A bioinformatics pipeline to detect imprinting from methyl-seq trio data

Jean-Noël Hubert, Mathilde Perret, Nathalie Iannuccelli, Eva Jacomet, Cédric Cabau, Julie Demars

A bioinformatics pipeline
to detect imprinting from methyl-seq trio data

Jean-Noël Hubert[1], Mathilde Perret[1], Nathalie Iannuccelli[1], Eva Jacomet[1], Cédric Cabau[2], Julie Demars[1]

[1]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France
[2]Sigenae, GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan, France

Methyl-seq pedigrees

Getting methyl-seq data is **minimally invasive**, which facilitates sampling, including from related individuals. Here, we consider usual trios as the primary pedigree unit, but other **familial structures subjected to methyl-seq** can be analyzed as well, even if incomplete. Such data contributes to jointly and reliably detect SNPs and DMRs, and ultimately **to untangle somatic from germline DMRs.**

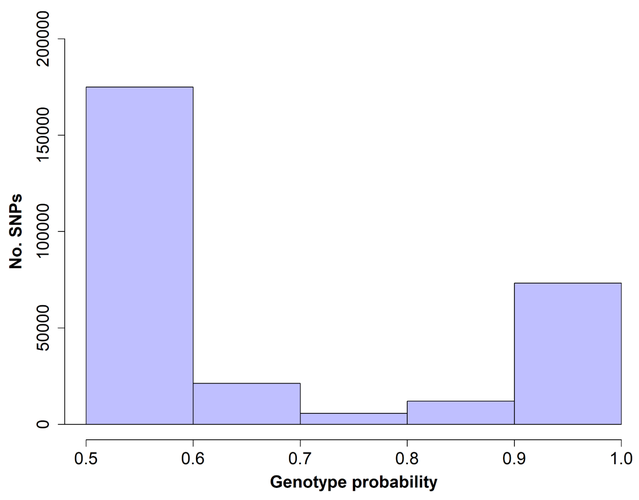

Fig. 1. Genotype probability distribution from our EM-seq RC experiment.
Due to increased bioinformatics processing complexity compared to classical sequence data, methyl-seq reads should preferentially be analysed through bisulfite-aware mapper and caller [1]. In addition, pedigree-based imputation [2] is advantageous since it improves both the completeness and the precision of the SNP calling set, relying on the upper fraction of imputed posterior probabilities. For subsequent analyses, we kept here only genotypes with posterior probability ≥ 0,9.

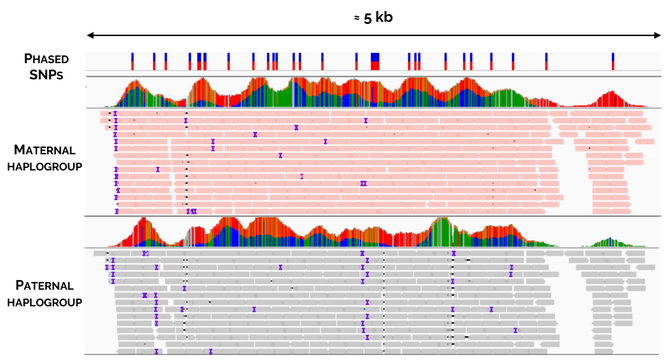

Fig. 2. IGV screenshot of haplotagged reads from our EM-seq RC experiment.
Combining read-based and pedigree-based phasing improves haplotype completeness and precision [3]. This results in the identification of SNP blocks transmitted together to offspring at the scale of imprinted regions. In addition, the parental origin of the inferred haplotypes can be observed at the read level after phasing. Such a feature is especially interesting for GI studies, since it paves the way to the systematic detection of allele-specific DMR patterns from short-read data, even if the physical distance between SNP alleles and their epigenetic regulators is large. IGV: Integrative Genome Viewer.

