nature biotechnology

Building the sequence map of the human pan-genome

Ruiqiang Li[1,2,7], Yingrui Li[1,7], Hancheng Zheng[1,3,7], Ruibang Luo[1,3,7], Hongmei Zhu[1], Qibin Li[1], Wubin Qian[1], Yuanyuan Ren[1], Geng Tian[1], Jinxiang Li[1], Guangyu Zhou[1], Xuan Zhu[1], Honglong Wu[1,6], Junjie Qin[1], Xin Jin[1,3], Dongfang Li[1,6], Hongzhi Cao[1,6], Xueda Hu[1], Hélène Blanche[4], Howard Cann[4], Xiuqing Zhang[1], Songgang Li[1], Lars Bolund[1,5], Karsten Kristiansen[1,2], Huanming Yang[1], Jun Wang[1,2] & Jian Wang[1]

Here we integrate the *de novo* assembly of an Asian and an African genome with the NCBI reference human genome, as a step toward constructing the human pan-genome. We identified ~5 Mb of novel sequences not present in the reference genome in each of these assemblies. Most novel sequences are individual or population specific, as revealed by their comparison to all available human DNA sequence and by PCR validation using the human genome diversity cell line panel. We found novel sequences present in patterns consistent with known human migration paths. Cross-species conservation analysis of predicted genes indicated that the novel sequences contain potentially functional coding regions. We estimate that a complete human pan-genome would contain ~19–40 Mb of novel sequence not present in the extant reference genome. The extensive amount of novel sequence contributing to the genetic variation of the pan-genome indicates the importance of using complete genome sequencing and *de novo* assembly.

The Human Genome Project[1] established the foundation for human genomics studies. Subsequent analyses unveiled genetic variations and identified their effects on phenotypic diversity and differences in disease susceptibility[2]. Guided by the National Center for Biotechnology Information (NCBI) reference genome, initial studies of human genetic variation focused largely on identifying[3] and cataloging[4,5] single-nucleotide polymorphisms (SNPs) and studying their association to human diseases[6]. Structural variation (which is thought to contribute more variant sequences than SNPs) has also been extensively identified and analyzed in the human genome[7–10].

The availability of a number of individual human genomes[11–15] has provided an unprecedented opportunity to investigate detailed genetic differences at the individual level. Preliminary analyses have revealed that these genomes contain sequences that could not be mapped onto the human reference genome (novel sequences), resulting in the proposal that the majority of these sequences likely belong to the gap regions in the current version of the human genome assembly[12]. When fosmid clones from HapMap samples were sequenced, 525 sequences were identified that mapped instead to highly polymorphic structural variant regions, among which 172 sequences appeared to be specific to the individual rather than to be sequences missing as a result of gaps in the reference genome[8]. Thus, the variable part of the sequence composition in the human genome (individual-specific sequences) may contribute considerable sequence divergence in addition to substitutions, base-pair level indels, rearrangements or copy number changes present in the commonly shared part of human genomes. Substantial effort has been taken to refine genomic reference sequences (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/), but thorough genome-wide characterization is still necessary to gain a comprehensive understanding of individual-specific sequences.

These intriguing findings and the limited amount of information about the extent and range of diversity that these individual-specific sequences contribute to human genetic variation prompted us to begin to build the human pan-genome—that is, the nonredundant collection of all human DNA sequence present in the entire human population. We assembled *de novo* the Asian and African complete individual genome sequences and compared them to the NCBI reference human genome. Our findings showed that human genomes contain a large amount of novel sequence that is both population and individual specific. Additional analyses allowed us to investigate the amount of sequence variation that is expected to exist between any two individuals as well as obtain information about the presence of potentially functional genetic elements within these novel sequences.

Our study also shows that combining individual-specific sequences with shared core human genome sequences will enable the creation of a human pan-genome that will be important for better understanding personal genomes and their use in medical genomics studies. Based on our findings here, it is also clear that establishing a complete human pan-genome will require using extensive sequencing data rather than relying primarily on array-based technologies that are dependent on the current reference genome.

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [3]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China. [4]Fondation Jean Dausset, Centre d'Étude du Polymorphisme Humain (CEPH), Paris, France. [5]Institute of Human Genetics, University of Aarhus, Aarhus, Denmark. [6]Genome Research Institute, Shenzhen University Medical School, Shenzhen, China. [7]These authors contributed equally to this work. Correspondence should be addressed to Jun Wang (wangj@genomics.org.cn) or Jian Wang. (wangjian@genomics.org.cn).

Received 21 October; accepted 30 November; published online 7 December 2009; doi:10.1038/nbt.1596

Table 1 Summary of YH and NA18507 novel sequence identification

	YH		NA18507	
Step	**Number**	**Total length (bp)**	**Number**	**Total length (bp)**
1. Genome assembly	185,086	2,874,204,399	314,877	2,682,734,144
2. Sequences nonexistent in the reference genome (NCBI build 36)	7,211	5,125,070	7,330	4,798,833
3. Individual-specific novel sequence	6,144	4,957,235	7,305	4,790,170
3.1. Aligned to the sequence of at least one of the following human genomes	5,193	4,072,922	6,556	4,280,560
a. On NA18507/YH genome	2,626	2,655,416	4,514	3,412,855
b. On Watson's genome (raw reads)	2,221	348,668	2,229	343,668
c. On Venter's genome	3,665	1,011,133	4,427	1,357,377
d. Aligned to the 363 kb gap regions	105	87,199	147	77,255
e. Aligned to the GenBank human clones	2,871	2,697,530	3,709	2,719,344
3.2. Aligned to other mammal genomes	507	311,467	362	176,535
3.3. Unknown	444	272,424	387	184,142

