Human Genetics (2019) 138:757–769
https://doi.org/10.1007/s00439-019-02032-6

ORIGINAL INVESTIGATION

Misassembly of long reads undermines de novo-assembled ethnicity-specific genomes: validation in a Chinese Han population

Zhibiao Mai[1] · Wanting Liu[1] · Wen Ding[1] · Gong Zhang[1]

Received: 7 November 2018 / Accepted: 21 May 2019 / Published online: 5 June 2019

Abstract

An ethnicity is characterized by genomic fragments, single nucleotide polymorphisms (SNPs), and structural variations specific to it. However, the widely used 'standard human reference genome' GRCh37/38 is based on Caucasians. Therefore, de novo-assembled reference genomes for specific ethnicities would have advantages for genetics and precision medicine applications, especially with the long-read sequencing techniques that facilitate genome assembly. In this study, we assessed the de novo-assembled Chinese Han reference genome HX1 vis-à-vis the standard GRCh38 for improving the quality of assembly and for ethnicity-specific applications. Surprisingly, all genomic sequencing datasets mapped better to GRCh38 than to HX1, even for the datasets of the Chinese Han population. This gap was mainly due to the massive structural misassembly of the HX1 reference genome rather than the SNPs between the ethnicities, and this misassembly could not be corrected by short-read whole-genome sequencing (WGS). For example, HX1 and the other de novo-assembled personal genomes failed to assemble the mitochondrial genome as a contig. We mapped 97.1% of dbSNP, 98.8% of ClinVar, and 97.2% of COSMIC variants to HX1. HX1-absent, non-synonymous ClinVar SNPs were involved in 140 genes and many important functions in various diseases, most of which were due to the assembly failure of essential exons. In contrast, the HX1-specific regions were scantly expressible, as shown in the cell lines and clinical samples of Chinese patients. Our results demonstrated that the de novo-assembled individual genome such as HX1 did not have advantages against the standard GRCh38 genome due to insufficient assembly quality, and that it is, therefore, not recommended for common use.

Introduction

Recent advancements in next-generation sequencing (NGS) have been making human whole-genome sequencing (WGS) and whole-exome sequencing (WES) increasingly affordable to the public, promoting applications in precision medicine (Cai et al. 2017; Wang et al. 2008), which depends on genetic variations, including single nucleotide polymorphisms (SNPs) and structural variants, identified by mapping the sequencing reads to a reference genome. Previous studies have revealed that genomes from different ethnicities have distinct and specific regions in their genomes, including some specific structural variants (Genomes Project et al. 2012).These form the genetic basis of their phenotypes. Therefore, the reference genome might have a remarkable influence in such data analyses.

To minimize the potential ethnicity bias in genomes, de novo assembly of ethnicity-specific individual human genomes has been proposed since long, especially for non-Caucasian races (Cho et al. 2016; Gnerre et al. 2011; Li et al. 2010). However, assembling a complete reference genome like the widely used and well-annotated human genome (GRCh37/38) with short-read NGS techniques remains cost intensive and time consuming. The short reads from NGS are limited to constructing short contigs for the repeat elements (Shi et al. 2016) and miss a considerable number of RefSeq transcripts (Schneider et al. 2017). Short contigs introduce N-gaps in scaffold assembly and thus lead to errors in the further analysis of genetic variations. Single-molecule real-time (SMRT) long-read sequencing (Eid et al. 2009) is a powerful tool to overcome this limitation and construct much longer contigs (Shi et al. 2016). With the aid of the

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-019-02032-6) contains supplementary material, which is available to authorized users.

✉ Gong Zhang
zhanggong-uni@qq.com; zhanggong@jnu.edu.cn

[1] Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, Institute of Life and Health Engineering, Jinan University, Guangzhou 510632, China

short-read sequencing data for error correction, a de novo-assembled Chinese Han individual genome (HX1) became a milestone for practical genome assembly with considerable completeness for a specific ethnicity (Shi et al. 2016). Compared to the short-read-based assemblies such as CHM1, the long-read assembly provided orders of magnitude larger than contig N50 and much fewer gaps and lower gap lengths (Shi et al. 2016). Obviously, the long-read single-molecule sequencing technique provides a cost-efficient solution for individual genome assembly.

However, whether the long-read assemblies could be used as better reference genomes for ethnicity-specific genetic and functional studies has not been systematically evaluated. In this study, we assessed HX1 as a reference genome for the Chinese Han population in terms of genetics and functional/disease-related applications, compared with the standard GRCh38 reference genome, using a series of WGS/WES/transcriptome sequencing data of Chinese Han people.

Results

WGS/WES reads of Asian individuals mapped to HX1 to a lesser extent than to GRCh38

We downloaded the WGS and WES datasets of 81 Chinese Han, 12 Chinese Dai (a minority in southwest China), 10 Vietnamese (Kinh), 10 Japanese, 10 Mongoloid, and 40 British individuals (see Supplementary Table S1 for details). All datasets were mapped to GRCh38 and HX1 reference genomes, respectively, with identical mapping parameters (Fig. 1a).

The reads of the British individuals mapped to GRCh38 better than to HX1 in terms of map rate and Levenshtein distance rate (LD rate, which is calculated as Levenshtein distance of a read to the reference divided by the read length), which is expected, given that the GRCh38 reference genome is based on Caucasians. Moreover, as expected, the sequencing datasets of Caucasians mapped to GRCh38 better than to HX1, with a higher map rate and lower LD rate (Fig. 1b). Theoretically, HX1 should be closer to the Chinese Han population, especially considering the 12.8 Mb-specific region that is absent in GRCh38 and is Han ethnicity-specific. Surprisingly, all Asian datasets mapped to GRCh38 better than to HX1 in terms of map rate and LD rate (Fig. 1b). The map rate against GRCh38 was on average 2.0% higher than that against HX1. Even the Illumina sequencing reads of HX1 were mapped to GRCh38 better than to the HX1 reference sequence itself (map rate: 87.4% against GRCh38 versus 86.2% against HX1).

The HX1 reference genome (2.93 Gb in total length) is 9.2% shorter than the GRCh38 reference genome (3.23 Gb in total length), indicating that some genomic fragment might be missing in HX1. However, this might not explain the lower mapping rate of HX1 owing to the following reasons. First, the advantage of the GRCh38 map rate against the HX1 map rate was significantly greater for WGS (0.020 ± 0.004) than for WES (0.017 ± 0.002) data ($P = 0.0006$, Mann–Whitney U test), and both map rate advantages were much lower for the missing regions (Fig. 1c). Exons in the human genome are mostly non-repetitive and highly conserved. Therefore, the greater advantage of the WGS map rate than the WES map rate indicated that the assembly of HX1 may be poorer in the highly repetitive regions rather than in the missing regions. Second, in all datasets, there are considerably more 'GRCh38-better' reads than 'HX1-better' reads (Fig. 1d, e).

