Supplementary Figures

Figure S1: G-banded karyotype of the DNA sample.

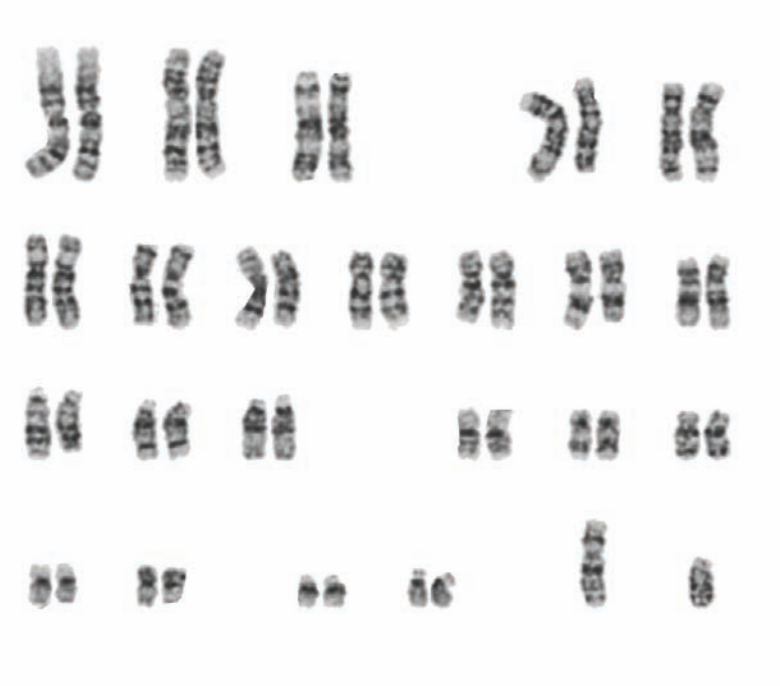


Figure S2: Overlap of (a) all SNPs, (b) non-synonymous SNPs, and (c) genes with non-synonymous SNPs among YH, Venter, and Watson genome.

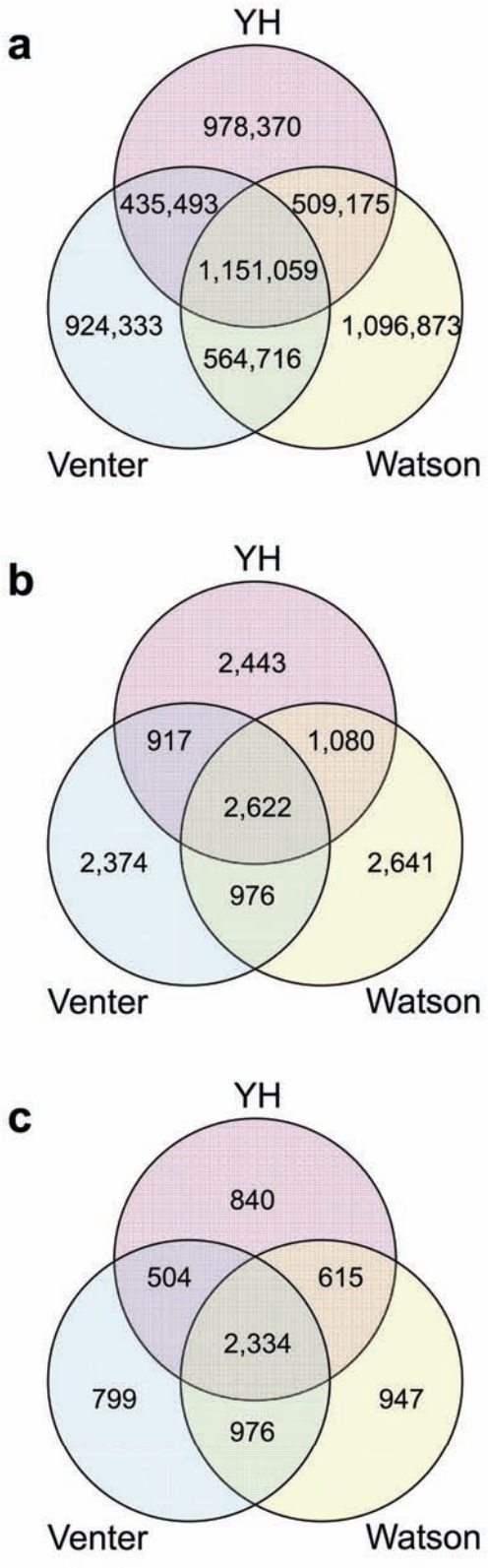


Figure S3: Allocation of the YH, Venter, and Watson genomes in phylogeny of ethno-geographic populations. A subset of 87,614 alleles with known genotyping in YH, Venter, Watson, and the 270 HapMap individuals were used for hierarchical clustering. CHB/JPT: Han Chinese in Beijing, and Japanese in Tokyo; CEU: Centre d'Etude du in Utah; YRI: Yoruba in Ibadan.

