DR GABRIELE USAI (Orcid ID : 0000-0001-9982-4883)

DR FLAVIA MASCAGNI (Orcid ID : 0000-0001-9747-8040)

Article type : Resource

TITLE

Epigenetic patterns within the haplotype phased fig (*Ficus carica* L.) genome

AUTHORS

G. Usai[1a]*, F. Mascagni[1a], T. Giordani[1], A. Vangelisti[1], E. Bosi[2], A. Zuccolo[3], M. Ceccarelli[4], R. King[5], K. Hassani-Pak[5], L.S. Zambrano[6], A. Cavallini[1] and L. Natali[1]*

AUTHOR AFFILIATIONS

[1]Department of Agriculture, Food and Environment, University of Pisa, Pisa, Italy; [2]Department of Biomedical Experimental and Clinical Sciences, University of Florence, Florence, Italy; [3]Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy; [4]Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy; [5]Rothamsted Research, Harpenden, England; [6]Facultad de Ciencias Zootécnicas, Universidad Técnica de Manabí, Portoviejo, Ecuador.

[a]These authors contributed equally to this study

***CORRESPONDING AUTHOR:** gabriele.usai@agr.unipi.it; lucia.natali@unipi.it

[1]Department of Agriculture, Food and Environment, University of Pisa, Pisa, Italy

SHORT RUNNING TITLE: Haplotype phased fig genome

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.14635

KEYWORDS: *Ficus carica* L., single-molecule, real-time sequencing, genome assembly, N^6-methyladenine, N^4-methylcytosine.

SUMMARY

Due to DNA heterozygosity and repeat content, assembly of non-model plant genomes is challenging. Herein, we report a high-quality genome reference of one of the oldest known domesticated species, fig (*Ficus carica* L.), using Pacific Biosciences single-molecule, real-time sequencing. The fig genome is ~333 Mbp in size, of which 80% has been anchored to 13 chromosomes. Genome-wide analysis of N^6-methyladenine and N^4-methylcytosine revealed high methylation levels in both genes and transposable elements, and a prevalence of methylated over non-methylated genes. Furthermore, the characterization of N^6-methyladenine sites led to the identification of ANHGA, a species-specific motif, which is prevalent for both genes and transposable elements. Finally, exploiting the contiguity of the 13 pseudomolecules, we identified 13 putative centromeric regions. The high-quality reference genome and the characterization of methylation profiles, provides an important resource for both fig breeding and for fundamental research into the relationship between epigenetic changes and phenotype, using fig as a model species.

INTRODUCTION

The emergence of third-generation sequencing (TGS) technologies (Ansorge *et al*., 2016) including single-molecule, real-time (SMRT) sequencing developed by Pacific Biosciences (PacBio), has led to the production of high-quality genome assemblies (Daccord *et al*., 2017; Vogel *et al*., 2018; Ye *et al*., 2018). Long-read sequencing has overcame the problems with short-read sequencing assembly, i.e. resolution of repetitive components of the genome (Veeckman *et al*., 2016) and phasing of heterozygous regions (Low *et al*., 2019).

Resolving heterozygosity by generating a haplotype phased reference genome, is necessary for a complete comprehension of a species biology and evolution, and for trait focused applications such as genomic selection (Meuwissen *et al*., 2013). Unphased genomes can have differing random haplotypes represented in a single genome assembly sequence, resulting in errors that effect its use (Korlach *et al*., 2017). The resulting unphased haplotypes, can lead to gene annotation issues due to insertion/deletion polymorphisms (indels) presence, causing incorrect frameshifts, and breaks in the genome sequence from repeat content differences in intergenic regions between haplotypes. TGS technologies provide a significant improvement towards a solution to these problems: taking advantage of long-reads that span haplotype differences to generate contigs that define one haplotype only and represent the other as a separate sequence, thus phasing

haplotypes from a single organism, without the need to generate clones or sequence parents (Kronenberg *et al*., 2018) and producing higher quality, contiguous genomes.

Additionally, SMRT sequencing can be used for the identification of DNA sequence modifications, including N^4-methylcytosine (4mC) and N^6-methyladenine (6mA) methylation (Rhoads and Au, 2015). Local DNA methylation variants can have major effects on the transcription of neighbouring genes and can be inherited over generations (Diez *et al*., 2014). DNA methylation of 6mA has been found to be a potentially epigenetic marker in several eukaryotes (Liang *et al*., 2018; Zhou *et al*., 2018). Nevertheless, the specific distribution of motifs and potential effects of such modifications in plant genomes, remain poorly understood.

The fig tree (*Ficus carica* L., *Moraceae*) is a heterozygous species (Mori *et al*., 2017) widely grown for its fruit throughout the temperate world. This crop is one of the oldest known domesticated species, whose cultivation has been practiced for 12 thousand years (Kislev *et al*., 2006).

Fig has valuable nutritional and nutraceutical characteristics (Vinson *et al*., 2005; Solomon *et al*., 2006; Veberic *et al*., 2008) and the ability to adapt to marginal soils and difficult environmental conditions (Vangelisti *et al*., 2019). However, rapid perishability of its fresh fruits, and deficiencies to its responses to abiotic stresses and new diseases, restricts its world distribution and commercial success. The availability of a high-quality reference genome would provide an important resource to genetic improvement and breeding programmes, to improve figs ability to be cultivated and distribute its fruit on a more extensive global scale. Recently, Mori *et al*. (2017) released a very preliminary genome sequence of a Japanese cultivar of *F. carica*, the cultivar Horaishi, which was affected by the typical deficiencies of short-read genome assemblies (Veeckman *et al*., 2016; Low *et al*., 2019).

Herein we present a haplotype phased reference genome and methylome of the most prominent Italian fig cultivar, Dottato, a cultivar of ancient origin, cultivated in Greece since 450 BC for fresh or dried fruit consumption and introduced into southern Italy prior to the 6th century BC (Mafrica *et al*., 2017).

Our *de novo* assembly of PacBio data using FALCON-Unzip (Chin *et al*., 2016) and published genetic marker information from Japanese cultivar Horaishi (Mori *et al*., 2017) resulted in 13 pseudomolecules which represent the 13 chromosomes of fig. This chromosome-scale assembly was complemented by a detailed *de novo* annotation of the genes and of the non-coding component of the genome. Furthermore, we produced a genome-wide DNA methylation map at single-base resolution, reporting global profiling of 6mA and 4mC sites. This analysis led to the identification of several DNA methylation motifs and patterns present in the fig genome, genes and mobile elements.

This work represents an attempt of sequencing the genome of a heterozygous plant by long-read assembly, providing a high-quality genomic resource for future genetic and epigenomic studies in fig, potentially useful for the breeding of this promising commercial crop.

RESULTS

Fig DNA sequencing

The genomic DNA of a wild-type diploid *F. carica* cultivar Dottato female plant was extracted from young leaf tissue and sequenced using the SMRT technology of PacBio. The process generated 2,140,959 long-reads (minimum length = 1,000 bp; maximum length = 111,426 bp; average length = 12,364 bp) with an N50 value of 18,419 bp, corresponding to ~74-fold genome coverage of the 356-Mbp fig haploid genome (as estimated by flow cytometry, Loureiro *et al*., 2007) (Figure 1a). The dataset read length distribution showed a typical PacBio sequencing profile with more than half of the data greater than 20 kbp (Figure 1b).

Genome assembly and curation

We carried out a preliminary analysis testing three different long-read assembly tools to verify which one is the most suitable to accurately and continually assemble the fig genome. We run our preliminary tests with the hybrid assembler DBG2OLC (Ye *et al*., 2016) and with two different non-hybrid assemblers, Canu (Koren *et al*., 2017) and FALCON/FALCON-Unzip (Chin *et al*., 2016). DBG2OLC and Canu produced assemblies showing a lower N50 value compared to FALCON, indicating a more fragmented assembly (Table S1). For this reason, we chose FALCON as the most appropriate assembler and decided to optimize the software parameters to further improve the overall assembly quality.

We setup a pipeline to take full advantage of the PacBio long-reads and construct a high-quality genome assembly (Figure 1c). The diploid FALCON-Unzip assembler produced a primary set of contigs and a set of linked haplotigs that represent the alternative genome structures of the primary contigs (i.e. single nucleotide polymorphisms and structural variations) (Table S2). The primary assembly was upgraded using FinisherSC (Lam *et al*., 2015) and then polished, together with the haplotigs, using Arrow (Chin *et al*., 2013) and Pilon (Walker *et al*., 2014). After excluding fungal and bacterial contamination, we measured the core gene completeness using BUSCO (Simão *et al*., 2015). BUSCO recovered 1,283 of the 1,375 (93.3%) highly conserved Embryophyta core genes, of which 1,177 (85.6%) were complete and single-copy and 106 (7.7%) were complete and duplicated, thirty-five genes (2.5%) were fragmented, and 57 (4.2%) missing. The BUSCO assessment of haplotigs is reported in Table S3.

A comparison with the assembly statistics and the annotation of the previously produced fig genome assembly (Mori *et al*., 2017) showed our assembly is more contiguous and has a higher percentage of annotated genes and repeats (Table 1). Overall, the process produced 905 primary contigs and 6,933 haplotigs. The primary contigs had a mean contig size of 368,398 bp (min size = 20,012 bp, max size = 5,010,936 bp) and N50 of 823,517 bp. We produced a total of 333,400,567 bp of the fig genome sequence, corresponding to ~95% of the estimated size. We observed that some haplotigs overlapped a part of their flanking regions to other adjacent-positioned haplotigs. For this reason, the haplotigs total size was larger than the size of primary contigs. This fact is probably due to heterozygosity and structural variations that

lead to complications during the assembly process. The mean contig size of the haplotigs was 58,872 bp (max size = 1,220,129 bp) and the N50 was 89,539 bp (Table S4).

Since the haplotigs represent the alternative genome structures of the primary contigs, they were used to estimate the heterozygosity of the Dottato cultivar. A genome-wide comparison between primary contigs and haplotigs allowed us to identify a total of 903,428 single nucleotide polymorphisms (SNPs) and indels, for an overall heterozygosity of approximately 2.7 polymorphisms per kilobase (0.27%). The heterozygosity was estimated also through SNP and indel calling from the aligned Illumina reads against the primary assembly, showing similar results. A total of 954,749 SNPs and indels were found, accounting for a heterozygosity of approximately 2.8 polymorphisms per kilobase (0.28%). Furthermore, we were able to identify the chloroplast and mitochondria genomes, represented by two circularized sequences of 160,594 bp and 230,028 bp, respectively.

Only the primary assembly was used in the downstream annotation and DNA modification analysis. We used the Illumina scaffolds of the cultivar Horaishi assembly as queries to order and orientate the contigs of our primary assembly according to the physical structure of the thirteen chromosomes of fig as established by Mori *et al*. (2017). A total of 407 primary contigs, corresponding to the 80% of the total assembly (266,522,563 bp), were associated to fig chromosomes (Table S5), producing a set of 13 pseudomolecules (Figure 2). Furthermore, this process allowed us to produce a genetic-to-physical comparative map (Table S6). It is important to point out that large structural variants may have been missed since a different cultivar was used for the anchoring process. Further information about the haplotigs distribution along the pseudomolecules is reported in Table S7.

Repeats prediction and annotation

For a comprehensive annotation of the repeat sequences, represented by transposable elements (TEs) and tandem repeats, the genome assembly was scanned with both structural- and homology-based predictive tools (Figure 1d). We identified a total of 123.8 Mbp of repeat sequences, representing 37.39% of the genome assembly (Figure 3a). TEs represent the most abundant repeats, covering the 33.57% of the assembly (111.06 Mbp), while tandem repeats represent 3.82% (12.74 Mbp). The most abundant TEs are retrotransposons or Class I elements, representing 84.95% of the repetitive content and 28.3% of the genome assembly (94.36 Mbp). Long terminal repeat retrotransposons (LTR-REs) are the most represented, accounting for 99.06% of this class and 28.03% of the total genome assembly (93.48 Mbp), whereas non-LTR REs (LINEs and SINEs) accounted for 0.94% (0.88 Mbp). Among the LTR-REs, *Gypsy* and *Copia* are the most abundant superfamilies, representing 16.36% (54.56 Mbp) and 8.74% (29.14 Mbp) of the genome assembly, respectively. DNA transposons or Class II elements represent 15.05% of the repetitive content and 5.01% of the genome assembly (16.7 Mbp). The complete list of identified TEs, their copy number, annotation and genome proportion can be found in Table S8.

