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] and Le Zhang[1,2, *]

[1] College of Computer Science, Sichuan University, Chengdu 610065, PR 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, PR China.
[3] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, PR China.
[4] University of Chinese Academy of Sciences, Beijing 100049, PR China.

*To whom correspondence should be addressed.
[#]The authors wish it to be known that these authors contributed equally.

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

Running title: *Comparative analysis of T2T-YAO and T2T-CHM13*

The total number of words in this manuscript is 6641, including 7 figures and 2 table.

Supplemental Information includes 8 figures, 1 table and 5 dataset files.

Abstract

Precise definition of every primary methylation sites and their secondary structural feature – CpG islands – are pivotal for subsequent genome-wide epigenetic studies of Chinese populations under different physiological and pathological conditions while the reference genome T2T-YAO is ready for use timely. We first align and examine the two high-quality genome references T2T-YAO and T2T-CHM13 from different ethnic backgrounds in a base-by-base fashion, computing their density-defined and position-defined CGIs. Second, we not only map some representative genome-wide methylation data of selective organs onto the T2T-YAO and T2T-CHM13 genomes but also scrutinize the difference of between these epigenetic parameters. Our major findings are: (1) There is about 4.7% (4.665%–5.751%) sequence difference between the two genomes, and the divergent sequences and related genes are closely correlated to neurological functions. (2) CGIs associated with the divergent sequences are statistically significantly different with respect to CpG density and O/E values. (3) Not only whole-genome-bisulfite-sequencing or WGBS data for T2T-YAO have greater coverage of CpG sites as compared to those of T2T-CHM13, but also T2T-YAO shows more hyper-methylated CpG sites than T2T-CHM13 does.

Keywords: T2T-YAO, Position-defined CGIs, DNA methylation, Genome analysis.

1. Introduction

Since the position-sensitive and co-methylated CpG island (CGI) is one of the key epigenetic mechanisms for gene expression regulation and chromosome integrity [1, 2], it becomes an important research direction for epigenomics [3-7] to accurately map the precise location of all CGIs and their methylation status onto the human genome.

Human reference genomes provide single-base-resolution coordinates for genome-wide genetic and epigenetic mapping with an ultimate precision [8].The release of the Telomere-to-Telomere CHM13 (T2T-CHM13) genome in 2022 has not only filled the previous 8% gap in the human genome [9], but also provides one of such benchmarks. Therefore, after the recent assembly of a Chinese reference genome T2T-YAO, we have attempted to detail some of the obvious inter-population deviations between T2T-YAO and T2T-CHM13 [10]. Here, we raise two questions; one is whether the two reference genomes are significantly different in epigenetic parameters, and the other whether T2T-YAO is more appropriate for the Chinese population. The latter is of importance for subsequent genome-wide epigenetic studies based on Chinese populations under different physiological and pathological conditions. We choose to accurately predict genome-wide CGIs for both references and validate their applications in methylation studies by testing them on a limited dataset.

According to the divergent sequences of the two reference genomes, Aganezov et al. [9] compared the sequence differences between the classical GRCH38 genome and T2T-CHM13 in detail, which found that the T2T-CHM13 genome have introduced nearly 200 million base pairs (Mbp) of sequence. Since Chinese T2T precision genome T2T-YAO has been assembled, and the sequence differences between T2T-YAO and T2T-CHM13 may represent the epigenetic characteristics for Chinese population, exploring the divergent sequences between T2T-YAO and T2T-CHM13 genomes and whether the divergent sequences are associated with specific biological

functions, became our first scientific questions.

According to the genome-wide CGIs prediction for the two reference genomes, our previous studies indicated that distance-based clustering method can localize more short CGIs (position-defined CGIs) for potential gene regulatory functions and are closely correlated with gene expression specificity [2, 11, 12]. Since we neither investigate the position-defined CGIs for T2T-CHM13, nor sure whether there is a difference between the CGI features of T2T-YAO and the T2T-CHM13, what is the position-defining CGI distribution of the T2T-YAO and T2T-CHM13 genomes, and whether there is any specificity between these two genomes according to CGI features, become our second scientific question.

According to the CGI methylation specificity investigation for these two reference genomes, several relevant studies have been carried out [2, 8, 11, 12]. For example, our previous studies already analyzed the relationship between position-defined CGIs associated with core promoters and the methylation level of the human genome [2, 11, 12]. Also, Gershman et al. [8] constructed a complete human genome methylation profiles using nanopore methylation data based on T2T-CHM13, and compared the methylation profiles with the methylation profiles using whole genome bisulfite sequencing (WGBS) data. However, since the differences in methylation patterns between the Chinese genome and T2T-CHM13, especially those CpG sites and CGIs associated with the divergent sequences, have not been well studied, using public genome-wide bisulfite sequencing data to analyze whether T2T-YAO and T2T-CHM13 are different at the CpG and CGI methylation levels become our third scientific question.

To address the above scientific questions, as shown in Figure 1, we firstly compared the difference of sequence between T2T-YAO and T2T-CHM13, and then analyzed the genes associated with the divergent sequences of these two reference genomes. Secondly, we computed the density-defined and position-defined CGIs of T2T-YAO and T2T-CHM13, and then compared the differences of their CGI features.

Finally, we not only map the genome-wide methylation WGBS data with representative organs from authoritative databases onto T2T-YAO and T2T-CHM13 genome respectively, but also analyze the differences from the methylation level of CpG sites and CGIs between these two reference genomes.

Figure 1. Workflow of the study.

The following is the major findings of the study:

1) There is about 4.5-5.5% (4.665%-5.751%) difference between the reference sequences of Chinese T2T-YAO and T2T-CHM13 genomes, and the divergent sequences related genes are closely correlated with neurological functions by GO enrichment analysis [13-18].

2) CGIs associated with Chinese T2T-YAO and T2T-CHM13 divergent sequences are statistically significantly different with respect to CpG density and O/E values [11, 19].

3) Not only WGBS data of Chinese T2T-YAO have greater coverage of CpG sites than theT2T-CHM13, but also T2T-YAO shows more hyper-methylated CpG sites than T2T-CHM13.

In general, our study demonstrated the statistically significant differences between T2T-YAO and T2T-CHM13 in genome sequence and epigenetic aspects, such as CGI and methylation patterns, which suggests that the establishment of CGI and methylation profiles based on Chinese T2T reference genome is crucial for the subsequent genome-wide epigenetic studies for the Chinese population.