Structural misassembly is a major flaw of de novo-assembled personal genomes

Since the LD rate when mapping to HX1 is higher than that of GRCh38, we first speculated that the HX1 reference genome contained a considerable number of single nucleotide variations (SNVs) of the PacBio primary assembly, which were not fully corrected by the Illumina short reads with higher accuracy. Therefore, we tried to further correct the HX1 assembly with the Illumina short reads using our previously established iterative genome correction strategy with the accurate FANSe algorithm; this strategy has been experimentally validated with negligible false positives and false negatives (Wu et al. 2014). We corrected 1.28 million SNVs in the first round. After four rounds of genome corrections, fewer than 100,000 SNVs were correctable, indicating that the SNV correction is almost saturated (Fig. 2a). The mapping rate of HX1 Illumina reads against the corrected genome increased marginally, by 0.3% (Fig. 2a). However, after four rounds of correction, the mapping rate against HX1 (86.5%) was still less than that of the mapping rate against GRCh38 (87.4%). We then mapped eight WGS datasets to the HX1 reference sequence before and after correction. We found an average 0.4% (0.3%~0.5%) increase in the mapping rate after the genome correction, but it was still less than that of GRCh38 in all cases (Supplementary Fig. S1). This suggested that the uncorrected SNVs of the HX1 assembly were not the major cause of the lower mapping rate.

We next speculated that the HX1 assembly contained a considerable number of misassembled large fragments due to the higher error rate of PacBio long reads. We mapped the WGS and WES datasets of Chinese Han individuals against the HX1 and GRCh38 reference genomes, respectively, and obtained reads that uniquely mapped to one reference genome. Intriguingly, the reads that uniquely mapped to GRCh38 reads were mostly enriched for mitochondrial genes (Fig. 2b and Supplementary Fig. S2), indicating that

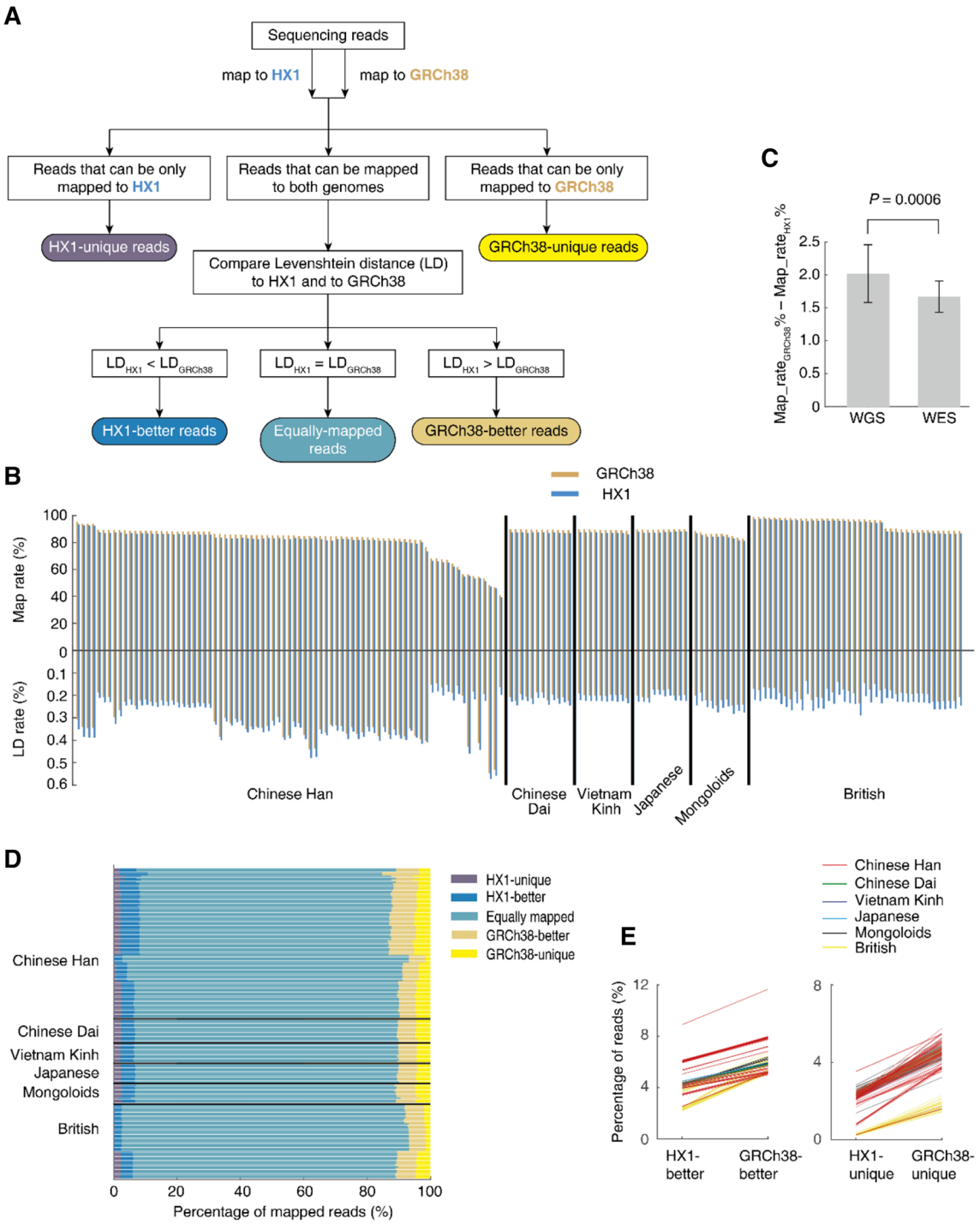

Fig. 1 Mapping WGS/WES datasets of various ethnic groups to GRCh38 and HX1 reference genome sequences. The accession numbers of these datasets are listed in Supplementary Table S1. **a** Analysis strategy. WGS/WES sequencing reads were mapped to HX1 and GRCh38 reference sequences, respectively. LD = Levenshtein distance. **b** The mapping rate and LD rate from the reference were compared for each dataset mapping to both reference genome sequences. Map rate = mapped reads/total reads. LD rate = LD/read length. **c** Difference in map rate between GRCh38 and HX1 for WGS and WES datasets. *P* value was calculated using the Mann–Whitney *U* test. **d** Percentages of the reads mapped to both reference genome sequences. **e** Comparison of the 'HX1-better' reads versus the 'GRCh38-better' reads, and 'HX1-unique' reads versus the 'GRCh38-unique' reads for all analysed WGS/WES datasets