We scanned the assembled sequences searching for the centromeric regions, typically represented by high repetitive tandem units. Overall, a site containing highly-repetitive 103 bp-long tandem repeats was identified once for each pseudomolecule and considered as the putative centromeric repeat. A consensus sequence of this putative centromeric tandem repeat is reported in Data S1. This observation was further supported by the high abundance of *Chromovirus* elements, which are typically found in the centromeric structures (Neumann *et al.,* 2011), in the above-mentioned sites. We isolated a total of 42 centromeric contigs, representing the thirteen putative centromere regions of fig. Each region was represented by an average number of 3.23 contigs and had an average length size of 2.86 Mbp. Further statistics on the centromeres can be found in Table S9. The distribution of the tandem repeat and the *Chromovirus* elements over the thirteen centromeric regions was profiled (Figure S1). No specific distribution was observed in these centromeric sequences for the LTR-REs and the centromeric tandem repeats, however a slightly greater abundance of LTR-REs was found in the central region of the putative centromeres.

Gene prediction and annotation

To produce a comprehensive gene annotation, we used a combination of *ab initio* and transcriptome-based strategies. The transcriptome assembled in this work consisted of 127,470 contigs (185,504,905 bp) with an average length of 1,455 bp and median length of 1,118 bp (Table S10). In total, we identified 37,840 protein-coding genes (Table S11), which represent 27.91% of the total genome assembly (93.05 Mbp), and 1,685 non-protein-coding genes (Table S12), representing 0.15% (0.5 Mbp) (Figure 3a). Gene average length was 2,460 bp and CDS average length was 956 bp. The average exon number per gene was 4.56, with an average length of 251 bp, while intron average length was 367 bp. In total, exon length was 43.5 Mbp, while intron length was 49.55 Mbp.

Functional annotation showed that 28,737 of the predicted protein-coding genes (76% of the total) had a BLAST hit (e-value < 0.001) in NCBI nr (Table S13). The number of protein-coding genes and their attributed Gene Ontology (GO) annotation is shown in Figure 3b, categorized by the three ontologies cellular component, molecular function and biological process.

We identified the sex determining region location by searching, in the Dottato assembly, for the female sex-associated *RESPONSIVE-TO-ANTAGONIST1* (*RAN1*) gene, encoding a copper-transporting ATPase, whose role in fig sex determination was identified by Mori *et al*. (2017) in several genotypes. Specifically, the sequence was located on the Chr1 pseudomolecule (contig FCD_7, coordinates: 401,567-407,422). The Dottato *RAN1* gene sequence showed 99.9% similarity with the same gene sequence characterized by Mori *et al*. (2017), confirming the female sex of the sequenced Dottato plant. As already discussed by Mori *et al*. (2017), the region of approximately 100 kbp upstream of the *RAN1* gene can be considered as the sex determining region.

DNA 4mC and 6mA modification analysis

The genome-wide base modification landscape of the fig genome was determined by SMRT sequencing. We focused on 4mC and 6mA modifications and their distribution patterns at both gene and TE levels (Figure 2d-e). The profile of both modifications shared a very similar distribution profile for all analysed regions.

Globally, 1,515,639 4mC (Table S14) and 984,245 6mA (Table S15) sites were observed, corresponding to 0.45% and 0.30% of the haploid assembly, respectively. As expected, the most significantly enriched 4mC motifs were CG (e-value = 1.9e-405), CHG (e-value = 3.3e-203) and CHH (e-value = 3.0e-180), where H is adenine, cytosine or thymine (Figure 4a). As regards to the 6mA modifications, we found that ANHGA (e-value = 3.2e-4103), GAGG (e-value = 3.5e-354), CAAG (e-value = 1.2e-302) and ACCT (e-value = 1.7e-600) were the most significantly enriched motifs, where N can be any nucleotide (Figure 4b). Most 6mA sites could not be associated to any specific motif (e-value = 3.7e-555).

We found a total of 41%, 38% and 21% of methylated sites in the CHG, CG and CHH sequence logos (Figure 4c, left), respectively. For the 6mA modifications, we found a total of 28%, 9%, 4% and 1% of methylation sites for ANHGA, GAGG, CAAG and ACCT sequence logos, respectively (Figure 4c, right). The 58% of 6mA modifications could not be associated with any motif. The complete distribution of 4mC and 6mA sites per motif and pseudomolecule can be found in Table S16 and Table S17, respectively.

Furthermore, we investigated the 4mC and 6mA sites on predicted genes (Figure 4d). Predicted gene sequences were composed of 30.69% of adenines and 19.31% of cytosines, respectively. 13.04% of the gene cytosines were methylated, while only 5.94% of the gene adenines were methylated. A total of 21,029 genes (55.57%) showed, at least once, both methylation sites, while 8,052 (21.27%) and 1,228 (3.24%) showed, at least once, a 4mC or 6mA site, respectively. The remaining 7,531 (19.9%) genes did not present any base modification. Among all the protein coding genes, we predicted a total of 11,528 5'-UTRs, of which 643 (5.57%) showed at least once, the occurrence of both 4mC and 6mA sites, 1,119 (9.7%) with at least a 4mC site and without 6mA sites, 1,365 (11.84%) with at least a 6mA site and without 4mC sites. In respect to the 3'-UTRs, 1,368 (11.55%) of the total 11,838 identified, presented both methylation sites, 1,714 (14.47%) presented at least a 4mC site and 1,594 (13.46%) presented at least a 6mA site. Interestingly, although both 5'-UTRs and 3'-UTRs presented less methylated sites compared to gene bodies, these regions seemed to be subjected to a 4-fold higher level of 6mA.

To better investigate the gene body methylation patterns, 20,925 genes were grouped by exon number classes, from two up to six, and their 4mC and 6mA profiles were plotted separately (Figure 5). A 1-kbp upstream and downstream window was taken from the position of transcription start site (TSS) and transcription end site (TES). Genes sorted by the total number of exons showed a negative trend between exon number and the overall methylation level. In particular, genes with two exons seem more prone to 4mC modification compared to genes with a higher exon number. In general, global methylation level of

4mC and 6mA within the exons was higher than in the introns. In addition, the profiles showed that 4mC modification levels were higher compared to 6mA modifications.

Methylated genes were analysed to investigate the distribution of GO terms and determine if a possible relationship between gene function and number of methylated sites was observed. As already shown, most genes presented both 4mC and 6mA modifications, and in all GO categories, genes with both modified bases were the majority (Figure S2), so there appeared no correlation.

Finally, we compared 4mC and 6mA profiles at gene and TE levels. We found that 15.22% and 17.13% of the 4mC and 6mA sites, respectively, were identified within genic regions, whilst most of the modified sites, 69.58% and 66.03% of 4mC and 6mA sites, were identified over the TE bodies. Since the gene bodies and TEs represented respectively 28.06% and 37.39% of the assembly, the levels of 4mC and 6mA modifications in TEs were 3-fold those of genes. We profiled the 4mC and 6mA distribution, investigating the different motif patterns that occur in both genes and TEs (Figure 6). The most represented 4mC motifs for both genes and TEs were CG and CHH, while the CHG motif was underrepresented (Figure 6, up). Interestingly, CG had higher methylation than CHH in genes, while the opposite trend was observed in TEs. For both sequence types, the TSS downstream and the TES upstream regions were less methylated compared to the body region. The 6mA profile showed that the most represented modifications could not be associated to any specific motif (Figure 6, down). The most represented specific 6mA motif for both genes and TEs was ANHGA. Similar levels of adenine modification were observed in gene bodies for the ANHGA and GAGG motifs. GAGG was less methylated than ANHGA in the TEs profile. The CAAG and ACCT motifs were the least represented.

DISCUSSION

A prerequisite to accelerate innovation via breeding or other means is the availability of a high-quality reference genome and gene annotations. In the last years, since the introduction of high-throughput sequencing technologies, the number of genome sequencing projects has increased at an exponential rate, leading to genome assemblies for organisms previously considered too low priority to be sequenced. For instance, high-quality genome assemblies were released for *Potentilla micrantha* (Buti *et al*., 2017), *Salvia splendens* (Dong *et al*., 2018), *Cuscuta campestris* (Vogel *et al*., 2018) and *Casuarina equisetifolia* (Ye *et al*., 2019).

Despite that *Ficus carica* L. has a predicted small genome size (Mori *et al*., 2017), heterozygosity represents a challenge to high-quality genome assembly. Long-read phased assemblies provide a suitable solution with minimal assembly and gene prediction errors (Korlach *et al*., 2017) versus alternate short-read assemblies. Phasing haplotypes from a single organism, without generating clones or sequencing the parents is one of the ultimate goals of current genome assembly projects. Diploid (haplotype phased) reference genomes have begun to increase due to leveraging the potential of third generation long-reads,

which can be used to generate contigs with enough sensitivity to define the alternative haplotypes (Schwessinger *et al*., 2018; Low *et al*., 2019). In plants, the first diploid reference genomes have been generated for *Durio zibethinus* (Teh *et al*., 2017), *Scutellaria baicalensis* (Zhao *et al*., 2018) and *Prunus* × *yedoensis* (Shirasawa *et al*., 2019) using Pacific Biosciences (PacBio) long-reads sequencing.

In this study, we generated a haplotype phased assembly for fig with a N50 of ~823 kbp, which in combination with the physical information released by Mori *et al*. (2017) for another genotype of *F. carica*, the Japanese cultivar Horaishi, has yielded highly contiguous pseudomolecules.

The estimated heterozygosity of fig (approximatively 0.27%) was similar to that of black cottonwood (0.26%) (Tuskan *et al*., 2006), but higher than those of papaya (0.06%) (Ming *et al*., 2008), pigeonpea (0.067%) (Varshney *et al*., 2012) and *Prunus mume* (0.03%) (Zhang *et al*., 2013). On the other hand, fig heterozygosity was lower than that of date palm (0.46%) (Al-Dous *et al*., 2011), pear (1.02%) (Wu *et al*., 2013.) and jujuba (1.90%) (Liu *et al*., 2014).

Comparing our results to the previous fig genome draft, it is evident that the use of long-read sequencing and high depth coverage produced a genome of high sequence contiguity and accuracy, and consequently improved gene and transposable element (TE) annotation (Table 1). A full characterization of the repetitive component is a crucial process to decipher the genome structure and, by using PacBio long-reads and our repeat identification pipeline (Figure 1d), we were able to significantly increase TE detection across the final assembly compared to the previous version of the fig genome (Mori *et al*., 2017). Furthermore, another goal in using long-read sequencing was to characterize the centromere regions, which remain largely unknown in short-read assemblies, typically difficult to assemble due to their highly repetitive composition. Exploiting the contiguity of the 13 pseudomolecules produced, we identified 13 putative centromeric regions (Data S1) and provided an evaluation into the structure of these chromosomal regions.

Finally, we performed a genome-wide analysis of base modification in relation to N^4-methylcytosine (4mC) and N^6-methyladenine (6mA). Modifications of DNA bases, such as methylation, can affect several cellular processes including gene regulation and DNA replication or repair (Bierne *et al*., 2012). Next-generation sequencing (NGS) technologies are generally limited in the identification of the majority of DNA modifications. PacBio sequencing uses an unusual approach to detect native epigenetic modifications by monitoring time between base incorporations in the read strand. Focusing on those modifications which are detected with high accuracy due to strong kinetic variation signals (Rhoads and Au, 2015), we identified 984,245 and 1,515,639 sites of 6mA and 4mC, respectively.

In bacterial genomes, it is established that 6mA and 4mC act as components of restriction-modification systems (Murray *et al*., 2012). In recent years, taking advantage of the latest technologies, 6mA has been found to be a potential epigenetic marker in multicellular eukaryotes (Fu *et al*., 2015; Zhang *et al*., 2015; Koziol *et al*., 2016; Zhou *et al*., 2018). In rice, for instance, the 6mA distribution profile is complementary to that of 5mC, reinforcing the hypothesis of a distinct role as epigenomic markers in plants (Zhou *et al*.,

2018). As already shown, 6mA consensus motifs are only partially conserved in eukaryotic species (Liang *et al*., 2018). Our genome-wide analysis of fig has identified the majority of eukaryotic species conserved motifs, and one species-specific motif, ANHGA. For those 6mA sites of which it was not possible to infer a defined motif, we retrieved some conserved motifs: GAGG, that is present in both plants and animals (Wu *et al*., 2016), ACCT, already identified in Arabidopsis (Liang *et al*., 2018), and CAAG, found in *Casuarina equisetifolia* (Ye *et al*., 2019).