2. RESULTS

2.1 Sequence comparative analysis results

To answer our first scientific question, we firstly calculated sequence differences between T2T-YAO-hp and T2T-CHM13, and then analyzed the categories and functional enrichment of divergent sequences related genes of T2T-YAO-hp.

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

Firstly, **Table 1** identifies the similar and divergent sequences between the two genomes by sequence alignment (the alignment results are detailed in Supplementary Data S1 and S2), as well as computes the proportion of divergent sequences for both T2T-YAO-hp and T2T-CHM13 genome.

Table 1. Sequence differences between T2T- CHM13 and T2T- YAO-hp

Table 1 demonstrates that T2T-CHM13 is about 54.6 mega-base pairs (Mbp) longer than T2T-YAO-hp. Meanwhile, the divergent sequences account for 5.751% of the full length for T2T-CHM13, whereas the divergent sequences account for 4.665% of the full length for T2T-YAO-hp. Thus, the difference of the reference sequences between Chinese T2T-YAO-hp and T2T-CHM13 genomes is about 4.5-5.5%.

Secondly, **Figure 2A** describes the mapping of similar sequences between Chinese T2T-YAO-hp and T2T-CHM13 genomes by using NGenomeSyn [20] to make a collinearity map. We observed that most of the T2T-CHM13 sequences are very similar to the sequences of the corresponding chromosomes of T2T-YAO, which is consistent with the expected chromosome correspondence. However, **Figure 2A** also shows some T2T-CHM13 sequences are very similar to the sequences of the different chromosomes of T2T-YAO. For example, chromosomes 21 and 22 of T2T-CHM13 have similar sequences to chromosomes 13, 14, and 15 of T2T-YAO-hp. The parts of sequences of chromosome 22 of T2T-YAO-hp are as same as that of chromosome 14 of T2T-CHM13.

Finally, **Figure 2B** visualizes the sequence comparison across chromosomes and **Figure 2C** counts the comparison of the length of the divergent sequences for different chromosomes.

As shown in **Figure 2B**, the divergent sequences between T2T-CHM13 and T2T-YAO genomes are mainly concentrated in specific regions of each chromosome.

For instance, the divergent sequences of chromosomes 15 and 22 are concentrated in the beginning region of the chromosome, while chromosome 9 is concentrated in the middle region of the chromosome.

The longest divergent sequence lengths were found on chromosomes 9 and Y. And the length of divergent sequence for T2T-CHM13 is longer than the T2T-YAO for chromosomes 9 and Y (3.31 vs 1.69 × 10^7 bp for chromosome 9, 2.15 vs 1.31 × 10^7 bp for chromosome Y).

Figure 2. Sequence alignment between T2T-CHM13 and T2T-YAO.

2.1.2 Divergent sequences related genes

Firstly, we calculated the divergent sequence related genes for the two reference genomes, and the specific gene are listed by Supplementary Data S3. By categorizing these divergent sequences related genes, **Figure 3A** indicates that most divergent sequence related genes are protein-coding genes accounting for 44.26% and IncRNA accounting for 25.93%.

Furthermore, we did GO enrichment analysis for protein-coding genes in the divergent sequences related genes. **Figure 3B** not only shows that the most significant terms were " homophilic cell adhesion via plasma membrane adhesion molecules", "ion channel complex", and "gated channel activity" in the biological process (BP), cellular component (CC), and molecular function (MF) [21], but also it demonstrates that most of the GO terms of divergent sequence related protein-coding genes are associated with the function of the nervous system.

Figure 3. Divergent sequences related genes analysis.

2.2 CGI prediction and analysis

To answer our second scientific question, we firstly carried out the statistical

analysis for the CGI prediction for these two genomes, and then we compared the CGI features between these two genomes.

2.2.1 CGI prediction results

Firstly, we employed rule-based method [12, 19] and distance-based clustering method [12, 19] to predict the number of CGIs and the average CGI length for T2T-CHM13 and T2T-YAO-hp, respectively, which are detailed by **Table 2** and Supplementary Data S4. We also predicted all CGIs related with divergent sequences of the two genomes, which are detailed by Supplementary Data S5.

Table 2 indicates that the number of density-defined CGIs and the average length predicted by T2T-YAO-hp is slightly smaller (84471 vs 87788) and shorter (391 vs 407 bp) than T2T- CHM13; the number of position-defined CGI predicted by position-defined CGIs for T2T-YAO-hp is greater than the T2T-CHM13 under the default parameter d=50 (220,736 vs. 222,421); the CGI number predicted by T2T-YAO-hp is less than the T2T-CHM13 not matter with density-defined CGIs or position-defined CGIs, but the average length of the position-defined CGIs predicted by T2T-YAO-hp is relatively longer than T2T-CHM13 (e.g., 60 vs. 58 bp with d=25) under the parameter d=25 and 12.

Table 2. Statistics of CGIs for T2T-CHM13 and T2T-YAO-hp

Secondly, to study the relationship between position-defined CGIs and density -defined CGIs for the same genome, we investigated the sequence intersection of density-defined CGIs and position-defined CGIs (d = 25bp) for both genomes.

Figure 4 lists the two consistence between T2T-YAO-hp and T2T- CHM13. One is that the total length of density-defined CGIs is much longer, approximately three times, than that of position-defined CGIs for each genome (e.g., (24,991,625+8,062,542) / (8,062,542 + 1,735,619) =3.37 for T2T-YAO-hp); The other is although most of the position-defined CGIs overlap with the density-defined CGIs,

there is still a fraction (16.5% and 17.7%, respectively) that is not predicted by the rule-based CGI predictive methods.

Figure 4. Sequence intersection of density-defined CGIs with position-defined CGIs (d = 25bp).

2.2.2 CGI features comparison

To compare the differences of CGIs between T2T-CHM13 and T2T-YAO-hp, we calculated four commonly used CGI features, which are CGI length, GC-content, CpG Density, and CGI Observed/Expected ratios [2, 22], and then **Figure 5** presents their distribution for all CGIs (**Figure 5A-5D**) and the CGIs related to the divergent sequences (**Figure 5E-5H**) for two reference genomes. In addition, we showed the features distribution of genome-wide and divergent sequence related position-defined CGIs (including d=12bp, 50bp) in Supplementary **Figure S1**.

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

As shown in **Figure 5**, firstly, no matter density-defined or position-defined CGI, there is no obvious difference for four commonly used CGI features between T2T-CHM13 and T2T-YAO-hp in terms of all CGIs (**Figure 5A-5D**). For example, **Figure 5A** indicates that the peak of length of density-defined CGIs for both reference genomes is around 210 bp, and the peak of length of position-defined CGIs is 30-50 bp.