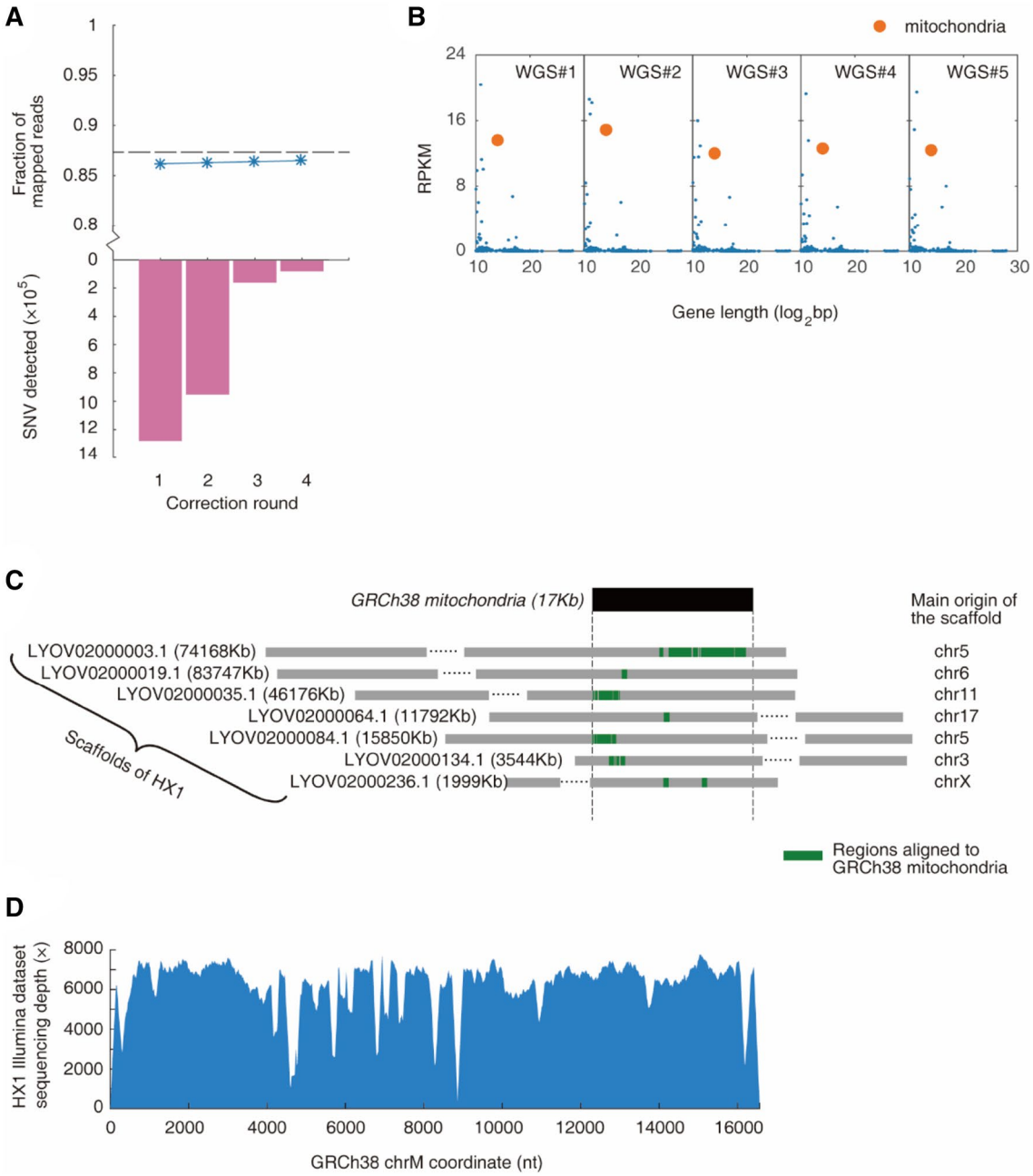

Fig. 2 Structural misassembly of the HX1 reference genome. **a** Correction of the HX1 reference genome using HX1 Illumina short reads. The upper part shows the fraction of mapped reads after each correction round. The dashed line shows the mapping rate against the GRCh38 reference genome. The lower part shows the number of detected and corrected SNVs after each correction round. **b** The read density of the GRCh38-specific reads, measured in read count per kilobase gene per million reads (RPKM). All the scaffolds, as well as the mitochondrial genome (17 kb), were considered as genes here. **c** The HX1 scaffolds containing homologous regions (green parts) against GRCh38 mitochondria sequence. The main origins of the scaffolds are indicated on the right. **d** The sequencing depth of the GRCh38 mitochondrial sequence when the HX1 Illumina reads were mapped to GRCh38

the mitochondrial sequence was largely misassembled in HX1. We aligned the HX1 scaffolds to the GRCh38 mitochondrial reference genome using BLAST and found that the mitochondrial genome was scattered in at least seven scaffolds (Fig. 2c). These scaffolds covered 69.0% of the mitochondrial genome. However, all mitochondrial fragments were assembled with other fragments belonging to the nuclear genome (Fig. 2c and Supplementary Fig. S3). The mitochondrial genome is 16 ~ 17 kb long and is highly conserved among human beings. We mapped the HX1 Illumina

WGS raw data to GRCh38 and obtained gapless full coverage (Fig. 2d), demonstrating that the HX1 genome contains a complete mitochondrial genome that is highly homologous to the GRCh38 genome. Therefore, these results confirmed that there are considerable structural misassembly events in the HX1 genome, which may cause lower mapping rates and higher LD rates when used as a reference. It should be noted that this is not an HX1-specific flaw. The other de novo-assembled individual genomes using various strategies and techniques (NA12878_ASM101398v1, Venter_HuRef, NA12878_ALLPATHS and AK1_v2) also contained severe fragmentation and misassembly of the mitochondrial genome (Supplementary Fig. S4). Importantly, not a single contig or scaffold in these de novo-assembled personal genomes represent a (near-)complete mitochondrial genome without nuclear fragments. These results indicate that there are likely to be many flaws in de novo-assembled personal genomes.

HX1 missed annotated and highly disease-relevant variations

For applications in genetics and medicine, a reference genome sequence should contain sufficient annotated variations. Therefore, we mapped the SNPs/SNVs in dbSNP (version 150), ClinVar (version 150), and COSMIC (v82) databases to the HX1 reference genome. We generated 1000 nt, 500 nt, and 100 nt reads centred at each SNP/SNV in these databases according to the GRCh38 coordinates and mapped them to HX1 using the FANSe3 algorithm. Since FANSe3 automatically neglects reads that are mapped to highly repetitive regions, the unmapped SNP/SNV reads were mapped again to HX1 using FANSe2. We then assessed the coverage of annotated variations in real-world WGS/WES datasets (PRJCA000246 and SRP050281) of the Chinese Han population. For all cases and all databases, mapping to GRCh38 covered 0.6–2.3% (1.7% on average) more annotated variations than mapping to HX1 (Fig. 3a), showing that HX1 missed a fraction of annotated variations.

For applications in genetics, 97.1% (315,746,941/325,174,796) annotated variation IDs from dbSNP can be found in the non-repetitive regions of HX1. An additional 1.09% dbSNP IDs were mapped to the repetitive regions of HX1. Compared with the 9.2% absent regions, the coverage of dbSNP in HX1 is near complete. Among the HX1-absent SNPs, 52.4% were included in dbSNP databases in version 150, and 79.8% of the HX1-absent SNPs were included later than in version 142 (Fig. 3b), indicating that these GRCh38-specific SNPs were largely recently recognized by the community due to technical improvements and the boom in worldwide WGS projects.