We compared the assembly of YH (Asian) and NA18507 (African) genome sequences against the human reference genome (NCBI build 36) and identified sequences in YH and NA18507 not present in the reference genome. Potential plant or microbial contaminations were filtered, and the remaining sequences that could not be aligned to the reference genome were defined as individual-specific sequences. The fraction of sequences that aligned to other available human sequences (>90% identity) or showed homology with other mammalian genomes (Blast, 1e-20) provided evidence that these were human sequences. The 26 gaps (363 kb) in the human reference genome sequence were previously closed as described[17].

RESULTS

Short-read assembly and novel sequence detection

For our analysis, we used the raw data of the Asian (YH) genome that we previously sequenced and the raw data of the African (NA18507) genome that we downloaded from NCBI. These data were both produced using the Illumina Genome Analyzer (GA) and consisted of 117.7 Gb and 135 Gb of sequencing reads, respectively, with read-lengths of ~35 bp[13,14]. We also used an additional 82.5 Gb paired-end reads that we recently generated from YH with library insert sizes ranging from 200 bp to 9.6 kb (**Supplementary Table 1**), raising the total amount of YH sequence data used to 200.2 Gb.

We carried out *de novo* short-read assembly (Online Methods) and obtained a total assembled sequence size of 2.87 Gb for YH and 2.68 Gb for NA18507 (**Table 1**). The N50 scaffold size of the two genomes was, respectively, 446.3 kb and 61.9 kb, and the N50 contig size was 7.4 kb and 6.0 kb.

We aligned the YH and NA18507 assembly scaffolds against the NCBI human reference genome[16] (Online Methods) and found that 5.1 and 4.8 Mb of the sequence (see Discussion), respectively, was absent in the reference genome, where absent sequences were defined to be those >100 bp long and with <90% identity. We filtered the novel sequence to eliminate possible contamination by comparing these sequences to all known plant and microbe genomes, which resulted in a final total of 5.0 Mb of novel sequence in the YH genome and 4.8 Mb in NA18507 (**Table 1** and **Supplementary Data Sets 1** and **2**). We also assessed whether using a lower (80%) identity cutoff would substantially alter the results and found little difference, indicating that the identified novel sequences have no close homologs in the reference genome (**Supplementary Fig. 1**). In this study, we define the term 'novel sequences' to denote sequences that are present in at least one human individual but not in the NCBI reference genome.

We then bootstrapped subsets of read data for assembly and checked the coverage of these novel sequences. At a sequencing depth of 40-fold or above, >95% of all novel sequences could be assembled with the subset data (**Supplementary Fig. 2**), which indicates that the two assemblies covered essentially the complete set of nonrepeat novel sequences of the donors' genomes, including the unique sequences and consensus sequences of repeats. The sum of novel sequences and the NCBI reference genome provided a human pan-genome for further analyses.

Initial characterization of novel sequences

We first validated the identified novel sequences by comparing them to all previously published human genome assemblies and human genome clone sequences to search for any matches or homologs (>90% identity). We found that ~2.7 Mb of YH (54%) and NA18507 (56%) novel sequences overlapped with each other; 348.7 kb of the YH and 343.7 kb of the NA18507 sequences could be aligned to Watson's raw sequencing reads; 1.0 Mb of the YH and 1.4 Mb of the NA18507 sequences could be aligned to Venter's (so-called HuRef)[11] genome; 2.7 Mb of each could be aligned to known human clones deposited in GenBank; and 87.2 kb of the YH and 77.3 kb of the NA18507 sequences could be aligned to the 26 recently closed gaps on the NCBI reference genome[17]. In all, 4.1 Mb (82.1%) of the YH and 4.3 Mb (89.4%) of the NA18507 novel sequences could be aligned to other human sequences, indicating that these are valid DNA sequences (**Table 1**).

We further compared the remaining unaligned novel sequences to all available sequenced mammalian genomes, and found that 311.5 kb of the YH and 176.5 kb of the NA18507 sequences had homologs in these genomes (E-value = 1e-20). In all, only 272.4 kb (5.5%) of the YH and 184.1 kb (3.8%) of the NA18507 novel sequences could not be aligned to any known human or mammalian genome sequence. Additionally, for each individual genome (Venter's, YH, NA18507), only part of the identified novel sequences could also be found in the others, which indicated there is also an individual-specific distribution of the novel sequences.

We then investigated the length distribution of novel sequences, which revealed that 1,171 (16.25%) of the YH and 1,201 (16.38%) of the NA18507 novel sequences had lengths >1 kb, and that 33 and 9 sequences had lengths >10 kb (**Supplementary Fig. 3**). (See **Supplementary Discussion** for more information about the length of novel sequences.)

