Hum Genet (2013) 132:899–911
DOI 10.1007/s00439-013-1300-9

ORIGINAL INVESTIGATION

# Comprehensively identifying and characterizing the missing gene sequences in human reference genome with integrated analytic approaches

**Geng Chen · Charles Wang · Leming Shi · Weida Tong · Xiongfei Qu · Jiwei Chen · Jianmin Yang · Caiping Shi · Long Chen · Peiying Zhou · Bingxin Lu · Tieliu Shi**

Received: 19 August 2012 / Accepted: 25 March 2013 / Published online: 10 April 2013


**Abstract** The human reference genome is still incomplete and a number of gene sequences are missing from it. The approaches to uncover them, the reasons causing their absence and their functions are less explored. Here, we comprehensively identified and characterized the missing genes of human reference genome with RNA-Seq data from 16 different human tissues. By using a combined approach of genome-guided transcriptome reconstruction coupled with genome-wide comparison, we uncovered 3.78 and 2.37 Mb transcribed regions in the human genome assemblies of Celera and HuRef either missed from their homologous chromosomes of NCBI human reference genome build 37.2 or partially or entirely absent from the reference. We further identified a significant number of novel transcript contigs in each tissue from de novo transcriptome assembly that are unalignable to NCBI build 37.2 but can be aligned to at least one of the genomes from Celera, HuRef, chimpanzee, macaca or mouse. Our analyses indicate that the missing genes could result from genome misassembly, transposition, copy number variation, translocation and other structural variations. Moreover, our results further suggest that a large portion of these missing genes are conserved between human and other mammals, implying their important biological functions. Totally, 1,233 functional protein domains were detected in these missing genes. Collectively, our study not only provides approaches for uncovering the missing genes of a genome, but also proposes the potential reasons causing genes missed from the genome and highlights the importance of uncovering the missing genes of incomplete genomes.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00439-013-1300-9) contains supplementary material, which is available to authorized users.

G. Chen · X. Qu · J. Chen · J. Yang · C. Shi · L. Chen · P. Zhou · B. Lu · T. Shi (✉)
Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, The Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China
e-mail: tieliushi01@gmail.com

C. Wang
Functional Genomics Core, Beckman Research Institute, City of Hope Comprehensive Cancer Center, Duarte CA 91010, USA

L. Shi · W. Tong
National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079, USA

## Introduction

The genome of an organism contains the hereditary information needed to build and maintain that organism. The genes at appropriate positions in the genome determine how the organism's phenotype will develop under a certain set of environmental conditions. Accordingly, the completeness and the accuracy of the genome are crucial for various functional genomic and transcriptomic studies. They have great impact on the experimental data processing and interpretation, especially for the high-throughput sequencing data which are growing exponentially. However, the accomplishments of International Human Genome Project (Lander et al. 2001; Consortium 2004) provide us the opportunities to study the human diseases and their genetics at genome-wide level which has affected biomedical research profoundly. However, two main types of heterochromatic and euchromatic gaps still exist in the constructed human genome and they occupy a significant portion of the total genomic sequence (Eichler

et al. 2004). Moreover, some genomic regions may have been mis-assembled. The quality of the assembled reference genome is influenced by various factors such as the complexity of genomic sequences, the limitations of assembly algorithms and the biases of sequencing technologies. It is still a long way to go building a fully completed human reference genome and annotate the fine structure of genome.

Several previous studies have revealed that a portion of human genomic sequences are absent from the previous versions of NCBI human reference assemblies and the absent genomic sequences could contain some novel genes (Khaja et al. 2006; Kidd et al. 2010; Li et al. 2010; Chen et al. 2011a). However, these studies do not combine different strategies and make full use of the human transcriptome sequencing data to comprehensively uncover the genes absent from the human reference genome. Moreover, the reasons behind the genomic sequences or genes missing from the human reference genome have not been fully explored yet. Obviously, these earlier findings show that the constructed human reference genome is still incomplete and the missing human genomic sequences could contain novel genes with unknown functions. Without the information of those missing gene sequences, it could be difficult to comprehensively interpret various analytic and experimental results.

The missing genes are an integral part of the human reference genome, thus it is necessary to further verify the quality of human reference genome and to uncover those missing genes. With the fast development of high-throughput sequencing technologies, a number of human genomes have been assembled besides the NCBI human reference genome, such as Celera (Venter et al. 2001), HuRef (Levy et al. 2007) and YH (Asian) (Wang et al. 2008). The genetic variations among those different individuals whose DNAs were used to construct the human genomes could cause variations of the genome assemblies. Furthermore, other factors, e.g., distinct assembly strategies and different sequencing depth or completeness of the genome could also further increase the differences between different genome assemblies. Unlike Celera and HuRef assemblies are independent of NCBI human reference genome, YH genome does not take all indels into its sequences and it has the same coordinates as NCBI build 36.1. Therefore, Celera and HuRef assemblies are valuable resources for genome comparison and for uncovering the missing genomic sequences. In addition, RNA-Seq technologies have the potential to capture all the expressed genes as a snapshot of cells at special spatiotemporal points by deep sequencing, which allows us to look at transcriptome at a digital level (Marioni et al. 2008; Wang et al. 2009; Marguerat and Bahler 2010; Nagalakshmi et al. 2010; Ozsolak and Milos 2011). Accordingly, the transcriptome sequencing data from various human tissues can provide us rich resources to reveal and investigate the possible missing genes of the human reference genome.

To more comprehensively identify and characterize the possible missing genes of the human reference genome, here we combined two different strategies to carry out this study. Specifically, we first performed genome-guided transcriptome assembling on Celera and HuRef assemblies using RNA-Seq data from 16 different human tissues (Cabili et al. 2011). Second, we conducted two groups of genome-wide comparisons by aligning the human genome assemblies of Celera and HuRef against the NCBI build 37.2. Third, we exploited the results from these two approaches and revealed a notable portion of gene sequences from Celera and HuRef that are either missing from their homologous chromosomes of NCBI build 37.2 or partially or entirely absent from the reference assembly. Considering the methodological differences between genome-guided and genome-independent transcriptome reconstruction (Chen et al. 2011b; Garber et al. 2011), we also conducted de novo assembling of the transcriptomes of the 16 different human tissues and further uncovered a significant number of transcript contigs missing from the human reference genome. We identified several potential factors that result in genes missed from the human reference genome. In addition, we also assessed the conservation of those missing genes among human, chimpanzee, macaca and mouse, and examined their potential protein products, including possibly known functional protein domains.

## Materials and methods

### Transcriptome sequencing datasets and other data used

The RNA-Seq data from 16 different human tissues used in this study belong to the Illumina Human Body Map 2 project and they were first used in a previous study (Cabili et al. 2011). We downloaded them from ArrayExpress (accession no. E-MTAB-513). They were generated using the Illumina HiSeq 2000 platform with the standard Illumina mRNA-Seq protocol. The reads are paired-end and 50 bp in length. On average, 159.84 million reads for each sample were collected. Those 16 tissues are thyroid, testes, ovary, white blood cells, skeletal muscle, prostate, lymph node, lung, adipose, adrenal, brain, breast, colon, kidney, heart and liver (see Supplementary File 1, Table S1).