For disease-relevant applications, 98.6% ClinVar (a disease-relevant subset of dbSNP) annotated SNP IDs were mapped to HX1. In contrast to the version distribution of the dbSNPs, among the 3821 HX1-absent ClinVar SNPs, only 8.4% were of version 150, and only 38.1% were included in ClinVar later than version 142 (Fig. 3c). These results demonstrated that most HX1-absent SNPs with clinical significance were recognized in early studies, demonstrating their high relevance to the disease phenotypes. In concordance, 69.8% of the HX1-absent ClinVar SNPs had minor allele frequencies (MAFs) less than 1% (Fig. 3d). Considering that the phenotypic impact is inversely correlated to the MAF (Hindorff et al. 2011; McCarthy et al. 2008; Rossier et al. 2015; Visscher et al. 2017), the low MAF of the HX1-absent ClinVar SNPs implied significant clinical impact. In addition, 2791 HX1-absent ClinVar SNPs (73.04%) were inside the annotated genes, and 50.69% appeared to alter amino acids in protein products, affecting 140 genes (Supplementary File 1). These results indicated that these SNPs should not be overlooked. We then validated this by performing ingenuity pathway analysis (IPA) of these 140 genes. These genes participate in a wide variety of important molecular and cellular functions, such as DNA replication, recombination and repair, cell cycle, nucleic acid metabolism, and cell-to-cell signalling and interactions (Fig. 3e). The top diseases related to these genes are enriched in hereditary disorders, developmental disorders, etc. (Fig. 3f). As the clinical significance of ClinVar SNPs has been demonstrated, the missing SNPs among these in the HX1 reference genome could be a drawback in disease-relevant studies.

Multiple housekeeping genes were misassembled in HX1

Among the highly disease-relevant SNPs, the missing ones reflected that the HX1 reference sequence did not contain homologous regions of these non-synonymous SNPs. Among the 140 affected genes, 87 genes were found to contain GRCh38-annotated exons, which could not be found in HX1-assembled scaffolds (Supplementary File 3). This result indicated that HX1 may fail to assemble these genes. Of note, many of these are housekeeping genes, which are ubiquitously expressed in most human tissues (data from the NCBI Gene Database).

To validate this hypothesis, we first mapped the RNA-seq datasets of the Chinese Han-derived cells, including HX1, MHCCLM3, MHCCLM6, and MHCC97H, to the RefSeq-RNA reference sequences of GRCh38. On average, 53 genes were detected in ≥ 10 RNA-seq reads, demonstrating their expression (Bloom et al. 2009). This showed that a considerable fraction of these genes was expressed in the Chinese Han RNA samples (Fig. 4a and Supplementary Table S2). Next, we mapped the HX1 Illumina WGS reads to GRCh38 reference genome sequences. We found that all these HX1-absent exons actually exist in the HX1 genomic DNA

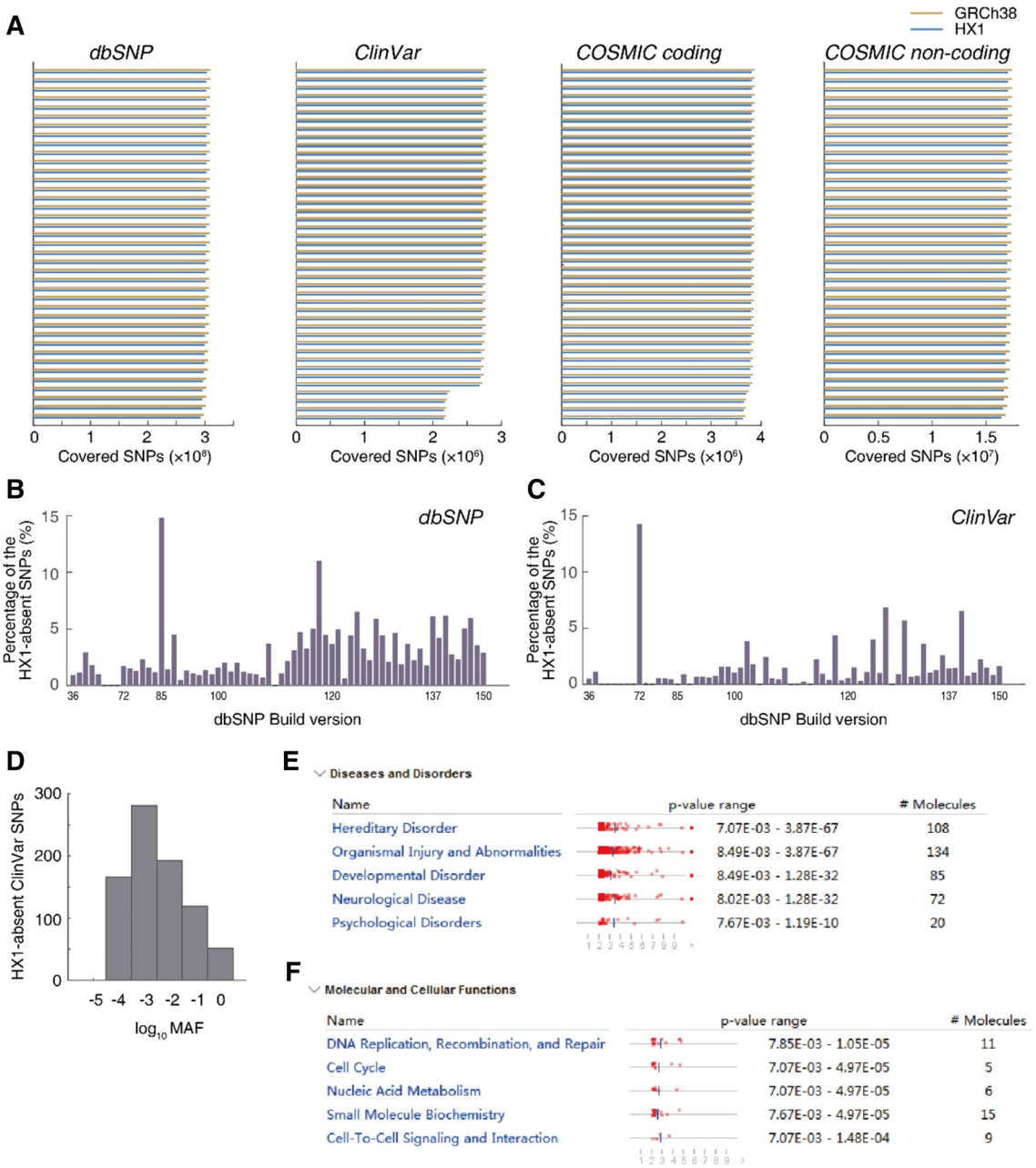

Fig. 3 HX1-missed disease-relevant variations. **a** WGS/WES datasets of Chinese Han populations mapped to GRCh38 and HX1 reference genomes, respectively. The SNPs/SNVs covered in dbSNP, ClinVar, and COSMIC databases of these datasets when mapped to the two reference genomes were compared. **b** The HX1-absent SNPs in the dbSNP database versus the database release versions. Detailed build versions are shown in Supplementary Fig. S5. **c** The HX1-absent SNPs in the ClinVar database versus the database release versions. Detailed build versions are shown in Supplementary Fig. S5. **d** The MAF distribution of the HX1-absent SNPs in ClinVar databases. **e–f** The top diseases and disorders (**e**) and the top molecular and cellular functions (**f**) of the 140 annotated genes containing HX1-absent and non-synonymous ClinVar SNPs analysed using ingenuity pathway analysis ingenuity pathway analysis (IPA)

sample, with sufficient sequencing depths (Fig. 4b shows three housekeeping genes as examples. See also Supplementary Fig. S6 and Table S3). For example, exons 2 ~ 5 of the *LMNB2* gene were missing in HX1 reference sequences, containing the ClinVar SNP rs57521499. More than 80% of the *TERT* exons and more than 70% of the *MAP3K15* exons were missing in HX1 reference sequences. These genes are crucial housekeeping genes. Therefore, massive exon deletion would be unrealistic. Sufficient coverage when mapping HX1 WGS data to GRCh38 showed that these exons were actually present and complete in HX1, but the flaws when assembling HX1 discarded these essential genomic