To assess how many of the YH and NA18507 novel sequences were likely to be insertions or deletions, we aligned flanking sequences at both ends of the novel sequences onto the NCBI human reference genome. We anchored 3.1 Mb (62%) of the YH and 2.9 Mb (61%) of the NA18507 novel sequence to the reference chromosomes (**Supplementary Table 2**). Among these, we found that about half (46% in YH and 43% in NA18507) were insertions in the individual genomes or deletions in the NCBI human reference. Only a small fraction (1.7% in YH and 1.4% in NA18507) appeared to be sequences that were highly divergent from sequences in the same location on the reference genome. About 878.9 kb of the YH and 833.3 kb of the NA18507 novel sequences mapped to gap regions in the

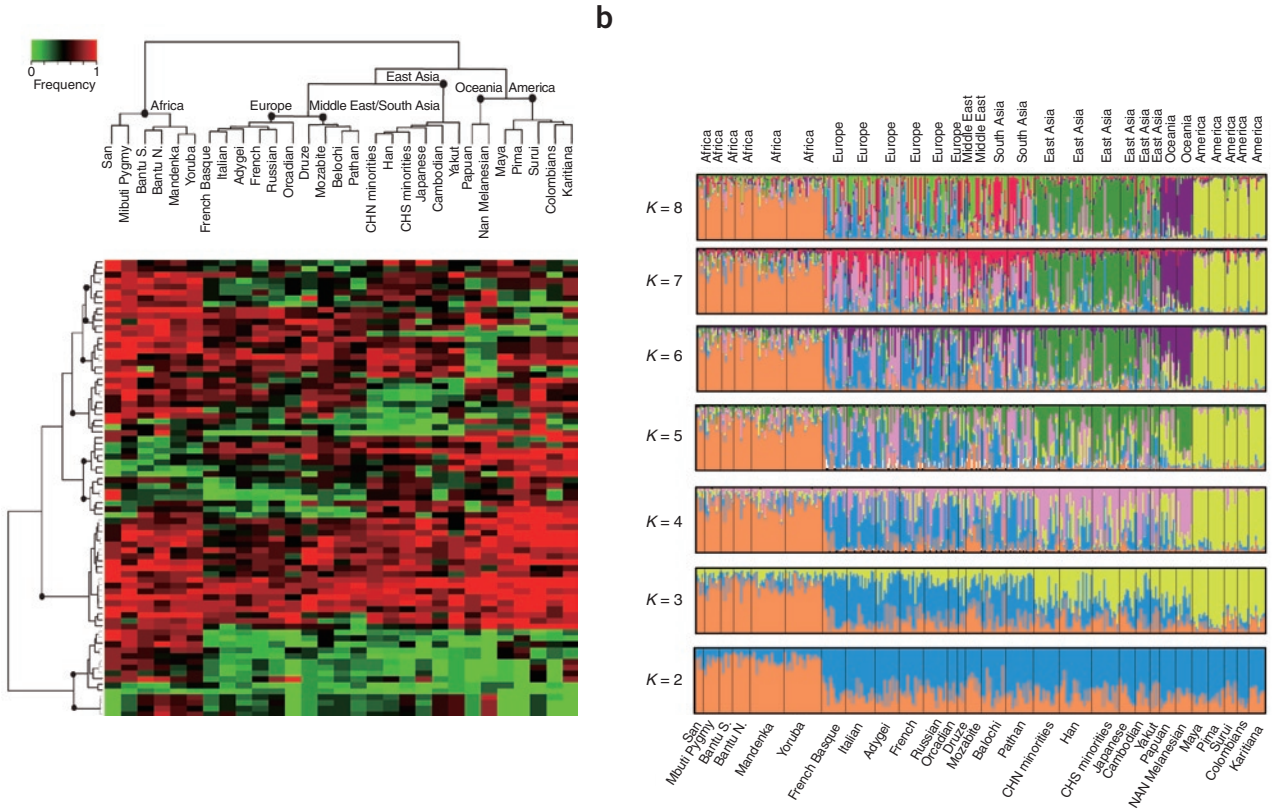

Figure 1 Population-specific patterns in novel sequences. (**a**) Frequency of individual-specific sequences (rows) in each population (columns) and neighbor-joining tree of the populations. PCR amplification was used to detect the presence or absence of each sequence in each individual. The novel sequence frequency in each population was calculated over multiple individuals belonging to the same population (on average ~8 individuals per population). For each sequence, the relative frequency in each population is represented by color intensity (red, higher frequency; green, lower frequency). Displayed here are the 83 novel sequences with <90% frequency variation over all samples. Groups of novel sequences that displayed different population-specific patterns are described in detail in **Supplementary Discussion**. (**b**) Population structure inferred by Bayesian clustering using novel-sequence frequency information. Each individual is shown as a thin vertical line, which is partitioned into *K* colored components representing estimated membership fractions in *K* genetic clusters. Population groups are separated by black lines. The population names are at the bottom of the figure and geographic locations are at the top.

reference chromosomes. The remaining anchored sequences were located in complex structural variant regions.

Population pattern of novel sequences

If these novel sequences constitute true variation in the sequence composition of the human genome, the expectation is that, as with SNPs and other types of sequence variation, these novel sequences will have population-specific characteristics. To determine the frequency variation of the novel sequences in humans in different worldwide populations, we randomly selected 164 novel sequences that did not overlap between YH and NA18507 (91 from YH and 73 from NA18507) (**Supplementary Table 3**). We used PCR to amplify these sequences in 351 individual samples from 41 worldwide populations of the HGDP-CEPH Human Genome Diversity cell line panel[18,19] (**Supplementary Tables 4** and **5**).

