CpG Island Definition and Methylation Mapping of the T2T-YAO Genome

Ming Xiao[1,#], Rui Wei[1,2,#], Jun Yu[3,4], Chujie Gao[1], Fengyi Yang[1], Le Zhang[1,2,*]

[1] *College of Computer Science, Sichuan University, Chengdu 610065, China*

[2] *Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China*

[3] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*

[4] *University of Chinese Academy of Sciences, Beijing 100049, China*

[#]Equal contribution.

[*]Corresponding author.

E-mail: zhangle06@scu.edu.cn (Zhang L).

Running title: *Xiao M, et al/ Comparative Analysis of T2T-YAO and T2T-CHM13*

Words: 5785

Figures: 7

Tables: 2

Supplementary figures: 9

Supplementary table: 6

Title: 60

Running title: 50

Keywords: 5

Abstract: 214

Abstract

Precisely defining and mapping all cytosine positions and their clusters, known as CpG islands (CGIs), as well as their methylation status are pivotal for genome-wide epigenetic studies, especially when population-centric reference genomes are ready for timely application. Here we first align the two high-quality reference genomes, T2T-YAO and T2T-CHM13, from different ethnic backgrounds in a base-by-base fashion and compute their genome-wide density-defined and position-defined CGIs. Second, mapping some representative genome-wide methylation data from selected organs onto the two genomes, we find that there are about 4.7–5.8% sequence divergency of variable categories depending on quality cutoffs. Genes among the divergent sequences are mostly associated with neurological functions. Moreover, CGIs associated with the divergent sequences are significantly different with respect to CpG density and observed CpG/expected CpG (O/E) ratio between the two genomes. Finally, we find that the T2T-YAO genome not only has a greater CpG site coverage than that of the T2T-CHM13 genome when whole-genome bisulfite sequencing (WGBS) data from the European and American populations are mapped to each reference, but also show more hyper-methylated CpG sites as compared to the T2T-CHM13 genome. Our study suggests that future genome-wide epigenetic studies of the Chinese populations rely on both acquisition of high-quality methylation data and subsequent precision CGI mapping based on the Chinese T2T reference.

Keywords: T2T-YAO; Position-defined CpG island; Density-defined CpG Island; DNA methylation; Genome analysis

Introduction

Position-sensitive and co-methylated CpG islands (CGIs) represent one of the several key epigenetic mechanisms for chromosome integrity and gene expression regulation [1, 2]. To define highly variable CGIs based on both position and density genome-wide

and to map their methylation status under different physiological conditions and in different cell types are prerequisite in most epigenomic studies [3–7]. Human reference genomes, especially those tailored to specific populations, provide single-base-resolution coordinates for genome-wide genetic and epigenetic mapping with ultimate precision [8]. The release of the telomere-to-telomere CHM13 (T2T-CHM13) genome in 2022 has provided such a benchmark by filling-up sequence gaps that contribute nearly 8% increase of the previous human genome assembly [9]. Therefore, it is of essence to compare it to the more recent human genome assembly of the Chinese population, the T2T-YAO genome [10], and to annotate all CGIs for both assemblies. Here, we simply raise three basic questions. First, are there significant differences in CGIs and their methylation between the two reference genomes? And if so, what are the relevant genes and their functions diverged between them? Second, does the T2T-YAO genome provide additional information or serve more appropriately as one of the reference genomes for the Chinese population? Both are of importance for subsequent genome-wide epigenetic studies and building methylation databases specifically for the Chinese populations under complex physiological and pathological conditions. We, therefore, chose to accurately predict genome-wide CGIs for both reference genomes and validate their applications in methylation studies by testing them on a limited number of datasets due to data availability and quality.

A genome-wide comparison of two reference genomes has been done recently between the classical GRCH38 and T2T-CHM13 genomes and the authors discovered nearly 200 million base pairs (Mbp) unique to the T2T-CHM13 genome[9]. This result supports our effort in mapping the T2T-YAO and T2T-CHM13 genomes to explore their divergent sequences, and determining whether the divergent sequences are associated with specific biological functions became our first scientific question. We have described previously a genome-wide inclusive CGI definition, *i.e.*, a distance-based clustering method yielding short CpG cluster, known as position-defined CGIs show potential gene regulatory functions and are closely correlated with gene expression

specificity [2, 11, 12]. Therefore, whether there was any specificity between T2T-YAO and T2T-CHM13 according to CGI features, became our second scientific question. It becomes feasible now to carry out a comparative analysis on the T2T-YAO and T2T-CHM13 genomes on CGI methylation profiles using public WGBS data [2, 8, 11, 12], and this is what we address as the third scientific question.

To this end, we align the two genomes, compute both density-defined and position-defined CGIs, separately, and map some genome-wide methylation WGBS data from representative organs to analyze differentially methylated CGIs between the two genomes. Our major findings are as follows: (1) There is an approximately 4.7–5.8% difference between the two genomes, which is composed of divergent sequence-associated genes correlated closely with neurological functions based on GO enrichment analysis [13–18]. (2) CGIs associated with the divergent sequences are significantly different with respect to CpG density and O/E ratio [11, 19]. (3) Not only do WGBS data mapped to the T2T-YAO genome have a greater CpG site coverage than those of the T2T-CHM13 genome, but also the T2T-YAO genome shows more hyper-methylated CpG sites than T2T-CHM13 does. Our study demonstrates statistically significant differences between T2T-YAO and T2T-CHM13 in genome sequence and epigenetic landmarks[20–22], such as CGI and methylation patterns, suggesting that the establishment of CGI and methylation profiles based on the Chinese T2T reference genome is crucial for subsequent genome-wide epigenetic studies based on Chinese populations.

Results

In order to address the three scientific questions (**Figure 1**), we analyzed the two genomic references, their CGI distribution, and genome-wide CGI-associated methylation.

Comparative sequence analysis

To address our first question, we aligned the T2T-YAO-hp and T2T-CHM13 genomes and functionally categorized the divergently enriched genes from the T2T-YAO-hp genome (*-hp* stands for the haploid sequence from the offspring of the triad family T2T-YAO assemblies; T2T-YAO-hp and T2T-YAO used here are equivalent unless otherwise specified).

Sequence differences between T2T-YAO-hp and T2T-CHM13

First, from the sequence alignment (the alignment results are detailed in Supplementary **Table S1 and S2**), as shown in **Table 1**, the T2T-CHM13 genome has approximately 54.6 Mbp more sequence as compared to the T2T-YAO-hp genome. The uniquely divergent sequences account for 5.751% and 4.665% of the full length T2T-CHM13 and T2T-YAO-hp genomes, respectively. Therefore, the difference between the two genomes is approximately 4.7%–5.8%. Second, using NGenomeSyn [23], we created a collinearity map of the two genomes and observed that the two sequences are very similar both in identity and chromosome length (**Figure 2)** in general [24]. The divergent sequences between the two genomes are mainly concentrated in specific regions of each chromosome (Figure 2B). For instance, some of the divergent sequences of chromosomes 15 and 22 are enriched at the their left telomeric ends, whereas others, such as chromosome 9, have those located near the centromeres. The examples of some longest diverge sequence segments include 3.31×10^7 bp *vs*. 1.69×10^7 bp of chromosome 9 and 2.15×10^7 bp *vs.* 1.31×10^7 bp of chromosome Y; both are longer in the T2T-CHM13 genome (Figure 2C).

Divergent-sequence-associated genes

We categorized the divergent-sequence-associated genes of the two genomes (**Figure 3A** and Supplementary **Table S3)** and show that most of them are protein-coding genes, accounted for 44.26%, and the rest include IncRNAs that are 25.93% of the total gene count. Our GO enrichment analysis for the protein-coding genes shows that the most significant terms are "homophilic cell adhesion via plasma membrane

138 adhesion molecules", "ion channel complex", and "gated channel activity" in the
139 biological process (BP), cellular component (CC), and molecular function (MF)
140 categories [25], respectively, and that most of these GO terms are associated with
141 certain functions of the nervous system (**Figure 3B)**.

142 **CGI prediction and analysis**

143 *CGI prediction results*

144 We employed the rule-based method [12, 19] and distance-based clustering
145 method [12, 19] to predict the number of CGIs and the average CGI length for the two
146 genomes (**Table 2** and Supplementary **Table S4)**, and show the distribution of CGIs on
147 each of the T2T-YAO-hp and T2T-CHM13 chromosomes (Supplementary **Figure S1**
148 compares). On one hand, the number of density-defined CGIs and the average length
149 predicted for the T2T-YAO-hp genome are slightly smaller (84,471 bp *vs.* 87,788 bp
150 and shorter (391 bp *vs.* 407 bp) than those of the T2T-CHM13 genome, respectively.
151 On the other, the number of position-defined CGIs predicted for the T2T-YAO-hp
152 genome is greater than those of the T2T-CHM13 genome under the default parameter
153 d=50 bp (222,421 bp *vs.* 220,736 bp). In addition, the CGI number predicted for the
154 T2T-YAO-hp genome is less than those of the T2T-CHM13 genome regardless if their
155 density-defined CGIs or position-defined CGIs, but the average length of the position-
156 defined CGIs predicted by the T2T-YAO-hp genome is collectively longer than those
157 of the T2T-CHM13 genome (*e.g.*, 60 bp *vs.* 58 bp with d = 25 bp) under the parameters
158 d = 25 bp and 12 bp. We also predicted all CGIs related to divergent sequences of the
159 two genomes (Supplementary **Table S5)**.