Similarly to what is observed in Arabidopsis (Liang *et al*., 2018), our results showed that 6mA sites within gene bodies are mostly located in exons (Figure 5). Furthermore, for TEs (Figure 6), 6mA sites exhibited a local depletion at the TSS followed by an intense peak, similarly to what was found in Arabidopsis and in *Chlamydomonas reinhardtii* (Fu *et al*., 2015; Liang *et al*., 2018).

Classical cytosine methylation of eukaryotic genomes has often been related to the repetitive component (Feng *et al*., 2010). TEs are predominantly located in the heterochromatin regions of the centromeres, usually showing high level of 5mC (Reinders *et al*., 2009). As shown in Figure 2, methylation profiles of both 4mC and 6mA indicate a prevalence of methylated bases in correspondence to regions which are the most abundant in repeats (as, for example, the putative centromeres), in accordance with what has been observed in the aforementioned studies.

Due to reducing PacBio sequencing costs, it would be practical to produce multiple assemblies, from different tissues. These assemblies could each have their own specific associated polymorphisms and epigenetic information. Therefore, it would be possible to infer multi-tissue epigenetic analyses in the future, using the work we have presented herein, as a foundation for such studies.

In conclusion, the production of a high-quality reference genome sequence for *F. carica* presents interesting insights into the genomic structure of a heterozygous plant diploid species. Furthermore, the characterization of the methylation profile of two unconventional DNA modifications provides a basis to investigate how specific epigenetic changes affect the phenotype of plants. Further analysis will be conducted to improve the representation of the alternative haplotype sequence and to investigate how heterozygosity and methylation levels effects allelic changes of gene expression.

MATERIALS AND METHODS

DNA extraction and genome sequencing

The sample for genome sequencing was collected from young leaf tissue (0.5-1 cm in diameter) obtained from a single female plant of the Italian *F. carica* cultivar Dottato. Genomic DNA was isolated using the CTAB method as modified by Mascagni *et al*. (2017; 2018). During the gel electrophoresis we used the Lambda Eco/Hind marker for sizing of the linear double-stranded DNA and isolating only the fragments with a molecular weight greater than 20,000 bp.

SMRT sequencing of the genomic library was performed on a PacBio Sequel system (V2 chemistry) at the Arizona Genomics Institute (Tucson, AZ, USA). ~300 µg high molecular weight DNA was submitted for sequencing using 6 SMRT cells, in accordance with the manufacturer protocols (Pacific Biosciences).

Genome assembly and curation

The assembly of the PacBio reads was performed using three different long-reads assemblers to find the best approach to resolve the heterozygosity of fig and provide a continuous genomic sequence. First, the hybrid assembler DBG2OLC v04062015 (Ye *et al*., 2016) was run using already available Illumina HiSeq 2000 paired-end reads of the same cultivar (Zambrano *et al*., 2017). The Illumina reads were trimmed with Trimmomatic v0.32 (Bolger *et al*., 2014) to remove adapters and low-quality regions (ILLUMINACLIP:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:25) and then assembled using SparseAssembler v20160205-4 (Ye *et al*., 2012) with default parameters. The Illumina assembly was submitted to DBG2OLC with the PacBio reads to produce the hybrid assembly. We ran the assembly process with default parameters and with different values of KmerCovTh (ranging from 2 to 8), MinOverlap (ranging from 20 to 150) and AdaptiveTh (ranging from 0.001 to 0.02). Next, we used Canu v1.5 (Koren *et al*., 2017) with default parameters and then with specific settings to avoid collapsing diverged repeats and haplotypes (corOutCoverage=200 correctedErrorRate=0.040 "batOptions=-dg 3 -db 3 -dr 1 -ca 500 -cp 50"). Finally, FALCON v0.5 (Chin *et al*., 2016) was run with default parameters. Errors in the PacBio reads were corrected within assembler pipelines. The DBG2OLC and Canu assemblies had a lower N50 value compared to the FALCON assembly, indicating a less contiguous assembly. For that reason, we decided to choose FALCON and to optimize the parameters to further reduce the assembly fragmentation. We compared different length_cutoff values for the raw reads length (ranging from 1,000 to 20,000) as well as for the length_cutoff_pr values for the preassembly stage (ranging from 1,000 to 20,000) and the seed_coverage values (ranging from 20 to 30). On the basis of the contig N50 results, we chose length_cutoff = 10000, length_cutoff_pr = 19000 and seed_coverage = 24. We additionally optimized pa_HPCdaligner_option (-T8 -b -v -B128 -e0.70 -M36 -l2500 -k18 -w8 -s100 -h1250), ovlp_HPCdaligner_option (-T8 -b -v -B128 -M36 -k24 -h1250 -l1500 -e.96 -s100), pa_DBsplit_option (-a -x500 -s100), ovlp_DBsplit_option (-s400), falcon_sense_option (--output_multi --min_idt 0.70 --min_cov 2 --max_n_read 400) and overlap_filtering_setting (--max_diff 120 --max_cov 200 --min_cov 2), obtaining specific parameters for our plant species. The resulting collapsed primary assembly and the correlated output files were used to perform the phasing step with FALCON-Unzip (Chin *et al*., 2016) using default parameters, obtaining a dataset of contigs, called primary contigs and haplotigs, the latter representing the alternative genome structure (structural variations and SNPs) of the primary assembly. The primary assembly was further scaffolded using the post-processing tool FinisherSC v2.0 (Lam *et al*., 2015), which uses MUMmer v3.23 (Kurtz *et al*., 2004). The primary assembly and the haplotigs were polished using

Arrow v2.2.2 (Chin *et al*., 2013) over three iterations, performing consensus and variant calls. The alignment was performed using BLASR v5.3 (Chaisson and Tesler, 2012) and then sorted and indexed with SAMtools v1.3.1 (Li *et al*., 2009). Finally, the primary assembly and the haplotigs were polished again using Pilon v1.13 (Walker *et al*., 2014) over three iterations using already available Illumina HiSeq 2000 paired-end reads of the same genotype (Zambrano *et al*., 2017), that were trimmed as described above. The alignment was performed using BWA v0.7.13-r1126 (Li and Durbin, 2009) and then sorted and indexed with SAMtools.

We performed a BLASTN search (v2.6.0) (Boratyn *et al*., 2012) using the primary contigs and the haplotigs against the whole NCBI nucleotide collection (word_size 11 extension_threshold 20 gapped_alignment banded nucleic_match 1 nucleic_mismatch -3 open_penalty -5 extend_penalty -2 -max_scores 1 -num_alignments 1 -threshold significance=1e-5). After that, MEGAN v6.5.8 (Huson *et al*., 2007) was used to perform a phylogenetic analysis on the BLASTN output to verify if some contigs end up in fungal or bacterial contamination branches. BUSCO v3.0.2 (Simão *et al*., 2015) was used to provide a quantitative measure of the genome assembly completeness regarding genes. We ran BUSCO both on primary assembly and haplotigs, and primary contig annotation using the Embryophyta v10 plant set (https://busco.ezlab.org) with default parameters.

We used MUMmer to align the haplotigs to the corresponding primary contigs in order to retrieve the correct position of every haplotig and remove duplicated haplotigs. The alignment was performed using the nucmer command with default parameters. After that, the alignment itself was used to estimate the heterozygosity of the Dottato cultivar by calling and counting the SNPs and indels positions across the primary assembly. The SNP and indel calling was performed using the show-snps command. Furthermore, the heterozygosity was estimated by using the previously produced BWA alignment between Illumina reads and primary assembly. We used FreeBayes v1.0.0 (Garrison and Marth, 2012) for SNP and indel calling (-C 5) and VcfFilter v1.0.0 (https://github.com/vcflib/vcflib) to filter by quality (QUAL >= 30).

To order and orientate the primary contigs, we used the Illumina scaffolds of the *F. carica* cultivar Horaishi assembly, ordered by a SSR and SNP maps (Mori *et al*., 2017), as queries to perform a BLASTN against our primary assembly in order to anchor our contigs to the thirteen chromosomes of fig. The BLASTN process was performed to retain only the top hits (word_size 11 extension_threshold 20 gapped_alignment banded nucleic_match 1 nucleic_mismatch -3 open_penalty -5 extend_penalty -2 -max_scores 1 -num_alignments 1 -threshold significance=1e-5). Furthermore, we considered significant only the hits with a percent alignment value greater or equal to 85% and both a query length and an alignment length values greater or equal to 1,000 bp. All assembly statistics were calculated using Assemblathon 2 (Bradnam *et al*., 2013).

Assembly of two organelle DNAs

The organellar contigs were isolated using the Map to Reference function of Geneious v10.2.3 (Kearse *et al*., 2012) using the chloroplast (NC_030299.1) and the mitochondria (NC_029809.1) sequences of *Ziziphus jujube* as references being the species evolutionary most related to fig with available complete organellar sequences. The process was run with default parameters. In doing so we isolated a single chloroplast contig and six mitochondrial contigs. We aligned the whole corrected PacBio reads to the chloroplast and mitochondrial contigs using BLASR with default parameters. Reads having a minimum alignment identity of 98% were isolated and reassembled using Canu with default parameters. The newly obtained organellar contigs were manually edited and circularized using Geneious, obtaining the two complete organellar sequences.

Repeats prediction and annotation

For a comprehensive annotation of the TEs, a pipeline incorporating several tools, including LTRharvest (Ellinghaus *et al*., 2008), LTRdigest (Steinbiss *et al*., 2009a), DANTE (https://repeatexplorer-elixir.cerit-sc.cz/galaxy/), SINE_scan (Mao and Wang, 2017), MITE-hunter (Han and Wessler, 2010), HelitronScanner (Xiong *et al*., 2014) and RepeatModeler (Smit and Hubley, 2008) was used.

The primary assembly was firstly scanned for Class I full length LTR-REs. The sequences were identified and annotated using LTRharvest, LTRdigest (GenomeTools v1.5.10) and the DANTE tool v1.0.0 available on the RepeatExplorer public Galaxy-based server. The primary assembly was indexed using the suffixerator tool of the GenomeTools package (-tis -suf -lcp -des -ssp -dna). LTRharvest was run on the indexed assembly to identify LTR-REs ranging from 1.5 kbp to 25 kbp, with two terminal repeats which are at least 85% similar, ranging from 100 bp to 6,000 bp. The elements must be flanked by a 5 bp TSD (target site duplication) within 10 bp from the end of the element (-minlenltr 100 -maxlenltr 6000 -mindistltr 1500 -maxdistltr 25000 -mintsd 5 -maxtsd 5 -similar 85 -vic 10).

The identified sequences were firstly annotated using LTRdigest. First, we created a tRNAs database for the identification of the PBS (primer binding site) regions by downloading all the available tRNA sequences of *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera* and *Zea mays* from the Genomic tRNA Database (http://gtrnadb.ucsc.edu). We then created a database of HMM profiles for the annotation of the elements by downloading the whole dataset of non-redundant protein domains obtained from distinct plant mobile elements from GyDB 2.0 (http://www.gydb.org/index.php/Collection_HMM). The whole dataset was converted in HMMER3 format using the hmmconvert tool of the HMMer package (Eddy S.R., 1998). Hence LTRdigest was run on the previously identified elements using the described dataset as references and default parameters. Elements with no internal domains were excluded.

Finally, the library of full length LTR-REs was submitted to the DANTE tool for further annotation. The process of annotation and filtering of the false positives was run with default parameters, obtaining an annotation and a classification of the domains via a phylogenetic approach. The annotations obtained from

the previously described approaches (i.e. LTRdigest and DANTE) were manually checked and compared to create a final annotation based on the most significant match. We used the classification scheme by Wicker *et al*. (2007) to provide a unified classification of the LTR-REs.

The Class I SINEs were identified using SINE_scan v1.1.1 with default parameters. The process produced a library of representative SINE family sequences and a library of all the identified SINE sequences.

The Class II MITEs were collected using MITE-Hunter v11-2011 with default parameters. The process created different consensus libraries of MITE sequences based on their similarity. The libraries were combined as the total putative MITE sequences.

The Class II Helitron elements were collected using HelitronScanner v1.0 (java v1.8.0_144) with default parameters, obtaining a library of single putative full-length elements.