We then carried out phylogenic and genetic structural analyses using the novel sequences. For the profiled populations, we built a neighbor-joining tree with the novel sequence frequencies between populations as distance without prior information of individual origins (**Fig. 1a**). The tree topology generally agreed with previously defined population relationships[20–24]. Six groups that clustered by genetic analysis fit well with the main geographic boundaries of Africa, Europe, Middle East/South Asia, East Asia, Oceania and America, and structural analysis[25] also displayed consistent results (**Fig. 1b**). The novel sequences also showed distinct frequency clustering in the African, European, Middle Eastern/South Asian, East Asian, Oceanian and Native American geographic populations (**Fig. 1a**).

Phylogenic analysis of the maternally inherited mitochondria genome and paternally inherited Y chromosome have previously demonstrated the out-of-Africa migration of modern human populations[26]. Interestingly, we found that several novel sequences showed a variety of different patterns of frequency change along these defined human migration paths[27] (**Fig. 2**). For example, **Figure 2a** shows a novel sequence that has a very high frequency in the South African San population, which becomes lower as one moves to the North African population, and becomes still lower in more distant populations, until it is completely absent in the most geographically distant populations in the European (Russian), Oceanian (NAN Melanesian), and Native American (Surui) populations. Another novel sequence (**Fig. 2b**) has the opposite frequency changes, showing increasing frequency along the migration path. In addition, **Figure 2c** shows a novel sequence where the frequency becomes lower as populations follow the out-of-Africa path and has a substantial frequency reduction through the Middle Eastern, South Asian, European and Oceanian populations, but still maintains a high frequency in the East Asian and Native American populations. **Figure 2d** shows a fourth example where the novel sequence has a rapid frequency reduction in East Asian and Oceanian populations as compared to that for the European populations.

(See **Supplementary Discussion** for more information of population studies carried out on the novel sequences.)

Estimating individual sequence differences and human pan-genome size

Our analysis of novel sequence variation between populations was concordant with SNP frequency differences between populations. Similarly, we would expect to see a concordance between SNP frequencies and novel sequence frequencies between individuals. We therefore compared the SNP differences identified in previous studies[20–24] against a PCR-validated sampling of novel sequences (**Supplementary Fig. 4**) and indeed observed a positive correlation.

Given these findings, we set out to assess the DNA sequence composition differences between two individuals. We used a test based on standard SNP variation rates between two individuals, and defined sequence composition differences as 'yes or no' calls that a novel sequence or SNP was present or absent between the two individuals; rearrangement of homologous sequences or copy number changes of repeat sequences were not