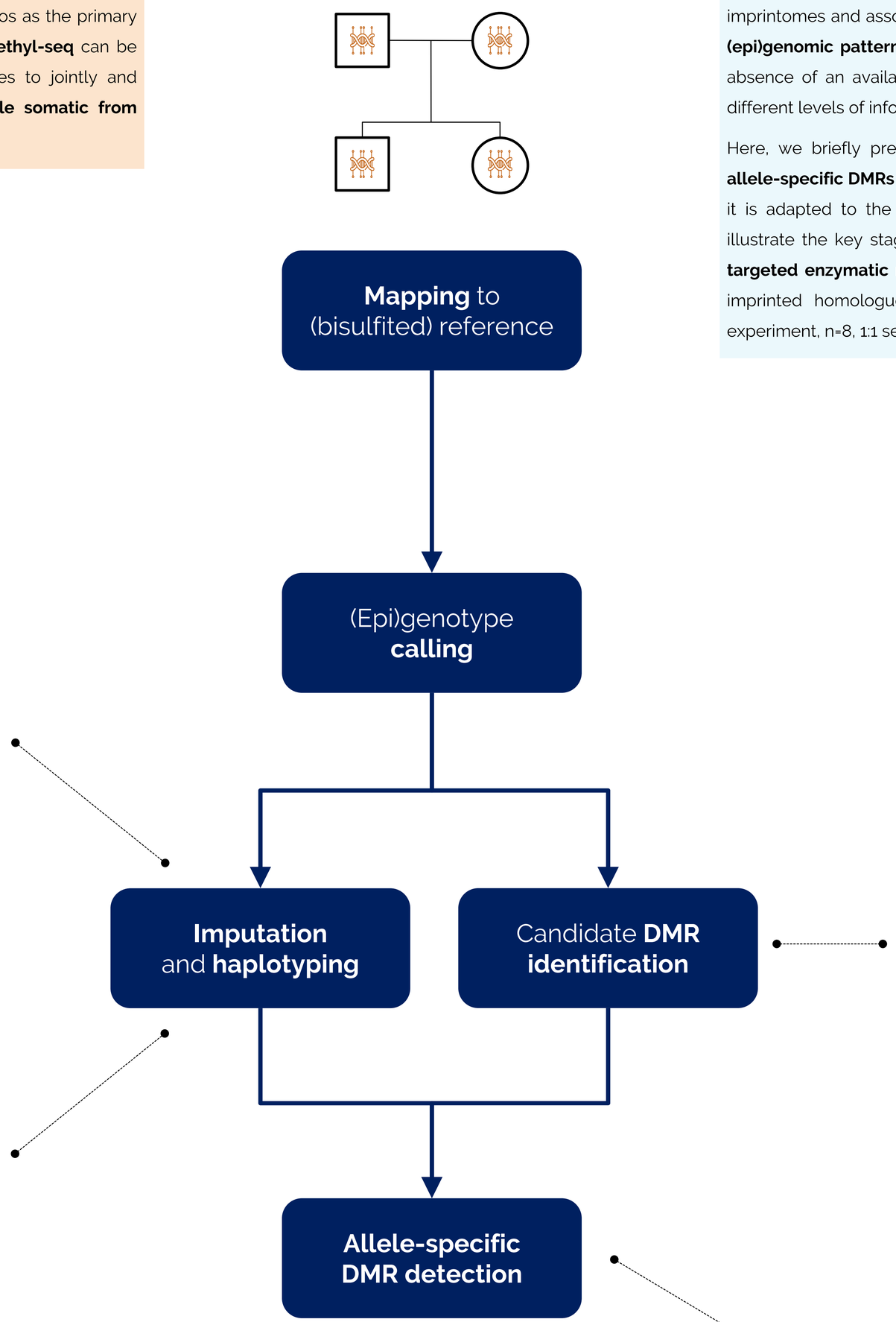

Fig. 3. Schematic breakdown of our pipeline for processing methyl-seq data tailored to the detection of molecular signatures of GI from trio setups.
Colored boxes represent the 5 major analysis steps from raw sequence data to allele-specific DMR detection. The pipeline is based on the combination of published models and software, including GemBS [1], Beagle [2] and Whatshap [3]. It is written primarily in Bash scripting language with some R code and is currently shared at our lab level. We plan for wider distribution via a public code repository and appropriate documentation. RRBS: Reduced Representation Bisulfite Sequencing ; WGBS: Whole-Genome Bisulfite Sequencing; EM-seq: Enzymatic Methyl-seq.

Overview

Coupling methyl-seq data with pedigree information is appealing to look for imprintomes and associated genes, since it allows to **track the transmission of (epi)genomic patterns** to the next generation. However, a difficulty lies in the absence of an available comprehensive procedure to make the most of the different levels of information provided from such setups.

Here, we briefly present a **full bioinformatics pipeline to detect parental allele-specific DMRs** from raw methyl-seq data from related individuals, so that it is adapted to the particular needs of genomic imprinting (GI) studies. To illustrate the key stages of our pipeline, we display results obtained with the **targeted enzymatic methyl sequencing** (EM-seq) of 165 regions enriched for imprinted homologues in a pig **reciprocal cross**-experiment (EM-seq RC experiment, n=8, 1:1 sex ratio).