The libraries of LTR-REs, SINEs, MITEs, and Helitrons were used to mask the whole assembly using RepeatMasker v4.0.3 (Smit *et al*., 1996-2004). To identify TEs missed by the previous searches, RepeatModeler v1.0.11 was run with default parameters on the repeat masked assembly. The library of repeated elements obtained by RepeatModeler was used to mask the assembly again.

Finally, Phobos v3.3.12 (Mayer, 2006-2010) was used to find tandem repeats on the whole assembly (-u 1 -U 500 --minScore 12 --mismatchScore -5 --indelScore -5 -M imperfect -r 5).

Centromeric analysis

We identified the assembled sequences of putative centromeric regions by searching for the characteristic tandem repetitions of these regions. The Phobos output was manually scanned to identify the most abundant tandem repeats in each putative pseudomolecules. We considered the tandem repeats with a tandem unit length ranging between 80 bp and 350 bp. The most abundant tandem repeats were used to mask the 13 pseudomolecules, previously divided in 100 kbp size intervals, using RepeatMasker with default parameters. Only the tandem repeat occurring in each pseudomolecule was retained. The regions highly masked by that tandem repeat were considered as putative centromeric regions and were masked with all the LTR *Chromovirus* elements previously identified in this work in order to obtain the profile of their abundance in these regions. To compare the centromere structure in the different chromosomes, each centromeric region was divided into 50 intervals.

Gene prediction and annotation

Structural prediction and annotation of protein-coding genes was carried out using MAKER v2.31.10 (Holt and Yandell, 2011) and Blast2GO v5 (Conesa *et al*., 2005).

First, Trinity v2.5.1 (Grabherr *et al*., 2011) was run with default parameters to obtain a *de novo* fig transcriptome using already available RNA sequencing (RNA-seq) data (Vangelisti *et al*., 2019). The Illumina HiSeq 2000 paired-end reads were trimmed with Trimmomatic to remove adapters and low

quality regions (ILLUMINACLIP:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:25). AUGUSTUS v3.3.1 (Stanke *et al*., 2006) was trained using the transcriptome as evidence. A repeat-masked assembly was used to train GeneMark-ES v4.32 (Ter-Hovhannisyan *et al*., 2008) using the self-training mode with default parameters.

The MAKER pipeline was run to generate a comprehensive set of protein-coding genes, integrating the trained AUGUSTUS and GeneMark-ES programs using the transcriptome as external evidence. We found that the default parameters were the most sensitive and effective.

Annotated genes were functionally annotated using Blast2GO using the protein predictions. The sequences were first searched for homologous sequences by a BLASTP analysis against the whole NCBI non-redundant database (nr, 2018), then GO terms were obtained following the Blast2GO pipeline (Conesa *et al*., 2005). The sequences were also submitted to InterPro within Blast2GO (Jones *et al*., 2014) to annotate with protein domain functional information, and classifying into families. These analyses were performed using the default parameters. An extensive manual inspection of the annotated genes was made to identify and discard sequences annotated as TEs.

To locate the sex-determining region in our assembly we used the female-associated *RESPONSIVE-TO-ANTAGONIST1* (*RAN1*) gene, encoding a copper-transporting ATPase and characterized by Mori *et al*. (2017) to identify the same sequence among our predicted gene sequences. We isolated all the predicted genes annotated as *RAN*-related and performed a BLASTN against the Horaishi cultivar *RAN1* gene sequence (-e value 1e-10).

tRNAscan-SE v2.0 (Lowe and Chan, 2016) and RfamScan v1.1.2 (Kalvari *et al*., 2017) were run with default parameters to annotate the non-coding RNA component of the assembly, including tRNA, rRNA, miRNA, snoRNA and others.

DNA 4mC and 6mA modification analysis

KineticsTools v2.3 was used (https://github.com/PacificBiosciences/kineticsTools) to perform an interpulse duration (IPD) analysis and produce the methylome of fig. The IPD corresponds to the time required for a new base to bind in the active site of the sequencing polymerase after the previous base has been incorporated. These values are automatically added to the PacBio reads during the sequencing process. The software loads IPDs observed at each position in the genome and outputs the localization of an unmodified or a modified base. We used this software to identify two modification types: 4mC and 6mA.

We used pbalign (https://github.com/PacificBiosciences/pbalign) to align and sort all PacBio reads to the primary assembly, generating an aligned BAM file (--concordant --hitPolicy=randombest --minAccuracy 70 --minLength 50 --algorithmOptions="--minMatch 12 --bestn 10 --minPctIdentity 70.0"). The primary assembly was indexed using the faidx function of SAMtools. KineticsTools was run with default parameters to produce the gff output with the information about the base modifications. MEME-ChIP

(Machanick and Bailey, 2011) was run to perform a motif discovery on a random sample of 500k sequences using a 4-bases window range upstream and downstream the identified modified bases (-time 300 -ccut 100 -order 1 -db db/ARABD/ArabidopsisDAPv1.meme -meme-mod zoops -meme-minw 6 -meme-maxw 30 -meme-nmotifs 3 -dreme-e 0.05 -centrimo-score 5.0 -centrimo-ethresh 10.0). The above analysis was performed twice, once for the 4mC and once for the 6mA motifs. Finally, the assembly was scanned to find and quantify the nucleotide motifs associated with the modified bases using an R script available on the PacBio github page (https://github.com/PacificBiosciences/motif-finding).

GO-annotated genes were further clustered to infer a possible correlation between functional activity and the distribution of methylation sites.

The complete gene body methylation profile was obtained by analysing all the predicted genes and extending their coordinates by 1,000 bp so to include flanking regions. Furthermore, genes from two to six exons were analysed separately for their methylation profile, both at exon and intron level, again extending gene coordinates by 1,000 bp. The methylated sites were intersected using bedtools v2.27.0 (Quinlan and Hall, 2010).

Transposon body methylation profiles were obtained using the previously obtained positions of TEs with 1,000 bp extensions. Overlapping results were collapsed. Data intersection between TEs and methylation sites was performed using bedtools. For the graph plots, genes, TEs and flanking regions were divided into 50 intervals.

Circos graph

A global representation of the genomic landscape was produced with Circos (Krzywinski *et al*., 2009). The pseudomolecules were divided into 200 kbp-window regions. For each region, heterozygosity, number of genes and TEs bases were calculated, as well as the total amount of 4mC and 6mA. Putative identified centromeric regions were also included for each pseudomolecule.

DATA AVAILABILITY

The data has been deposited in the NCBI repository under BioProject PRJNA565858. The PacBio reads has been deposited in the Sequence Read Archive (SRA) with accession numbers SRR10127734, SRR10127735, SRR10127736, SRR10127737, SRR10127738, SRR10127739. The fig genome has been deposited at DDBJ/ENA/GenBank under the accession VYVB00000000. The version described in this paper is version VYVB01000000. The collection of TEs is available at the repository sequence page of the Department of Agriculture, Food, and Environment of the University of Pisa (http://pgagl.agr.unipi.it/sequence-repository/). Intermediary files and other miscellaneous information are available from the corresponding author upon request.

ACKNOWLEDGEMENT

The authors acknowledge funding from University of Pisa, Italy, project "Progetto di Ricerca di Ateneo 2017: Genomica, fisiologia e difesa del fico (*Ficus carica* L.), una specie antica con grandi prospettive" and from Department of Agriculture, Food and Environment (DAFE) project "Plantomics". Thanks are due to Siro and Mario Petracchi - Associazione Produttori Fichi Secchi di Carmignano (Prato, Italy).

AUTHOR CONTRIBUTIONS

L.N., F.M., T.G., A.Z. and A.C. planned and designed the project. G.U., T.G., A.V. and L.S.Z. performed nucleic acid extractions. M.C. performed the cytological analyses. G.U., F.M., A.V., E.B., L.S.Z., A.Z., R.K. and K.H.P. performed the genome and transcriptome assembly and annotation. G.U., E.B. and F.M. performed genome-wide methylation analysis. F.M., G.U. and A.C. wrote the manuscript with contributions from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

SUPPORTING MATERIAL LEGENDS

Figure S1. Centromeric region distributions.

Figure S1. Gene Ontology distribution of methylated and not-methylated genes.

Table S1. Comparison between assembly statistics.

Table S2. Resulting statistics of the FALCON-unzip assembler.

Table S3. Haplotigs gene content completeness assessment.

Table S4. Haplotig assembly statistics.

Table S5. Ordered and orientated scaffold statistics.

Table S6. Number of contigs per genetic map bin.

Table S7. Haplotigs distribution.

Table S8. Transposable elements annotation.

Table S9. Centromeres statistics.

Table S10. Transcriptome statistics.

Table S11. Protein-coding genes structural prediction.

Table S12. Non-protein-coding genes functional annotation.

Table S13. Protein-coding genes functional annotation.

Table S14. Genome-wide 4mC modifications.

Table S15. Genome-wide 6mA modifications.

Table S16. 4mC distribution.

Table S17. 6mA distribution.

Data S1. Putative centromeric tandem unit.

REFERENCES

Al-Dous, E.K., George, B., Al-Mahmoud, M.E. et al. (2011) *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.* **29**, 521.

Ansorge, W. J. (2016) Next-generation DNA sequencing (II): techniques, applications. *Next Generat. Sequenc. & Applic*. **1**, 1-10.

Bierne, H., Hamon, M. and Cossart, P. (2012) Epigenetics and bacterial infections. *Cold. Spring. Harb. Perspect. Med*. **2**, a010272.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Boratyn, G.M., Camacho, C., Cooper, P.S. et al. (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29-W33.

Bradnam, K.R., Fass, J.N., Alexandrov, A. et al. (2013) Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* **2**, 10.

Buti, M., Moretto, M., Barghini, E. et al. (2017) The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry). *GigaScience* 7, giy010.

Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238.

Chin, C.S., Alexander, D.H., Marks, P. et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563.

Chin, C.S., Peluso, P., Sedlazeck, F.J. et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676.

Daccord, N., Celton, J.M., Linsmith, G. et al. (2017) High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genet*. **49**, 1099.

Diez, C.M., Roessler, K. and Gaut, B.S. (2014) Epigenetics and plant genome evolution. *Curr. Opin. Plant. Biol.* **18**, 1-8.

Dong, A.X., Xin, H.B., Li, Z.J. et al. (2018) High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *GigaScience* **7**, giy068.

Eddy, S. HMMER: profile HMMs for protein sequence analysis (2003). https://hmmer.org/

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18.

Feng, S., Cokus, S.J., Zhang, X. et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci.* **107**, 8689-8694.

Fu, Y., Luo, G.Z., Chen, K. et al. (2015) N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* **161**, 879-892.

Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv*, 1207.3907.

Grabherr, M.G., Haas, B.J., Yassour, M. et al. (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnol*. **29**, 644.

Han, Y. and Wessler, S. R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res*. **38**, e199-e199.

Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491.

Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377-386.

Jones, P., Binns, D., Chang, H.Y. et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D. and Petrov, A.I. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*. **46**, D335-D342.

Kearse, M., Moir, R., Wilson, A. et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649.

Kislev, M.E., Hartmann, A. and Bar-Yosef, O. (2006) Early domesticated fig in the Jordan Valley. *Science* **312**, 1372-1374.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. **27**, 722-736.

Korlach, J., Gedman, G., Kingan, S.B., Chin, C.S., Howard, J.T., Audet, J.N., Cantin, L. and Jarvis, E.D. (2017) *De novo* PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience* **6**, 1-16.

Koziol, M.J., Bradshaw, C.R., Allen, G.E., Costa, A.S., Frezza, C. and Gurdon, J.B. (2016) Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat. Struct. Mol. Biol*. **23**, 24.

Kronenberg, Z.N., Hall, R.J., Hiendleder, S., Smith, T.P., Sullivan, S.T., Williams, J.L. and Kingan, S.B. (2018) FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*, 327064.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*. **19**, 1639-1645.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol*. **5**, R12.

Lam, K.K., LaButti, K., Khalak, A. and Tse, D. (2015) FinisherSC: a repeat-aware tool for upgrading *de novo* assembly using long reads. *Bioinformatics* **31**, 3207-3209.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Liang, Z., Shen, L., Cui, X. et al. (2018) DNA N6-adenine methylation in *Arabidopsis thaliana*. *Dev. Cell*. **45**, 406-416.

Liu, M.J., Zhao, J., Cai, Q.L. et al. (2014) The complex jujube genome provides insights into fruit tree biology. *Nat. Commun.* **5**, 5315.

Loureiro, J., Rodriguez, E., Doležel, J. and Santos, C. (2007) Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Ann. Bot.* **100**, 875-888.