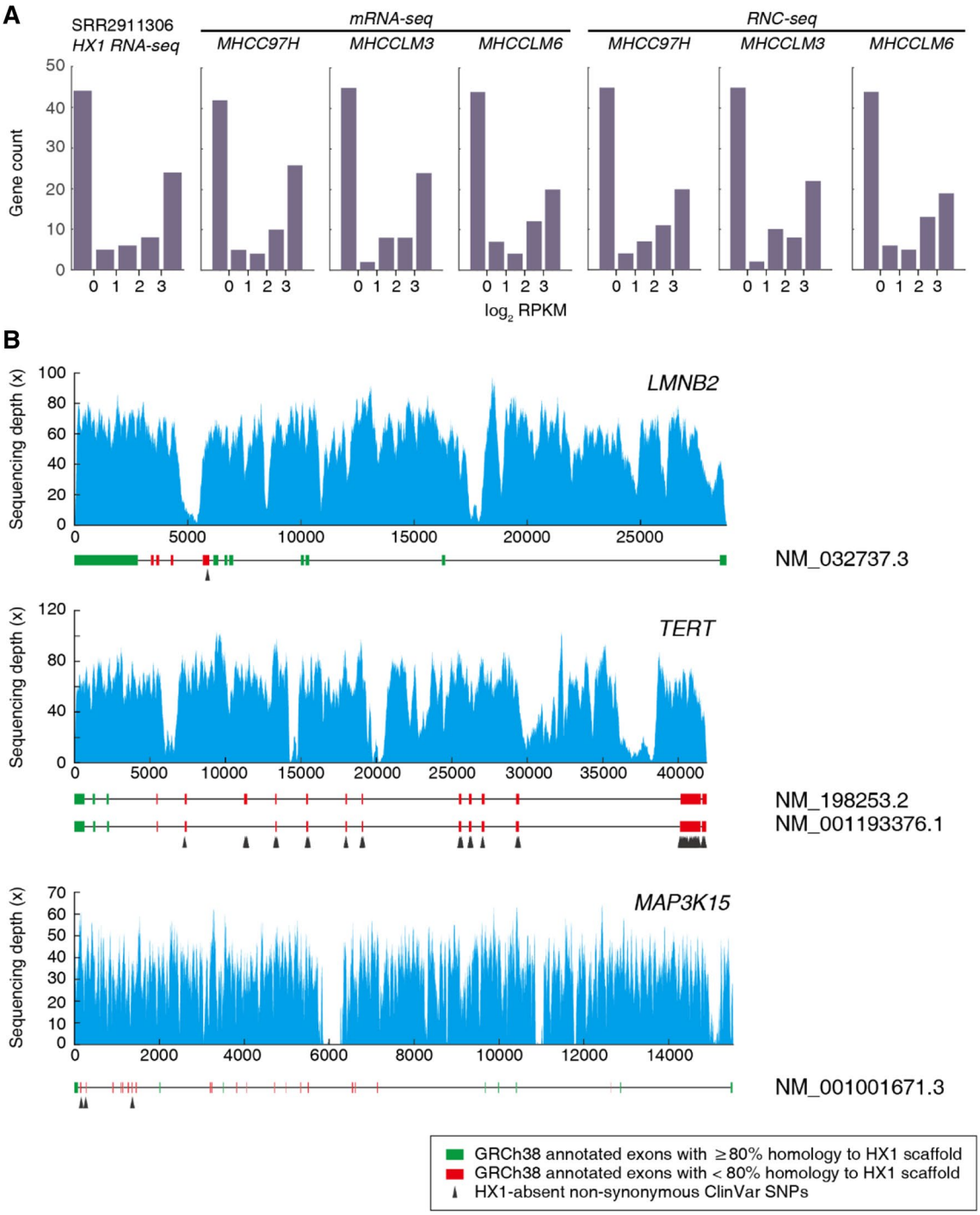

Fig. 4 Misassembled disease-relevant genes in the HX1 reference genome. **a** The gene expression level (transcription and translation) of the 87 misassembled genes containing non-synonymous ClinVar SNPs. The RNA-seq reads were mapped to GRCh38 RefSeq-RNA reference sequences and quantified using the RPKM method. **b** The sequencing depth of the three misassembled housekeeping genes as examples. HX1 Illumina WGS short reads were mapped to the GRCh38 reference genome. The sequencing depths of the uniquely mapped reads were plotted. The exon annotation in GRCh38 was plotted below the sequencing depth diagram. Green exons are GRCh38 annotated exons with > 80% homology to the HX1 reference genome, while the red exons have no homology (< 80%) to the HX1 assembled scaffolds. Black triangles denote the HX1-absent non-synonymous ClinVar SNPs. The complete list of 87 such genes is shown in Supplementary Fig. S6)

segments. These genes were not enriched in any pathways nor GO terms, according to PANTHER calculation (www.pantherdb.org), indicating that the gene misassembly in the HX1 reference sequence is ubiquitous and random.

HX1-specific regions are almost not expressed in RNA-seq datasets

HX1 was thought to have its own advantages in ethnicity-specific disease research: 12.8 Mb HX1-specific sequences (0.44% of the genome size) have been reported (Shi et al. 2016). These regions contain 533 contigs with lengths between 10 and 160 kb. However, we found that 54 (10.13%) of these regions could be aligned to GRCh38 reference sequences with > 80% sequence coverage and > 80% identity using BLAST, indicating that these contigs are still highly homologous to GRCh38, just with a higher number of SNPs or sequencing errors. Only 479 contigs and 11.99 Mb regions (0.41% of the genome size) were genuinely HX1 specific (Supplementary File 2). Considering that only 2% of the human genome comprises coding genes, this tiny fraction of HX1-specific regions has little chance to express proteins.