Secondly, there is a greater difference for four commonly used CGI features of the divergent sequences CGIs of T2T-CHM13 and T2T-YAO-hp than all CGIs, (**Figure 5 E-5H**). For example, **Figure 5F** indicates that the density-defined CGI of T2T-CHM13 (light gray line) has two peaks in the GC content range from 50% to 60%, while the density-defined CGI of T2T-YAO-hp (dark gray line) has only one

peak in this range. Also as shown in **Figure 5E-5H** and Supplementary **Figure S1**, the position-defined CGIs of the divergent sequences, except for a small difference in the distribution of CGI lengths, have significantly statistically different distributions for GC content, O/E ratios, and CpGDensity (Supplementary **Table S1**).

Finally, **Figure 5** indicates that no matter for all CGIs or divergent sequences CGIs, distance-based clustering method is more likely to find such CGIs that have shorter length, but greater GC content, O/E values, and CpGDensity than rule-based CGI predictive method.

2.3 Methylation analysis and comparison

To answer our third scientific question, we not only compared the number of cytosine sites (C sites, Supplementary **Figure S2**) and the methylation levels of all CpG sites, but also the methylation levels of genome-wide CGIs for the two reference genomes.

2.3.1 CpG methylation level comparison

We compare the difference at different CpG methylation level for similar sequences (Supplementary **Figure S3**) and divergent sequences (**Figure 6**) between T2T-CHM13 and T2T-YAO-hp, **Figure 6** maps the WGBS data of two representative tissues (ovary and testis) onto the divergent sequences of the two genomes, respectively. In addition, based on the WGBS data of these two tissues, we also performed the methylation profiles of whole genome and the divergent sequences of T2T-YAO-hp and T2T-CHM13 which shows in Supplementary **Figure S4.**

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

Figure 6 demonstrates that the number of divergent sequences CpG sites that failed to be mapped by the same WGBS methylation data is much less for

T2T-YAO-hp than the T2T-CHM13 in both tissue Ovary and testis. For example, the number of divergent sequences CpG sites for Ovary that failed to be mapped to T2T-YAO-hp is 2.16×10^6, whereas the T2T-CHM13 is 3.49×10^6 (**Figure 6**). The number of divergent sequences CpG sites that can be mapped for T2T-YAO-hp is greater than the T2T-CHM13 at all three methylation levels. And in terms of proportions, the most notable difference is in hyper-methylation (10.59% vs. 5.77% for Ovary, 10.99% vs. 5.89% for Testis).

2.3.2 CGI methylation level comparison

We used Eq. 6 to calculate the methylation level of all density-defined CGIs and position-defined CGIs for the two reference genomes (**Figure 7** and Supplementary **Figure S5**).

Figure 7. CGI methylation level of two reference genomes.

Firstly, **Figure 7** turns out that the percentage of CGIs that failed to be mapped by WGBS data of T2T-CHM13 for all CpGs (invalid) is greater the T2T-YAO-hp for both density-defined CGI or position-defined CGI methods. For example, there are 10.46% vs 7.27% for density-defined CGI, and 23.63% vs 14.87% for position-defined CGI in Ovary (**Figure 7A**)。

Secondly, the proportions of hyper-methylated, intermediate methylated, and hypo-methylated of CGIs, predicted by density-defined CGIs and position-defined CGIs, are slightly greater for T2T-YAO-hp than the T2T-CHM13. For example, **Figure 7B** describes that the proportions of hyper-methylated, intermediate methylated, and hypo-methylated for T2T-YAO-hp predicted by position-defined CGI are respectively 19.73%, 13.29%, and 53%, whereas the corresponding proportions for T2T-YAO-CHM13 are 17.32%, 11.77%, and 48.02%, respectively.

Finally, **Figure 7** shows that density-defined CGIs have the greatest proportion of hyper-methylation for both T2T-YAO-hp and T2T-CHM13, while position-defined

CGIs highlight the proportion of hypo-methylation. For example, **Figure 7A** shows that the density-defined CGIs of both two reference genomes have a higher proportion of hyper-methylated (50.02% and 48.40%, respectively) than that of intermediate methylated and hypo-methylated, whereas the position-defined CGIs have the greatest proportion of hypo-methylation, which are 53.65% and 48.57% for T2T-YAO-hp and T2T-CHM13, respectively.

3. Discussion and conclusion

This research aims to compare the differences of sequence, CGI, and methylation patterns between two reference genomes, T2T-YAO and T2T-CHM13. And then, we can further determine the necessity of using T2T-YAO as a reference for future genome-wide epigenetic studies for Chinese populations.

For this purpose, we firstly compared the similar and divergent sequences of the high-quality T2T-YAO-hp haploid reference genome with T2T-CHM13 genome, respectively. And then, we analyzed the categories and functions of the divergent sequences related genes. Secondly, we computed position-defined CGIs and density-defined CGIs for two reference genomes and compared the differences for four commonly used CGI features between these two genomes. Finally, we compared the methylation level of CpGs and CGIs for these two genomes by public WGBS data.

In terms of sequence divergence of the two reference genomes, firstly, sequence alignment indicated that there is about 4.5-5.5% (4.665%-5.751%) difference between the reference sequences of Chinese T2T-YAO and T2T-CHM13, which further validated the great interracial genomic difference since T2T-CHM13 cells are from Europe [9, 10]. Secondly, although a few sequences of certain chromosomes appeared to be like those of other chromosomes possibly due to mutations (**Figure 2A**), most of the similar sequences are consistent with the expected chromosome correspondence. And **Figure 2B** turns out that most of the differential sequences of the two genomes

are mainly concentrated on a few chromosomes, such as specific regions of chr9 and chrY (**Figure 2C**), which implies that different races represented by these two genomes have retained ancestral genetic traits during their evolution. Meanwhile, the GO enrichment analysis demonstrated (**Figure 3B**) that divergent sequences related genes of T2T-YAO show a strong preference for several specific molecular mechanisms of biological functions of neuronal cells and synapses in the nervous system, which provides a series of clear information for the future study.

