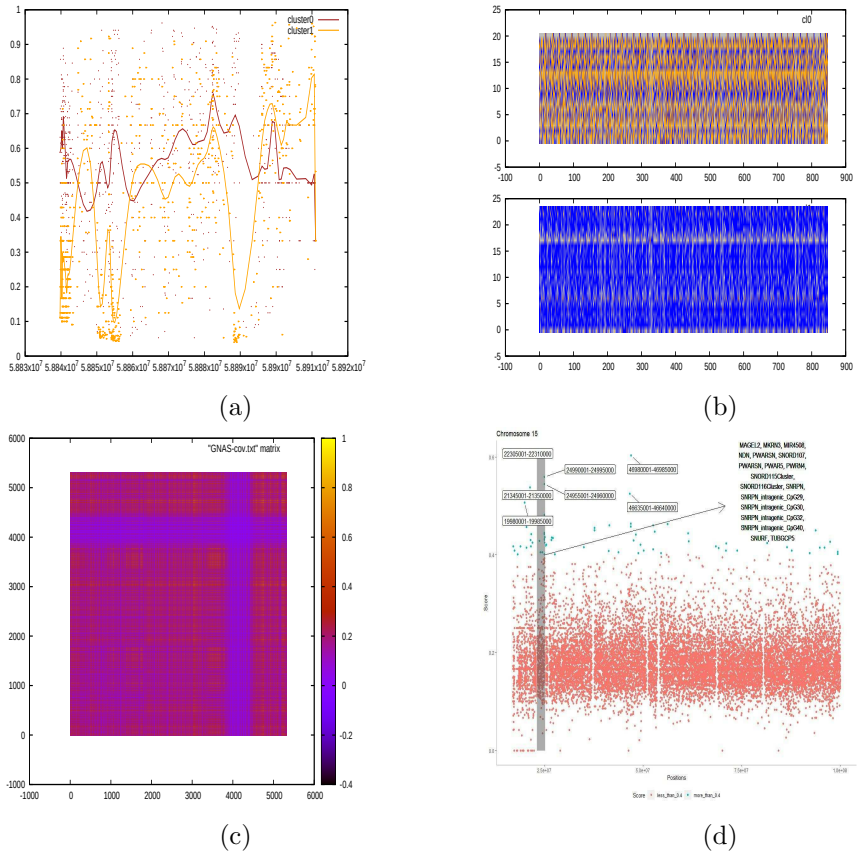

Figure 1: a:Mean values of methylation in the 2 clusters over the GNAS region. b:The two corresponding subpopulations of reads in the subregion $5.88545e7 - 5.88547e7$. (blue=not methylated, orange=methylated, gray=not known). See the main test for the concordance of these data with phasing information. c: covariance matrix of the methylation built using reads mapping on the GNAS gene. d:To build this plot, we scanned chromosome 15 in windows of 5Kb. For each window we computed the clustering ($k = 2$) of the reads and the median absolute difference in methylation across clusters, and plotted it along the chromosome. Some of the peaks overlap with annotated genes as indicated in the plot.

reads mapping on each window; as a measure of goodness-of-fit we compute the median absolute difference in methylation between the clusters (figure 1d). The plot contains a number of peaks which correspond to loci where reads can be well separated into two groups; some of these peaks correspond to known imprinted regions for example the SNRPN gene region.

3 Methods

More details about the methods are given in the Supplementary material. We organize the measurements into a binary matrix (with missing data) X, with n rows representing reads and d columns corresponding to genomic positions. Each position X_{ij} can be 0 (unmethylated) 1 (methylated) or -1 (unknown). Reads are clustered (into k clusters) via a mixture of multivariate Bernoulli distributions. To this purpose, we use an EM algorithm allowing for incomplete observations; the output of the algorithm is :

- a cluster assigment for each read
- a $k \times d$ matrix μ containing the average methylation for each cluster at each position
- the $d \times d$ covariance matrix of the methylation

The goodness of fit of the clustering is defined as the median absolute difference in methylation between all pairs of clusters across all the positions; a post-processing script also computes the posterior probability (position by position) than each cluster is different from the others (in pairs) using the method described in [8]. The methylation status is where possible read directly from CRAM files produced by `Medaka` which appends a methylation tag to each alignment; if methylation is stored in other formats (eg in the output format generated by `Nanometh`[10]) users are responsible for creating a `TAB` separated input file; the command line also allows to specify the number of clusters, a seed for the random initialization of the algorithm and the maximum number of iterations of the EM algorithm. The binary matrix computed as described above has often $> 90\%$ missing positions.

References

[1] J Abante, Y Fang, AP Feinberg, and J Goutsias. Detection of haplotype-dependent allele-specific dna methylation in wgbs data. *Nature communications*, 11(1):1–13, 2020.

[2] Vahid Akbari, Jean-Michel Garant, Kieran O'Neill, Pawan Pandoh, Richard Moore, Marco A Marra, Martin Hirst, and Steven JM Jones. Megabase-scale methylation phasing using nanopore long reads and nanomethphase. *Genome biology*, 22(1):1–21, 2021.

[3] James E Barrett, Andrew Feber, Javier Herrero, Miljana Tanic, Gareth A Wilson, Charles Swanton, and Stephan Beck. Quantification of tumour evolution and heterogeneity via bayesian epiallele detection. *BMC bioinformatics*, 18(1):1–10, 2017.

[4] Fang Fang, Emily Hodges, Antoine Molaro, Matthew Dean, Gregory J Hannon, and Andrew D Smith. Genomic landscape of human allele-specific dna methylation. *Proceedings of the National Academy of Sciences*, 109(19):7332–7337, 2012.

[5] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338–345, 2018.

[6] Gilad Landan, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, Daniela Amann Zalcenstein, Naomi Goldfinger, Adi Zundelevich, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature genetics*, 44(11):1207–1214, 2012.

[7] Marcel Martin, Murray Patterson, Shilpa Garg, Sarah Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schöenhuth, and Tobias Marschall. Whatshap: fast and accurate read-based phasing. *BioRxiv*, page 085050, 2016.

[8] Emanuele Raineri, Marc Dabad, and Simon Heath. A note on exact differences between beta distributions in genomic (methylation) studies. *PLoS One*, 9(5):e97349, 2014.

[9] Wolf Reik and Jörn Walter. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32, 2001.

[10] Jared Simpson. Signal-level algorithms for minion data. https://github.com/jts/nanopolish.

[11] Jared T Simpson, Rachael E Workman, PC Zuzarte, Matei David, LJ Dursi, and Winston Timp. Detecting dna cytosine methylation using nanopore sequencing. *Nature methods*, 14(4):407–410, 2017.