160 To study the relationship between position-defined CGIs and density-defined
161 CGIs for the same genome, we investigated sequences that intersect the two groups (for
162 the position-defined CGIs, d = 25 bp) for both genomes **(Figure 4)**. First, the total
163 length of density-defined CGIs is much longer, approximately three times that of
164 position-defined CGIs for each genome *(e.g.*, $(24{,}991{,}625 + 8{,}062{,}542)/$
165 $(8{,}062{,}542 + 1{,}735{,}619) = 3.37$ for the T2T-YAO-hp genome). Second, although

166 most of the position-defined CGIs overlap with the density-defined CGIs, there is still
167 a fraction (16.5% and 17.7%, respectively) that is not predicted by the rule-based CGI
168 predicting method.

169 *CGI feature comparison*

170 First, to compare the differences in CGIs between the two genomes, we calculated
171 four commonly used CGI features, *i.e.*, CGI length, GC content, CpG density, and CGI
172 O/E ratio [2, 26] (**Figure 5**) and show the feature distribution of genome-wide and
173 divergent-sequence-associated position-defined CGIs, including d = 12 bp, 50 bp
174 (Supplementary **Figure S2)**. Other than the grouping of density-defined or position-
175 defined CGIs, there is no obvious difference among the four commonly used CGI
176 features between the two genomes based on all CGI counts (Figure 5A–5D). For
177 example, in both genomes, the peak lengths of the density-defined CGIs and the
178 position-defined CGIs are approximately 210 bp and 30 – 50 bp, respectively (Figure
179 5A).

180 Second, there is a greater difference in the four commonly used CGI features of
181 the divergent sequence CGIs between the two genomes than those of all CGI counts
182 (Figure 5E–5H). For instance, the density-defined CGIs of the T2T-CHM13 genome
183 have two peaks in high GC content regions (from 50% to 60%), whereas the density-
184 defined CGIs of the T2T-YAO-hp genome have a single peak in the same regions. In
185 addition, the position-defined CGIs in the divergent sequences, except for a small
186 difference in the distribution of CGI lengths, have significantly different distributions
187 for GC content, O/E ratio, and CpG density (Figure 5E–5H, Supplementary **Figure S2**,
188 and Supplementary **Table S6**).

189 Third, for all CGIs or divergent-sequence-associated CGIs, the distance-based
190 clustering method is more likely to identify CGIs with shorter lengths and greater GC
191 content, O/E ratio, and CpG density than the rule-based CGI predicting method (Figure
192 5).

193 **Methylation analysis and comparison**

To address our third question, we compared not only the number of cytosine sites (C sites, Supplementary **Figure S3**) and methylation levels of all CpG sites, but also methylation levels of genome-wide CGIs for the two genomes.

CpG methylation levels

We classified CpG methylation levels of similar (Supplementary **Figure S4**) and divergent sequences between the two genomes (**Figure 6**). In addition, based on the WGBS data of two germline tissues, the ovary and testis, we also determined methylation profiles of whole genomes and divergent sequences for the two genomes (Supplementary **Figure S5).** The number of CpG sites of the divergent sequences failed to be mapped based on the same WGBS methylation data is much lower in the T2T-YAO-hp genome than in the T2T-CHM13 genome in both germline tissues. For example, the number of CpG sites in the divergent sequences of the ovary failed to be mapped to T2T-YAO-hp is 2.16×10^6 bp, whereas that in the T2T-CHM13 genome is 3.49×10^6 bp. Inversely, the number of CpG sites in the divergent sequences mapped for T2T-YAO-hp is greater than that for T2T-CHM13 at all three methylation levels. In terms of quantity, the most notable difference is what found in the hypermethylated sequences (2.8×10^5 *vs*. 2.2×10^5 for ovary; 3.1×10^5 *vs.* 2.5×10^5 for testis).

CGI methylation levels

We used Equation 6 to calculate the methylation level of all density-defined and position-defined CGIs for the two reference genomes (**Figure 7** and Supplementary **Figure S6**). First, after both definitions, the proportions of CGI-invalid data (WGBS data failed to be mapped to any CGIs) are higher in the T2T-CHM13 than in T2T-YAO-hp genomes (Figure 7); there are 10.46% *vs*. 7.27% for the density-defined CGIs and 23.63% *vs.* 14.87% for position-defined CGIs not mapped in the ovary for the two genomes (Figure 7A). Second, the proportions of hyper-methylated, intermediate methylated, and hypo-methylated of CGIs, predicted by both definitions, are slightly greater for the T2T-YAO-hp genome than those of the T2T-CHM13 genome; the

proportions of hyper-methylated, intermediate methylated, and hypo-methylated for the T2T-YAO-hp genome predicted as position-defined CGIs are 19.73%, 13.29%, and 53%, respectively, as opposed to those of the T2T-CHM13 genome, 17.32%, 11.77%, and 48.02% (Figure 7B). Third, the density-defined CGIs appear to have the greatest proportion of hypermethylation for both genomes, whereas position-defined CGIs have a greater proportion of hypomethylation (Figure 7). The density-defined CGIs of both genomes have higher proportions of hyper-methylated CGIs (50.02% *vs.* 48.40% for the T2T-YAO and T2T-CHM13 genomes, respectively) than those of intermediate methylated and hypo-methylated CGIs, whereas the position-defined CGIs have the greatest proportion of hypomethylation (53.65% *vs.* 48.57% for the T2T-YAO and T2T-CHM13 genomes, respectively).

Discussion and conclusion

The aim of this study is to emphasize merits and specific applications of a population-centric reference genome by comparing the sequence divergence, as well as various epigenetic parameters, between the T2T-YAO and T2T-CHM genomes. We have further provided the evidence suggesting the necessity of using T2T-YAO as a reference genome for future genome-wide genetic and epigenetic studies in Chinese populations. We have started with the two high-quality genome sequences to compute the genome-wide cytosine and CpG sites and further categorized them into two groups of CGIs, position-defined and density-defined. After the three-round stratification, we have presented the divergent-sequence-associated genes and concluded that these genes are mostly neurological in function. In addition, we have mapped the methylation levels of CpGs and CGIs for the two reference genomes using selected public WGBS data.

The overall sequence divergence between the two reference genomes are significant and interesting, albeit only 4.7–5.8% (minimal alignment length of 5,000bp and identity cutoff of 90%). Such a loosely-defined difference, generally speaking, is mostly population-associated variations between the northern Chinese and European

populations [9, 10]. Future studies will have to expand empirical data into within-population variations based on big data for their detailed categorization of functional relevance. The current divergent sequences appear to non-uniformly spread out into certain chromosomal regions within a few chromosomes and their specific regions, such as the cases involving chromosomes 9 and Y; such a distribution also suggests population-associated selection and recombination. Furthermore, the GO enrichment of divergent-sequence-associated genes specific to the T2T-YAO genome indicates a strong preference for specific molecular mechanisms related to neuronal cells and synapses, providing clues for future in-depth investigation.

Our genome-wide CGI study between the two genomes indicates that two-fold CGI stratifications are necessary for precision methylation mapping. It is important to start methylation mapping from all potential methylation sites that are ensured only by T2T population-specific genomes. The two CGI definitions and their grouping schemes, such as the distance (12 bp, 25 bp, and 50 bp) and the density (high, moderate, and low), provide matrices for pattern and rule recognitions, which are used to identify genes and other genomic elements and features for further functional scrutinization. In addition, after the application of the two CGI definitions, there is still a significant fraction of the C and CpG sites to be further annotated, leveraging on the complete genomes, as some of these sites are actually methylated.

From the methylation mapping of the two datasets representing the two germline tissues, we have learnt several lessons. First, we expected more CpG sites to be mapped in the T2T-CHM13 reference from the methylation data of the European origins since its DNA is from a single homozygous complete hydatidiform mole (46, XX) and the Y sequence is from another assembly of similar origin. However, to our surprise, the number of CpG methylation sites to the divergent sequence associated to the T2T-YAO genome is greater than that of the T2T-CHM13 reference. Second, the CpG sites that failed to be mapped by the gender-specific methylation data are also biased toward the same direction, *i.e.*, there are more to be mapped in the T2T-YAO reference. These

subjective mapping results lead to an across-the-board conclusion that methylation patterns are reference sensitive regardless the methylation levels. Third, we have also shown apparent differences of CGI methylation between the two CGI definitions and the rate varies among different methylation levels.

After all, there are still several questions that need further exploration. First, the success of large-scale genome-wide epigenomic studies rely on hign-quality and population-specific reference genomes and there are more much more references to be constructed for human methylation atlas of different populations. Second, novel and more thorough CpG and CGI definitions are needed for understanding the relatedness of data to be mapped and their references. Since artificial intelligence (AI) has been gradually used in genome-wide methylation-associated studies [27, 28], determining how to use AI techniques [29–32] such as neural networks to investigate the relationship between sequence features, CGI, methylation, and expression specificity of the human genome has become a future research direction. Finally, genome-wide CGI prediction and methylation analysis are not only data intensive but also time-consuming. However, there is currently no such web service that collects and manages CGI and methylation annotation data for the Chinese T2T genome. Therefore, we will build a web service for online computation, data sharing and visualization based on high-performance computing[33, 34], such as GPU acceleration and distributed computing, to provide a data foundation and an online platform for further study.