The NCBI human reference genome build 37.2, Celera (Venter et al. 2001) and HuRef (Levy et al. 2007) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). The human expressed sequence tags (ESTs) and the non-redundant protein sequence database (nr) and the genomes

of chimpanzee (NCBI build 3.1), macaca (NCBI build 1.2) and mouse (NCBI build 37.2) were also obtained from NCBI. These human genome assemblies and the genomes of chimpanzee, macaca and mouse were the unmasked versions [without repeat masking by RepeatMasker (Saha et al. 2008)]. The protein family database Pfam (Punta et al. 2012) (Pfam 26.0 and 13,672 families) was downloaded from http://pfam.sanger.ac.uk/ and we only used the high-quality, manually curated family entries (Pfam-A) for analyses. These diverse data were used in the corresponding analyses as shown in Fig. 1.

## Genome-guided transcriptome reconstruction

We carried out transcriptome reconstruction on the human genome assemblies of Celera and HuRef. First, all the RNA-Seq reads from the 16 different human tissues were separately aligned onto Celera and HuRef using the spliced read aligner TopHat (Trapnell et al. 2009) (V1.3.3) with default parameter setting. Then, we separately assembled the transcriptome of each tissue based on the TopHat output with Cufflinks (Trapnell et al. 2010) (V1.2.1) using default parameters (and 'min-frags-per-transfrag = 0') on Celera and HuRef. After assembling, each human genome assembly of Celera and HuRef has 16 transcriptome assemblies from different human tissues. Next, to remove those transcript fragments that are probably artifacts, we merged the 16 transcriptome assemblies for Celera and HuRef using Cuffmerge (Trapnell et al. 2010) with default parameters. Gene fragments (including the exons and introns assembled by Cufflinks) shorter than 100 bp were removed from the further analyses. Finally, a merged transcriptome from all 16 different human tissues was obtained for each human genome assembly of Celera and HuRef. The constructed transcriptome on Celera and HuRef were then used in the subsequent analyses.

## Genome-wide comparison

In order to identify the genes that are present in Celera and HuRef but are missed from the NCBI human reference genome build 37.2, we conducted two groups of genome-wide comparison by comparing NCBI build 37.2 with Celera and HuRef. We separately aligned the chromosomes of NCBI build 37.2 onto their homologous chromosomes in Celera and HuRef using LAST (Kielbasa et al. 2011) (version r189). The LAST software uses adaptive seeds approach and copes more efficiently with repeat-rich genomes than other tools like BLAST (Altschul et al. 1990; Altschul et al. 1997), YASS (Noe and Kucherov 2005) and LASTZ (Harris 2007). Therefore, in order to more comprehensively reveal the sequence variations between Celera, HuRef and NCBI Build 37.2, we used these genomes without repeat masking for comparison. We used the option of '-f0 -r10 -q90 -a0 -b90 -e300' for LAST to only export those match hits with ≥90 % identity and a minimum score of 300 for gapped alignments. The parameter setting was similar and comparable with previous studies

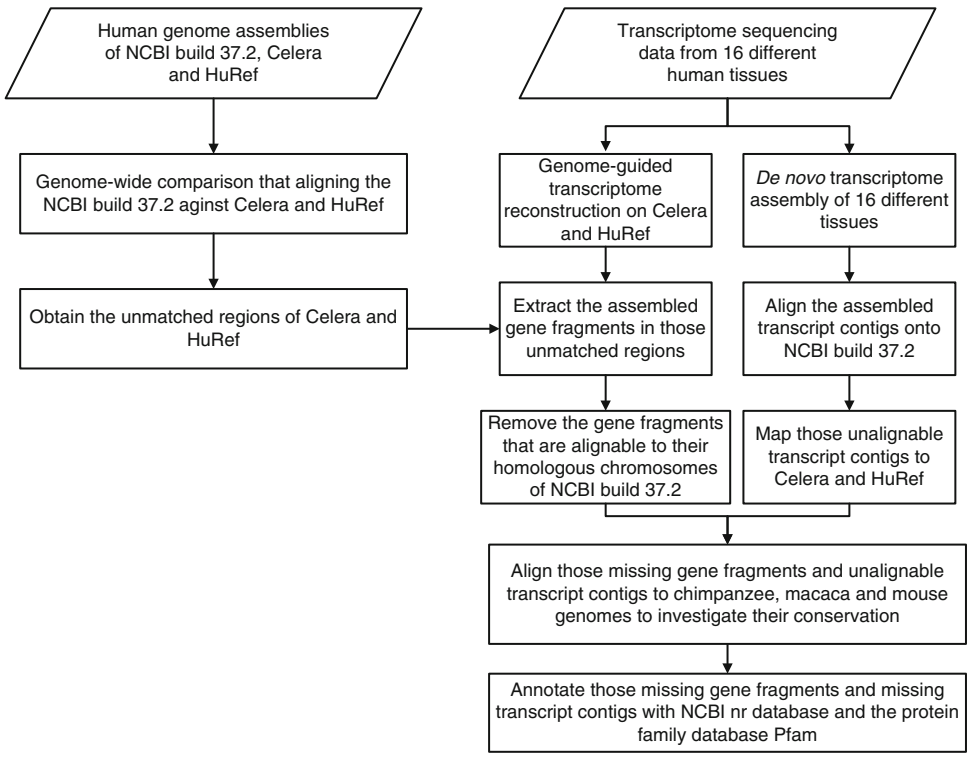


**Fig. 1** Strategies of identifying and characterizing genes missing from the human reference genome. To more comprehensively identify and characterize the missing gene sequences of the NCBI human reference genome build 37.2, we integrated two different strategies: (1) genome-wide comparison coupled with genome-guided transcriptome reconstruction; (2) de novo transcriptome assembly and then align the assembled transcript contigs onto the genomes

conducted on human genome-wide comparison (Khaja et al. 2006; Kidd et al. 2010; Li et al. 2010). To generate non-redundant and non-overlapping alignments for each chromosome, we sorted the alignment results of each chromosome in descending order and iteratively compared each hit with the lower scoring alignments. Those alignments that entirely or partially overlap with the coordinates of a higher scoring hit were removed. Through these steps, we obtained the unmatched regions that have no satisfactory match hits from the chromosomes of NCBI build 37.2 for their homologous ones of Celera and HuRef.

### Obtaining gene fragments missing in the NCBI human reference genome

After conducting genome-guided transcriptome reconstruction and genome-wide comparison, we scanned the results and extracted those gene fragments that partially or entirely overlapped with the unmatched regions of Celera and HuRef. In consideration of the limitation of LAST and to minimize the mapping artifacts, we realigned these extracted gene fragments onto NCBI build 37.2 using BLAT (Kent 2002) (version 34) with -fastMap enabled. We further excluded those extracted gene fragments that could be aligned to their homologous chromosomes of NCBI build 37.2 with the threshold of 90 % identity and 90 % coverage. Moreover, those gene fragments that contain unknown bases 'N' were also removed. After filtering, the remaining extracted gene fragments were used as the candidates of the missing gene fragments of NCBI build 37.2 for further analyses.