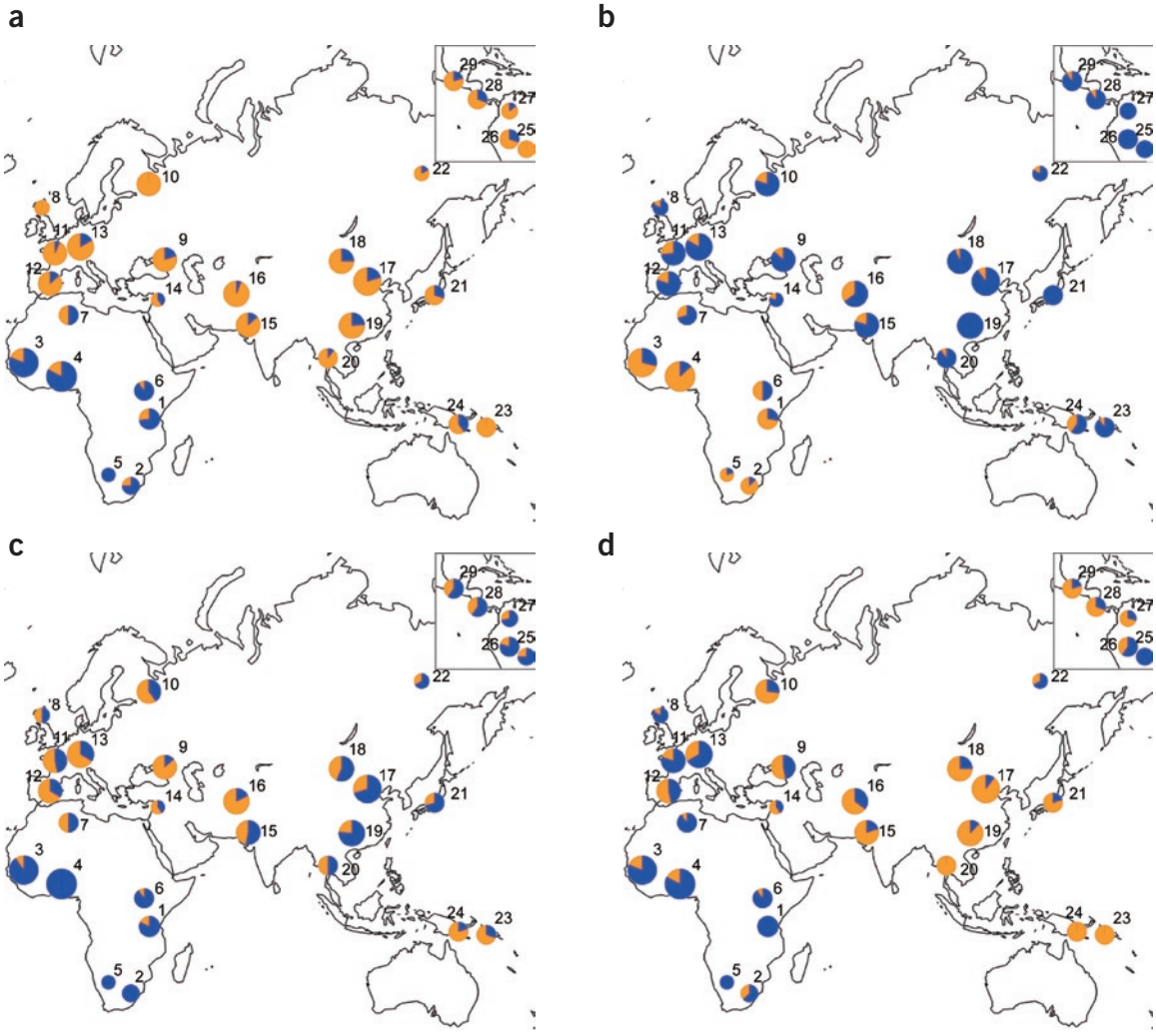

Figure 2 Examples of novel sequences with variant frequencies across populations. (**a**) NA18507, Scaffold_13185, shows a very high frequency in African populations and declines as populations grow more geographically distant. (**b**) YH, Scaffold_1781, shows a very low frequency in African populations and increases as populations grow more geographically distant. (**c**) YH, Scaffold_14717, shows a very low frequency in European populations. (**d**) NA18507, Scaffold_80603, shows a very low frequency in Asian populations. Each pie represents a single population; pie position on the map denotes the approximate geographical location of the population; pie size represents the number of DNA samples analyzed. Blue in each pie indicates the frequency of the novel sequence in the population. Key shows name of population corresponding to each number; number of samples for each population is given in parentheses.

included. Alignment of the YH and NA18507 genomes and comparison of the SNP data sets from these two genomes showed that the individual-specific sequence differences were at least 8 Mb in total and that the SNPs differed by 0.155% (**Supplementary Fig. 5**). The 8-Mb difference included 4 Mb SNPs and 4 Mb individual-specific sequences.

To assess variation between individuals that are more related based on the closeness of their populations, we compared the YH assembly to a preliminary assembly of the Korean (SJK[28]) genome with 28-fold coverage (data not shown). Our analysis revealed a 1.6-Mb individual-specific sequence difference and a 0.092% SNP difference between YH and SJK. The current sequence for the SJK genome was not sufficient for complete assembly, thus we used a calculation based on **Supplementary Figure 2** from which we could infer that the individual-specific sequence difference would cover ~1.8 Mb. These two analyses provide an individual-specific sequence increase from 1.8 Mb to 4 Mb and a SNP difference increase from 0.092% to 0.155%, indicating SNP rate and individual-specific sequence differences are positively correlated. We therefore estimated that the length of individual-specific sequences between a random pair of human individuals would range between 1.8 Mb and 4 Mb, and with the inclusion of the composition differences from SNPs, it would be in the range of 4.2 Mb to 8 Mb.

To estimate the size of the pan-genome, we used the above range for individual sequence differences and (given the correlation between the individual-specific sequences and SNP differences) used a transformation of Watterson's θ_w[27,29] (Online Methods), to evaluate the sequence differences in the population. From this we calculated a population mutation parameter for individual-specific sequences of 3.5e-4 to 7.4e-4 per base. Given a world population of over six billion people, we estimated that a complete human pan-genome would include an additional 19–40 Mb of novel sequences over the reference genome.

To assess the accuracy of this estimate, we carried out a preliminary assessment of the pan-genome size by sequentially adding Venter's HuRef[11], YH and NA18507 to the NCBI human reference genome. This preliminary pan-genome had a cumulative length that fell within our expected range (**Fig. 3**). We also estimated that common polymorphic, individual-specific sequences (those having >1% frequency in the human population) would be about 5–10 Mb in total length, and that these should be able to be defined after complete sequencing of about 100–150 individuals randomly selected from the world population.

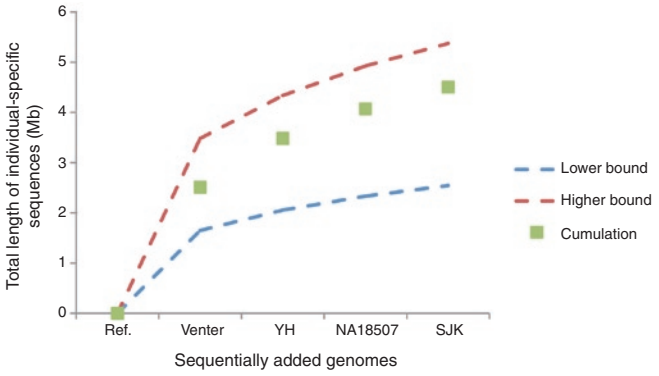

Figure 3 Cumulative length of individual-specific sequences resulting from sequentially adding genomes to the pan-genome. When adding a new genome, the novel sequences >100 bp and with <90% identity to previously added sequences were considered new individual-specific sequences and added to the data set. The real length of all novel sequences after adding each genome (green point) is in the range from the lower-bound (blue dashed line) and higher-bound (red dashed line) of the proposed individual-specific sequence population model.