Materials and methods

Data source

We downloaded the human GRCh38 and T2T-CHM13 (v2.0) genome sequences from the NCBI databases [35] and the Chinese T2T-YAO-hp genome sequence from the BIGD (BIG Data Center, Beijing) databases [10]. We used CGI data of the T2T-CHM13 genome, which were predicted based on the Gardiner-Garden and Frommer (GGF) [36] standard in UCSC [37]. We downloaded the WGBS data for two

representative tissues from the ENCODE database [38], which are ovary (ENCSR417YFD) from a European donor and testis (ENCSR806NNG) from an American individual.

Methods

The workflow of this study in three main aspects is described in Figure 1, and our key algorithms used are described as follows.

Comparative sequence analysis

Statistics of all cytosine C sites

We counted and compared the number and distribution of cytosine sites (C sites) that can be methylated among the T2T-YAO-hp, T2T-CHM13, and GRCh38 genomes. Since most DNA methylation sites found in mammals involve CpG dinucleotides [39], we categorized C sites into three different contexts: CpG, CHG, and CHH, where H represents A, C or T.

Genome alignment and analysis

We used nucmer, a sequence alignment algorithm of the MUMmer system [40], to carry out genome-wide alignment and analysis for the T2T-YAO-hp and T2T-CHM13 genomes (**Figure S7)**. We set the key parameters for nucmer to calculate similar and divergent sequences as follows: minimum alignment length is 5,000 bp, minimum alignment similarity is 90%. The similar sequences are those regions identified by nucmer as aligning between the two genomes. Conversely, divergent sequences are those regions that do not meet these above-mentioned alignment criteria, thus representing unique sequences present in each genome. The GO enrichment of divergent-sequence-associated genes were analyzed by using the clusterProfiler algorithm [41].

CGI definition and analysis

Our previous studies denoted the CGIs predicted by the classical rule-based

method [36] as density-defined CGIs, and the CGIs predicted by the distance-based clustering method [26] as position-defined CGIs, which are able to localize more short CGIs of potential gene regulatory functions and are closely correlated with gene expression specificity [2, 11, 12] in the human genome. Therefore, in order to comprehensively compare CGI features of the two reference genomes, we calculated not only all the density-defined and position-defined CGIs for the T2T-YAO-hp and T2T-CHM13 genomes but also the density-defined and position-defined CGIs related to divergent sequences of the two genomes (Supplementary **Figure S8**).

Density-defined CGI

Based on the GGF standard of CGIs [36], we computed the sequence segments that meet Equations. 1–3 as density-defined CGIs by the CpGPlot algorithm [42].

$$\text{length} \geq 200\text{bp} \tag{1}$$

$$\text{GC content} \geq 50\% \tag{2}$$

$$O/E\ ratio = \frac{Number\ of\ CpG}{Number\ of\ C\ \times\ Number\ of\ G} \times \text{length} \geq 0.6 \tag{3}$$

where length is the nucleotide number of the analysed gene sequence region, GC content is the proportion of G+C in the analysed gene sequence region, and O/E ratio represents the ratio of observed CpG to expected CpG.

Position-defined CGI

We first calculated the CpG clusters in the genome by Equation 4, and then considered these CpG clusters with small p values [26] (Equation 5) as position-defined CGIs.

$$d_i = x_{i+1} - x_i - 1 \tag{4}$$

$$P(d) = (1-p)^{d-1}\text{p} \tag{5}$$

where x represents the coordinates of CpG and i represents the index of CpG. P(d) represents the probability of finding a distance d between neighboring CpGs. p corresponds to the probability of CpGs in the sequence. Since the default value of d in the CpGcluster algorithm is 50 and our previous study found that LAUPs (lineage-

associated underrepresented permutations) are strongly associated with CGIs, the length of the shortest LAUPs of multiple species ranges from 10 to 14 [19], and the length of 25 bp is the statistical proportionality demarcation point for all CpGs and core-promoter-associated CpG regions in the human genome [2], we computed CGIs with d = 50 bp, 25 bp, and 12 bp (the average of 10 and 14).

Genome-wide methylation analysis

WGBS data mapping

We applied gemBS [43], a large-scale WGBS data analysis method, to carry out methylation analysis for the two representative tissues (ovary and testis) by setting T2T-YAO-hp and T2T-CHM13 as references. The details of the procedure are described in Supplementary **Figure S9**.

CGI methylation ratio computing

After obtaining the CGI prediction for the reference genome and the methylation rate for each CpG site, Equation 6 calculates the methylation ratio for each CGI.

$$\text{mean_methylation}(cgi) = \frac{\sum_{cpg \in cgi} reads \times \text{methylation_ratio}(cpg)}{\sum_{cpg \in cgi} reads} \quad (6)$$

where *reads* represent the sequencing depth of WGBS for each CpG site in the CGI, and methylation_ratio(*cpg*) represents the methylation ratio of each CpG site derived from the gemBS parallelization algorithm.

Methylation level classification

According to the previous studies [2, 44], Equation 7 was used to classify the methylation ratio for each CpG site and CGI into three methylation levels, "hyper-methylated", "hypo-methylated", and "intermediate methylated".

$$\text{Methylation level}(chr, p) = \begin{cases} -1, \text{ invalid} & \text{no reads map to this cpg or cgi} \\ 1, \text{ hyper-methylated} & \text{methylation}_{\text{ratio(chr,p)}} > 0.75 \\ 2, \text{hypo-methylated} & \text{methylation}_{\text{ratio(chr,p)}} < 0.1 \\ 3, \text{intermediate methylated} & \text{otherwise} \end{cases} \quad (7)$$

where chr and p represent the chromosome and the coordinates of the CpG site,

respectively.

CRediT author statement

Ming Xiao: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Rui Wei:** Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Jun Yu:** Conceptualization, Writing - review & editing, Visualization, Supervision, Project administration. **Chujie Gao:** Formal analysis, Investigation, Resources, Data curation. **Fengyi Yang:** Formal analysis, Investigation, Resources, Data curation. **Le Zhang:** Conceptualization, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China [Grant No. 62372316], National Science and Technology Major Project (Nos. 2021YFF1201200 and 2018ZX10201002, China), China Postdoctoral Science Foundation (2020M673221, China), Fundamental Research Funds for the Central Universities (2020SCU12056, China), Sichuan Science and Technology Program (2022YFS0048), Chongqing Technology Innovation and Application Development Project (CSTB2022TIAD-KPX0067), China. We thank Dr. Yu Kang from Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation for her assistance in data analysis.

ORCID

ORCID 0000-0001-8608-5903 (Ming Xiao)

ORCID 0009-0006-8597-5924 (Rui Wei)

ORCID 0000-0002-2702-055X (Jun Yu)

ORCID 0009-0000-3797-7040 (Chujie Gao)

ORCID 0009-0008-4105-030X (Fengyi Yang)

ORCID 0000-0002-3708-1727 (Le Zhang)

References

[1] Pongor CI, Bianco P, Ferenczy G, Kellermayer R, Kellermayer M. Optical trapping nanometry of hypermethylated CPG-Island DNA. Biophys. J 2017;112:512.

[2] Xiao M, Yin R, Gao P, Yu J, Ma F, Dai Z, et al. Position-defined CpG islands provide complete co-methylation indexing for human genes. International Conference on Intelligent Computing (2022).2022:319-33.

[3] Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. Lancet 2018;392:777-86.

[4] Wu H, Zhang Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell 2014;156:45-68.

[5] Acton R, Yuan W, Gao F, Xia Y, Bourne E, Wozniak E, et al. The genomic loci of specific human tRNA genes exhibit ageing-related DNA hypermethylation. Nat. Commun 2021;12:2655.

[6] Liu B, Du Q, Chen L, Fu G, Li S, Fu L, et al. CpG methylation patterns of human mitochondrial DNA. Sci. Rep 2016;6:23421.

[7] El-Maarri O, Olek A, Balaban B, Montag M, van der Ven H, Urman B, et al. Methylation levels at selected CpG sites in the factor VIII and FGFR3 Genes, in mature female and male germ cells: implications for male-driven evolution. Am. J. Hum. Genet 1998;63:1001-8.

[8] Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, et al. Epigenetic patterns in a complete human genome. Science 2022;376:eabj5089.

[9] Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. Science 2022;376:eabl3533.

[10] He Y, Chu Y, Guo S, Hu J, Li R, Zheng Y, et al. T2T-YAO: a telomere-to-telomere assembled diploid reference genome for Han Chinese. Genomics Proteomics Bioinformatics 2023;S1672-0229(23)00100-6.

[11] Zhang L, Dai Z, Yu J, Xiao M. CpG-island-based annotation and analysis of human housekeeping genes. Brief. Bioinformatics 2021;22:515-25.

[12] Xiao M, Yang X, Yu J, Zhang L. CGIDLA：developing the web server for CpG island related density and LAUPs (Lineage-associated Underrepresented Permutations) Study. IEEE/ACM Trans Comput Biol Bioinform 2019;17:2148-54.

[13] Lai X, Zhou J, Wessely A, Heppt M, Maier A, Berking C, et al. A disease network-based deep learning approach for characterizing melanoma. Int J Cancer 2022;150:1029-44.

[14] Zhang L, Liu G, Kong M, Li T, Wu D, Zhou X, et al. Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. Bioinformatics 2021;37:1554-61.

[15] Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Comput Struct Biotechnol J 2020;18:417-26.