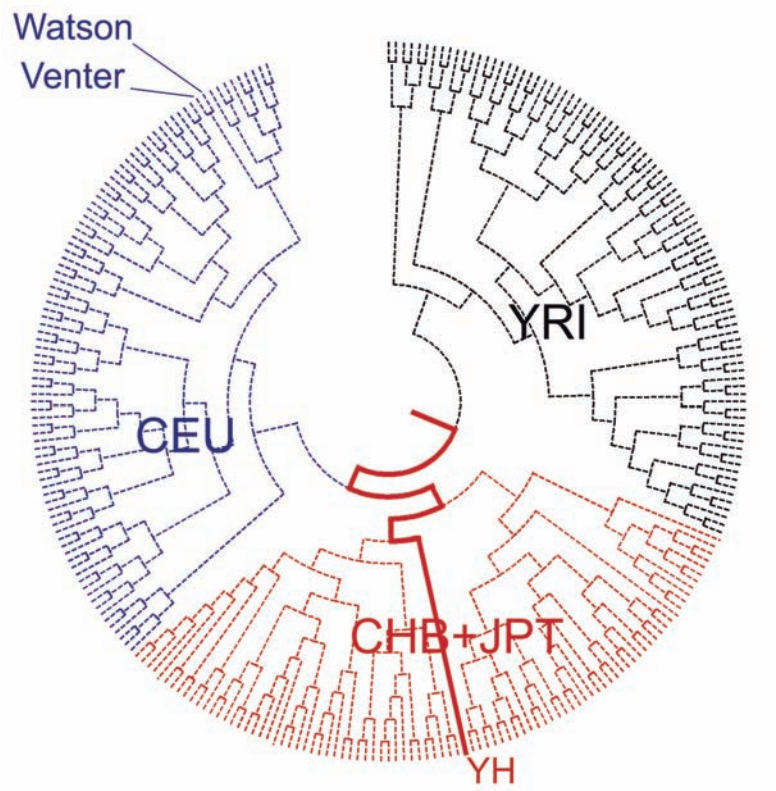


Figure S4: Distribution of sequence depth of YH genome autosomes and sex chromosomes. Depth was calculated for each base from the aligned reads. Reads with multiple equal best placements were randomly assigned to one best-hit location. Sequencing depth exhibited a Poisson-like distribution with a median depth of 34-fold and 19-fold on autosomes and sex chromosomes, respectively. The variance in the sequencing depth of the experimental data for the autosomes and sex chromosomes, respectively, was 2.5 and 2 times larger than that of their theoretical Poisson distribution.

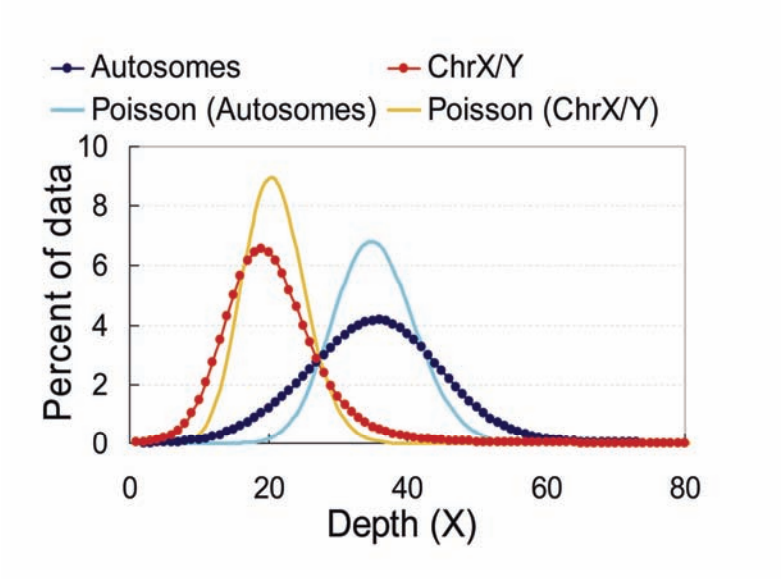


Figure S5: GC content and median sequence depth of each human chromosome. The data show that sequence depth is negatively correlated with GC content. Chromosomes with a higher GC content had a significantly lower sequencing depth. For example, chromosome 4, which has the lowest GC content (38.2%), had the greatest sequence depth (36-fold), while chromosome 19, with the highest GC content (48.4%), had the lowest sequence depth (28-fold)