In terms of genome-wide CGI differences between the two reference genomes, firstly, since **Table 2** indicates that distance-based clustering method with default d = 50bp (Eq. 4) predicts opposite trends in the number and average length of CGIs for two reference genomes compared to the other three methods, it is necessary for us to optimize the parameter d (Eq. 4) for different genomes. Secondly, when we compare the density-defined CGI with position-defined CGI (d=25), **Figure 4** shows that density-defined CGI may not be able to predict such CGIs that are shorter but have potential regulatory functions. Third, we compare four commonly used CGI features for two reference genomes (**Figure 5**). **Figure 5A** turns out that the length distributions of all CGIs for two genomes are very similar, whereas the GC content, O/E value and CpGDensity predicted by distance-based clustering method shows less sensitivity than the rule-based method, which suggest that highly homologous T2T-YAO and T2T-CHM13 have certain diversity for the shorter position-defined CGI. Also, we found the position-defined CGIs related to the divergent sequences of the two reference genomes demonstrate more obvious differences of features rather than all CGIs (**Figure 5E-5H**), and since the genes related to the divergent sequences show strong preference for several specific gene functions (**Figure 3B**), it is worthwhile to further investigate whether these divergent sequence related CGIs are associated with the regulation of corresponding gene functions.

In terms of methylation pattern specificity of the two reference genomes, we firstly found that the number of CpG methylation sites of T2T-YAO that can be

mapped onto the divergent sequence of two reference genomics is greater than the T2T-CHM13 for the same public WGBS data (**Figure 6**). In addition, the methylation level of the CpG sites of the divergent sequences shows a prominent proportion of hyper-methylation (**Figure 6**), which not only indicates that the methylation patterns of two reference genomics are quite different, but also further demonstrates that T2T-YAO sequenced by nanopore has greater sequencing accuracy and assembly precision than the T2T-CHM13 [10]. Secondly, by comparing the methylation level (Eq.6) of all CGIs for two reference genomes, **Figure 7** shows that the methylation level predicted by position-defined CGIs has greater difference than the density-defined CGIs due to the high sensitivity of position-defined CGIs. Also, we found that the hypo-methylated CGIs has the highest proportions predicted by position-defined CGIs for two genomes, which are completely different from the density-defined CGIs with great hyper-methylated proportions. This not only suggests that CGI methylation patterns are very different for the two reference genomes, but we also urgently need a more precise and comprehensive annotation for all CpG sites and CGIs in the human genome.

In general, our study demonstrated that T2T-YAO and T2T-CHM13 have statistically significant differences (**Figure 6, Figure 7** and Supplementary **Table S1**) in genome sequence and epigenetic aspects such as CpG and CGI methylation patterns, which suggests that it is important for the subsequent genome-wide epigenetic studies of Chinese population to establish CGI and methylation profiles according to Chinese T2T reference genome.

However, there are still several questions that need further exploring. Firstly, it is crucial for genome-wide epigenetic regulation study to know how to use long-read sequencing technology to sequence genome-wide methylation data for more people and construct complete human methylation atlas of different populations. Secondly, since artificial intelligence (AI) has been gradually used in genome-wide methylation-related studies [23, 24], how to use AI techniques such as neural

networks to investigate the relationship between sequence features, CGI, methylation, and expression specificity of T2T-YAO becomes one of the future research directions. Lastly, genome-wide CGI prediction and methylation analysis are not only data-intensive, but also time-consuming. However, there is currently no such a web service that collects and manages the CGI and methylation annotation data for Chinese T2T genome. Therefore, we will build up a web service for online computation, data sharing and visualization based on high-performance computing such as GPU acceleration and distributed computing to provide a data foundation and an online platform for the further study.

4. Materials and methods

4.1 Data source

According to the divergent sequences of the two reference genomes, we not only downloaded the sequences of the classical human genomes GRCh38 and T2T-CHM13 (v2.0) from the NCBI database [25], but also obtained the Chinese T2T-YAO-hp genome sequenced by He et al. [10] using Nanopore sequencing and the corresponding genome annotation data in GFF3 format.

According to genome-wide CGIs prediction for both genomes, we used the T2T-CHM13 genomic CGI data, which are predicted by the Gardiner-Garden and Frommer (GGF) [26] standard in UCSC [27].

According to the CGI methylation specificity investigation for two reference genomes, we downloaded the WGBS (whole genome bisulfite sequencing) data for two representative tissues, which are ovary (ENCSR417YFD) and testis (ENCSR806NNG) from the ENCODE database [28].

4.2 Methods

Figure 1 describes the workflow of the study with three main aspects, which are sequence comparative analysis, CGI detection and analysis, and CpG methylation

analysis. Next, we will introduce their key algorithms.

4.2.1 Sequence comparative analysis

（1）Statistics of all cytosine C sites

We counted and compared the number and distribution of cytosine sites (C sites) that can be methylated in T2T-YAO-hp, T2T-CHM13, and GRCh38 genomes. Since majority of DNA methylation in mammals is found in CpG dinucleotides [29], we categorized C sites into three different contexts: CpG, CHG, and CHH, where H represents A, C or T.

（2）Genome alignment and analysis

Supplementary **Figure S6** details the process of genome alignment and analysis. Firstly, we used nucmer, a sequence alignment algorithm of MUMmer system [30], to carry out genome-wide alignment and analysis for T2T-YAO-hp and T2T-CHM13 genome. We set the key parameters for nucmer as follows: minimum alignment length is 5000 bp, minimum alignment similarity is 90%. And then, we calculated the similar and divergent sequences for these two genomes.

Secondly, we calculated the genes related to the divergent sequences between T2T-YAO-hp and T2T-CHM13 based on the genome annotation data of T2T-YAO-hp.

Finally, we used the clusterProfiler algorithm [31] to carry out GO enrichment analysis for the protein-coding genes in the divergent sequences related genes.

4.2.2 CGI detection and analysis

Our previous studies denoted those CGIs predicted by classical rule-based method [26] as density-defined CGIs, and those CGIs predicted by distance-based clustering method [22] 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] on the human genome. Therefore, in order to comprehensively compare the CGI features of the two reference genomes, we

calculated not only all the density-defined and position-defined CGIs for T2T-YAO-hp and T2T-CHM13, but also the density-defined and position-defined CGIs related to the divergent sequences of two reference genomes, respectively (Supplementary **Figure S7**).