Genes contained in novel sequences

To gain insight into the presence or absence of genes between the novel sequences and reference genomes, we aligned the 162 human NCBI RefSeq genes that could not be mapped to the NCBI reference genome onto the assembled YH and NA18507 novel sequences. We found that 72 and 69 of these genes could be fully or partially (>100 bp) found in the YH and NA18507 genomes, respectively, and that 55 of those individual genes overlapped between the two genomes (**Supplementary Tables 6** and **7**). Functional analysis showed that about a third of the RefSeq genes present in YH and NA18507 are members of highly variable gene families (such as mucin 2, major histocompatibility complex HLA-DQA1 and non-coding RNA SNORA66), whereas the majority of the remaining genes (57 (79%) of YH and 53 (77%) of NA18507) are currently considered hypothetical genes and have unknown functions.

Analysis of these novel sequences at the protein level by aligning the novel sequences to all human RefSeq proteins using tBlastN (E value = 1e-5) gave 1,151 and 1,087 hits in the YH and NA18507 novel sequences, respectively (**Supplementary Tables 8** and **9**). As with the RefSeq gene mapping results, the majority (915 (79%) in YH and 809 (74%) in NA18507) of the hits were to hypothetical proteins of unknown function, indicating that these genes have not been well studied. Among the hits that matched functionally classified proteins, the most abundant were members of the double homeobox protein (*DUX*) family (113 hits in YH and 58 hits in NA18507), which are known to be associated with heterochromatin[30] and also to include a number of pseudogenes[31]. Additional abundant protein categories were made up of gene families that are known to be quickly evolving and have many variant copies or may have copy number differences between genomes, which reflects the findings in the above gene analysis. These protein categories included mucin[32], zinc finger[33] and olfactory receptor proteins[34] (**Supplementary Fig. 6**).

To check whether the genes predicted by homology are likely to be functional, we investigated the conservation level of these genes across species. In total, 200 YH novel sequences (35 of which are predicted genes) and 155 NA18507 novel sequences (14 of which are predicted genes) have identified homologous regions present in all three of the chimpanzee, macaque and mouse genomes. Using 'intergenic' sequences of ~2–5 kb in length that are at least 5 kb distant from genes annotated by Ensembl as a neutral control and the well-annotated (with "NM-" prefix) RefSeq genes coding sequence present in the NCBI human reference genome as a positive control, we saw a bimodal distribution in the sequence identity of the homologous novel sequences (**Fig. 4**): the left peak of the distribution conformed to the neutral control and the right peak conformed to the positive control. The predicted coding sequences were clearly enriched at the high cross-species identity level (>90%), which is consistent with the sequence identity distribution of known annotated coding sequences. This strongly indicates that at least a portion of the homology-predicted genes might be functional and biologically important.

DISCUSSION

This study provided a genome-wide quantitative exploration of novel sequences in different individual genomes, and initiated an effort to construct the human pan-genome. We identified an extensive amount of novel sequences, which were found to be common variant sequences with different frequencies across populations. We also estimated the extent of sequence variation between two human individuals.

Cross-species conservation analysis revealed that some genes contained in these novel sequences are conserved among mammalian genomes, suggesting that these genes might be biologically functional and thus may be related to differences in gene networks between human individuals. Our

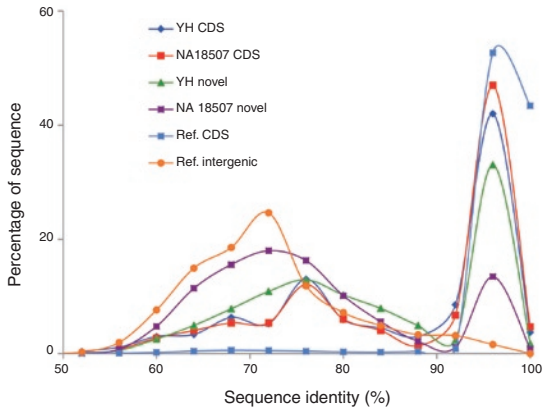

Figure 4 Distribution of sequence identity (in percentage) calculated from multiple alignments between human, chimpanzee, macaque and mouse genomes. YH novel sequences (green triangles) and NA18507 novel sequences (purple squares) had a bimodal distribution of sequence identity in the multiple alignments, whereas the distribution peak with between 90% and 100% identity is enriched in both YH novel coding sequence (dark blue diamonds) and NA18507 novel coding sequence (red squares). Coding sequences of the reference human genome (light blue squares) and the intergenic region (orange circles) were used as positive controls and neutral controls of conservation level, respectively. CDS, coding sequence.

finding that individual genomes contain a considerable amount of novel sequence indicates that similar analyses may be useful for medical genomics studies to augment array-based technologies that rely on the reference genome. Complete sequencing and assembly of personal genomes may allow larger numbers of various types of genetic variation to be identified that lead to more complete information about the genetic determinants of phenotype and disease.

Hence, it is important and practical to sequence and carry out *de novo* assembly on more human genomes to discover the common polymorphic sequences in the human population and to obtain a complete human pan-genome. Theoretically, our current pan-genome built from four individual genomes has already covered >90% of novel sequences that had a frequency >0.5 in the human population and about a half of those at a 0.1 frequency. With continuous innovation in sequencing technology, sequencing is becoming a practical and affordable method for analyzing a large number of complete human genomes, making it feasible to establish a more comprehensive understanding of the human genome, to make discoveries in medical genomics and to develop new applications for personalized medicine.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Data access. DDBJ/EMBL/GenBank: ADDF000000000 (YH) and DAAB000000000 (NA18507). The versions described in this paper are the first versions, ADDF010000000 (YH) and DAAB010000000 (NA18507). NCBI: sequencing reads of YH genome, NCBI Short Read Archive SRA009271. The assembled genomes and all of the associated analyses are freely available at http://yh.genomics.org.cn.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This project is supported by the Chinese Academy of Science (GJHZ0701-6), the National Natural Science Foundation of China (30725008; 30890032), Shenzhen local government, the Danish Platform for Integrative Biology, the Ole Rømer grant from the Danish Natural Science Research Council. L. Goodman edited the manuscript. J. Sun, M. Zhao, Y. Liu, Y. Zheng and H. Wang helped on designing the primers. W. Jin helped on experimental validation. San A, J. Wang, Y. Huang, M. Jian, M. Chen, Y. Huang, Xiaoli Ren, H. Liang, H. Zheng, S. Lin helped on the data production.