Low, W.Y., Tearle, R., Bickhart, D.M. et al. (2019) Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat. Commun*. **10**, 260.

Lowe, T.M. and Chan, P.P. (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res*. **44**, W54-W57.

Machanick, P. and Bailey, T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696-1697.

Mafrica, R., Marchese, A., Bruno, M., Costa, F., Fretto, S., Marra, F.P., Pangallo, S., Quartararo, A. and Caruso, T. (2015) Morphological and molecular variability within the fig cultivar 'Dottato' in the Italian protected designation origin area "Fichi di Cosenza". *Acta Hortic.* **1173**, 29-34.

Mao, H. and Wang, H. (2016) SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. *Bioinformatics* **33**, 743-745.

Mascagni, F., Cavallini, A., Giordani, T. and Natali, L. (2017) Different histories of two highly variable LTR retrotransposons in sunflower species. *Gene* **634**, 5-14.

Mascagni, F., Vangelisti, A., Giordani, T., Cavallini, A. and Natali, L. (2018) Specific LTR-retrotransposons show copy number variations between wild and cultivated sunflowers. *Genes* **9**, 433.

Mayer, C. Phobos 3.3.11 (2006-2010). https://www.rub.de/ecoevo/cm/cm_phobos.htm

Meuwissen, T., Hayes, B. and Goddard, M. (2013) Accelerating improvement of livestock with genomic selection. *Annu. Rev. Anim. Biosci*. **1**, 221-237.

Ming, R., Hou, S., Feng, Y. et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991.

Mori, K., Shirasawa, K., Nogata, H., Hirata, C., Tashiro, K., Habu, T., Kim, S., Himeno, S., Kuhara, S. and Ikegami, H. (2017) Identification of RAN1 orthologue associated with sex determination through whole genome sequencing analysis in fig (*Ficus carica* L.). *Sci. Rep*. **7**, 41124.

Murray, I.A., Clark, T.A., Morgan, R.D., Boitano, M., Anton, B.P., Luong, K., Fomenkov, A., Turner, S.W., Korlach, J. and Roberts, R.J. (2012) The methylomes of six bacteria. *Nucleic Acids Res*. **40**, 11450-11462.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J. and Macas, J. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA* **2**, 4.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.

Reinders, J., Wulff, B.B., Mirouze, M., Marí-Ordóñez, A., Dapp, M., Rozhon, W., Bucher, E., Theiler, G. and Paszkowski, J. (2009) Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. *Genes Dev*. **23**, 939-950.

Rhoads, A. and Au, K.F. (2015) PacBio sequencing and its applications. *Genomics, Proteomics Bioinformatics* **13**, 278-289.

Schwessinger, B., Sperschneider, J., Cuddy, W.S., Garnica, D.P., Miller, M.E., Taylor, J.M., Dodds, P.N., Figueroa, M., Park, R.F. and Rathjen, J.P. (2018) A near-complete haplotype-phased genome of the dikaryotic wheat stripe rust fungus *Puccinia striiformis* f. sp. *tritici* reveals high interhaplotype diversity. *Am. Soc. Microbiol*. **9**, e02275-17.

Shirasawa, K., Esumi, T., Hirakawa, H., Tanaka, H., Itai, A., Ghelfi, A., Nagasaki, H. and Isobe, S. (2019) Phased genome sequence of an interspecific hybrid flowering cherry, Somei-Yoshino (*Cerasus* × *yedoensis*). *BioRxiv*, 573451.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212.

Smit, A.F.A. and Hubley, R. RepeatModeler Open-1.0. (2008). https://www.repeatmasker.org/RepeatModeler/

Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker Open-4.0. (2013-2015). http://www.repeatmasker.org

Solomon, A., Golubowicz, S., Yablowicz, Z., Grossman, S., Bergman, M., Gottlieb, H.E., Altman, A., Kerem, Z. and Flaishman, M.A. (2006) Antioxidant activities and anthocyanin content of fresh fruits of common fig (*Ficus carica* L.). *J. Agric. Food Chem*. **54**, 7717-7723.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res*. **34**, W435-W439.

Steinbiss, S., Willhoeft, U., Gremme, G. and Kurtz, S. (2009) Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Res*. **37**, 7002-7013.

Teh, B.T., Lim, K., Yong, C.H. et al. (2017) The draft genome of tropical fruit durian (*Durio zibethinus*). *Nature Genet*. **49**, 1633.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. and Borodovsky, M. (2008) Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res*. **18**, 1979-1990.

Tuskan, G.A., Difazio, S., Jansson, S. et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-1604.

Vangelisti, A., Zambrano, L.S., Caruso, G. et al. (2019) How an ancient, salt-tolerant fruit crop, *Ficus carica* L., copes with salinity: a transcriptome analysis. *Sci. Rep*. **9**, 2561.

Varshney, R.K., Chen, W., Li, Y. et al. (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83.

Veberic, R., Colaric, M. and Stampar, F. (2008) Phenolic acids and flavonoids of fig fruit (*Ficus carica* L.) in the northern Mediterranean region. *Food Chemistry* **106**, 153-157.

Veeckman, E., Ruttink, T. and Vandepoele, K. (2016) Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759-1768.

Vinson, J.A., Zubik, L., Bose, P., Samman, N. and Proch, J. (2005) Dried fruits: excellent *in vitro* and *in vivo* antioxidants. *J. Am. Coll. Nutr*. **24**, 44-50.

Vogel, A., Schwacke, R., Denton, A.K. et al. (2018) Footprints of parasitism in the genome of the parasitic flowering *plant Cuscuta campestris*. *Nat. Commun*. **9**, 2515.

Walker, B.J., Abeel, T., Shea, T. et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963.

Wicker, T., Sabot, F., Hua-Van, A. et al. (2007) A unified classification system for eukaryotic transposable elements. *Nature. Rev. Genet.* **8**, 973.

Wu, T.P., Wang, T., Seetin, M.G. et al. (2016) DNA methylation on N 6-adenine in mammalian embryonic stem cells. *Nature* **532**, 329.

Wu, J., Wang, Z., Shi, Z. et al. (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* **23**, 396-408.

Xiong, W., He, L., Lai, J., Dooner, H.K. and Du, C. (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci*. **111**, 10263-10268.

Ye, C., Hill, C. M., Wu, S., Ruan, J. and Ma, Z.S. (2016) DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep*. **6**, 31900.

Ye, C., Ma, Z.S., Cannon, C.H., Pop, M., and Douglas, W.Y. (2012) Exploiting sparseness in *de novo* genome assembly. *BMC Bioinformatics* **13**, S1.

Ye, G., Zhang, H., Chen, B., Nie, S., Liu, H., Gao, W., Wang, H., Gao, Y. and Gu, L. (2019) *De novo* genome assembly of the stress tolerant forest species *Casuarina equisetifolia* provides insight into secondary growth. Plant J. 97, 779-794.

Zambrano, L.S., Usai, G., Vangelisti, A. et al. (2017) Cultivar-specific transcriptome prediction and annotation in *Ficus carica* L. *Genom. Data* **13**, 64-66.

Zhang, Q., Chen, W., Sun, L. et al. (2012) The genome of *Prunus mume*. *Nat. Commun.* **3**, 1318.

Zhang, W., Spector, T.D., Deloukas, P., Bell, J.T. and Engelhardt, B.E. (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol*. **16**, 14.

Zhao, Q., Yang, J., Liu, J. et al. (2018) A draft reference genome sequence for *Scutellaria baicalensis* Georgi. *bioRxiv*, 398032.

Zhou, C., Wang, C., Liu, H. et al. (2018) Identification and analysis of adenine N 6-methylation sites in the rice genome. Nat. Plants 4, 554.

FIGURE LEGENDS

Figure 1 (a) Statistics and (b) length distribution of PacBio reads (Length >= 1,000). (c) Flow chart of genome assembly using the PacBio dataset. (d) Flow chart of the annotation pipeline.

Figure 2 Distribution of genomic and epigenomic features of the fig genome. (a) The thirteen pseudomolecules. The black label on each pseudomolecule represents the putative centromeric region. Pseudomolecules are divided into 1 Mbp intervals. (b) Heterozygosity. (c) Histogram representing gene density. (d) Histogram representing TEs density. (e) Heat map representing 4mC modification levels. (f) Heat map representing 6mA modification levels.

Figure 3 Annotation of the fig genome. (a) Pie chart showing the percentage of genes (i.e. protein-coding genes, rRNAs, tRNA and ncRNAs) (left), repeats (i.e. LTR Retrotransposons, Non-LTR Retrotransposons, DNA transposons and tandem repeats) (right) and unknown sequences. (b) Distribution of GO terms including the number of genes attributed to the GO terms (i.e. cellular component, molecular function and biological process).

Figure 4 Overall characterization of fig methylated sites. (a) Consensus motif for 4mC sites. (b) Consensus motif for 6mA sites. (c) Pie charts showing the percentage of DNA methylation distributions of the three cytosine methylation sequence logos (CG, CHG or CHH) (left) and the five adenosine methylation sequence logos (A, ANHGA, GAGG, CAAG and ACCT) (right) (N: any nucleotide, H: adenine, cytosine or thymine). (d) Count of the different quantities of 6mA and 4mC sites in 5'-UTRs (left), genes (centre) and 3'-UTRs (right).

Figure 5 Gene body 4mC and 6mA profiles in genes grouped by exon (from two up to six) including a 1-kbp window upstream of the transcription start site (TSS) and downstream of the transcription end site (TES).

Figure 6 Global profile of fig methylated sites. (a) Distribution of 4mC (up) and 6mA (down) modification levels at protein-coding gene (left) and TE (right) levels, including a 1-kbp window upstream of the transcription start site (TSS) and downstream of the transcription end site (TES).

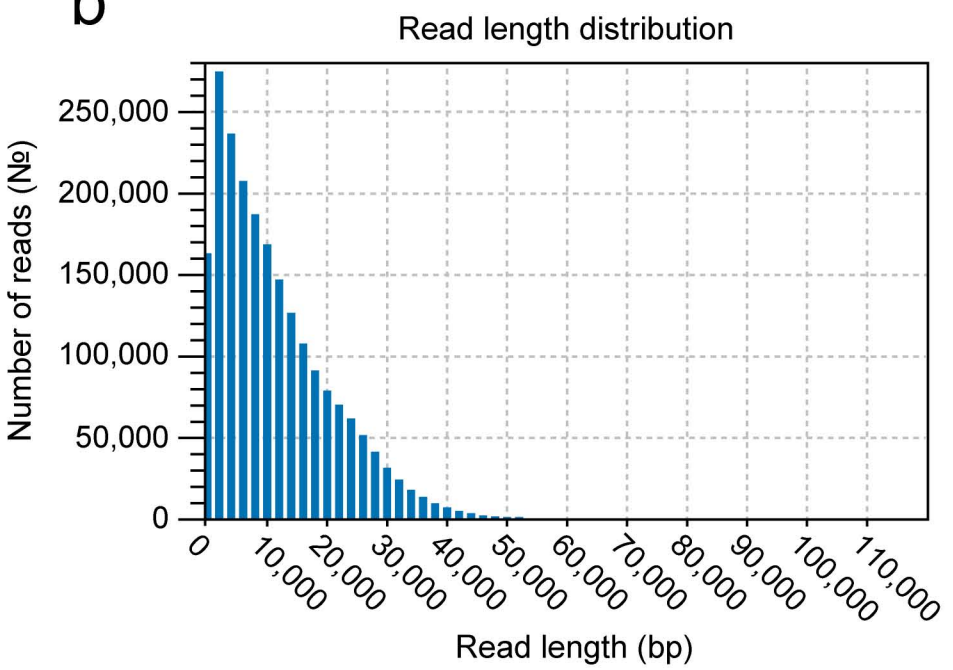
Table 1. Comparisons between the Dottato and Horaishi assemblies

	Dottato cultivar	**Horaishi cultivar (Mori *et al*., 2017)**
Genome representation (%)	95	70
Number of sequences (№)	905	27,995
Total size of the assembly (bp)	333,400,567	247,090,738
Longest sequence (bp)	5,010,936	1,764,766
Shortest sequence (bp)	20,012	479
Number of sequences > 10 kbp (№)	905 (100%)	2,081 (7.4%)
Number of sequences > 100 kbp (№)	595 (65.7%)	671 (2.4%)
Number of sequences > 1 Mbp (№)	81 (9.0%)	8 (0.0%)
Mean sequence size (bp)	368,398	8,826
Median sequence size (bp)	167,241	893
N50 sequence length (bp)	823,517	166,092
L50 sequence length (№)	121	374
N sequence content (%)	0.00	14.72
BUSCO assessment (%)	93.3	90.5
Predicted protein-coding genes (№)	37,840	36,138
Annotated protein-coding genes (№)	28,737	25,011
Rate of annotated protein-coding genes (%)	76	69
TE proportion (%)	33	16