### Characteristic analyses of the missing gene fragments

To investigate how many of these missing gene fragments could find supporting NCBI ESTs (expressed sequence tags), we then mapped the human ESTs onto these missing gene fragments by BLAT with -trimT option enabled. Considering that these missing gene fragments may not be fully reconstructed by Cufflinks, we used the criteria of $\geq$90 % identity and >50 % coverage. In order to further verify the authenticity of these sequences, we also aligned the missing gene fragments from one human genome assembly of Celera and HuRef to another using BLAT with -fastMap option enabled. To consistent with genome-wide comparison, we used the threshold of 90 % identity and 90 % coverage for this step. We investigated the interspersed repeats and low complexity sequences in all those missing gene fragments using RepeatMasker (Saha et al. 2008) (version: open-3.3.0). RepeatMasker uses a sequence search engine (one of Cross_Match, RMBlast and WUBlast/ABBlast) to perform sequence comparison for searching repeats. This analysis was carried out on the RepeatMasker website http://www.repeatmasker.org/ using default parameter setting.

### De novo transcriptome assembly

Because the genome-guided transcriptome reconstruction method is different from the genome-independent (de novo) transcriptome assembly approach, we also separately de novo assembled the transcriptome of these 16 different human tissues using Velvet (Zerbino and Birney 2008) (version 1.2.03). For de novo transcriptome assembly, a single value for *k*-mer (*k* consecutive nucleotides of a read) is not enough to yield a good assembly (Robertson et al. 2010; Surget-Groba and Montoya-Burgos 2010; Chen et al. 2011c), therefore we used multiple *k*-mer lengths and combined these assemblies together to increase the contiguity of the transcript contigs and improve the overall assembly results. We used the contributed program AssemlyAssembler in the Velvet package to assist the transcriptome assembly. AssemlyAssembler is a wrapper script designed to automate a directed series of assemblies using Velvet assembler. We assembled a series of *k*-mer values ranging from 23 to 39 on each RNA-Seq data set and then combined different *k*-mer assemblies together to yield a final assembly for each tissue. To identify those missing transcript contigs in the NCBI human reference genome, we aligned all these assembled transcript contigs onto NCBI build 37.2 using BLAT with -trimT option enabled. Transcript contigs that could be matched to NCBI build 37.2 with the thresholds of 90 % identity and 90 % coverage were eliminated. Then, we aligned those unalignable transcript contigs onto Celera and HuRef to identify those truly missing transcript contigs using the same criteria of $\geq$90 % identity and $\geq$90 % coverage.

### Conservation analyses

To examine the conservation of those missing gene fragments and whether the identified missing transcript contigs are conserved between human and other mammalian genomes, we aligned them onto the genomes of chimpanzee, macaca and mouse using BLAT (with -fastMap option enabled for the missing gene fragments and with -trimT option enabled for the unalignable transcript contigs). For this step, we used three different coverage cut-offs of 70, 80 and 90 % at a fixed identity threshold of 90 % to survey the conservation of those missing gene fragments and missing transcript contigs. Consequently, we further identified more missing transcript contigs that are unalignable to NCBI build 37.2 but are alignable to chimpanzee, macaca or mouse.

### Annotation of the missing genes

We annotated these missing gene fragments and the identified missing transcript contigs with the NCBI nr database and the protein family database Pfam (only the Pfam-A entries) (Punta et al. 2012). For the missing gene fragments, we retrieved their transcript fragments for annotation. We aligned these transcript sequences to nr using BLASTX and set the cut-offs as $E$ value $<10^{-5}$, bit score $>50$ and identity $>50$ %. We removed the duplicate records and only retained the best records with the smallest $E$ value and then the highest percent identity. We then translated each transcript fragment and transcript contig into all six possible frames and used HMMER (Finn et al. 2011) to identify whether they encoded any of the 13,672 protein families cataloged in Pfam. HMMER used the probabilistic models called profile hidden Markov models to search sequence databases for protein homologs. We required that both the $E$ value of the target and the best-scoring domain found in the sequence should be $<10^{-5}$. In addition, the bias (refer to the biased sequence composition or the single best-scoring domain) for the sequence and best-scoring domain should be lower by at least one order of magnitude compared to their bit score.

## Results

### Genome-guided transcriptome assemblies on genomes of Celera and HuRef

To detect genes that are present in the human genome assemblies of Celera and HuRef, but absent from the NCBI human reference genome build 37.2, we first carried out genome-guided transcriptome assemblies (see Fig. 1) using Celera and HuRef as references to separately reconstruct the transcriptomes of 16 different human tissues (Supplementary File 1, Table S1). First, we separately aligned the transcriptome sequencing reads from each tissue to the human genomic sequences of Celera and HuRef using TopHat (Trapnell et al. 2009). Then we used Cufflinks (Trapnell et al. 2010) to reconstruct the transcriptome of each tissue on Celera and HuRef (see "Materials and methods"). Assembled fragments shorter than 100 bp were eliminated. For each of the Celera and HuRef assemblies, hundreds of thousands of transcribed regions were detected (Supplementary File 1, Table S2), because the bias of sequencing technologies could cause gaps in sequencing coverage which will lead to breaks in transcript reconstruction (Trapnell et al. 2012). Since each transcribed region might be one fragment of a gene, these transcribed regions were denoted as "gene fragments" in the text.

### Comparisons of the Celera and HuRef genomes with the NCBI human reference genome

To obtain the gene sequences present in Celera and HuRef but completely or partially missing from the corresponding chromosomes of NCBI human reference genome build 37.2, we compared the NCBI build 37.2 against Celera and HuRef. We aligned the chromosomes of NCBI build 37.2 onto their corresponding ones in Celera and HuRef using LAST (Kielbasa et al. 2011) (see "Materials and methods"). Although a number of preceding studies have investigated the structural variations of the human genome in different ways (Istrail et al. 2004; Tuzun et al. 2005; Feuk et al. 2006; Khaja et al. 2006; Redon et al. 2006; Korbel et al. 2007; Kidd et al. 2008, 2010; Conrad et al. 2010; Li et al. 2011), our study mainly focuses on those gene sequences missing from the human reference genome. We adopted a series of measures to reduce the possibility of false positive alignments caused by the alignment errors of the aligner and the intricacy of human genomic sequences (see "Materials and methods"). Through comparison, those unmatched regions that have no satisfactory LAST alignments from the chromosomes of NCBI build 37.2 to their homologous counterparts in Celera and HuRef were obtained. We then extracted the assembled gene fragments that completely or partially overlapped with those unmatched regions of each chromosome for Celera and HuRef.

After filtering, 6,788 (total length 3.78 Mb; mean length 557 bp; median length 222 bp) gene fragments from Celera and 5,590 (total length 2.37 Mb; mean length 424 bp; median length 207 bp) from HuRef that are unalignable to their homologous chromosomes of NCBI build 37.2 were obtained (Table 1 and Supplementary File 2). Only a small portion of those missing gene fragments from Celera (234 out of 6,788) and HuRef (177 out of 5,590) contain splice sites, others are single fragments. We then aligned the NCBI human ESTs onto these missing gene fragments to evaluate the percentage of NCBI ESTs supporting. Considering that these gene fragments might not be full length genes, we used the thresholds of $\geq$90 % identity and $\geq$50 % coverage as matching criteria. We observed that 2,956 gene fragments from Celera and 2,348 from HuRef were supported by at least one EST. Since these EST data are limited and all these missing gene fragments were identified with the approach of genome-guided transcriptome assembly and genome-wide comparison, and they were from the regions of the chromosomes in Celera and HuRef but not exist in their homologous chromosomes in NCBI build 37.2, they can be regarded as candidate missing gene sequences of NCBI build 37.2.