(1) Density-defined CGI

Based on the rule of CGI [26], we computed these sequence segments that meet Eq.1-3 as density-defined CGIs by CpGPlot algorithm [32].

$$\text{length} \geq 200\text{bp}\#(1)$$

$$\text{GC content} \geq 50\%\#(2)$$

$$\frac{\text{ObsCpG}}{\text{ExpCpG}} = \frac{Number\ of\ CpG}{Number\ of\ C\ \times\ Number\ of\ G} \times \text{length} \geq 0.6\#(3)$$

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

(2) Position-defined CGI

We firstly calculated the CpG clusters in the genome by Eq. 4, and then consider these CpG clusters with small p-values [22] (Eq. 5) as position-defined CGIs.

$$d_i = x_{i+1} - x_i - 1\#(4)$$

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

where x represents the coordinates of CpG and i represents the index of CpG. P(d) represents the probability to find a distance d between neighboring CpGs. p corresponds to the probability of CpGs in the sequence. It is noted 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 and the length of the shortest LAUPs of multi-species ranges from 10 to 14 [19], and the length of 25bp 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 of 50bp, 25bp, and 12bp (the average of 10 and 14), respectively.

4.2.3 Genome-wide methylation analysis

(1) WGBS data mapping

We applied gemBS [33], a large-scale WGBS data analysis method, to carry out methylation analysis for two representative tissues(ovary and testis) by setting T2T-YAO-hp and T2T-CHM13 as reference genomes, respectively. The details of the procedure are described in Supplementary **Figure S8**.

(2) CGI methylation ratio computing

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

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

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.

(3) Methylation level classification

According to the previous studies [2, 34], Eq. 7 classifies the methylation ratio of each CpG site and CGI into three methylation levels, "Hyper-methylated", "Hypo-methylated", and "Intermediate methylated".

$$\text{Methylation level}(chr, p) = \begin{cases} -1,\ Invalid & no\ reads\ map\ to\ this\ cpg\ or\ cgi \\ 1,\ \text{Hyper} - \text{methylated} & \text{methylation}_{\text{ratio(chr,p)}} > 0.75 \\ 2,\ \text{Hypo} - \text{methylated} & \text{methylation}_{\text{ratio(chr,p)}} < 0.1 \\ 3,\ \text{Intermediate methylated} & \text{otherwise} \end{cases} \#(7)$$

where chr and p represent 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.

Acknowledgements

This work was supported by grants from National Natural Science Foundation of China [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).

References

[1] Pongor CI, Bianco P, Ferenczy G, Kellermayer R, Kellermayer M. Optical Trapping

Nanometry of Hypermethylated CPG-Island DNA. Biophysical Journal 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 Genes2022:319-33.

[3] Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. The 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. Nature Communications 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. Scientific Reports 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. The American Journal of Human Genetics 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.

[11] Zhang L, Dai Z, Yu J, Xiao M. CpG-island-based annotation and analysis of human housekeeping genes. Briefings In 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 Transactions on Computational Biology & Bioinformatics 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 Genomics 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] 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.

[21] 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 Research 2015;43:1057-63.

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

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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics Tig 2000;16:276-7.

[33] 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.

[34] 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.

[35] 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 Science and Technology 2023;28:673-85.

[36] 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.

[37] 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.

[38] You Y, Zhang L, Tao P, Liu S, Chen L. Spatiotemporal Transformer Neural Network for Time-Series Forecasting. Entropy (Basel) 2022;24:1651.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

Figure legends

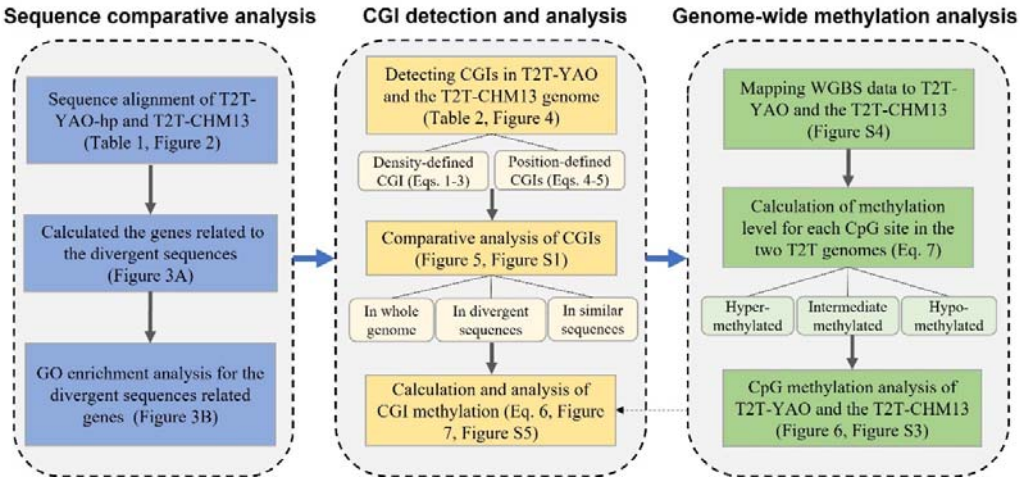

Figure 1. Workflow of the study.