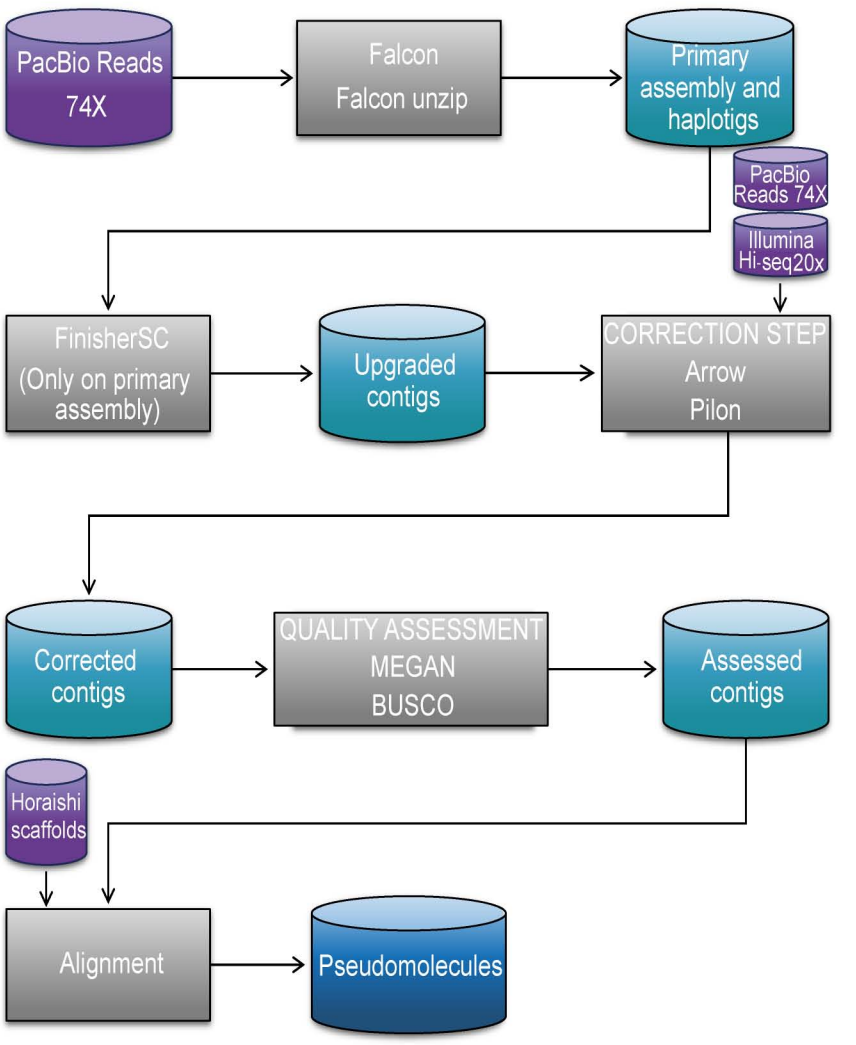
a

PacBio dataset
Sequences (№): 2,140,959
Sum (bp): 26,472,809,399 (74x)
Min size (bp): 1,000
Max size (bp): 111,426
Average size (bp): 12,364
N50 (bp): 18,419