We mapped the HX1 short-read RNA-seq data to the HX1 reference genome using the FANSe2Splice algorithm, which can map spliced reads (Mai et al. 2017). Since there was no clear gene annotation in the HX1-specific contigs, we presume that 31.3% of the contigs were exons, which equals the average fraction of the exons in all annotated human genes. Thus, we calculated the fragments per kilobase exons per million reads (FPKM) of these HX1-specific regions. A typical mammalian cell contains 200,000 mRNA molecules (Shapiro et al. 2013). A transcript with an average length of 2.2 kb and with an abundance of one copy per cell is equivalent to 2.27 FPKM (Li et al. 2018). In the HX1 RNA-seq dataset, as well as the transcriptome and translatome sequencing datasets of the three Chinese hepatocellular carcinoma cell lines, we found that no HX1-specific contig yielded more than 2.27 FPKM (Fig. 5a), suggesting that they were expressed at an extremely low and fluctuating level, with less than a single copy per cell on average—if they were expressed at all. The FPKM values of these contigs at the translation level were even lower (Fig. 5b), demonstrating that these contigs could hardly be translated into functional proteins in cells. We further examined 28 RNA-seq datasets of Chinese prostate cancer patients (accession number ERP000550). Only one contig was expressed at > 2.27

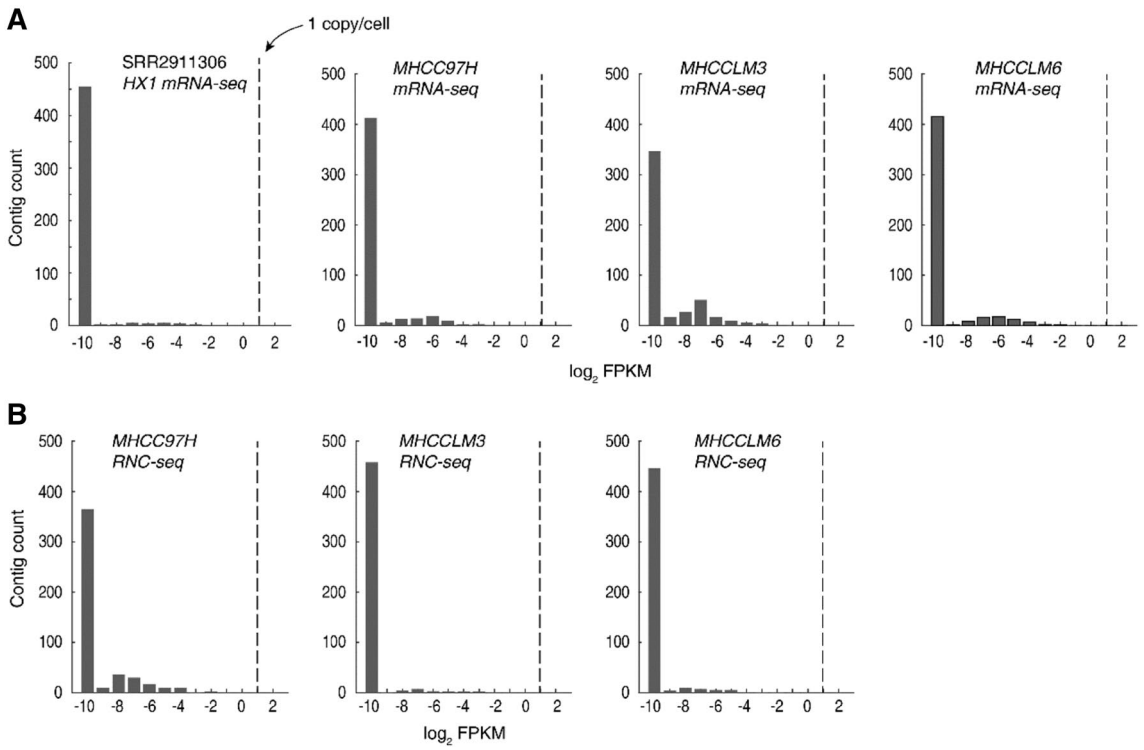

Fig. 5 Expression level of HX1-specific contigs, measured in fragments per kilobase exons per million reads (FPKM). HX1 RNA-seq, RNA-seq (transcriptome sequencing), and RNC-seq (translatome sequencing) datasets of three hepatocellular carcinoma cell lines (MHCC97H, MHCCLM3, and MHCCLM6), which were derived from Chinese Han individuals, were analysed. The dashed line denotes the minimum FPKM of one copy per cell transcript

FPKM in only one sample (Supplementary Fig. S7 and Supplementary Table S4). Therefore, the HX1-specific regions could hardly be expressed at least in the 35 RNA samples we analysed, implying a minimal chance to be useful as protein-coding genes.

Discussion

The 'standard' human reference genomes, such as GRCh38 and GRCh37, have been confirmed to miss functionally important variants in populations other than those with European ancestry (Genomes Project et al. 2015). De novo assembly of personal genomes should be helpful to recover these missed variants and to analyse the genetic evolution. The development of long-read single-molecule sequencing technology, such as PacBio SMRT and Nanopore sequencers, has enabled the achievement of more than 10× longer reads compared to second-generation sequencers, which facilitated the de novo assembly of an individual human genome with a rather large N50. In the enthusiasm over long contigs, the accuracy of the assembly has been ignored. A high error rate (12–20%) is a common drawback of the current single-molecule sequencing technology. However, many regions in human genomes are similar. For example, an 8.8 kb region in chromosome 5 is 88.88% homologous to the mitochondrial genome (detailed alignments are shown in Supplementary File 4), i.e. the deviation of these two regions is less than the LD rate of the single-molecule sequencing reads. Therefore, the assembly algorithms would easily misassemble these reads. Such events are spread all along the human genome. The genome-wide spread NUMTs (nuclear mitochondrial DNA segments) occur hundreds of times in a human genome and variable for each individual (Dayama et al. 2014), setting additional challenges to the assembly due to their high homology up to 94% (Mishmar et al. 2004). As a consequence, all the current de novo-assembled individual genomes failed to assemble the complete mitochondrial genome sequence as a distinct contig (Supplementary Fig. S4), reflecting a severe misassembly problem. Another consequence of the high error rate introduced by the single-molecule sequencing is the possibility of assembly failure of some reads due to high error incidences, causing exclusion by the assembly algorithms. This leads to a complete loss of some gene fragments. For example, most fragments of the housekeeping genes *TERT* and *MAP3K15* were missing in HX1 reference sequences (Fig. 4b); however, they actually exist in the sample. These problems could not be solved by the scaffolding, in which the contig sequences would be generally untouched. In contrast, GRCh38 involved considerable effort in terms of genomic clones and error correction, thereby circumventing many misassembly problems.

To reduce the errors in assembly, researchers use short but more accurate reads generated from Illumina sequencers (error rate 0.5–2%) to correct the contigs assembled using single-molecule long reads. In some assemblers like MECAT, this correction step takes up to 80% computational time, longer than the long-read assembly step (Xiao et al. 2017). However, neither the misassembly nor the fragment loss during the assembly could be corrected by these short reads. A minor reason is that the mainstream algorithms used in the genome correction provide poor performance in the cases of high error rates (Carlsson et al. 2013; Storer et al. 2012; Yang et al. 2013). Nevertheless, even with corrections using the hyper-accurate FANSe algorithm (Zhang et al. 2012), which perfectly corrected 20% error rates and achieved 100% accuracy with nearly 2000 experimentally validated sites (Wu et al. 2014), we could not elevate the mapping rate against HX1 to the same level as that of GRCh38 (Fig. 2a and Supplementary Fig. S1). This result echoed the widespread assembly problem of the de novo-assembled individual genome. Before a fundamental improvement in the sequencing accuracy of the long-read single-molecule sequencers, it seems that no thorough solution could effectively rectify the structural misassembly and fragment loss.