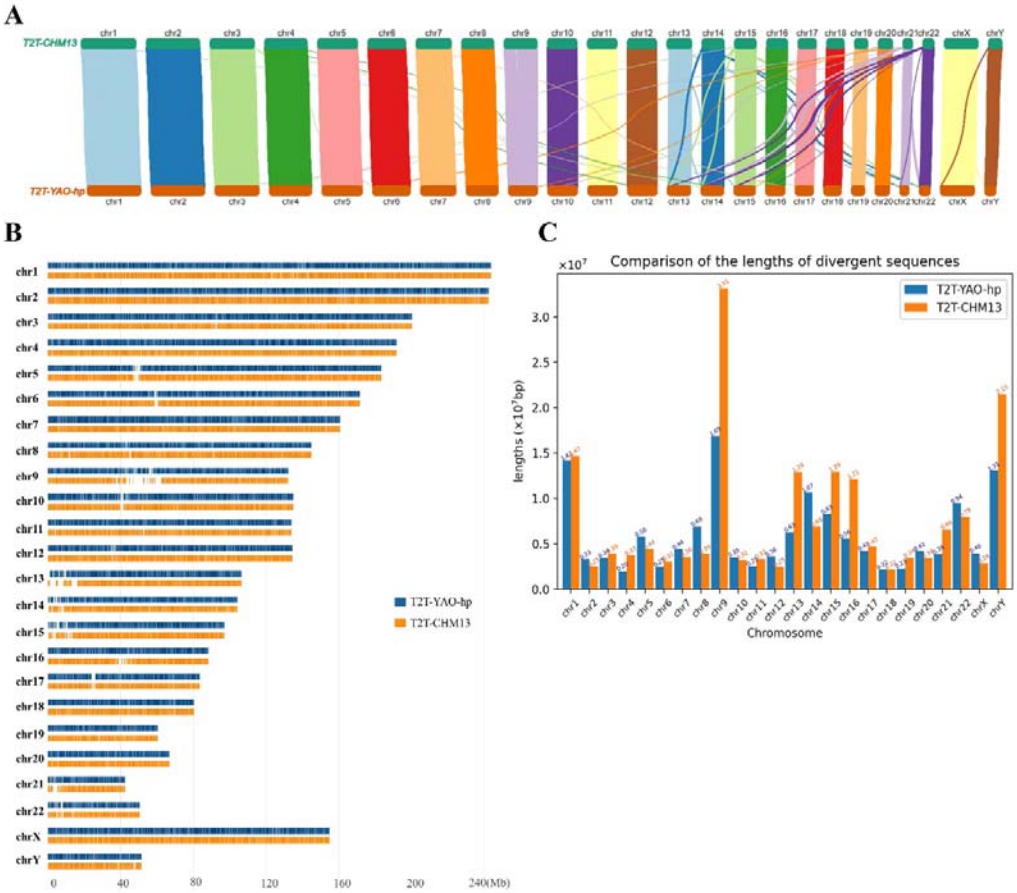

Figure 2. Sequence alignment between T2T-CHM13 and T2T-YAO. (A) Sequence alignment collinearity map. Here, colored links between two genomes indicate genomic fragments with similar sequences, and the thickness of the link represents the similarity between two genomic fragments. (B) Visualization of sequence comparison across chromosomes, where blue and orange sequences represent similar sequences of T2T-CHM13 and T2T-YAO-hp in the current chromosome, respectively, and the blank sections represent the divergent sequences of this reference genome relative to the other. (C) Divergent sequence length statistics of T2T-CHM13 and T2T-YAO-hp genomes, where the blue and orange bars represent the divergent sequence lengths of T2T-YAO-hp and T2T-CHM13 in different chromosomes, respectively.

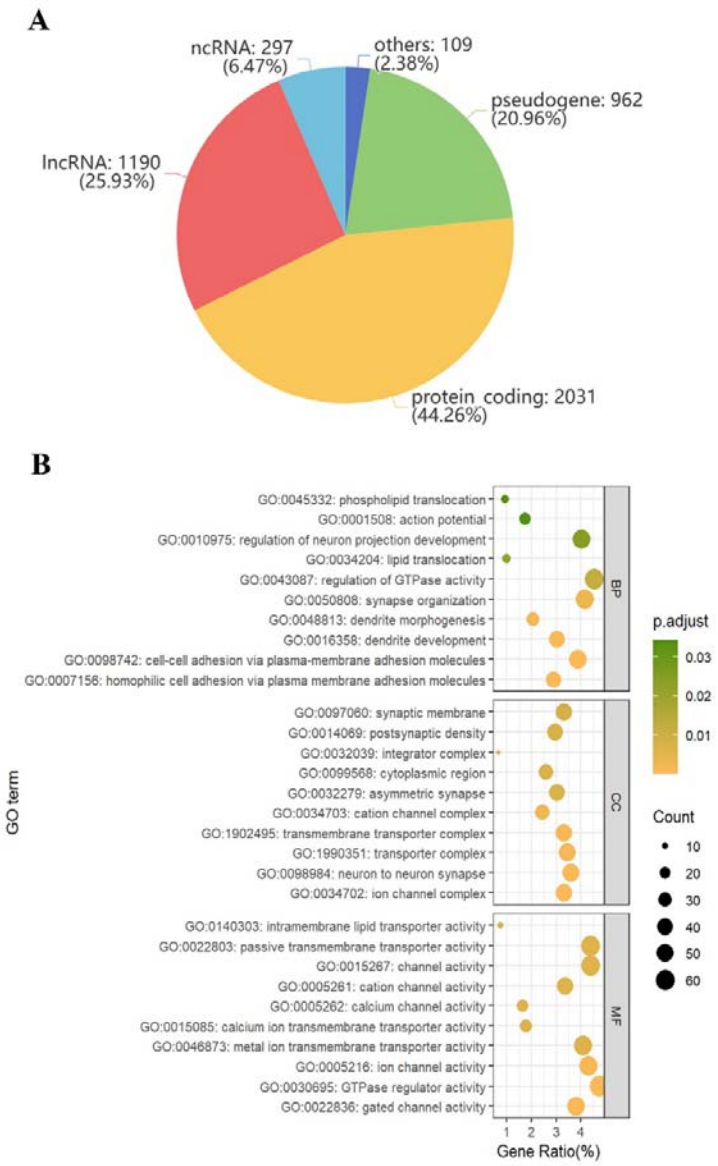

Figure 3. Divergent sequences related genes analysis. (A) Proportion of the types of divergent sequences related genes in these two reference genomes. (B) GO enrichment analysis of divergent sequence related protein-coding genes, where the size and color of the circle represent the number of enriched genes and the statistical significance value (p.adjust) [31, 35-43] of the current GO term, respectively.

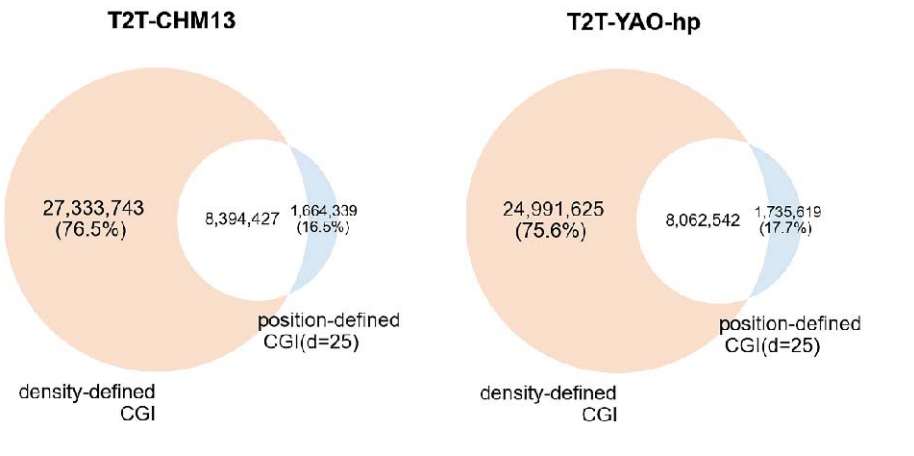

Figure 4. Sequence intersection of density-defined CGIs with position-defined CGIs (d = 25bp). Here, the white parts represent the sequence lengths of the intersection for position-defined CGI and density-defined CGI, while the orange and blue parts represent the sequence lengths unique to density-defined CGI and position-defined CGI, respectively.