[16] Zhang L, Zheng C, Li T, Xing L, Zeng H, Li T, et al. Building up a robust risk mathematical platform to predict Colorectal Cancer. Complexity 2017;2017:1-14.

[17] Zhang L, Liu Y, Wang M, Wu Z, Li N, Zhang J, et al. EZH2-, CHD4-, and IDH-linked epigenetic perturbation and its association with survival in glioma patients. J Mol Cell Biol 2017;9:477-88.

[18] Xia Y, Yang C, Hu N, Yang Z, He X, Li T, et al. Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model. BMC Genom 2017;18:950.

[19] Zhang L, Xiao M, Zhou J, Yu J. Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA). Bioinformatics 2018;34:3624-30.

[20] Zhang L, Zhao J, Bi H, Yang X, Zhang Z, Su Y, et al. Bioinformatic analysis of chromatin organization and biased expression of duplicated genes between two poplars with a common whole-genome duplication. Hortic Res 2021;8:62.

[21] Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol 2019;15:e1007069.

[22] Zhang Q, Zhang H, Zhou K, Zhang L. Developing a physiological signal-based, mean threshold and Decision-Level Fusion Algorithm (PMD) for emotion recognition. Tsinghua Sci Technol 2023;28:673-85.

[23] He W, Yang J, Jing Y, xu L, Yu K, Fang X. NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. Bioinformatics (Oxford, England) 2023;39.

[24] Guarracino A, Buonaiuto S, de Lima LG, Potapova T, Rhie A, Koren S, et al. Recombination between heterologous human acrocentric chromosomes. Nature 2023;617:335-43.

[25] Huntley RP, Tony S, Prudence MM, Aleksandra S, Carlos B, Martin MJ, et al. The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res 2015;43:1057-63.

[26] Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL. CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinform 2006;7:446.

[27] You Y, Lai X, Pan Y, Zheng H, Vera J, Liu S, et al. Artificial intelligence in cancer target identification and drug discovery. Signal Transduct Target Ther 2022;7:156.

[28] Tian H, He Y, Xue Y, Gao YQ. Expression regulation of genes is linked to their CpG density distributions around transcription start sites. Life Science Alliance 2022;5:e202101302.

[29] Zhang L, Fan S, Vera J, Lai X. A network medicine approach for identifying diagnostic and prognostic biomarkers and exploring drug repurposing in human cancer. Comput Struct Biotechnol J 2023;21:34-45.

[30] You Y, Zhang L, Tao P, Liu S, Chen L. Spatiotemporal transformer neural network for time-series forecasting. Entropy (Basel) 2022;24:1651.

[31] Song H, Chen L, Cui Y, Li Q, Wang Q, Fan J, et al. Denoising of MR and CT images using cascaded multi-supervision convolutional neural networks with progressive training. Neurocomputing 2022;469:354-65.

[32] Liu G-D, Li Y-C, Zhang W, Zhang L. A brief review of artificial intelligence applications and algorithms for psychiatric disorders. Engineering 2020;6:462-7.

[33] Zhang L, Badai J, Wang G, Ru X, Song W, You Y, et al. Discovering hematoma-stimulated circuits for secondary brain injury after intraventricular hemorrhage by spatial transcriptome analysis. Front Immunol 2023;14:1123652.

[34] Zhang L, Zhang L, Guo Y, Xiao M, Feng L, Yang C, et al. MCDB: A comprehensive curated mitotic catastrophe database for retrieval, protein sequence alignment, and target prediction. Acta Pharm Sin B 2021;11:3092-104.

[35] Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005;33:D501-4.

[36] Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol 1987;196:261-82.

[37] Lee BT, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC genome browser database: 2022 update. Nucleic Acids Res 2022;50:D1115-d22.

[38] Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the encyclopedia of DNA elements (ENCODE) data portal. Nucleic Acids Res 2020;48:D882-d9.

[39] He X-J, Chen T, Zhu J-K. Regulation and function of DNA methylation in plants and animals. Cell Res 2011;21:442-65.

[40] Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol 2018;14:e1005944.

[41] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics-a Journal of Integrative Biology 2012;16:284-7.

[42] Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet 2000;16:276-7.

[43] Merkel A, Fernández-Callejo M, Casals E, Marco-Sola S, Schuyler R, Gut IG, et al. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. Bioinformatics 2019;35:737-42.

[44] Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature 2013;500:477-81.

Figure legends

Figure 1 Workflow of the study

GO, Gene Ontology; T2T, telomere-to-telomere; WGBS, whole-genome bisulfite sequencing; CGI, CpG island.

Figure 2 Sequence alignment between the T2T-CHM13 and T2T-YAO-hp genomes.

A. A collinearity map of the genome alignment. Colored lines link the high-identity or collinear sequence segments of the two compared genomes and the line thickness as a similarity measure. **B.** Visualization of sequence comparison across chromosomes. Blue and orange highlight similar sequences as vertical bars of the T2T-CHM13 and T2T-YAO-hp genomes, respectively. Blank bars represent divergent sequences of the two genomes. **C.** Divergent sequence length statistics of the two genomes. Blue and orange bars represent the divergent sequence lengths of the T2T-CHM13 and T2T-YAO-hp genomes in different chromosomes, respectively. Chr, chromosome.

Figure 3 GO enrichment of divergent-sequence-associated genes

A. Proportion of the types of divergent-sequence-associated genes in these two reference genomes. **B.** GO enrichment analysis of divergent sequence related protein-coding genes. Size and color of the circle represent the number of enriched genes and the statistical significance value (adjusted *P* value) [22, 41] of the current GO term, respectively. ncRNA, non-coding RNA; lncRNA, long non-coding RNA; BP, biological process; MF, molecular function; CC, cellular component.

Figure 4 The sequence intersecting the density-defined and position-defined CGIs (d = 25 bp)

The white part represents the sequence length intersecting position-defined and density-defined CGIs. The orange and blue parts represent the sequence length unique to the density-defined and position-defined CGIs, respectively.

Figure 5 Comparison of CGI features between the T2T-CHM13 and T2T-YAO-hp genomes

A. Length of all CGIs; **B.** GC content of all CGIs; **C.** O/E ratio of all CGIs; **D.** CpG density of all CGIs; **E.** Length of divergent-sequence-associated CGIs; **F.** GC content of divergent-sequence-associated CGIs; **G.** O/E ratio of divergent-sequence-associated CGIs; **H.** CpG density of divergent-sequence-associated CGIs. O/E, observed CpG/expected CpG.

Figure 6 CpG methylation level of divergent sequences between the T2T-CHM13 and T2T-YAO-hp genomes

Green, blue, and orange highlight the number of hyper-methylated, intermediate methylated, and hypo-methylated, respectively. Gray highlights the number of CpG sites that failed in mapping with WGBS data. Dark and light colors are used to distinguish the T2T-YAO-hp and T2T-CHM13 genomes, respectively.

Figure 7 CGI methylation levels of the two genomes

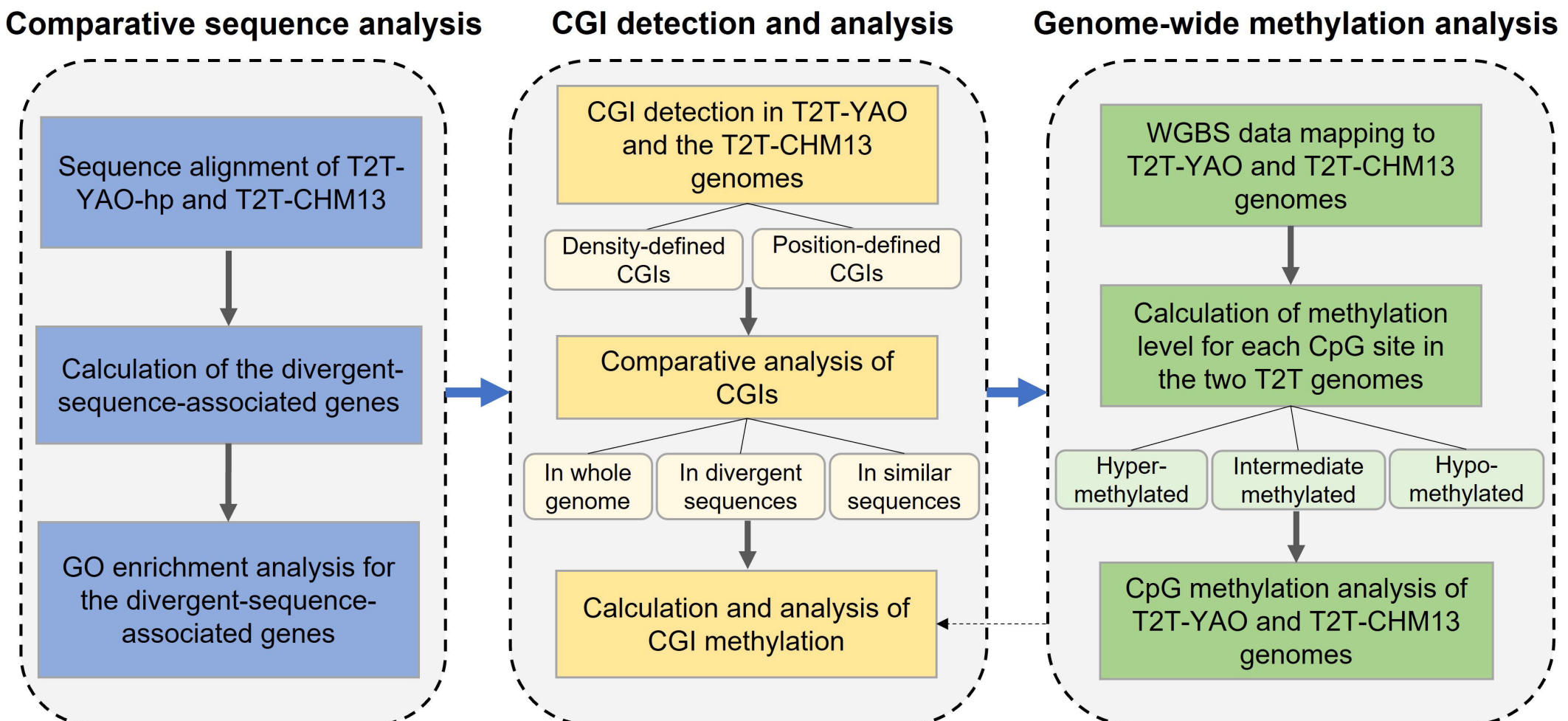
A. CGI methylation levels annotated by using the ovary data. **B.** CGI methylation levels annotated by using the testis data. Green, blue, and orange highlight the proportion of CGIs that are hyper-methylated, intermediate methylated, and hypo-methylated, respectively. Gray highlights the proportion that failed from all CGIs mapping, denoted as *invalid*. Darker and lighter colors are used to distinguish the T2T-YAO-hp and T2T-