**Table 1** Statistics for the possible missing gene fragments in NCBI build 37.2

The missing gene fragments identified from Celera and HuRef were aligned to the human genome assemblies using Blat with the thresholds of 90 % identity and 90 % coverage

| Item | Celera | HuRef |
|---|---|---|
| Count of gene fragments | 6,788 | 5,590 |
| Mean length (bp) | 557 | 424 |
| Median length (bp) | 222 | 207 |
| Total length (bp) | 3778,866 | 2369,277 |
| Count of gene fragments alignable to non-homologous chromosome of NCBI build 37.2 | 1,616 | 535 |
| Count of gene fragments partially and entirely unalignable to NCBI build 37.2 | 5,172 | 5,055 |
| Count of gene fragments alignable to another human genome assembly | 5,822 | 3,403 |

## Characteristics of the missing gene sequences

We found that 1,616 gene fragments from Celera and 535 from HuRef could not be aligned to their homologous chromosomes in the NCBI build 37.2 but were alignable to other non-homologous chromosomes in the NCBI build 37.2. Transposable elements (TEs) could duplicate and insert themselves randomly around the human genome and alter the genome size (Lorenc and Makalowski 2003; Wicker et al. 2007; Conrad et al. 2010). Besides, the chromosome translocation that arises from the rearrangements between non-homologous chromosomes also could result in extra or missing genetic information (Mackie Ogilvie and Scriven 2002; Oliver-Bonet et al. 2002). Thus, the fact that these missing gene fragments are absent from their homologous chromosome of NCBI build 37.2 but exist in other non-homologous chromosomes indicates an intriguing phenomenon that might have been a result of transposition and/or translocation.

The remaining gene fragments are partially or entirely unalignable (<90 % identity and/or <90 % coverage) to NCBI build 37.2. We observed that 4,454 gene fragments from Celera and 4,283 from HuRef could be partially (<90 % coverage) matched onto the NCBI build 37.2. However, 718 gene fragments from Celera and 772 from HuRef still could not find any significant matches on NCBI build 37.2. The results suggest that the majority of these remaining gene fragments (86.12 % for Celera and 84.73 % for HuRef) may only miss their partial sequences from NCBI build 37.2 and they may result from the deletions of the genomic sequence of NCBI build 37.2. To check whether these partially alignable and entirely unmatchable missing gene fragments could be located onto NCBI build 37.2, we extended 5 kb at the both ends of them using their corresponding chromosomes of Celera and HuRef as a reference and then mapped the extended sequences onto NCBI build 37.2 (Fig. 2a). We found that 1,104 extended missing gene fragments from Celera could be uniquely remapped onto the chromosomes of NCBI build 37.2 and 1,310 for HuRef. As shown in Fig. 2b, one example of our identified missing gene fragment from Celera is that its sequence also exist in HuRef but still absent from the updated human reference genome of GRCh37.p5. Whereas others still could not find appropriate match positions on NCBI build 37.2 and they might be caused by large sequence deletion or other structural variations between NCBI build 37.2 and Celera or HuRef.

We then screened the interspersed repeats and low complexity sequences in all those missing gene fragments using RepeatMasker (Saha et al. 2008). We found that 1,829,351 bp (48.41 %) of the missing gene fragments from Celera were annotated as repeats by RepeatMasker and 1,112,770 bp (46.83 %) for HuRef. Those annotated repeats could be classified into diverse classes and families (Supplementary File 1, Table S3). The transposable elements are the most abundant type (34.67 % of the sequence was annotated as having significant match with known TEs for Celera and 31.21 % for HuRef) among those annotated repeats. The annotated TEs include SINE (Short Interspersed Elements), LINE (Long Interspersed Element) and LTR (Long Terminal Repeat). The repeat annotations also showed that these missing gene fragments may harbor dozens of small RNAs. Others are the satellites, simple repeats, DNA elements (such as hAT-Charlie, TcMar-Tigger), low complexity sequences, etc. Consequently, a significant portion of those missing gene fragments from Celera and HuRef contain transposons. The results further verified our above inference that transposition is a reason that causes the genes missed from their homologous chromosomes between different human genome assemblies.

In order to investigate the existence of the missing gene fragments identified from Celera and HuRef, we separately aligned the missing gene fragments from one human genome assembly to another with threshold of 90 % identity and 90 % coverage. For the 6,788 missing gene fragments from Celera, 5,822 of them (90.99 % of the sequence) are alignable to HuRef. For the 5,590 missing gene fragments from HuRef, 3,403 of them (70.20 % of the sequence) could be matched to Celera. Consequently, most of the gene fragments from one human genome assembly are contained in the other one, further validating that they are

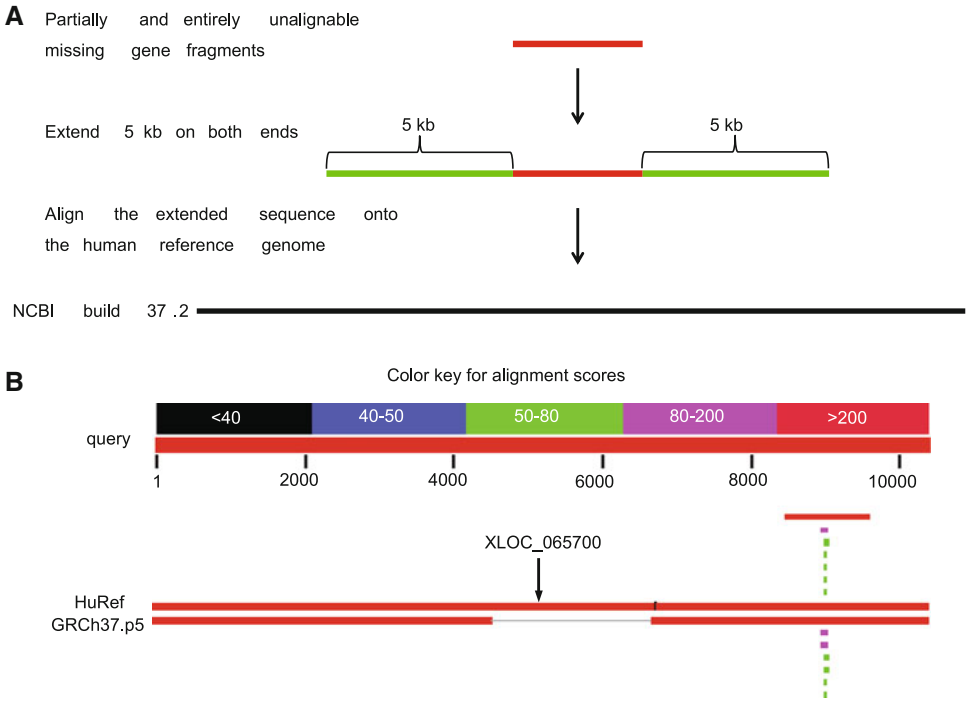


**Fig. 2** Realignment of the extended missing gene fragments to the human reference genome. **a** The partially and entirely unalignable missing gene fragments were first extended 5 kb on their both ends using their corresponding chromosomes of Celera and HuRef as a reference. Then these extended missing gene fragments were realigned onto the NCBI human reference genome build 37.2. **b** An example of a gene fragment missing from the NCBI human reference genome. We extended 5 kb on both sides of XLOC_065700 and then searched it in the NCBI blast webserver. This graph shows that the sequence of gene fragment XLOC_065700 (363 bp) on Celera is also present in HuRef but is absent from the updated NCBI human reference genome GRCh37.p5

likely the missing sequences of NCBI build 37.2. Furthermore, both Celera and HuRef have their own gene fragments that are unmatchable to each other. The missing gene fragments from Celera (or HuRef) that are unalignable to HuRef (or Celera) also belong to the missing part of HuRef (or Celera). Hence the results also illustrate that none of NCBI build 37.2, Celera and HuRef is fully completed, each of them misses a number of human genes in different degrees.