Since most genomic and clinical genetic studies have been based on GRCh37/GRCh38, the annotated variations were all based on these 'standard' reference genomes. Although most of the annotated variations in dbSNP, ClinVar, and COSMIC databases could be mapped to HX1 reference sequences, a considerable fraction of the SNPs, especially those with very small MAFs and thus with significant disease relevance, were missing in HX1, primarily due to misassembly and fragment loss. In contrast, the genuine HX1-specific region, which is a negligible fraction (0.41%) of the genome, hardly expresses any gene, at least, no housekeeping genes (Fig. 5). Therefore, it would be easier and less risky for genetic and functional studies to use GRCh38 instead of HX1 as the reference genome in general.

Schneider et al. systematically evaluated the other de novo personal genome assemblies based on short reads against GRCh38 (Schneider et al. 2017). They concluded that those assemblies failed to compete with GRCh38 for quality due to the complex and repetitive regions. Longer read inputs were necessary to enhance the assembly. However, our present study demonstrated that much longer single-molecule sequencing reads and the consequent super-long N50 did not remarkably improve the assembly quality, illustrating a dilemma. This highlights the importance of quality control of long-read sequencing data. Furthermore, it raises a fundamental question on the applicability of the de novo assembly of personal genomes, as it is very costly and labour intensive and may not have any practical advantage either. A personal genome almost free of assembly error and complete

coverage achieved by extremely intensive investment and effort (at least comparable to the assembly of GRCh38) would be ideal for personalized precision medicine. However, improving the error rate of long-read sequencers may still take a long time. Therefore, the most practical way for now would be to continue with the 'standard' reference genome sequences to maximize the availability of genetic and genomic knowledge. The HX1-specific regions may be more precisely re-sequenced and re-assembled and then patched to the 'standard' reference genome sequence if these regions really matter in ethnicity-specific studies.

To minimize the problem of the error-prone long-read single-molecule sequencing and the highly fragmented short reads of the second-generation sequencing, an efficient solution would be the BAC (bacterial artificial chromosome) clones. Typical BAC libraries contain insert DNA of 150–350 kb, which is much longer than the current single-molecule sequencing read lengths. The complexity of BAC clones is far lower than the entire genome; therefore, the highly similar fragments are unlikely to occur in one BAC clone, which makes the assembly of a complete and accurate BAC insert DNA much easier. This would avoid the misassembly of the mitochondrial genome and the missing exons of housekeeping genes. The NUMTs would no longer be a problem since the lengths of mitochondrial genome and NUMTs are considerably shorter than the BAC inserts. Assembling the highly accurate BAC insert sequences into entire chromosomes would be also much easier and more accurate than assembling much shorter reads. To be noted, this is one of the key strategies applied in the initial sequencing of human genome (Lander et al. 2001) and also the GRCh38 to ensure the long-range assembly accuracy (Schneider et al. 2017). However, the community was later amazed by the convenience brought by the BAC-free direct human genome assembly solely with NGS short reads (Li et al. 2010), but overlooked the limitation of this strategy.

Based on the abovementioned rationales, we propose a preliminary guideline for a high-quality assembly of a personal genome:

- Use sequencing technology that delivers raw error rate no more than 5%.
- Make BAC libraries of the target genome. Assemble the BAC clones at high accuracy and then assemble the chromosomes.
- For an ethnic-specific assembly, the map rate and LD rate of the same ethnic NGS dataset should be, in general, better than that of distant ethnic.
- The assembly should contain a complete and gapless mitochondrial genome sequence as an individual contig.
- The assembly should be examined for the integrity of housekeeping genes.

Based on a well-assembled personal genome, the other individual-specific segments of the same ethnic can be then added as patches like GRCh38 to build an ethnic-specific population reference genome.

Methods

Public data downloads

The raw sequencing data in FASTQ format were downloaded from the SRA/ENA database (accession number listed in Supplementary Table S1). GRCh38, HX1, Venter_HuRef, NA12878_ALLPATHS, NA12878_ASM101398v1 and AK1_v2 reference genome sequences were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov). The HX1-related files were downloaded from http://hx1.wglab.org/. The dbSNP (v.150) database and the ClinVar (v.150) database were downloaded from the NCBI database. The mutations of COSMIC (v.82) were downloaded from https://cancer.sanger.ac.uk/cosmic.

Transcriptome and translatome sequencing of Chinese Han-derived cell lines

Total RNA of three hepatocellular carcinoma cell lines derived from Chinese patients—MHCC97H, MHCCLM3, and MHCCLM6—were extracted using the standard Trizol protocol. RNC-RNA of MHCCLM6 was also extracted as described before (Li et al. 2018; Wang et al. 2013). PolyA + mRNA was extracted from the total RNA and RNC-RNA using RNA purification beads (Illumina). Then library construction was performed using the Vazyme mRNA-seq v2 Library Kit with insert sizes of 200 ~ 300 bp. mRNA and RNC-mRNA libraries were sequenced in an Illumina HiSeq X Ten sequencer under PE150 mode. The data were deposited to the Gene Expression Omnibus database under the accession number GSE121013.

NGS data processing

The WGS/WES datasets were mapped to reference sequences using the FANSe3 algorithm (Liu et al. 2018) (http://www.chi-biotech.com/fanse3/) with the parameters -E3% –indel. For the quantification of known transcripts, the RNA-seq reads were mapped to RefSeq-RNA reference sequences using FANSe3 with the parameters -E5% –indel. The RNA expression level was quantified using the RPKM method (Mortazavi et al. 2008). In the process of iterative genome correction of HX1, the HX1 Illumina short reads were mapped to the HX1 genome using the FANSe3 algorithm with the parameters -E7 –indel.

The map rate of a dataset is calculated as number of mapped reads/number of total reads. The Levenshtein

distance rate (LD rate) is calculated as: Levenshtein distance of a read to reference/read length.

quantified using the FPKM method according to the following formula:

$$\mathrm{FPKM}_{\mathrm{specific}} = \frac{\text{total exon reads}}{\text{mapped reads (millions)} \times \text{sequences length (KB)} \times 31.3\%},$$

Mapping the GRCh38 mitochondrial genome to HX1

The GRCh38 mitochondrial genome sequence was aligned to HX1 and other assembled personal genome sequences using BLAST with a threshold of 90% identity. The top seven scaffolds with mitochondrial homology in all assembled personal genome sequences were aligned to GRCh38 nuclear chromosomes using BLAST with a threshold identity of 90%.

Mapping GRCh38 exons to HX1

The annotated GRCh38 exons were aligned to the HX1 reference genome sequence using BLAST. Exons with > 80% sequence coverage and > 80% identity (with > 80% homology to HX1) were considered to have been found in HX1.