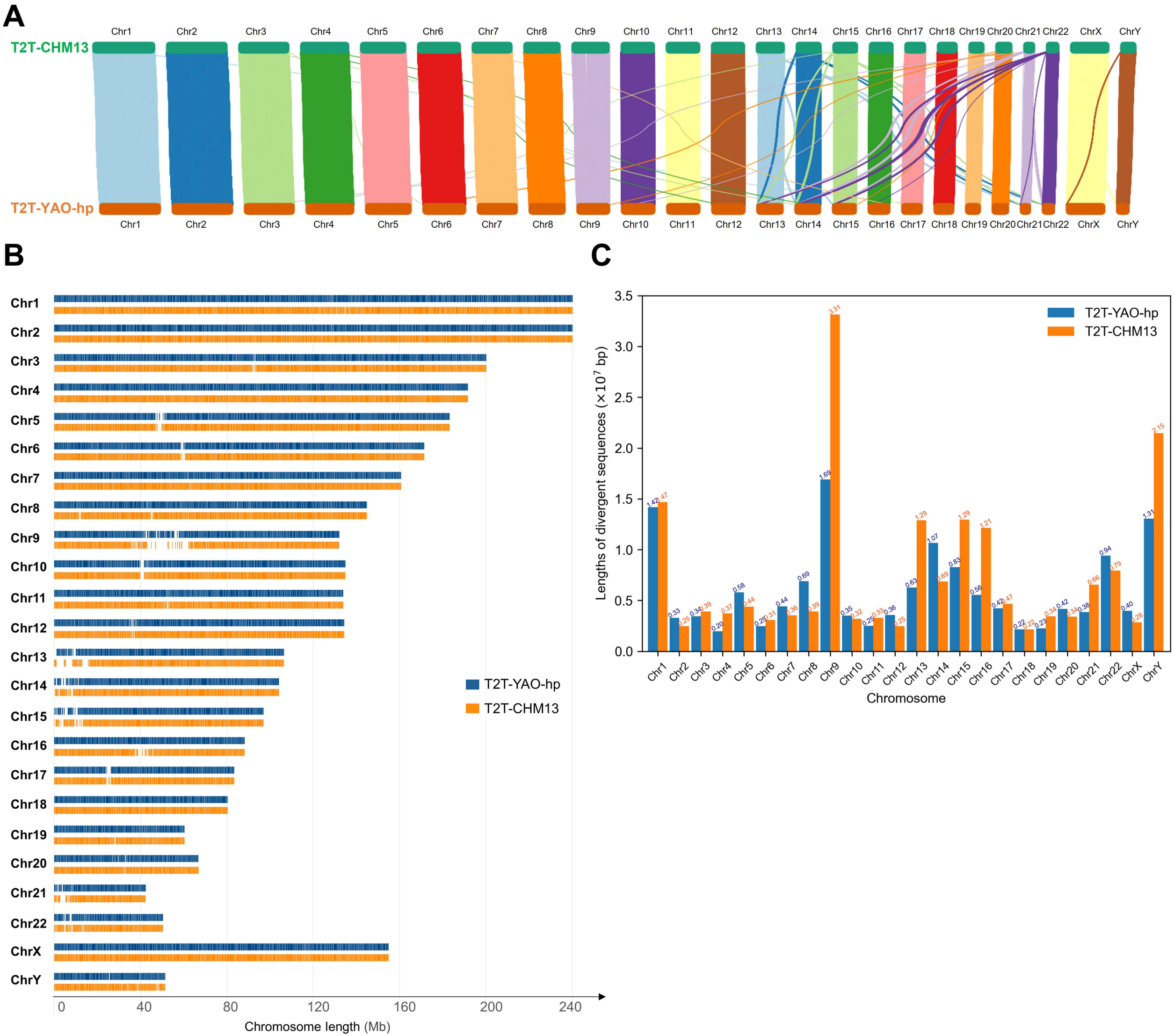
CHM13 genome, respectively.