## De novo assembly of the transcriptomes from 16 different human tissues

To more comprehensively investigate the missing genes in the NCBI human reference genome, we also separately de novo assembled the transcriptome of 16 different tissues with Velvet (Zerbino and Birney 2008) (see "Materials and methods"). Gene expression usually exhibits temporal and spatial specificities, meaning that each tissue type expresses a specific set of genes in addition to those ones commonly expressed across tissue types or development stages. Moreover, genome-guided transcriptome assembly and de novo transcriptome assembly are two different methods for transcriptome reconstruction, and each has its own advantages and disadvantages (Chen et al. 2011b; Garber et al. 2011). Besides, none of these three human genome assemblies of NCBI build 37.2, Celera and HuRef are fully complete and there may still be novel genes missing from these three human genomes. Accordingly, it is necessary to carry out de novo transcriptome assembly to further uncover those missing genes of the human genome.

We conducted assembly with a series of *k*-mer values ranging from 23 to 39 and then combined assemblies from different *k*-mer values together to yield a final assembly for each tissue. The assembled transcript contigs shorter than 100 bp were excluded. As expected, the combinational strategy significantly improved the transcriptome assembly results of each tissue which are better than the results of any single *k*-mer (Supplementary File 1, Table S4). Then, we aligned all the assembled contigs from the 16 different human tissues to NCBI build 37.2 and removed those that are alignable. With the criteria of $\geq$90 % identity and $\geq$90 % coverage, we found that on average 3.45 % of these assembled transcript contigs are unalignable (Table 2). Most of the unalignable transcript contigs could be partially matched to NCBI build 37.2, but for a small portion of them there is no any significant match to the NCBI build 37.2. Similar to above-identified missing gene fragments, the results also suggest that the majority of unmatchable transcript contigs may only be partially absent from NCBI

**Table 2** Statistics for the assembled transcript contigs from 16 different human tissues

| Tissues | Number of contigs | Unmapped to NCBI build 37.2 | Mapped to Celera | Mapped to HuRef | Mapped to Chimpanzee | Mapped to Macaca | Mapped to Mouse | Sum of mapped (non-redundant) | Mapped length (bp) |
|---|---|---|---|---|---|---|---|---|---|
| Thyroid | 151,886 | 5,226 | 610 | 601 | 563 | 401 | 51 | 1,184 | 209,073 |
| Testes | 146,936 | 5,673 | 665 | 707 | 636 | 457 | 40 | 1,398 | 222,778 |
| Ovary | 190,959 | 5,270 | 624 | 591 | 602 | 486 | 42 | 1,236 | 233,258 |
| White blood cells | 72,989 | 2,475 | 377 | 344 | 329 | 268 | 39 | 723 | 122,150 |
| Skeletal muscle | 65,447 | 2,495 | 378 | 345 | 346 | 258 | 60 | 743 | 113,899 |
| Prostate | 114,055 | 3,757 | 481 | 500 | 457 | 351 | 56 | 1,015 | 161,290 |
| Lymph node | 118,219 | 5,436 | 647 | 633 | 571 | 423 | 62 | 1,289 | 205,726 |
| Lung | 107,748 | 4,253 | 602 | 550 | 538 | 397 | 69 | 1,161 | 181,375 |
| Adipose | 110,897 | 4,202 | 527 | 521 | 491 | 370 | 49 | 1,020 | 171,081 |
| Adrenal | 219,048 | 6,046 | 689 | 746 | 691 | 559 | 70 | 1,417 | 257,609 |
| Brain | 194,886 | 5,547 | 685 | 709 | 627 | 516 | 55 | 1,368 | 259,981 |
| Breast | 138,175 | 4,755 | 573 | 555 | 540 | 402 | 55 | 1,128 | 210,896 |
| Colon | 104,451 | 4,755 | 539 | 479 | 475 | 373 | 67 | 1,047 | 170,488 |
| Kidney | 118,851 | 5,086 | 505 | 532 | 468 | 411 | 53 | 1,072 | 174,814 |
| Heart | 120,460 | 3,216 | 432 | 453 | 412 | 319 | 28 | 891 | 150,982 |
| Liver | 87,893 | 3,062 | 471 | 451 | 403 | 329 | 54 | 894 | 147,076 |

The assembled transcript contigs from the 16 different human tissues were first aligned onto the NCBI human reference genome build 37.2 using Blat with the criteria of ≥90 % identity and ≥90 % coverage. Those unalignable transcript contigs were then aligned to the genomes of Celera, HuRef, chimpanzee, macaca and mouse by employing Blat with the thresholds of 90 % identity and 90 % coverage

build 37.2 due to misassemblies/polymorphisms. We then aligned all those unmatchable transcript contigs onto the human genome assemblies of Celera and HuRef. Using the same cut-off values of ≥90 % identity and ≥90 % coverage, averagely, 12.36 and 12.24 % of the unalignable transcript contigs were successfully mapped onto Celera and HuRef (Table 2). We also observed that a portion of the unmatchable transcript contigs could be aligned to multiple loci of the chromosomes of Celera and HuRef, suggesting that they might result from copy number variations between NCBI build 37.2 and Celera and HuRef.

To facilitate the comparison of those missing gene sequences identified by the strategy of genome-guided transcriptome assembly coupled with genome-wide comparison and the strategy of de novo transcriptome assembly with mapping contigs to the genome, we merged those uniquely alignable transcript contigs from 16 different tissues mapped onto Celera and HuRef according to their matching coordinates. We found that the great majority of the transcript contigs that mapped onto Celera and HuRef are not overlapped with the genomic coordinates of those missing gene fragments, except for a very small portion that is partially or entirely overlapped. We reasoned that this is owing to the methodological differences (strengths and weaknesses) between these two different identification strategies. Our results show that these two different strategies are complementary to each other and combing them together enabled us to more comprehensively uncover those missing genes in the human reference genome.