AUTHOR CONTRIBUTIONS

Ruiq. L., Y.L., Ha. Z. and Ruib. L. contributed equally to this work. H.Y., Ju. W. and Ji. W. managed the project. Ju. W., Ruiq. L., L.B. and Y.L. designed the analyses. Ju. W., Ruiq. L., Y.L., Ha. Z., Ruib. L., Ho. Z., Q.L., W.Q., G.Z., H.W., J.Q., X.J., D.L., Hon. C., S.L. and K.K. performed the data analyses. H.B. and How. C. contributed the DNA samples. Y.R., X.H. and Xu. Z. performed PCR validation. G.T., J. L., Xi. Z. performed sequencing. Ju. W., Ruiq. L., Y.L. and Ruib. L. wrote the paper.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
3. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
4. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
5. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
6. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
7. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1418 (2006).
8. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
9. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
10. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
11. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
12. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
13. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
14. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
15. Li, Y. & Wang, J. Faster human genome sequencing. *Nat. Biotechnol.* **27**, 820–821 (2009).
16. Li, R. *et al.* De novo assembly of the human genomes with massively parallel short read sequencing. *Genome Res.* (in the press).
17. Bovee, D. *et al.* Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.* **40**, 96–101 (2008).
18. Cann, H.M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
19. Cavalli-Sforza, L.L. The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* **6**, 333–340 (2005).
20. Tishkoff, S.A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
21. Wang, S. *et al.* Genetic variation and population structure in native Americans. *PLoS Genet.* **3**, e185 (2007).
22. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
23. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
24. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008).
25. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
26. Underhill, P.A. & Kivisild, T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* **41**, 539–564 (2007).
27. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
28. Ahn, S.M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629 (2009).
29. Wong, G.K. *et al.* A population threshold for functional polymorphisms. *Genome Res.* **13**, 1873–1879 (2003).
30. Beckers, M. *et al.* Active genes in junk DNA? Characterization of DUX genes embedded within 3.3 kb repeated elements. *Gene* **264**, 51–57 (2001).
31. Holland, P.W., Booth, H.A. & Bruford, E.A. Classification and nomenclature of all human homeobox genes. *BMC Biol.* **5**, 47 (2007).
32. Dekker, J., Rossen, J.W., Buller, H.A. & Einerhand, A.W. The MUC family: an obituary. *Trends Biochem. Sci.* **27**, 126–131 (2002).
33. Krishna, S.S., Majumdar, I. & Grishin, N.V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **31**, 532–550 (2003).
34. Young, J.M. *et al.* Extensive copy-number variation of the human olfactory receptor gene family. *Am. J. Hum. Genet.* **83**, 228–242 (2008).

ONLINE METHODS

Data availability. The data described in this study are freely available in the YH genome database (http://yh.genomics.org.cn/download.jsp). The full data set includes: (i) previous and newly generated Illumina Genome Analyzer (GA) sequencing reads in FASTQ format; (ii) *de novo* genome assembly of both YH and NA18507 genome sequences (contigs, scaffolds); (iii) identified novel sequences in the two genomes with corresponding alignment information; (iv) PCR gel figures and validation results in HGDP-CEPH panels. Sequencing reads of YH genome are also available at the NCBI Short Read Archive (SRA009271).

Public data used. The human genome reference assembly (NCBI Build 36.3), HuRef assembly, RefSeq mRNA and protein sequences, EST sequences and core nucleotide database were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov). Protein sequences and annotations were downloaded from the UniProt database (http://www.uniprot.org/downloads). Read sequences of Watson's genome were provided by the Baylor College of Medicine. Read sequences of sample NA18507 were provided by Illumina, Inc., which is also publicly available in NCBI Short Read Archive (accession number SRA000271).

Data production. Library construction and read sequencing by Illumina Genome Analyzer II platform followed the manufacturer's instructions. Fluorescent images and base-calling were performed by Illumina data process pipeline (IlluminaPipeline-1.3.2).

Genome assembly. Whole genome short read *de novo* assemblies of the YH and NA18507 genomes were performed by SOAPdenovo[16] software (**Supplementary Data**, http://soap.genomics.org.cn), which is based on the De Brujin graph algorithm. The algorithm details and step-by-step YH and NA18507 assembly results have been described in the assembler manuscript[16]. Here is a brief summary of the algorithm.

First, sequencing errors that were primarily accumulated at the 3′-end of reads were corrected according to 17-mer frequency. In this step, all the raw read sequences from short insert-sized libraries (<500 bp paired-end insert size or single-ended) were broken down into 17 mers, and the frequency of each 17 mer was counted. Low frequency (<5 in this study) 17 mers that were likely to be sequencing errors were edited to the closely similar high-frequency 17 mers.