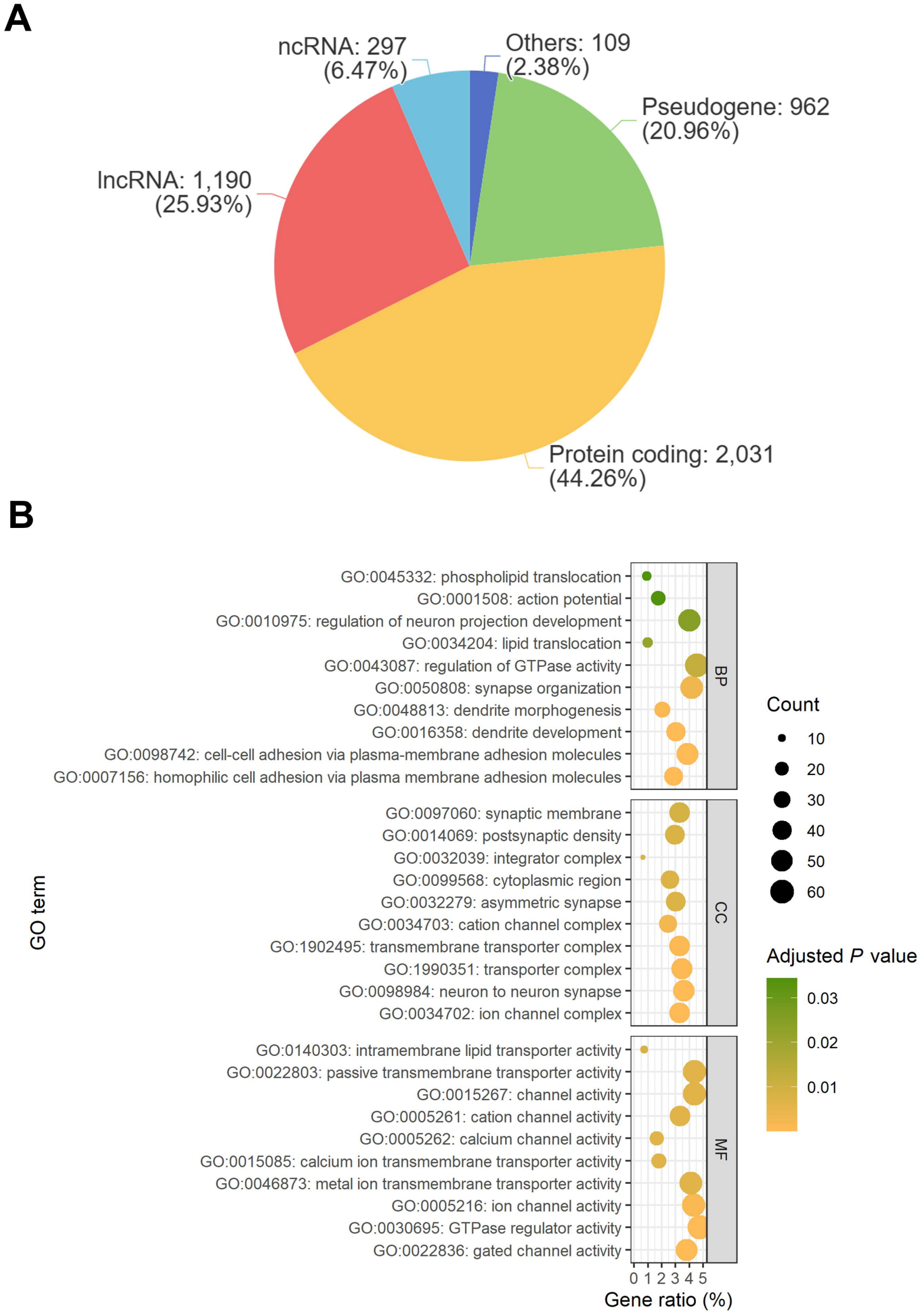
Tables

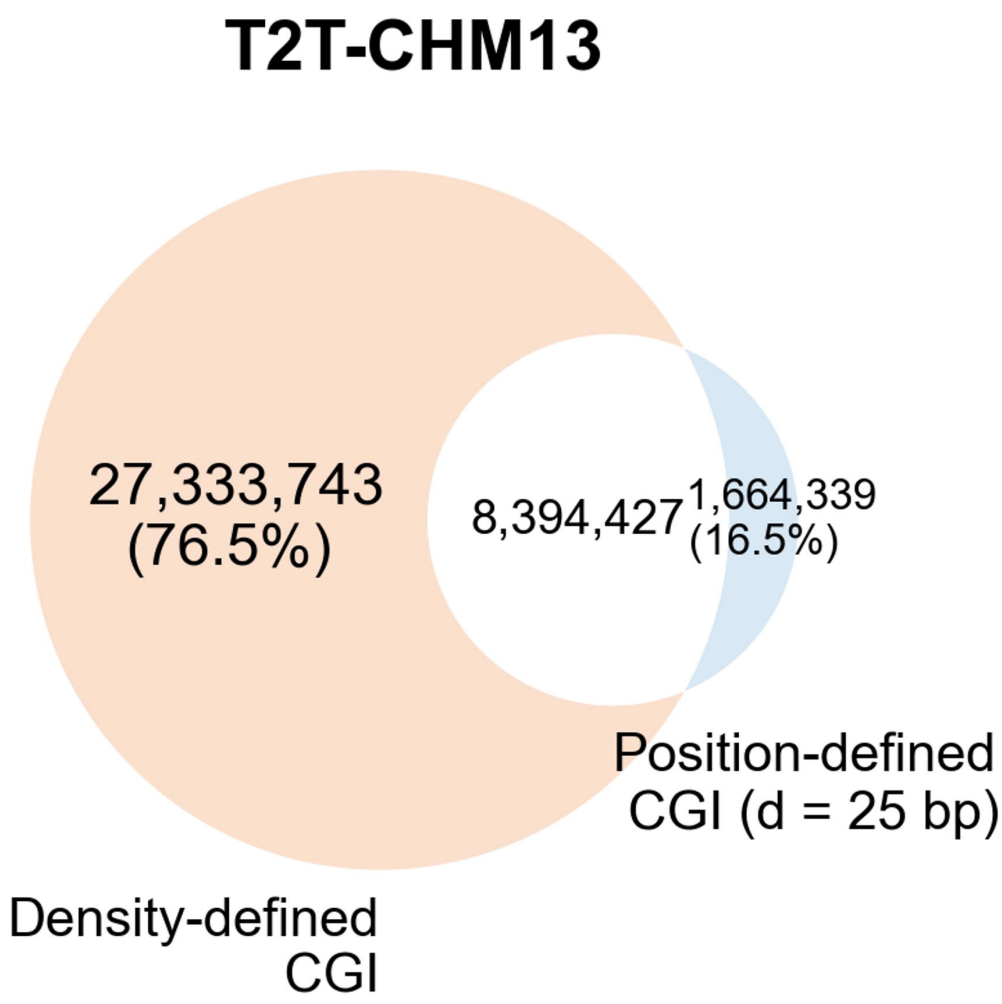
Table 1 Sequence differences between the T2T-CHM13 and T2T-YAO-hp genomes

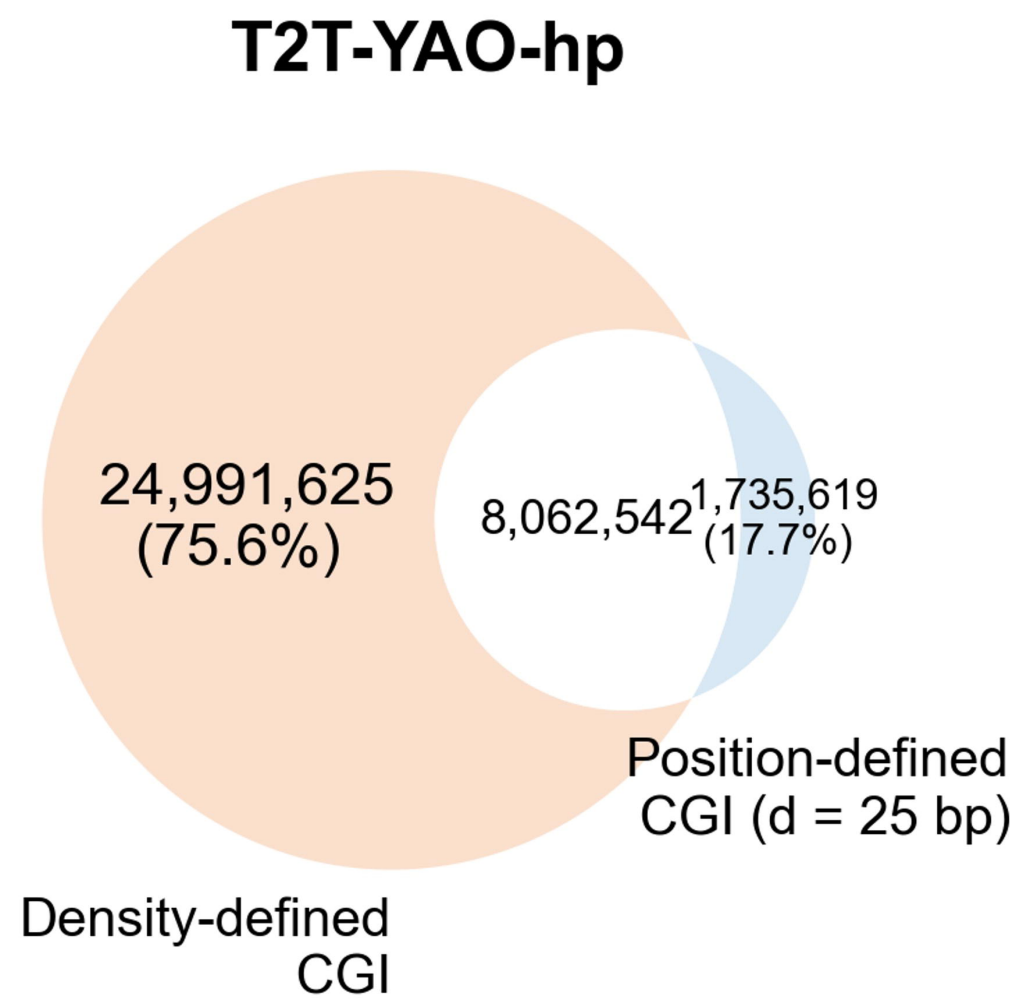
Table 2 CGIs Statistics of the T2T-CHM13 and T2T-YAO-hp genomes