## Conservation of the missing genes

To investigate the conservation of these identified missing gene sequences, we aligned the missing gene fragments identified from Celera and HuRef and the de novo assembled transcript contigs that are unalignable to NCBI human build 37.2 onto the genomes of chimpanzee, macaca and mouse. Considering that most of these missing gene fragments and transcript contigs are partially alignable to NCBI build 37.2, we used the fixed identity threshold of 90 % and three different coverage cut-offs to survey their conservation (Fig. 3). The number of missing gene fragments identified from Celera and HuRef that could be mapped to chimpanzee and macaca decreased sharply with the increased coverage threshold of 70, 80 and 90 %. Whereas very few of the missing gene fragments from Celera and HuRef could be aligned to mouse using these three coverage cut-offs. Employing different criteria, a

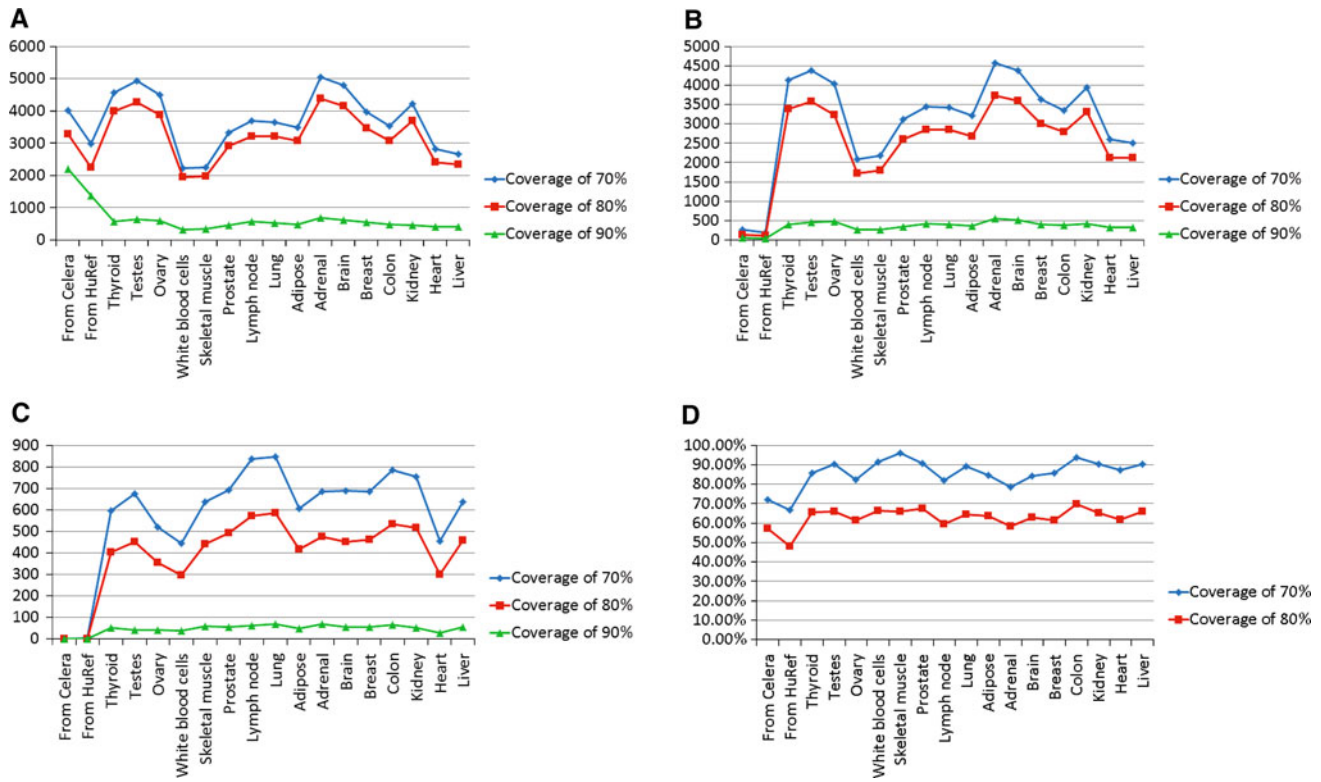


**Fig. 3** Different coverage cut-offs for the conservation analysis of the missing genes. **a** Number distribution of the missing gene fragments and transcript contigs that could be aligned to chimpanzee. Three different coverage cut-offs (70, 80 and 90 %) were used with the fixed identity threshold of 90 %. "From Celera" and "From HuRef" represent the missing gene fragments identified from the human genome assemblies of Celera and HuRef, respectively. Others were the assembled transcript contigs from 16 different human tissues that are unalignable to NCBI build 37.2. **b** Number distribution of the missing gene fragments and transcript contigs that could be mapped to macaca. **c** Number distribution of the missing gene fragments and transcript contigs that are alignable to mouse. **d** Percentage distribution of the identified missing gene fragments and transcript contigs that could be aligned to NCBI build 37.2 with the coverage cut-offs of 70 and 80 %

large number of the missing gene fragments from Celera and HuRef are alignable to chimpanzee, suggesting that they are conserved between human and chimpanzee. And fewer of them are conserved between human and macaca. However, in contrast, the great majority of them seem to divergent from mouse.

To examine the conservation of the assembled transcript contigs that are unmatchable to NCBI build 37.2, we also separately used the coverage cut-offs of 70, 80 and 90 %. A notable reduction was observed in the number of these contigs that are alignable to chimpanzee, macaca and mouse between using the coverage threshold of 80 and 70 %, but much greater decline was found between employing the criteria of 90 and 80 % coverage (Fig. 3a, b, c). Although lower coverage cut-off can enable more contigs align to these mammalian genomes, we finally chose the stringent coverage of 90 % as the criterion for identifying the truly missing transcript contigs to minimize the false positives. Interestingly, the majority of these identified missing gene fragments and missing transcript contigs could be mapped to NCBI build 37.2 with the coverage cut-offs of 70 and 80 % (Fig. 3d), suggesting that they may only miss partial nucleotides of their sequences from NCBI build 37.2. For each tissue, on average 11.44 % of them could be aligned to the chimpanzee genome; 8.87 % to the macaca genome and only 1.19 % to the mouse genome (Table 2). Overall, 4.07 % of those unmatchable transcript contigs from 16 different human tissues are alignable to chimpanzee and macaca, indicating that they are potentially conserved among human, chimpanzee and macaca. Only a small part of these transcript contigs are alignable to the mouse genome, suggesting that most of them are divergent between human and mouse. Through this process, we finally identified that an average 24.68 % (187.03 kb per tissue) of the transcript contigs that are unalignable to NCBI build 37.2 could be mapped onto at least one of the five genomes of Celera, HuRef, chimpanzee, macaca and mouse (Table 2 and Supplementary File 3).

The missing gene fragments and missing transcript contigs that are alignable to chimpanzee, macaca, or mouse suggest that they are conserved between human and other mammals. Consequently, these conserved missing gene sequences should not be the individual or population-

specific genes and they probably have important unknown functions. The misassemblies and/or the genetic variations of the genome could be the main reasons that caused them missed from the human reference genome. About 75.32 % of those unmatchable transcript contigs remain unalignable to any one genome of Celera, HuRef, chimpanzee, macaca or mouse using the threshold of 90 % identity and 90 % coverage. These remaining unevaluated contigs may result from those aspects, including, but not limited to, the rigorous threshold we used, the limited genomic sequences to validate, genetic variations (such as SNPs, insertions, deletions, or other variations), the modifications of RNAs in the post-transcriptional process leading to the sequence changed (e.g. RNA-editing), some transcriptome assembly errors and contamination.

### Functions of the missing genes

To gain insights into the functions of these missing gene fragments and missing transcript contigs, we examined their possible derived proteins and the known domains contained in those proteins (see "Materials and methods"). We observed that 29.88 % of the missing gene fragments for Celera and 27.08 % for HuRef could find satisfied hits in the NCBI nr database, whereas 70.65 % of the identified missing transcript contigs from the 16 different human tissues have found acceptable records in nr database. These matched proteins are primarily from human, chimpanzee, macaca, gibbon and mouse. However, we found that most of these matched protein records were marked as hypothetical, predicted, or unnamed proteins, illustrating that their biological functions are still unclear.