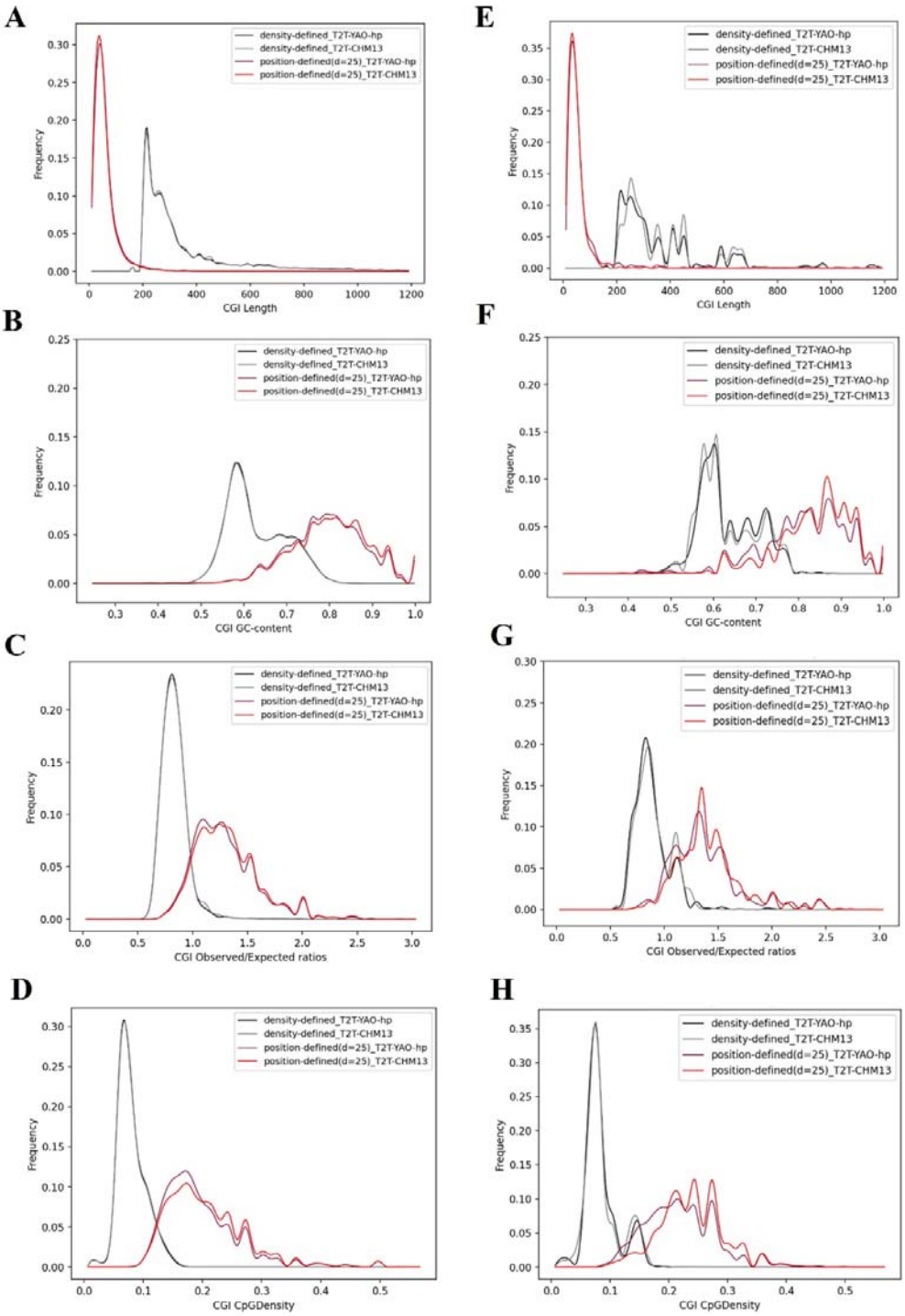

Figure 5. Comparison of CGI features between T2T-CHM13 and T2T-YAO-hp. (A) Length of all CGIs; **(B)** GC content of all CGIs; **(C)** O/E ratios of all CGIs; **(D)** CpGDensity of all CGIs; **(E)** Length of divergent sequences CGIs; **(F)** GC content of divergent sequences CGIs; **(G)** O/E ratios of divergent sequences CGIs; **(H)** CpGDensity of divergent sequences CGIs.

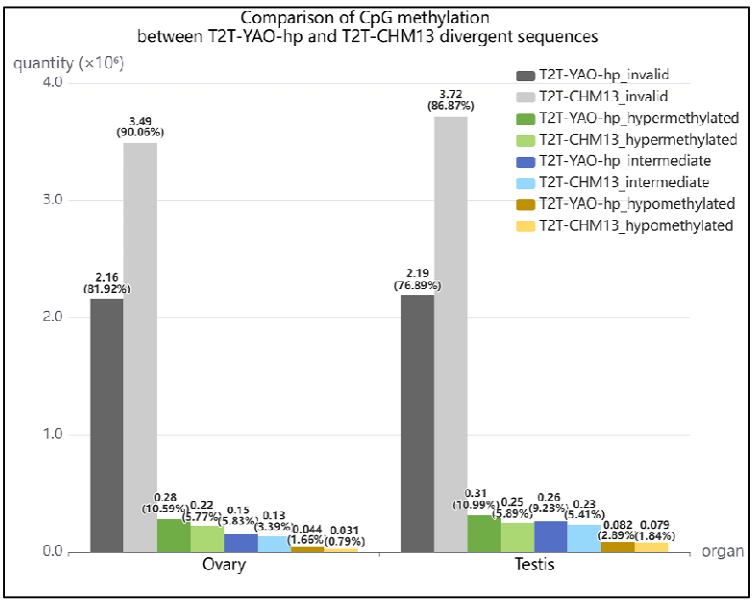

Figure 6. CpG methylation level of divergent sequences between T2T-CHM13 and T2T-YAO-hp. Here, green, blue, and orange represent the number of hyper-methylated, intermediate methylated and hypo-methylated, respectively; gray represents the number of CpG sites that failed to be mapped by WGBS data; dark and light colors are used to distinguish T2T-YAO-hp and T2T-CHM13.