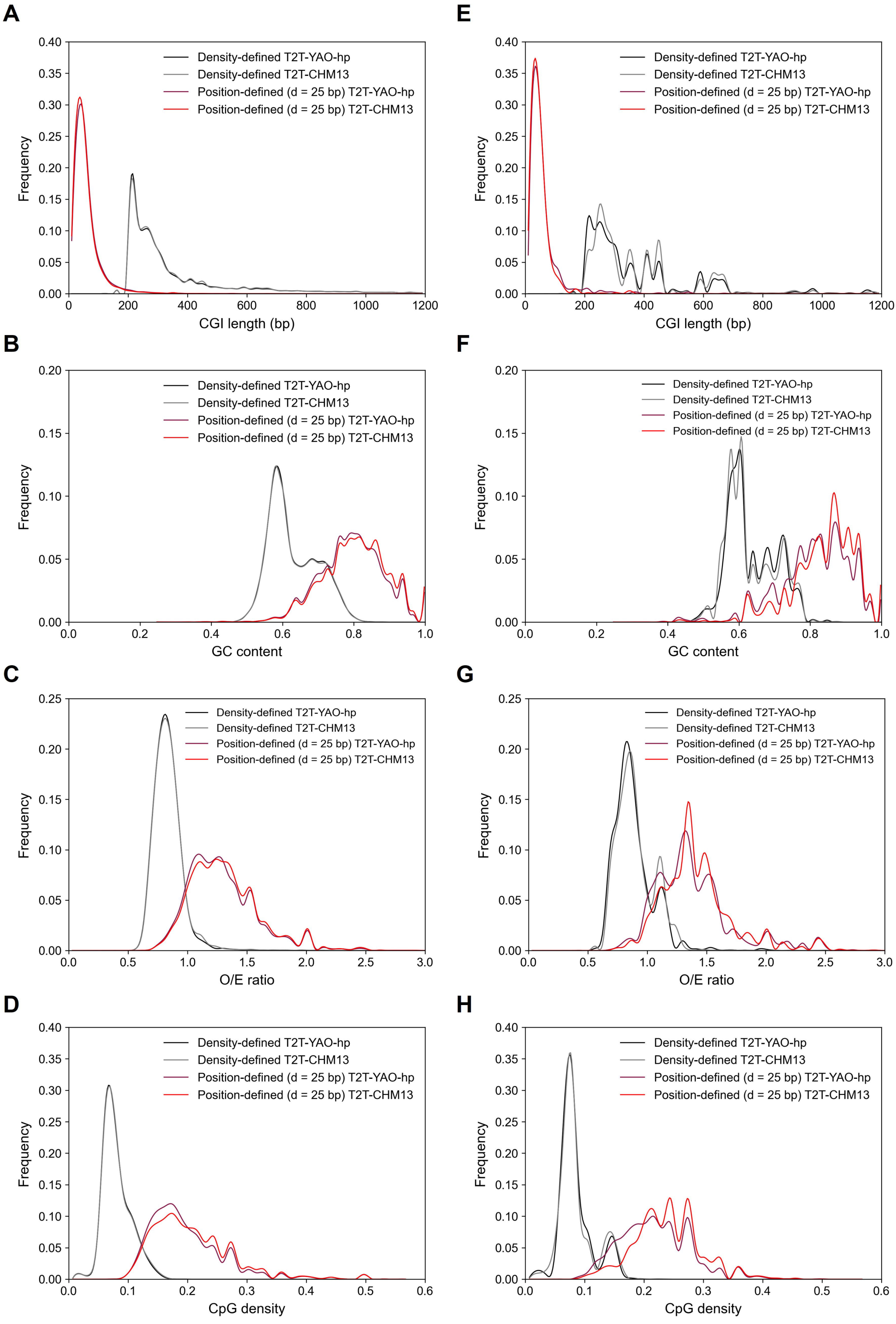

Comparative sequence analysis
Sequence alignment of T2T-YAO-hp and T2T-CHM13
Calculation of the divergent-sequence-associated genes
GO enrichment analysis for the divergent-sequence-associated genes
CGI detection and analysis
CGI detection in T2T-YAO and the T2T-CHM13 genomes
Density-defined CGIs
Position-defined CGIs
Comparative analysis of CGIs
In whole genome
In divergent sequences
In similar sequences
Calculation and analysis of CGI methylation
Genome-wide methylation analysis
WGBS data mapping to T2T-YAO and T2T-CHM13 genomes
Calculation of methylation level for each CpG site in the two T2T genomes
Hyper-methylated
Intermediate methylated
Hypo-methylated
CpG methylation analysis of T2T-YAO and T2T-CHM13 genomes

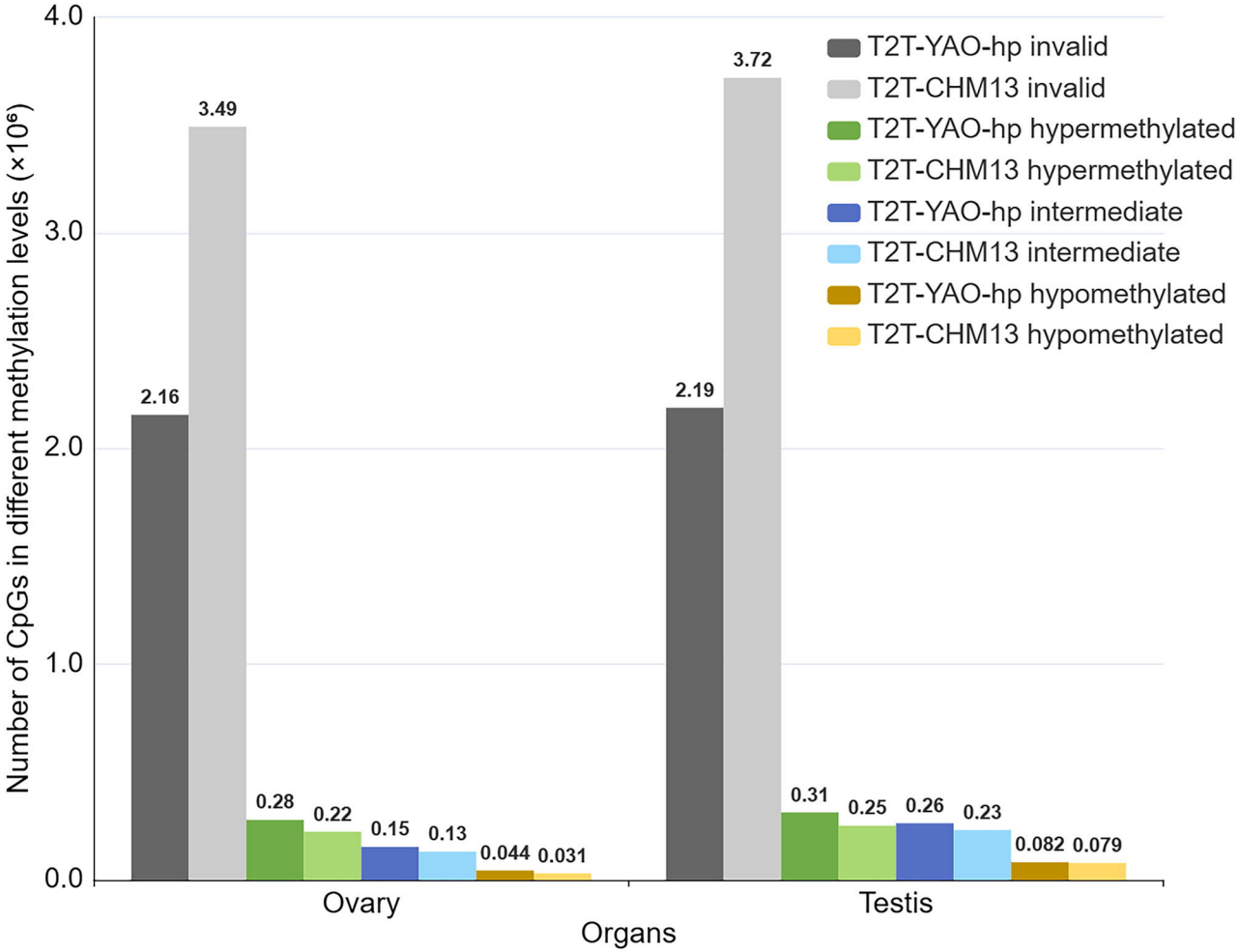

A
T2T-CHM13
T2T-YAO-hp
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 Chr9 Chr10 Chr11 Chr12 Chr13 Chr14 Chr15 Chr16 Chr17 Chr18 Chr19 Chr20 Chr21 Chr22 ChrX ChrY
B
T2T-YAO-hp
T2T-CHM13
0 40 80 120 160 200 240
Chromosome length (Mb)
C
T2T-YAO-hp
T2T-CHM13
Lengths of divergent sequences (×10^7 bp)
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Chromosome


A
ncRNA: 297
(6.47%)
Others: 109
(2.38%)
Pseudogene: 962
(20.96%)
lncRNA: 1,190
(25.93%)
Protein coding: 2,031
(44.26%)
B
GO:0045332: phospholipid translocation
GO:0001508: action potential
GO:0010975: regulation of neuron projection development
GO:0034204: lipid translocation
GO:0043087: regulation of GTPase activity
GO:0050808: synapse organization
GO:0048813: dendrite morphogenesis
GO:0016358: dendrite development
GO:0098742: cell-cell adhesion via plasma-membrane adhesion molecules
GO:0007156: homophilic cell adhesion via plasma membrane adhesion molecules
GO:0097060: synaptic membrane
GO:0014069: postsynaptic density
GO:0032039: integrator complex
GO:0099568: cytoplasmic region
GO:0032279: asymmetric synapse
GO:0034703: cation channel complex
GO:1902495: transmembrane transporter complex
GO:1990351: transporter complex
GO:0098984: neuron to neuron synapse
GO:0034702: ion channel complex
GO:0140303: intramembrane lipid transporter activity
GO:0022803: passive transmembrane transporter activity
GO:0015267: channel activity
GO:0005261: cation channel activity
GO:0005262: calcium channel activity
GO:0015085: calcium ion transmembrane transporter activity
GO:0046873: metal ion transmembrane transporter activity
GO:0005216: ion channel activity
GO:0030695: GTPase regulator activity
GO:0022836: gated channel activity
BP
CC
MF
GO term
0 1 2 3 4 5
Gene ratio (%)
Count
10
20
30
40
50
60
Adjusted P value
0.03
0.02
0.01