In order to detect whether the derived proteins from those missing genes contain known functional domains, we inspected whether they could encode any of the 13,672 protein families cataloged in the protein family database Pfam (Punta et al. 2012). The missing transcript fragments and missing transcript contigs were translated into related protein sequences in all six possible frames and then used HMMER (Finn et al. 2011) to search against Pfam. For the missing gene fragments, 9.91 % for Celera and 9.93 % for HuRef could find statistically significant hits with Pfam-A families (Fig. 4a, b). But a large portion of these matched hits were the putative binding domains. Only 89 and 67 unique protein domains (total 120 unique domains) were, respectively, identified in the missing gene fragments from Celera and HuRef. While on average 17.66 % of the identified missing transcript contigs for each of the 16 different human tissues could be annotated with Pfam-A families (Fig. 4c) and totally 1,174 unique protein domains were detected (Fig. 4d). Together, 1,233 functional protein domains were found in these missing gene fragments and missing transcript contigs. However, the functions of the majority of these missing gene fragments and missing transcript contigs still cannot find statistically significant match records with Pfam-A entries. One main reason could be that these missing gene sequences are partially or entirely absent from the human genome and thus have rarely been studied. Consequently, only a small part of them could be annotated according to their sequence homologs with known protein domains.

## Discussion

We combined two different strategies together to identify and characterize the missing gene sequences of the human reference genome. Using the genome-guided transcriptome reconstruction method and the genome-wide comparison approach, we identified that megabases of gene fragments from Celera (3.78 Mb) and HuRef (2.37 Mb) are either missed from their homologous chromosomes of the NCBI human reference genome build 37.2 or partially or completely absent from the human reference. With the de novo transcriptome reconstruction strategy, we revealed that on average 187.03 kb per tissue from 16 different human tissues are unalignable to NCBI build 37.2 but can be matched onto at least one of the five genomes of Celera, HuRef, chimpanzee, macaca and mouse. Our findings imply that the majority of these missing gene fragments and missing transcript contigs may only miss part of their sequences from NCBI build 37.2 due to the genome misassemblies and polymorphisms. Furthermore, we found that most of the missing gene fragments from Celera and HuRef are not overlapped with those identified missing transcript contigs that are uniquely alignable to Celera and HuRef. This is likely due to the distinct differences between these two different strategies to identify the missing genes of the human reference genome. Our results show that integrating these approaches together can enable us to more comprehensively uncover the missing genes of the human genome.

Genome-wide comparison between different genome assemblies can enable us to obtain those unmatched regions of each assembly. These unmatched regions are the candidates for further verifying whether they are the truly missing part of the genome. Moreover, the genome-guided transcriptome reconstruction method provides an efficient way to detect the genes in those unmatched genomic sequences. However, genome-guided transcriptome reconstruction method is directly related to the short-read mapping step. The RNA-Seq reads may fail to map onto those genomic regions that have sequence variations and misassemblies. In contrast, de novo transcriptome assembly is independent of the genome, so it is not limited by the quality of the genome. It could help us to further uncover

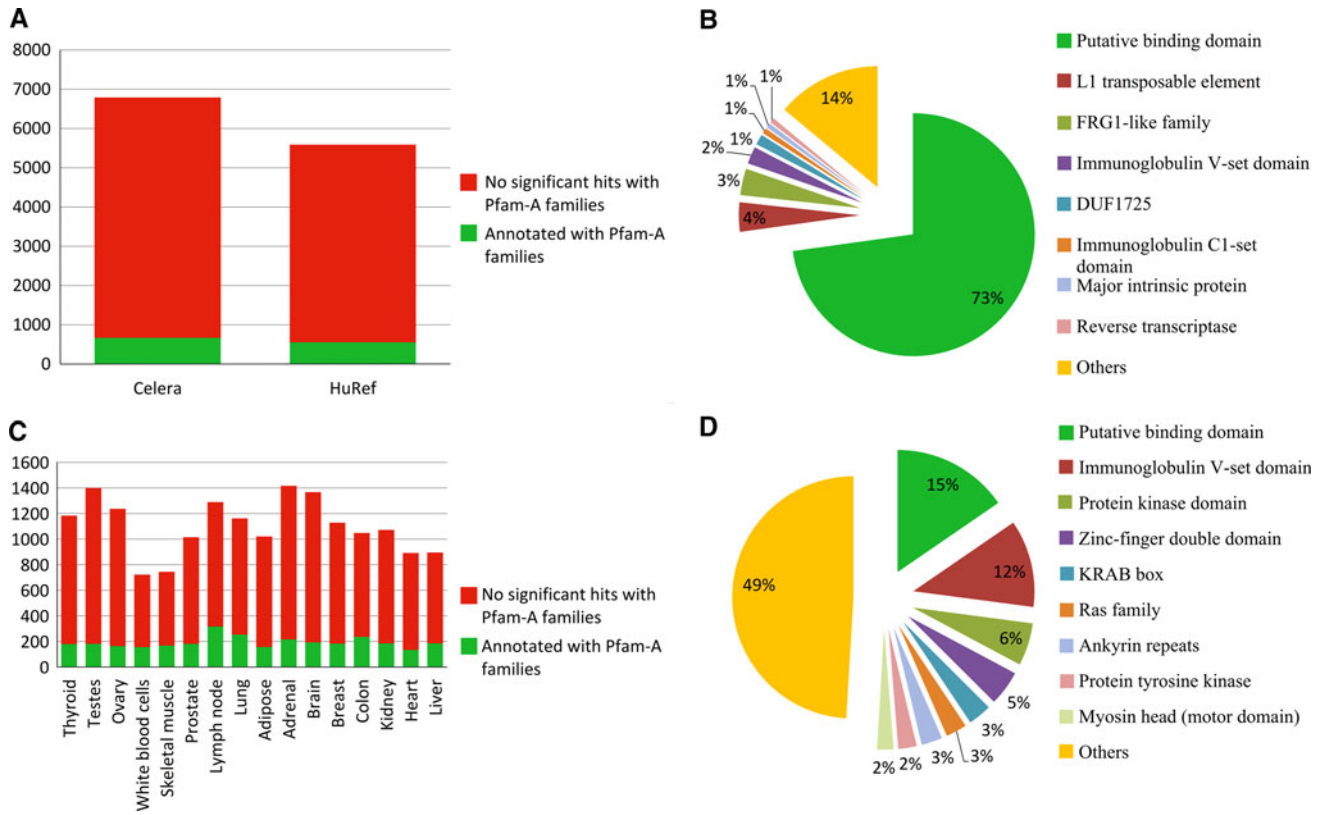


**Fig. 4** Annotation of missing gene fragments and missing transcript contigs with Pfam. **a** Annotation distribution of the missing gene fragments identified from Celera and HuRef with Pfam-A families. **b** Percent of the protein domains detected in the missing gene fragments from Celera and HuRef. The counts of the identified protein domains were sorted in descending order and displayed in turn. A total of 120 unique protein domains were detected in those missing gene fragments identified from Celera and HuRef. **c** Annotation distribution of the missing transcript contigs in 16 different human tissues. **d** Percent of the protein domains detected in the missing transcript contigs of all 16 tissues. Totally, 1,174 unique protein domains were identified in all those missing transcript contigs

the genes missing from related genome. However, de novo assembly is very sensitive to repetitive sequences and cannot successfully assemble those repetitive regions longer than the length of single-end sequencing reads or the insert-length for paired-end reads. In general, the first strategy is suitable for the organisms that have at least two available genome assemblies, while the second strategy is not subjected to the genome and can complement the first strategy.