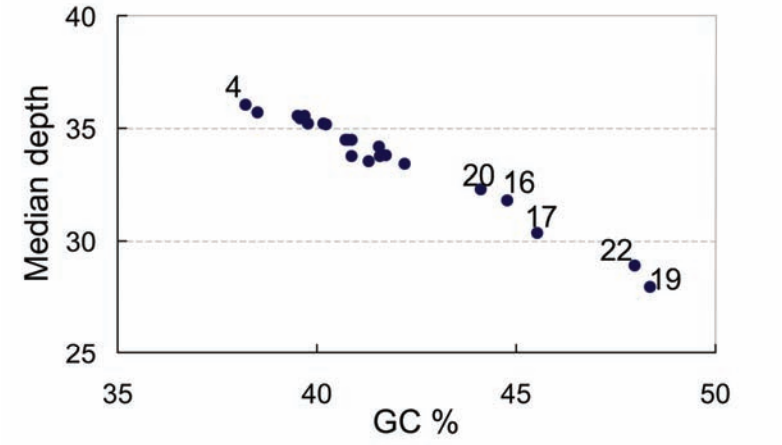


Figure S6: Occurrence of indels with different sizes in the whole genome and in the coding region.

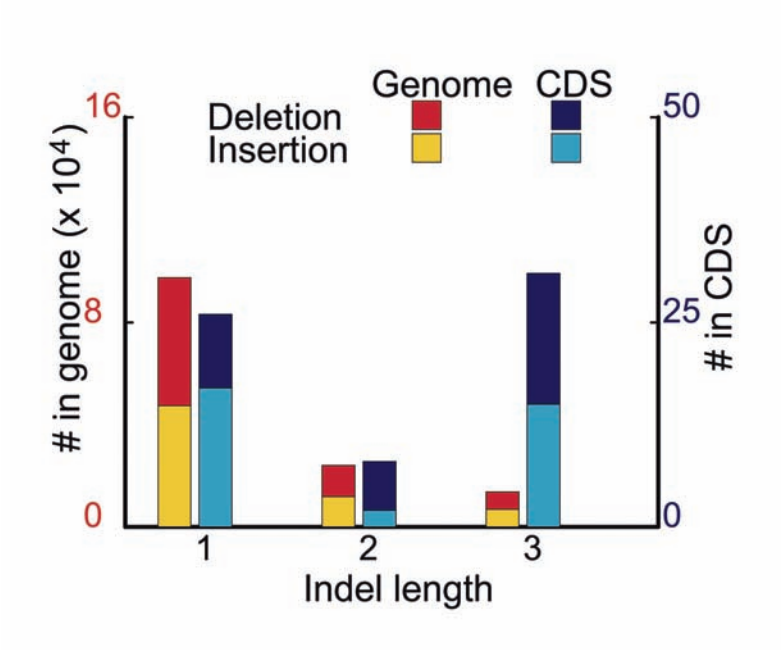
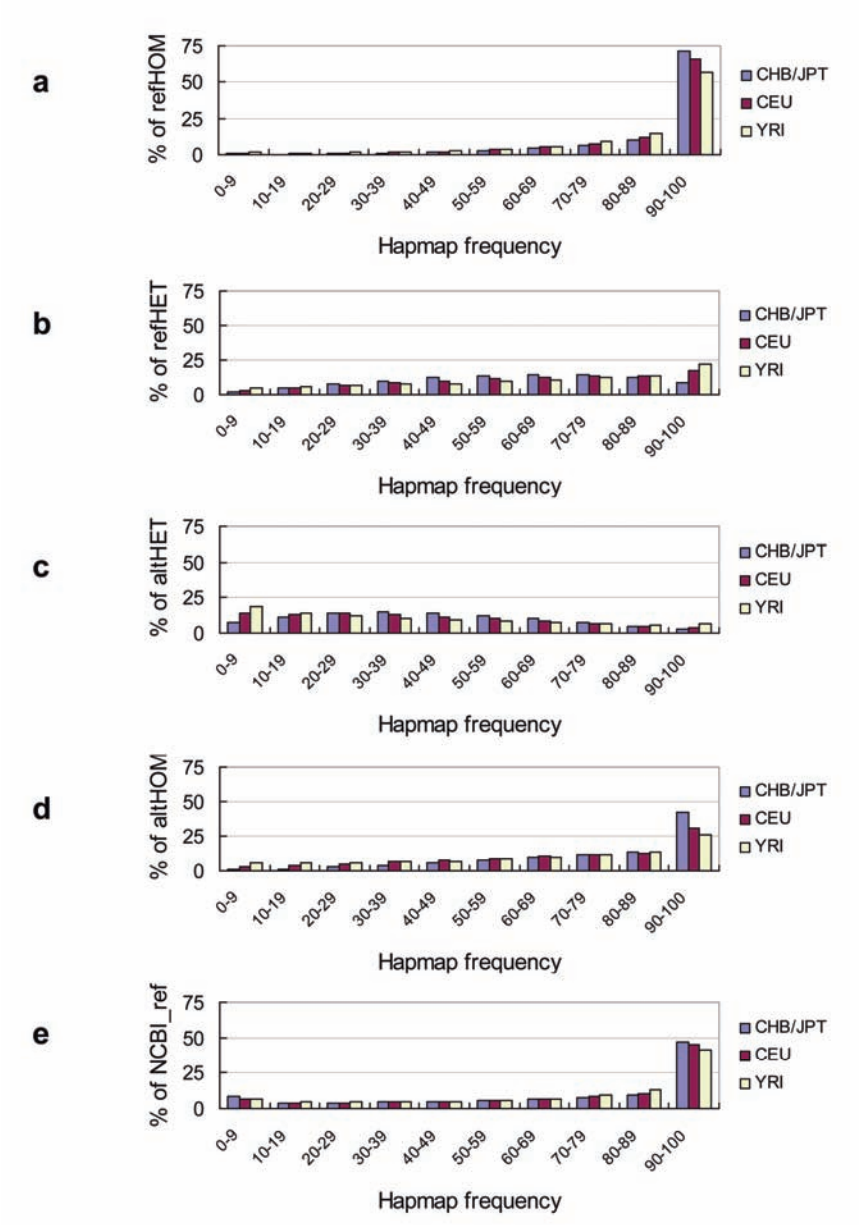