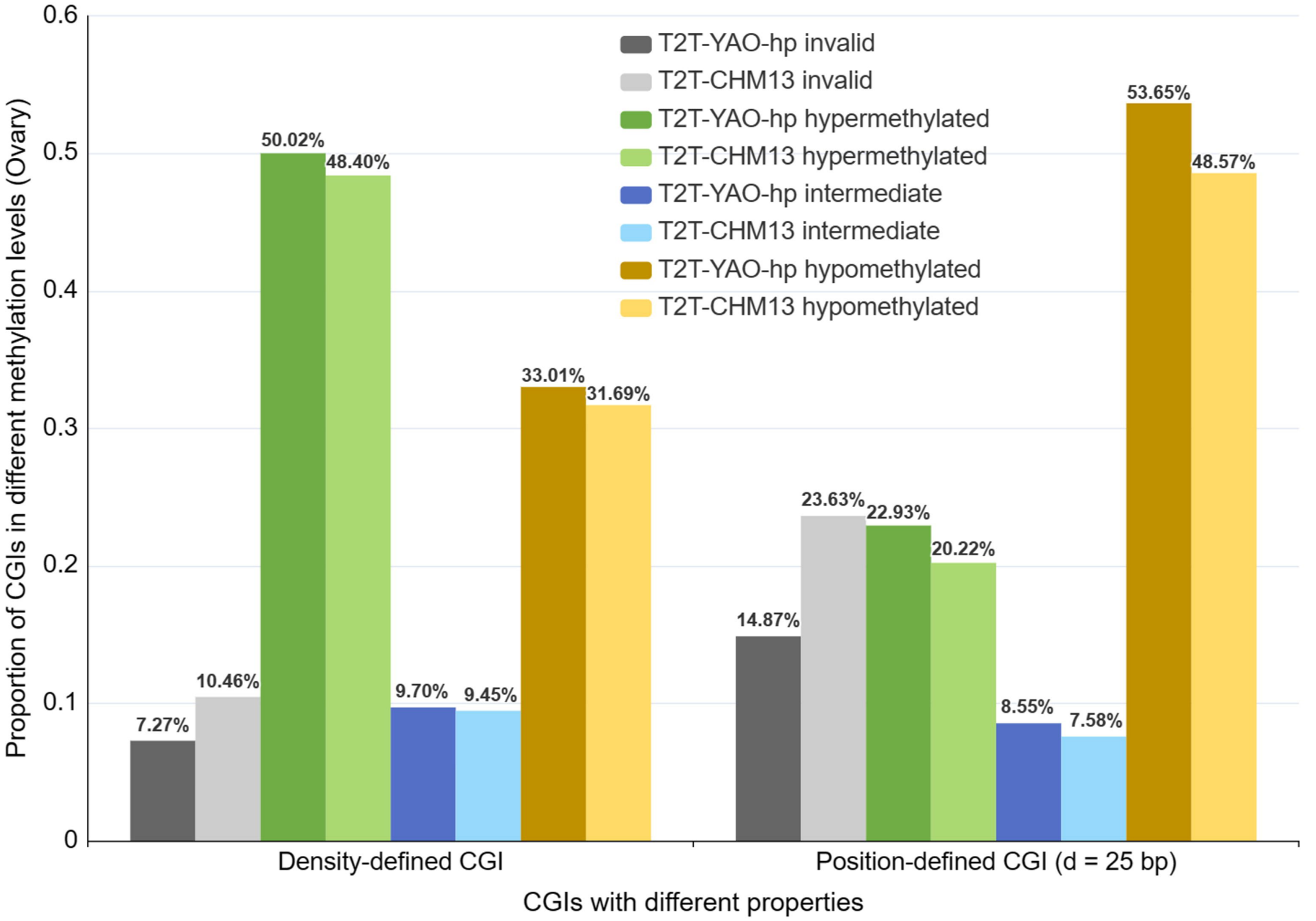

T2T-CHM13
27,333,743
(76.5%)
8,394,427
1,664,339
(16.5%)
Position-defined
CGI (d = 25 bp)
Density-defined
CGI


T2T-YAO-hp
24,991,625
(75.6%)
8,062,542
1,735,619
(17.7%)
Position-defined
CGI (d = 25 bp)
Density-defined
CGI


A
B
C
D
E
F
G
H
Density-defined T2T-YAO-hp
Density-defined T2T-CHM13
Position-defined (d = 25 bp) T2T-YAO-hp
Position-defined (d = 25 bp) T2T-CHM13
Frequency
CGI length (bp)
GC content
O/E ratio
CpG density


4.0
3.0
2.0
1.0
0.0
Number of CpGs in different methylation levels (×10⁶)
2.16
3.49
0.28
0.22
0.15
0.13
0.044
0.031
2.19
3.72
0.31
0.25
0.26
0.23
0.082
0.079
Ovary
Testis
Organs
T2T-YAO-hp invalid
T2T-CHM13 invalid
T2T-YAO-hp hypermethylated
T2T-CHM13 hypermethylated
T2T-YAO-hp intermediate
T2T-CHM13 intermediate
T2T-YAO-hp hypomethylated
T2T-CHM13 hypomethylated

A


0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion of CGIs in different methylation levels (Ovary)
T2T-YAO-hp invalid
T2T-CHM13 invalid
T2T-YAO-hp hypermethylated
T2T-CHM13 hypermethylated
T2T-YAO-hp intermediate
T2T-CHM13 intermediate
T2T-YAO-hp hypomethylated
T2T-CHM13 hypomethylated
7.27%
10.46%
50.02%
48.40%
9.70%
9.45%
33.01%
31.69%
14.87%
23.63%
22.93%
20.22%
8.55%
7.58%
53.65%
48.57%
Density-defined CGI
Position-defined CGI (d = 25 bp)
CGIs with different properties

B

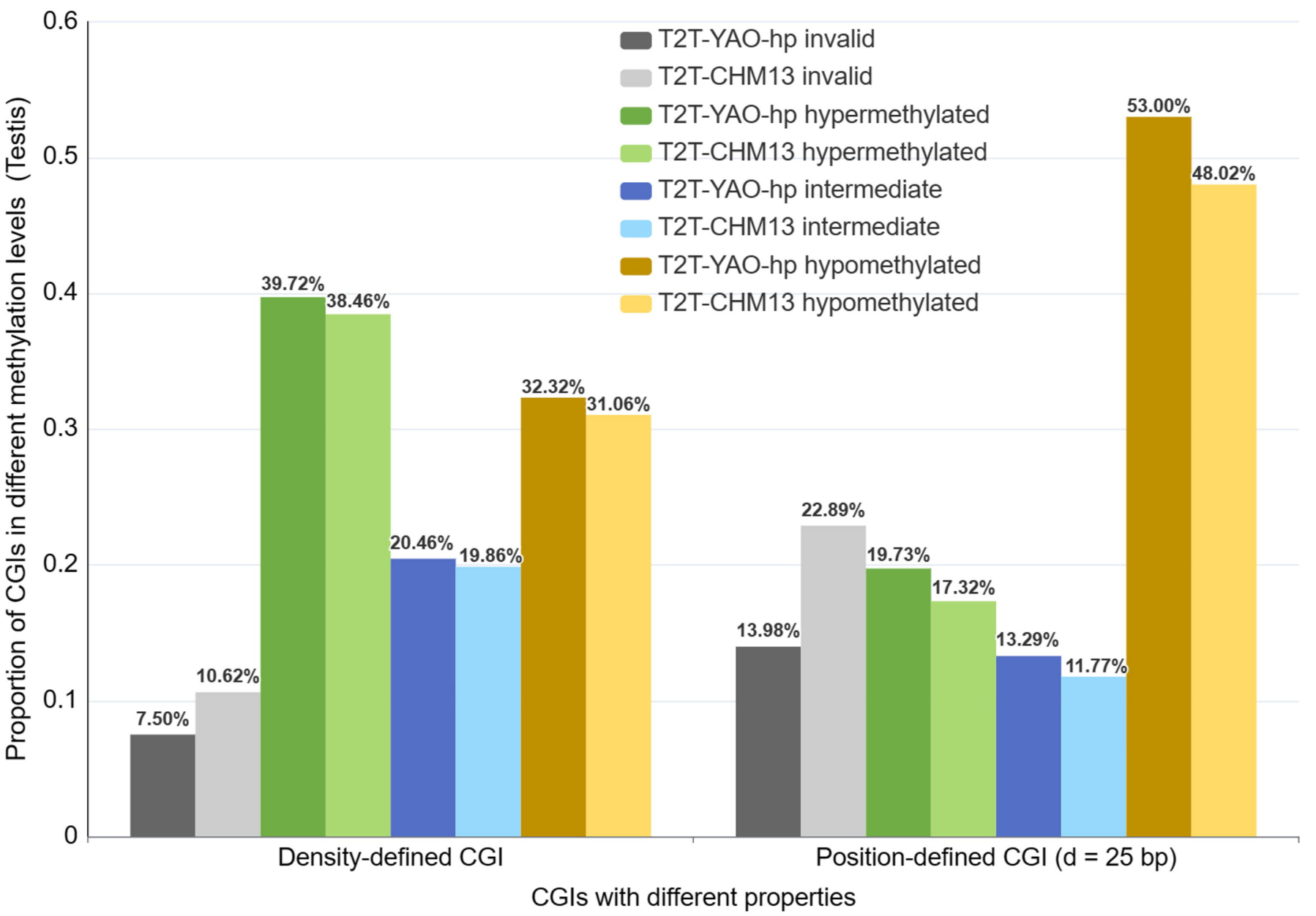

0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion of CGIs in different methylation levels (Testis)
T2T-YAO-hp invalid
T2T-CHM13 invalid
T2T-YAO-hp hypermethylated
T2T-CHM13 hypermethylated
T2T-YAO-hp intermediate
T2T-CHM13 intermediate
T2T-YAO-hp hypomethylated
T2T-CHM13 hypomethylated
7.50%
10.62%
39.72%
38.46%
20.46%
19.86%
32.32%
31.06%
13.98%
22.89%
19.73%
17.32%
13.29%
11.77%
53.00%
48.02%
Density-defined CGI
Position-defined CGI (d = 25 bp)
CGIs with different properties

Table 1 Sequence differences between the T2T-CHM13 and T2T-YAO-hp genomes

	Total length (bp)	Divergent sequence length (bp)	Divergent sequence proportion	Similar sequence length (bp)
T2T-CHM13	3,117,275,501	179,281,068	5.751%	2,937,994,433
T2T-YAO-hp	3,062,707,975	142,866,485	4.665%	2,919,841,490

Note: We use nucmer to compute similar sequences between the two reference genomes and call other remaining sequences as divergent sequences.

Table 2 CGIs Statistics of the T2T-CHM13 and T2T-YAO-hp genomes

Category		Density-defined CGI	Position-defined CGI		
			d = 50 bp	d = 25 bp	d = 12 bp
T2T-CHM13	CGI count	87,788	220,736	174,594	101,972
	Average CGI length (bp)	407	258	58	23
T2T-YAO-hp	CGI count	84,471	222,421	161,997	96,449
	Average CGI length (bp)	391	261	60	24

Note: d denotes the maximum distance between neighboring CpGs in the position-defined CGIs.