We found that a notable portion of the missing gene fragments from Celera and HuRef are absent from their homologous chromosomes of NCBI build 37.2 but are alignable to other non-homologous chromosomes. Moreover, the missing gene fragments have been shown to contain many transposons. Consequently, our results indicate that transposition is one reason that causes the genes missing from the homologous chromosomes between different human genome assemblies. On the other hand, translocation could be another potential factor because it can result in extra or missing genetic information between non-homologous chromosomes. We also observed that a significant part of the missing gene fragments and missing transcript contigs could be aligned to chimpanzee, macaca and mouse, indicating that they are conserved between human and other mammals. Therefore, they should not be the individual- and population-specific genes. Furthermore, a significant number of those missing gene fragments that are partially alignable or entirely unmatchable to NCBI build 37.2 could be uniquely relocated onto NCBI build 37.2 with their extended sequences (extended 5 kb on both ends of their sequence using Celera or HuRef as reference). These results illustrate that genome misassembly is one of the reasons that caused genes partially or entirely missed from the human genome. In addition, some transcript contigs that are unalignable to NCBI build 37.2 could find multiple mapping loci on Celera or HuRef, suggesting that they are caused by the copy number variations. Other genomic rearrangements can also result in the changes of the genomic structure, they may also involve in the event that lead to some genes missing from the human genome as well (Zerbino et al. 2012).

A total of 1,233 known protein domains could be detected in these missing gene fragments and missing transcript contigs. Our findings indicate that the majority of

those missing genes of the human genome have not been studied and little is known about their biological functions. Our analyses also suggest that many of those missing gene fragments and missing transcript contigs are conserved. Therefore, these missing genes probably have important biological functions. In order to reveal the biological functions of these missing genes, great efforts are still needed to reach the objective.

Our results also indicate that none of the human genome assemblies of NCBI build 37.2, Celera and HuRef are fully completed, with each of which missing a number of human genes. Moreover, most of the missing gene fragments from Celera or HuRef are present in another, further validating the authenticity of these missing sequences. Using the threshold of 90 % identity and 90 % coverage, about 75.32 % of the transcript contigs that are unmatchable to the NCBI build 37.2 still cannot be aligned to any one of Celera, HuRef, chimpanzee, macaca and mouse. These unevaluated unmatchable transcript contigs may include a portion of true ones not supported by the genomes we used. With the dramatic decrease in cost of DNA sequencing, it can be anticipated that more and more human genomes will be de novo assembled in the future. The high-quality human genome assemblies will be valuable resources for researchers to continually identify and characterize genes missing from the human genome annotations (Baker 2012).

The missing genes of the human genome are an integral part of the complete set of human genes. Their absence in the genome annotation will hinder us from comprehensively carrying out various related studies and interpret the biological meaning. In addition, the human genome is one of the most studied genomes, so its incompleteness implies that other assembled complex eukaryotic genomes are also most likely incomplete. For example, a 1001 Genome Project to sequence multiple *Arabidopsis thaliana* populations also revealed the gene gain and loss between genomes (Cao et al. 2011). Our study provides guidance for future similar researches to explore the genes missing from other eukaryotic genomes. With the innovation of both sequencing technologies and bioinformatics approaches, it can be anticipated that the missing genes could be eventually identified, the gaps in the genomes will be filled and the mis-assembly problems in the genomes will be solved. All of the progresses will help us reach the goal to construct a fully completed gene set for each organism and better understand the genome structure.

**Acknowledgments** We thank Danielle and Jean Thierry-Mieg from NCBI, Kangping Yin, Hui Liu and Jiang Li for helpful discussions. This work was supported by the National 973 Key Basic Research Program (Grant Nos. 2010CB945401 and 2012CB910400), the National Natural Science Foundation of China (Grant No. 31171264, 31071162 and 31240038), the Science and Technology Commission of Shanghai Municipality (11DZ2260300) and the Graduate School of East China Normal University.

**Conflict of interest** None.

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Baker M (2012) De novo genome assembly: what every biologist should know. Nat Method 9:333–337

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25:1915–1927

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C et al (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956–963

Chen G, Li R, Shi L, Qi J, Hu P, Luo J, Liu M, Shi T (2011a) Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. BMC Genomics 12:590

Chen G, Wang C, Shi T (2011b) Overview of available methods for diverse RNA-Seq data analyses. Sci China Life Sci 54: 1121–1128

Chen G, Yin K, Wang C, Shi T (2011c) De novo transcriptome assembly of RNA-Seq reads with different strategies. Sci China Life Sci 54:1129–1133

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P et al (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712

Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

Eichler EE, Clark RA, She X (2004) An assessment of the sequence gaps: unfinished business in a finished human genome. Nat Rev Genet 5:345–354

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Method 8:469–477

Harris RS (2007) Improved pairwise alignment of genomic DNA. PhD Thesis, The Pennsylvania State University, Pennsylvania

Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, Walenz B, Shatkay H, Dew I, Miller JR et al (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci USA 101:1916–1921

Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664

Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, Rafiq MA, Qian C, Shago M, Pantano L et al (2006) Genome assembly comparison identifies structural variants in the human genome. Nat Genet 38:1413–1418

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F et al (2008)

Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64
Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G et al (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Method 7:365–371
Kielbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21: 487–493
Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921
Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al (2007) The diploid genome sequence of an individual human. PLoS Biol 5:e254
Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J et al (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63
Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H et al (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat Biotechnol 29:723–730
Lorenc A, Makalowski W (2003) Transposable elements and vertebrate protein diversity. Genetica 118:183–191
Mackie Ogilvie C, Scriven PN (2002) Meiotic outcomes in reciprocal translocation carriers ascertained in 3-day human embryos. Eur J Hum Genet 10:801–806
Marguerat S, Bahler J (2010) RNA-seq: from technology to biology. Cell Mol Life Sci 67:569–579
Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517
Nagalakshmi U, Waern K, Snyder M (2010) RNA-Seq: a method for comprehensive transcriptome analysis. In: Frederick M Ausubel et al (eds) Current protocols in molecular biology. Chaps 4: Unit 4 11, pp 11–13
Noe L, Kucherov G (2005) YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res 33:W540–W543
Oliver-Bonet M, Navarro J, Carrera M, Egozcue J, Benet J (2002) Aneuploid and unbalanced sperm in two translocation carriers: evaluation of the genetic risk. Mol Hum Reprod 8:958–963
Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12:87–98
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J et al (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301
Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W et al (2006) Global variation in copy number in the human genome. Nature 444: 444–454
Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ et al (2010) De novo assembly and analysis of RNA-seq data. Nat Method 7:909–912
Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36:2284–2294
Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res 20:1432–1440
Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515
Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578
Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D et al (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–732
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291:1304–1351
Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982
Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829
Zerbino DR, Paten B, Haussler D (2012) Integrating genomes. Science 336:179–182