The sequencing-error corrected reads were then loaded into memory, and De Brujin graph data format was used to build the overlap graph, with 25-mers as vertex and read paths across the 25-mers as edges. Thus, two reads will be joined if they have at least a 25 bp overlap. The repeat sequences and sequencing errors would make the graph very complex. To reduce the complexity of the graph and filter noise connections, we removed the 'tips' which are short (<50 bp) and low-coverage dead ends in the graph, removed the low-coverage connections that nodes were linked by only one or a few reads in the graph, and merged the bubbles where redundant paths having the same input and output nodes while with minor differences (polymorphisms or difference between homologous sequence copies). After error correction, we broke the graph at repeat boundaries and outputted the unambiguously continuous sequence fragments as contigs.

Next, we realigned the short reads onto the contigs, and transferred the read paired-end information into contig linkage information. The unreliable linkages between two contigs that have equal or less than three read-pairs were filtered. The contig linkage graph was linearized by masking repeat contigs, which have multiple conflicted connections to the other contigs. And the remaining contigs with compatible connections to each other were constructed into scaffolds. The paired-end information was used step by step that started from short paired-ends to longer paired-ends.

The final step of *de novo* assembly is to close the gaps inside constructed scaffolds. We collected read pairs with one end located at the edge of contigs and another end located in the gaps, and performed local assembly to extend the contig sequence into the gaps. The final gap closed scaffolds were used for all analysis in the project.

Identification of novel sequences. We aligned all assembled contigs to the human reference genome (NCBI Build 36.3) using BLAT[35] with -fastmap option enabled. Alignment position of each contig indicates a candidate location of the scaffold to which the contig belongs. For contigs with multiple hits, the top ten hits with highest sequence identity and >90% coverage of the contig remained as candidate locations. Then we checked candidate locations of contigs within a scaffold to build scaffold-reference alignment to maintain orientations in as linear a fashion as possible between scaffolds and the NCBI reference genome. The alignment with the longest length in linear orientation between a scaffold and the reference was picked as 'best-hit' of the scaffold. We then aligned the scaffolds against the located regions on the NCBI reference genome by LASTZ.

The unmapped sequences derived from LASTZ alignment were treated as candidate novel sequences. We then aligned the scaffolds with unmapped sequences to the whole NCBI reference genome again using BLASTn[36]. The scaffold fragments with <90% identity to any region of the NCBI reference genome was defined as novel sequences. Novel sequences with <100 bp were filtered.

The identified novel sequences were first aligned to gap-closure fosmids[17] and the genomes of HuRef, Watson, YH (if novel sequences are from NA18507), and NA18507 (if novel sequences are from YH). Novel sequences that had alignments with >90% identity to any these genomes were kept apart, and the remaining novel sequences were then aligned to the other mammalian genomes. Hits with an E-value < 1e-20 were retained as valid alignment in classification. Novel sequences that aligned to human and mammalian genomes were retained, and those aligned to non-mammalian genomes were treated as potential contaminations and were filtered in this analysis. The remaining novel sequences that had no alignment to any sequences in GenBank database were classified as unknown.

Population profiling of novel sequences by PCR. To validate novel sequences identified in our assemblies and survey their frequency in human populations, we extracted novel sequences with lengths >500 bp and randomly selected 233 novel sequences for PCR validation in 347 human DNA samples from a worldwide population that were provided by HGDP-CEPH (**Supplementary Fig. 7**) and four HapMap CHB samples. The appropriate temperature of these PCR experiments was 58 ± 2 °C. The high-quality amplification of 164 novel sequences was used in this analysis.

Phylogeny tree construction. The frequency of novel sequences in each ethnic group was calculated. We then used the frequency information to cluster the ethnic groups and the novel sequences by hierarchical clustering in heatmap function implemented in R scripts. The distance between objects was measured by standard complete linkage clustering (farthest neighbor method) by comparing the frequency of novel sequences between two clusters.

Genetic structure. The genetic structures of world ethnic groups were calculated using STRUCTURE[37] (version 2.2) by *K*-means partitional clustering. The monoploid model was used to adapt for novel sequence present/absent information. The PCR validation results in 288 individuals in this study were transformed to a 0/1 matrix as the data input of STRUCTURE, where 0 denoted for absence of the novel sequence and 1 for presence. STRUCTURE calculated membership coefficients to place all the individuals to *K* clusters, where *K* value was set from 2 to 8 in our study.

Evaluate the sequence differences in the population. To estimate the pan-genome size, we used the range of individual sequence difference and transformed Watterson's θ, which was primarily designed for SNP divergence, to suit this estimation. First, we estimated the average composition difference in DNA sequence between two individuals was 1.8Mb to 4Mb, where the two boundaries were defined by SJK-YH (both are Asians) and YH-NA18507 differences. Second, by definition of Watterson's θ, the total amount of identified individual-specific sequences would approximately conform the following formula:

$$K = \theta \times L \times a, a = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$$

Where n is the sample size (1 for a haploid, 2 for a diploid), L is the size of a single human genome and θ is the averaged individual specific sequence rate among all samples. We therefore estimate the range of $(\theta \times L)$ to be ~0.9–1.9 Mb. Third, by extrapolating the above results to the whole human population with a size of ~6.5 billion (a is calculated to be about 23.9), we estimated the total amount of human individual-specific sequences to be ~19–40 Mb.

35. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
36. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
37. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574–578 (2007).