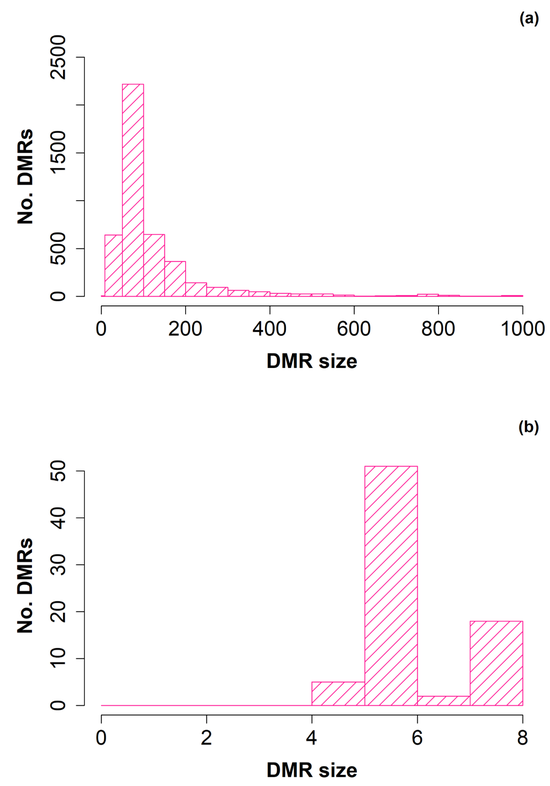

Fig. 4. DMR size distribution from our EM-seq RC experiment.
When it comes to GI, DMR detection stands outside the scope of classical differential methylation analysis and requires a specific approach taking into account hemi-methylation patterns. Here, we met the strictest criteria proposed in [4], looking for groups of at least 5 hemi-methylated CpGs in the range 40%-60% within (a) 100 bp- and (b) 5 bp-sliding windows. Our implementation is easily adjustable, which makes it possible to define a set of personalized cutoffs (e.g., for methylation range, number of CpGs and window size) adapted to the search of particular methylation patterns.

A bioinformatics tool for GI studies

Our pipeline relies on a set of recent and adapted approaches to provide an **all-in-one tool** for the detection of molecular signatures of GI. In addition, our example implementation through EM-seq reciprocal cross data show the **benefit of methyl-seq pedigrees for GI studies**, which allow improving the range of possible inferences and their precision, from SNP calling to parental allele-specific DMR discovery and prioritization.

In particular, we were able to identify known and novel molecular signatures of GI, suggesting the relevance of our bioinformatics tool for the ***de novo* identification of imprinted alleles and epigenomic regulatory patterns** with limited prior information. This type of analysis can be considered from **methyl-seq patient data**, as well as in populations which have been the subject of **little or no GI study**. In addition, it can be a **starting point for more integrated characterization analyses** of imprinted clusters, using multi-level and multi-type omics data, when available.

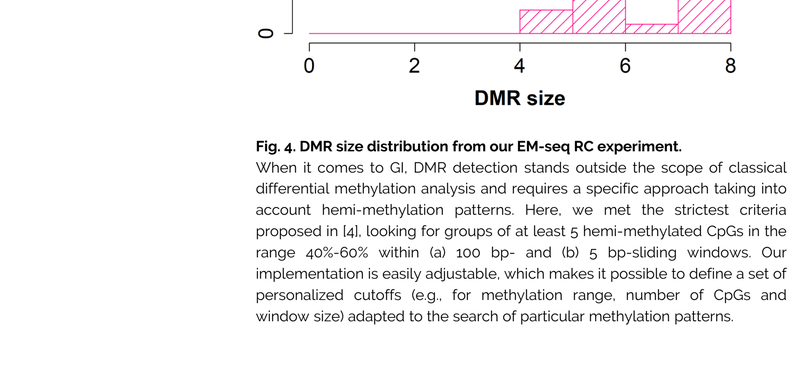

Fig. 5. Example of allele-specifc DMR pattern from our EM-seq RC experiment.
Combining information from the haplotype and methylation detection branches of our pipeline makes it possible to identify local patterns suggesting the presence of allele-specific DMRs. Here, we show a candidate region for GI with a clear pattern of allele hyper/hypo-methylation dependent on the parental origin in a F1 individual. The plot depicts the chromosome, variant and methylation levels (at the DMR site) of maternal (in red, bottom half) and paternal (in blue, top half) origins.

[1] Merkel, Angelika, et al. "gemBS: high throughput processing for DNA methylation data from bisulfite sequencing." *Bioinformatics* 35.5 (2019): 737-742.
[2] Browning, Brian L., and Sharon R. Browning. "Genotype imputation with millions of reference samples." *The American Journal of Human Genetics* 98.1 (2016): 116-126.
[3] Garg, Shilpa, Marcel Martin, and Tobias Marschall. "Read-based phasing of related individuals." *Bioinformatics* 32.12 (2016): i234-i242.
[4] Jima, Dereje D., et al. "Genomic map of candidate human imprint control regions: the imprintome." *Epigenetics* 17.13 (2022): 1920-1943.

Projet cofinancé par le Fonds Européen de Développement Régional

Genomic Imprinting – from Biology to Disease
22-24 March 2023
Wellcome Genome Campus, UK