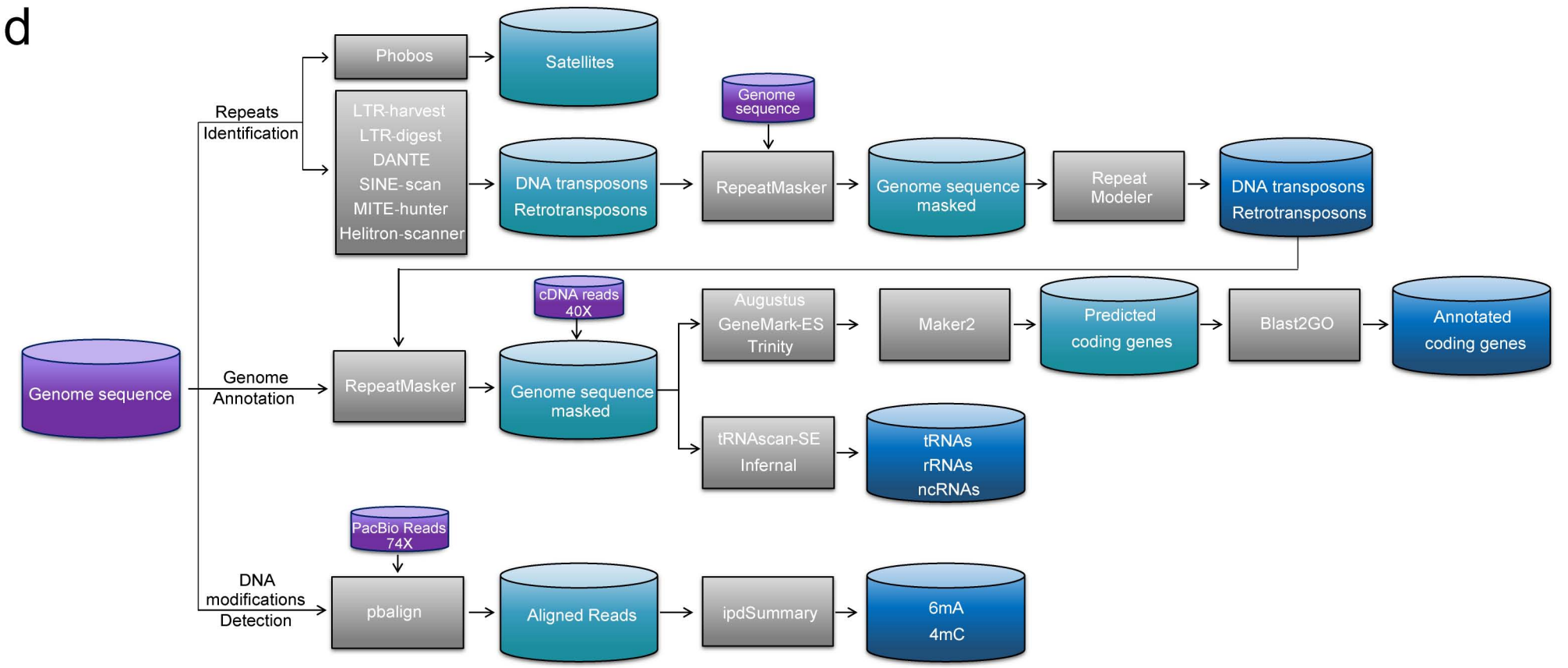
b

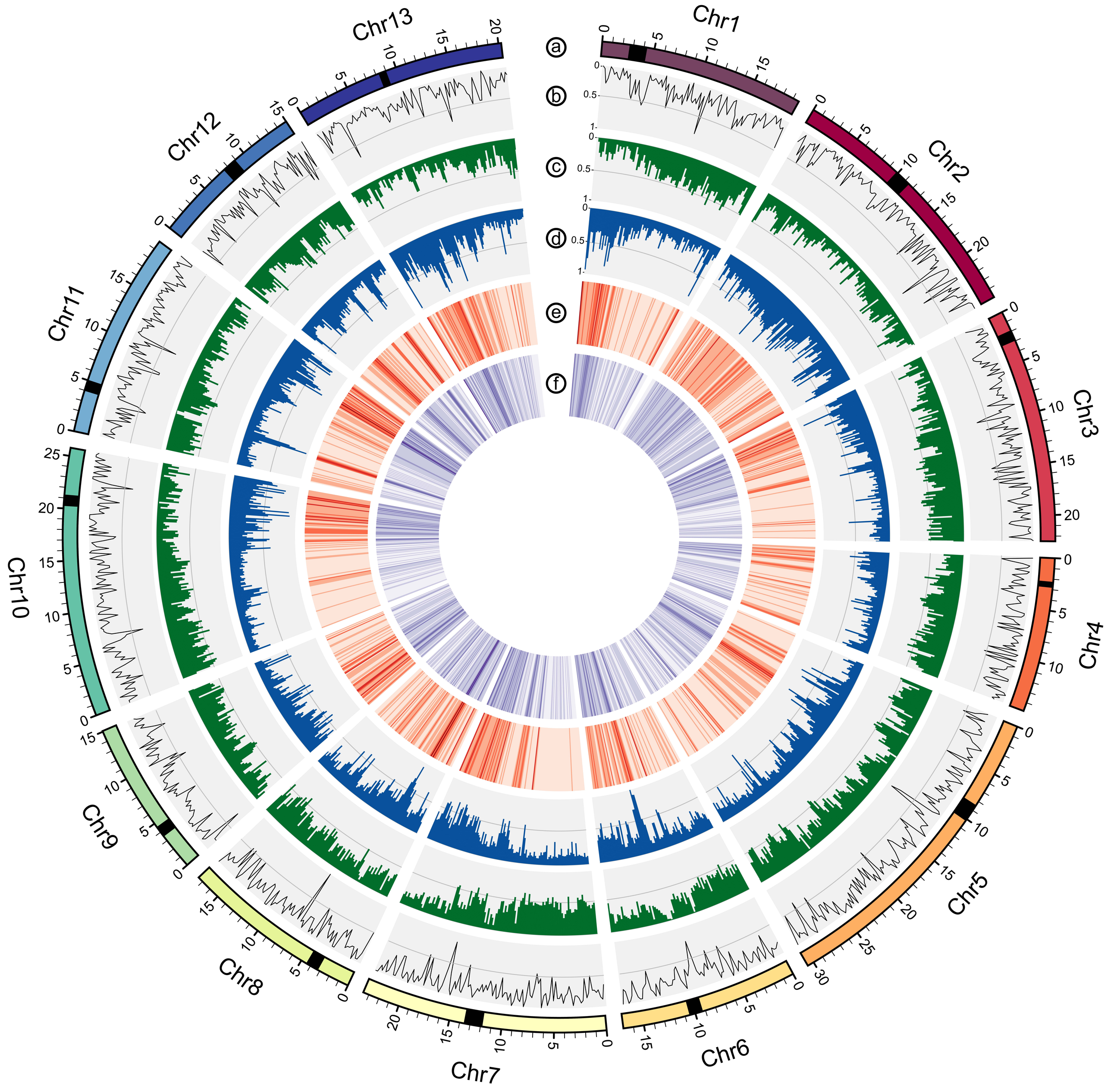


c

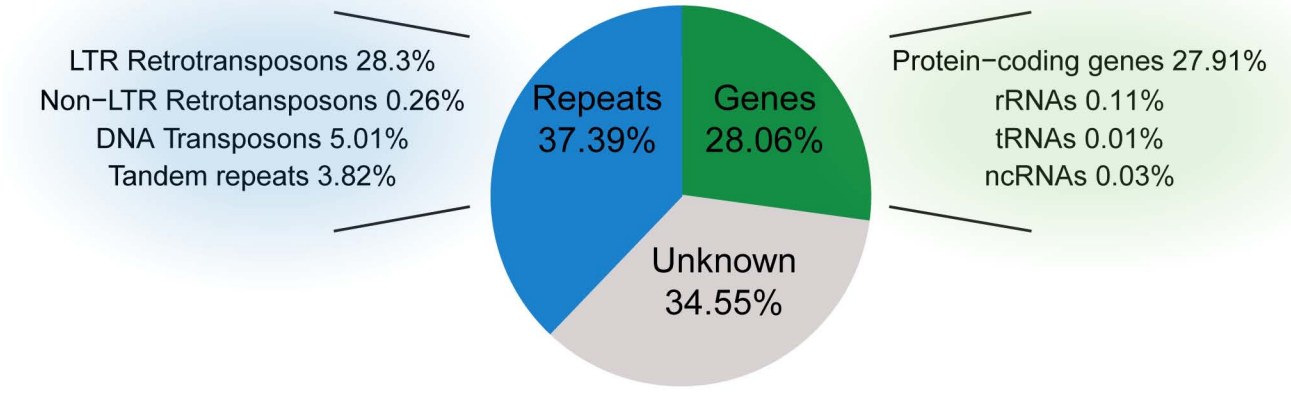


d

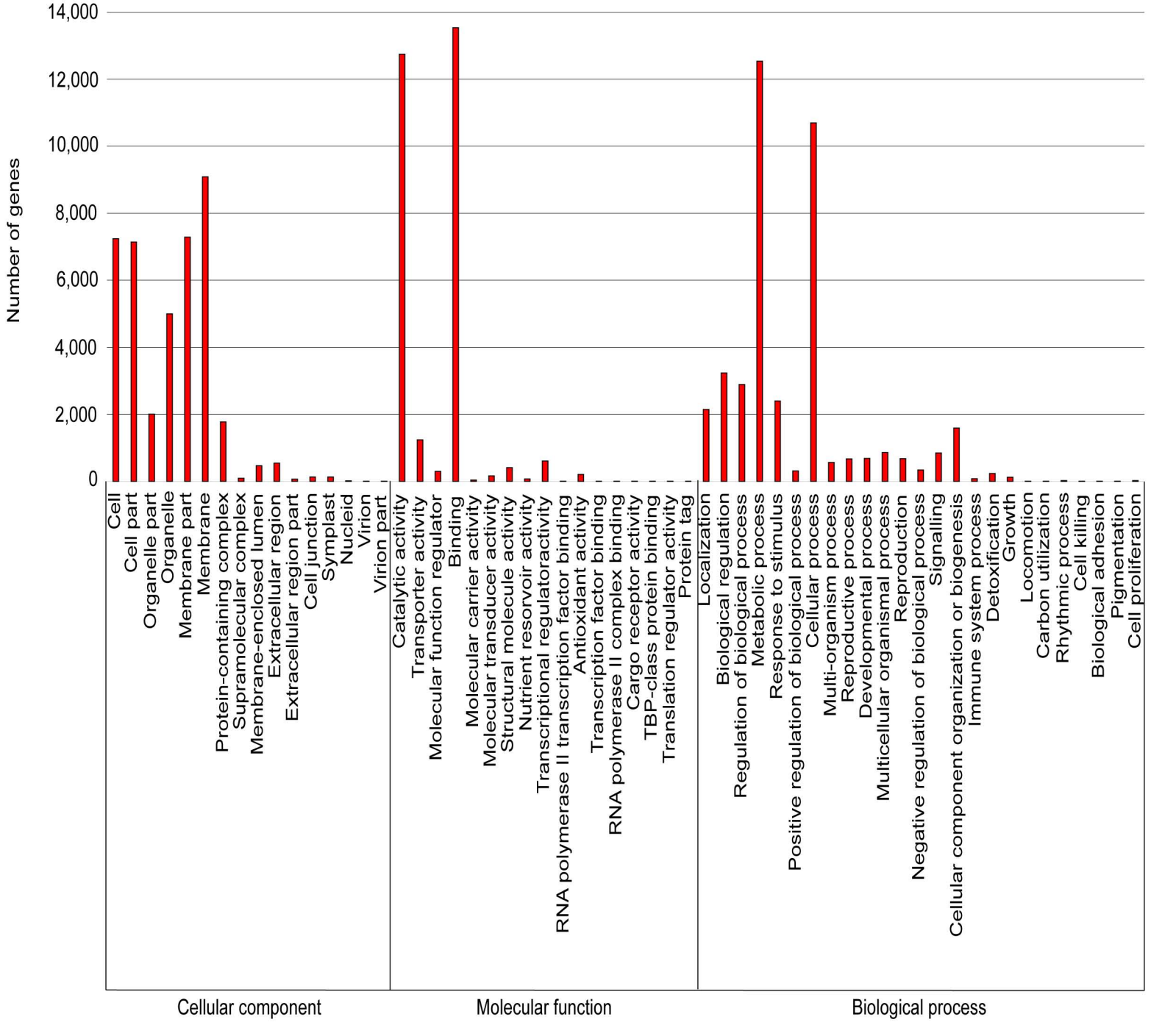


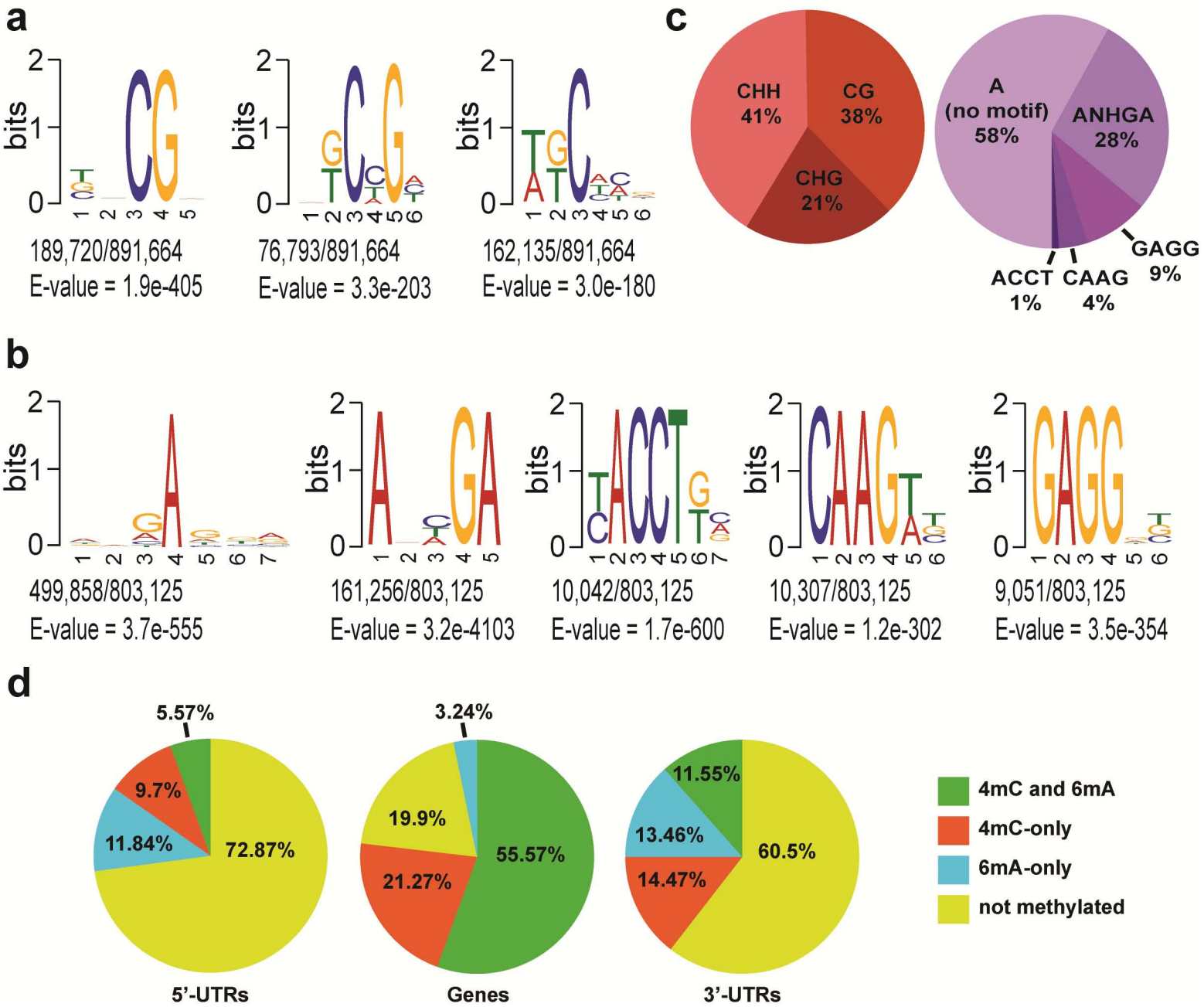

Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
a
b
c
d
e
f

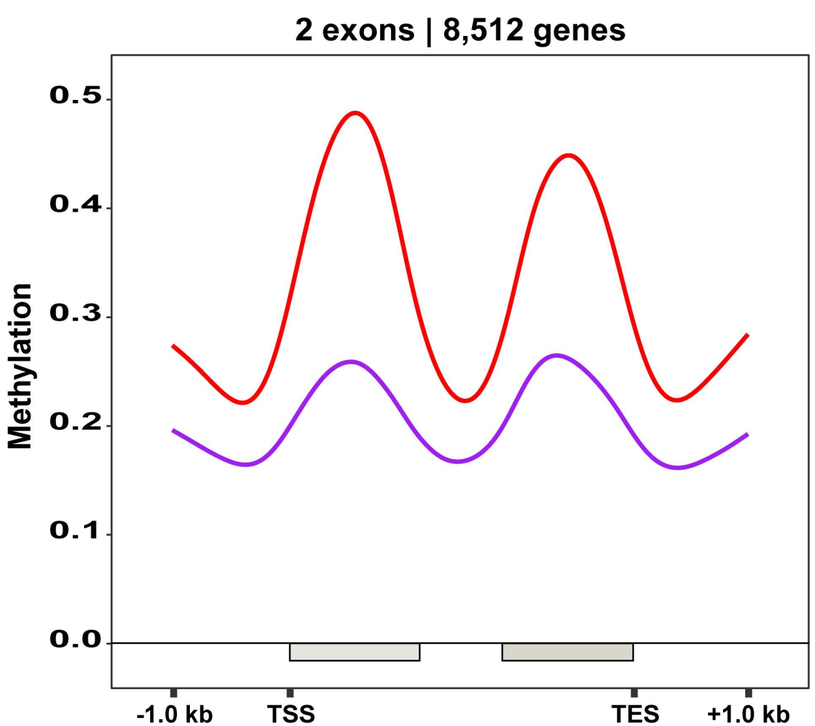
a

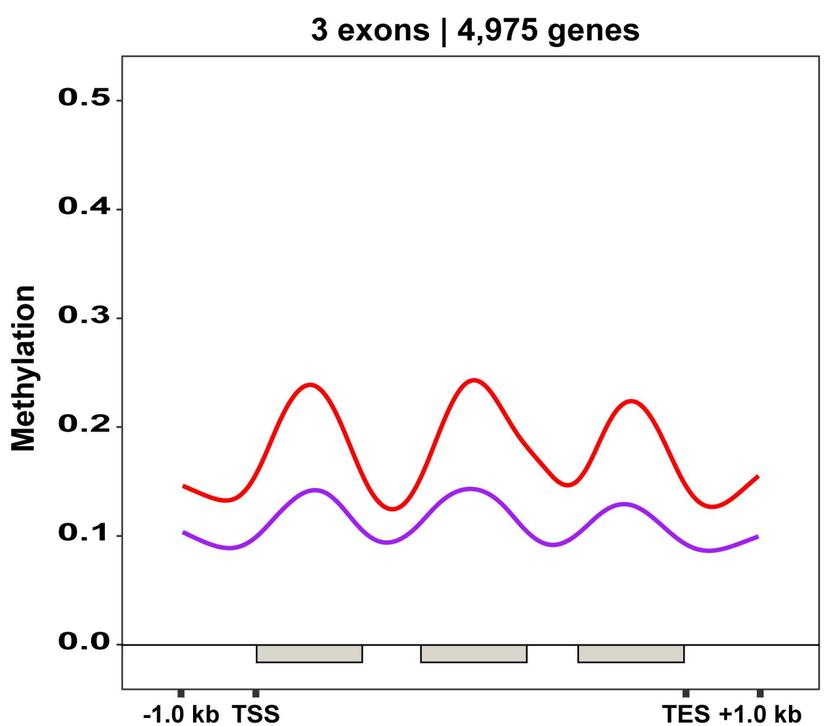

LTR Retrotransposons 28.3%
Non−LTR Retrotransposons 0.26%
DNA Transposons 5.01%
Tandem repeats 3.82%
Repeats 37.39%
Genes 28.06%
Unknown 34.55%
Protein−coding genes 27.91%
rRNAs 0.11%
tRNAs 0.01%
ncRNAs 0.03%