Mapping annotated SNPs/SNVs of GRCh38 to HX1

According to the positions of the SNPs/SNVs in dbSNP (v. 150), ClinVar (v. 150), and COSMIC (v.82), artificial sequencing reads 1000 nt long and centred at the SNPs/SNVs were generated. These artificial reads were mapped to HX1 using FANSe3 (Liu et al. 2018) with up to 5% mismatch. The unmapped artificial reads were again mapped to HX1 using FANSe2 (Xiao et al. 2014) with up to 5% mismatch again to map the reads mapping to repetitive sequences. The artificial reads that were mapped to the HX1 reference genome sequence determined the position of these SNPs/SNVs in the HX1 reference genome. For the SNPs/SNVs whose corresponding artificial reads failed to map to HX1, shorter reads (500 nt and 100 nt) were generated, and the mapping process was repeated. The SNPs/SNVs whose corresponding 100 nt artificial reads failed to map to HX1 were considered HX1-absent SNP/SNVs.

Quantification of HX1-specific contigs

First, 12.8 Mb of HX1-specific sequences were aligned to GRCh38 using BLAST. In the alignment, HX1-specific sequences with a total query coverage of > 80% and identity of > 80% were removed. To find and quantify possible transcripts from the HX1-specific regions, the RNA-seq reads were mapped to the HX1 reference genome sequence using a technically accelerated version of FANSe2splice (Mai et al. 2017) with up to three errors. The possible transcripts were

where factor 31.3% is the average fraction of the exon lengths in all annotated human genes.

Data access

The datasets are explicitly described in the manuscript. For details of all accession numbers, please refer to Supplementary Table S1.

Acknowledgements This work was supported by the Ministry of Science and Technology of China 'National Key Research and Development Program' (2017YFA0505001/2018YFC0910201/2018YFC0910202) and the Distinguished Young Talent Award of National High-level Personnel Program of China.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA, Genomes Project C (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. https://doi.org/10.1038/nature11632

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR, Genomes Project C (2015) A global reference for human genetic variation. Nature 526:68–74

Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics 10:221. https://doi.org/10.1186/1471-2164-10-221

Cai N, Bigdeli TB, Kretzschmar WW, Li Y, Liang J, Hu J, Peterson RE, Bacanu S, Webb BT, Riley B, Li Q, Marchini J, Mott R, Kendler KS, Flint J (2017) 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. Sci Data 4:170011. https://doi.org/10.1038/sdata.2017.11

Carlsson J, Gauthier DT, Carlsson JE, Coughlan JP, Dillane E, Fitzgerald RD, Keating U, McGinnity P, Mirimin L, Cross TF (2013) Rapid, economical single-nucleotide polymorphism and microsatellite discovery based on de novo assembly of a reduced representation genome in a non-model organism: a case study of Atlantic cod Gadus morhua. J Fish Biol 82:944–958. https://doi.org/10.1111/jfb.12034

Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, Chae KS, Kim CG, Kim S, Eriksson A, Edwards JS, Lee S, Kim BC, Manica A, Oh TK, Church GM, Bhak J (2016) An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun 7:13637. https://doi.org/10.1038/ncomms13637

Dayama G, Emery SB, Kidd JM, Mills RE (2014) The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res 42:12640–12649. https://doi.org/10.1093/nar/gku1038

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138. https://doi.org/10.1126/science.1162986

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518. https://doi.org/10.1073/pnas.1017351108

Hindorff LA, Gillanders EM, Manolio TA (2011) Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis 32:945–954. https://doi.org/10.1093/carcin/bgr056

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272. https://doi.org/10.1101/gr.097261.109

Li D, Lu S, Liu W, Zhao X, Mai Z, Zhang G (2018) Optimal settings of mass spectrometry open search strategy for higher confidence. J Proteome Res 17:3719–3729. https://doi.org/10.1021/acs.jproteome.8b00352

Liu W, Xiang L, Zheng T, Jin J, Zhang G (2018) TranslatomeDB: a comprehensive database and cloud-based analysis platform for translatome sequencing data. Nucleic Acids Res 46:D206–D212. https://doi.org/10.1093/nar/gkx1034

Mai Z, Xiao C, Jin J, Zhang G (2017) Low-cost, low-bias and low-input RNA-seq with High experimental verifiability based on semiconductor sequencing. Sci Rep 7:1053. https://doi.org/10.1038/s41598-017-01165-w

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369. https://doi.org/10.1038/nrg2344

Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Hum Mutat 23:125–133. https://doi.org/10.1002/humu.10304

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226

Rossier BC, Baker ME, Studer RA (2015) Epithelial sodium transport and its control by aldosterone: the story of our internal environment revisited. Physiol Rev 95:297–340. https://doi.org/10.1152/physrev.00011.2014

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood JM, Clarke L, Koren S, Boitano M, Peluso P, Li H, Chin CS, Phillippy AM, Durbin R, Wilson RK, Flicek P, Eichler EE, Church DM (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 27:849–864. https://doi.org/10.1101/gr.213611.116

Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14:618–630. https://doi.org/10.1038/nrg3542

Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, Lintner KE, Ding Q, Wang Z, Hu J, Wang D, Wang F, Wang L, Lyon GJ, Guan Y, Shen Y, Evgrafov OV, Knowles JA, Thibaud-Nissen F, Schneider V, Yu CY, Zhou L, Eichler EE, So KF, Wang K (2016) Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun 7:12065. https://doi.org/10.1038/ncomms12065

Storer CG, Pascal CE, Roberts SB, Templin WD, Seeb LW, Seeb JE (2012) Rank and order: evaluating the performance of SNPs for individual assignment in a non-model organism. PLoS One 7:e49018. https://doi.org/10.1371/journal.pone.0049018

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101:5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65. https://doi.org/10.1038/nature07484

Wang T, Cui Y, Jin J, Guo J, Wang G, Yin X, He QY, Zhang G (2013) Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. Nucleic Acids Res 41:4743–4754. https://doi.org/10.1093/nar/gkt178

Wu X, Xu L, Gu W, Xu Q, He QY, Sun X, Zhang G (2014) Iterative genome correction largely improves proteomic analysis of nonmodel organisms. J Proteome Res 13:2724–2734. https://doi.org/10.1021/pr500369b

Xiao CL, Mai ZB, Lian XL, Zhong JY, Jin JJ, He QY, Zhang G (2014) FANSe2: a robust and cost-efficient alignment tool for

quantitative next-generation sequencing applications. PLoS One 9:e94250. https://doi.org/10.1371/journal.pone.0094250

Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z (2017) MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods 14:1072–1074. https://doi.org/10.1038/nmeth.4432

Yang X, Chockalingam SP, Aluru S (2013) A survey of error-correction methods for next-generation sequencing. Brief Bioinform 14:56–66. https://doi.org/10.1093/bib/bbs015

Zhang G, Fedyunin I, Kirchner S, Xiao C, Valleriani A, Ignatova Z (2012) FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads. Nucleic Acids Res 40:e83. https://doi.org/10.1093/nar/gks196