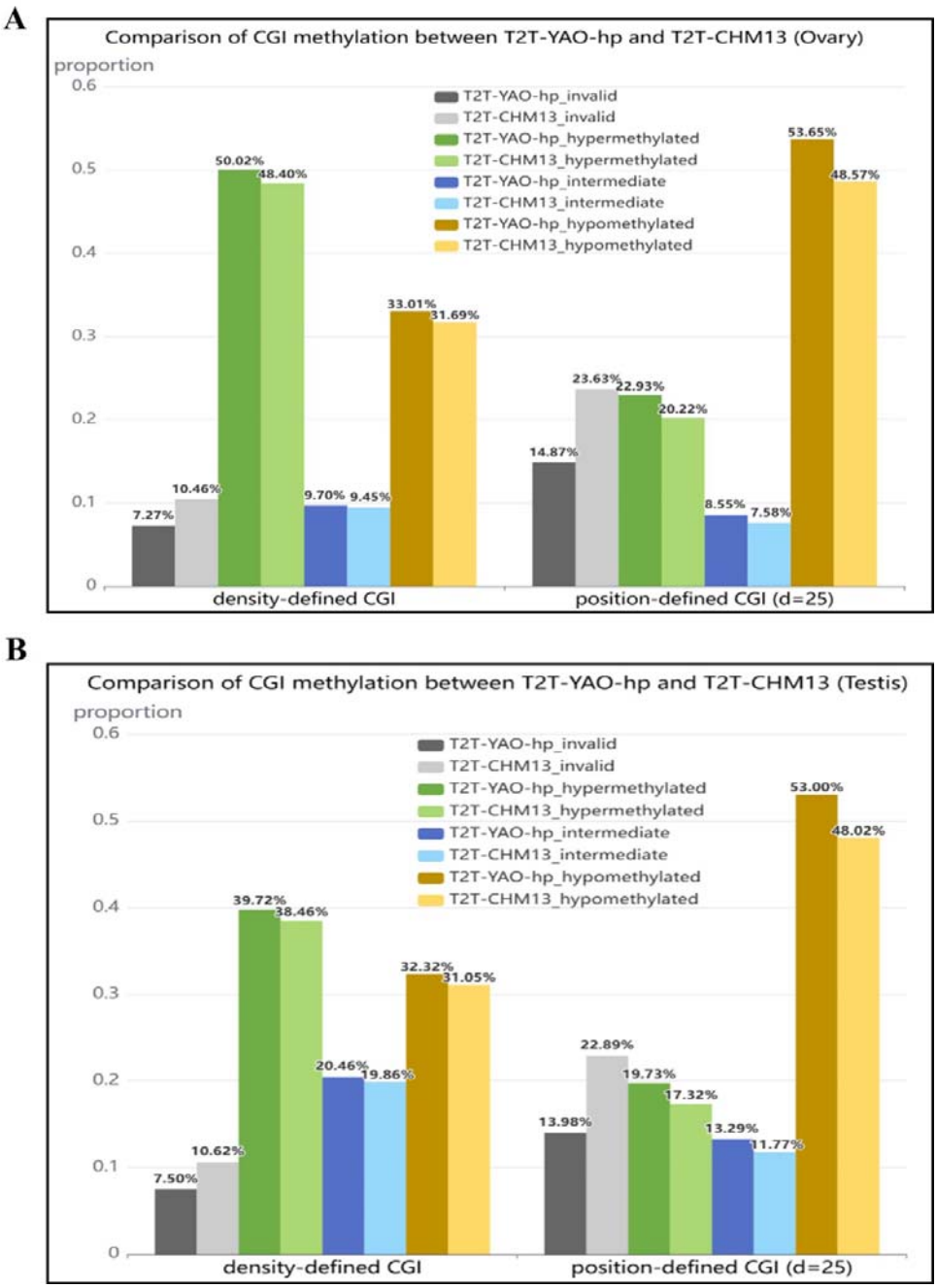

Figure 7. CGI methylation level of two reference genomes. (A) CGI methylation level of two reference genomes using the WGBS data of Ovary. **(B)** CGI methylation level of two reference genomes using the WGBS data of Testis. Here, green, blue, and orange represent the proportion of CGIs that are hyper-methylated, intermediate methylated, and hypo-methylated, respectively. Gray represents the proportion of CGIs where all CpGs failed to be mapped by WGBS data which named "invalid". Darker and lighter colors are used to distinguish the T2T-YAO-hp and T2T-CHM13.

Tables

Table 1. Sequence differences between T2T- CHM13 and T2T- YAO-hp

	Total length	Divergent sequence length	Divergent sequence proportion	Similar sequence length
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

Table 2. Statistics of CGIs for T2T-CHM13 and T2T-YAO-hp

Category		density-defined CGI	position-defined CGI		
			d=50	d=25	d=12
T2T-CHM13	CGI count	87788	220736	174594	101972
	CGI Avg. length(bp)	407	258	58	23
T2T-YAO-hp	CGI count	84471	222421	161997	96449
	CGI Avg. length(bp)	391	261	60	24

Supplementary material

Supplemental Information includes:

Supplementary figures

Figure S1. Comparison of CGI features between T2T-CHM13 and T2T-YAO-hp CGIs (including position-defined with d=12bp, 50bp).

Figure S2. Comparison of C sites between T2T-YAO-hp, T2T-CHM13 and GRCh38.

Figure S3. CpG methylation level of similar sequences between T2T-YAO-hp and T2T-CHM13.

Figure S4. Whole genome and the divergent sequences methylation profiles of T2T-YAO-hp and T2T-CHM13.

Figure S5. Comparison of CGl methylation between T2T-YAO-hp and T2T-CHM13 (including position-defined with d=12bp, 50bp).

Figure S6. Flow chart of genome alignment and analysis.

Figure S7. Whole-genome and divergent sequences CGI detection of T2T-YAO-hp and T2T-CHM13.

Figure S8. Flow chart of WGBS data mapping and calculation.

Supplementary Tables

Table S1. Differential statistics of the distribution of CGI features between the two genomes.

Supplementary data

Data S1. Similar sequences between T2T-YAO-hp and T2T-CHM13

Data S2. Divergent sequences between T2T-YAO-hp and T2T-CHM13.

Data S3. Divergent sequences related genes.

Data S4. All CGIs for T2T-YAO-hp and T2T-CHM13.

Data S5. All divergent sequences related CGIs for T2T-YAO-hp and T2T-CHM13.