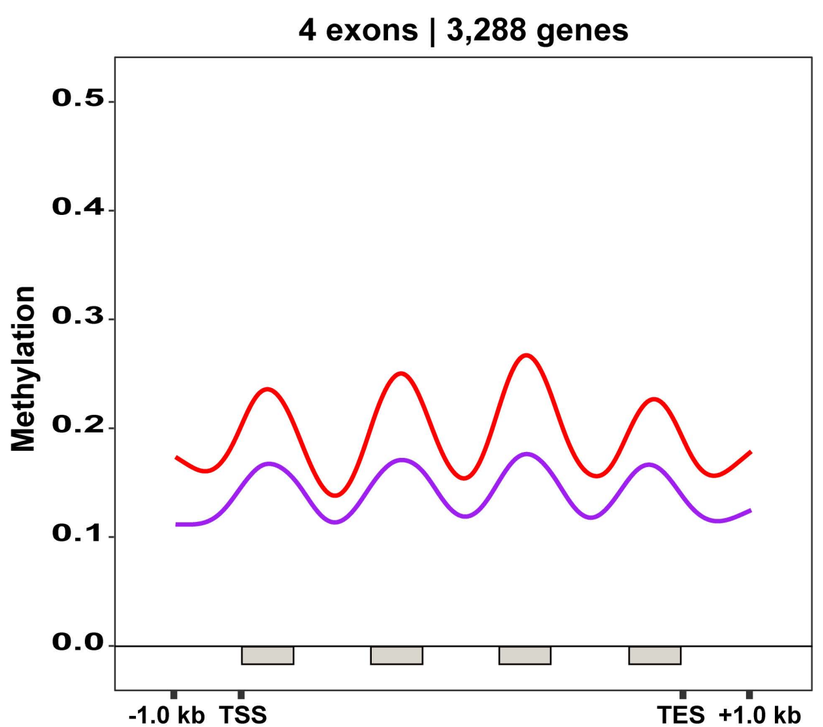
b

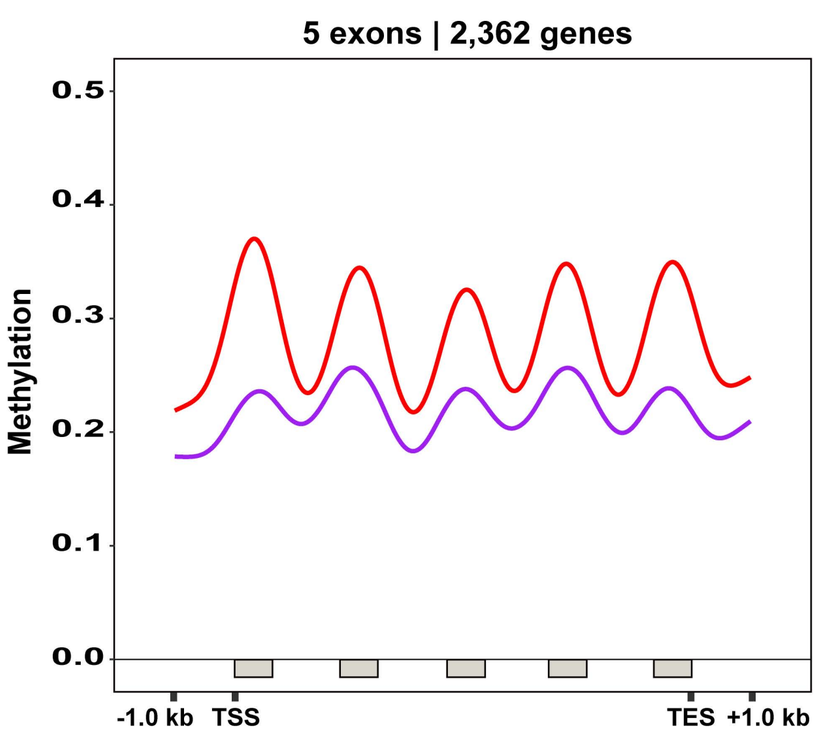

Number of genes
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Cell
Cell part
Organelle part
Organelle
Membrane part
Membrane
Protein-containing complex
Supramolecular complex
Membrane-enclosed lumen
Extracellular region
Extracellular region part
Cell junction
Symplast
Nucleid
Virion
Virion part
Catalytic activity
Transporter activity
Molecular function regulator
Binding
Molecular carrier activity
Molecular transducer activity
Structural molecule activity
Nutrient reservoir activity
Transcriptional regulatoractivity
RNA polymerase II transcription factor binding
Antioxidant activity
Transcription factor binding
RNA polymerase II complex binding
Cargo receptor activity
TBP-class protein binding
Translation regulator activity
Protein tag
Localization
Biological regulation
Regulation of biological process
Metabolic process
Response to stimulus
Positive regulation of biological process
Cellular process
Multi-organism process
Reproductive process
Developmental process
Multicellular organismal process
Reproduction
Negative regulation of biological process
Signalling
Cellular component organization or biogenesis
Immune system process
Detoxification
Growth
Locomotion
Carbon utilization
Rhythmic process
Cell killing
Biological adhesion
Pigmentation
Cell proliferation
Cellular component
Molecular function
Biological process

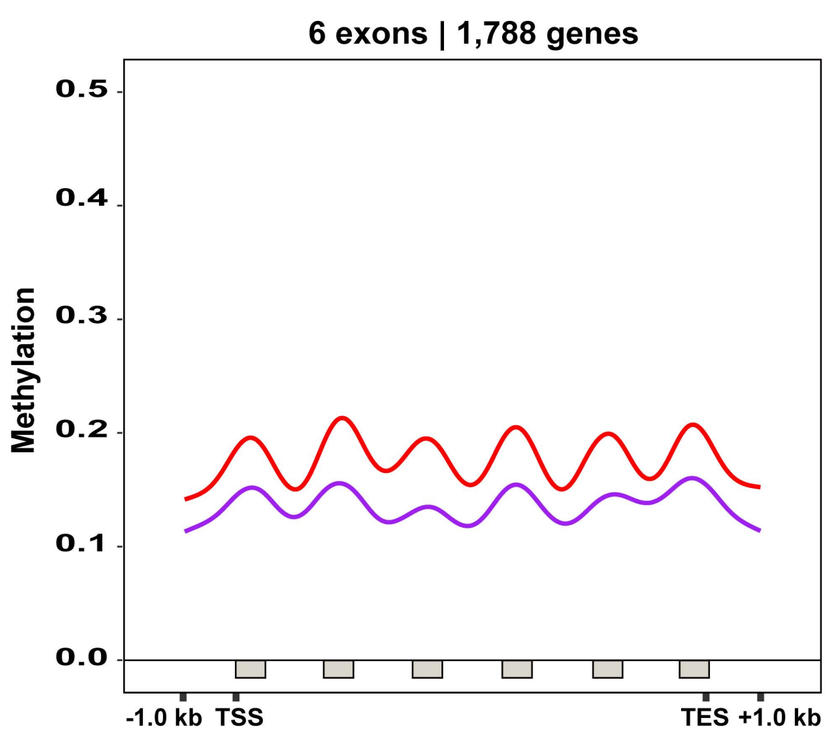

a
bits
189,720/891,664
E-value = 1.9e-405
76,793/891,664
E-value = 3.3e-203
162,135/891,664
E-value = 3.0e-180
c
CHH
41%
CG
38%
CHG
21%
A
(no motif)
58%
ANHGA
28%
GAGG
9%
ACCT
1%
CAAG
4%
b
499,858/803,125
E-value = 3.7e-555
161,256/803,125
E-value = 3.2e-4103
10,042/803,125
E-value = 1.7e-600
10,307/803,125
E-value = 1.2e-302
9,051/803,125
E-value = 3.5e-354
d
5.57%
9.7%
11.84%
72.87%
5'-UTRs
3.24%
19.9%
55.57%
21.27%
Genes
11.55%
13.46%
60.5%
14.47%
3'-UTRs
4mC and 6mA
4mC-only
6mA-only
not methylated

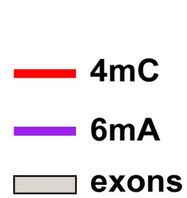

2 exons | 8,512 genes
Methylation
0.5
0.4
0.3
0.2
0.1
0.0
-1.0 kb
TSS
TES
+1.0 kb

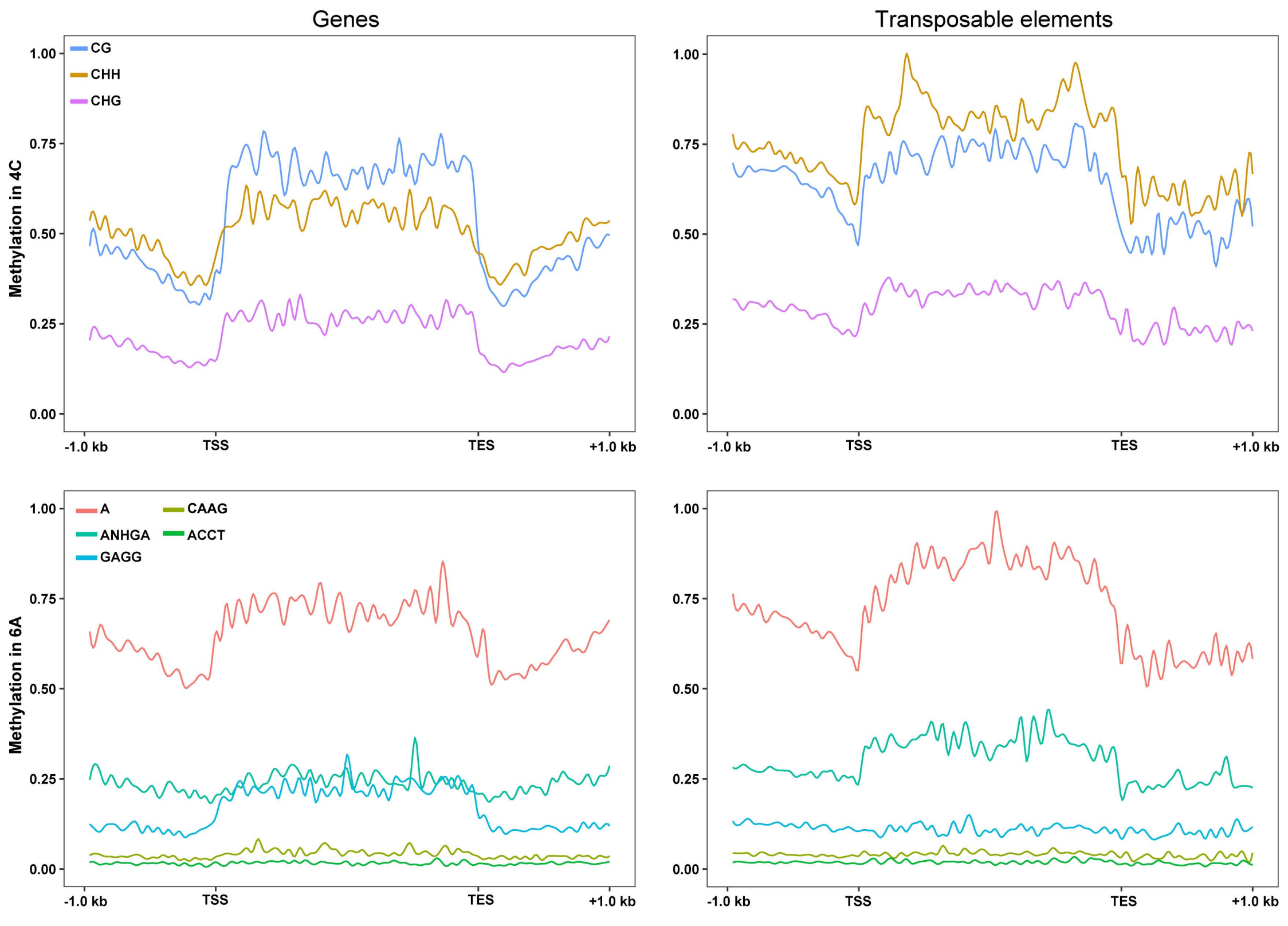

3 exons | 4,975 genes
Methylation
0.5
0.4
0.3
0.2
0.1
0.0
-1.0 kb
TSS
TES
+1.0 kb


4 exons | 3,288 genes
Methylation
0.5
0.4
0.3
0.2
0.1
0.0
-1.0 kb
TSS
TES
+1.0 kb


5 exons | 2,362 genes
Methylation
0.5
0.4
0.3
0.2
0.1
0.0
-1.0 kb
TSS
TES
+1.0 kb


6 exons | 1,788 genes
Methylation
0.5
0.4
0.3
0.2
0.1
0.0
-1.0 kb
TSS
TES
+1.0 kb


4mC
6mA
exons


Genes
Transposable elements
Methylation in 4C
CG
CHH
CHG
1.00
0.75
0.50
0.25
0.00
-1.0 kb
TSS
TES
+1.0 kb
Methylation in 6A
A
CAAG
ANHGA
ACCT
GAGG