Figure S7: Example of a frameshift on the NCBI reference genome where the YH genome had the same allele type as that in the sequences of all other organisms examined. The deletion on the NCBI human reference genome is present at the 580th amino acid of protein NP_001103669. That this gene can be transcribed is supported by 4 cDNAs (NCBI accession numbers: BC071856, BC096755, AK092590 and AX747617).

Peptide	Ala	Arg	Lys	Arg	Pro	Ile
NCBI	GCC	-GG	AAG	CGT	CCT	ATT
YH	GCC	CGG	AAG	CGT	CCT	ATT
Rhesus	GCC	CGG	AAG	CGT	CCT	ATT
Mouse	GCC	CGG	AAG	CGT	CCC	ATT
Dog	GCC	CGG	AAG	CGT	CCC	ATC
Opossum	GCC	CGG	AAG	CGT	CCC	ATC
Chicken	GCT	CGC	AAG	CGT	CCC	ATC
Stickleback	GCT	CGG	AAG	CGG	CCG	ATA

Figure S8: Allele frequencies in CHB (Han Chinese in Beijing, China)/JPT (Japanese in Tokyo, Japan), CEU (in Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection), and YRI (Yoruba in Ibadan, Nigeria) populations of the HapMap phase II map (3.1 million SNPs). Since the HapMap mainly included common variations, most of the alleles have high frequency. (a) NCBI reference alleles; (b) YH homozygous alleles that are identical to the NCBI reference genome; (c) YH homozygous alleles that are different from the NCBI reference genome; (d) YH heterozygous alleles that are identical to the NCBI reference genome; (e) YH heterozygous alleles that are different from the NCBI reference genome.



Supplementary Tables

Table S1: Experimental parameters of the (a) 8 single-end and (b) two paired-end libraries for GA sequencing.

(a)

Library #	Starting DNA amount (ug)	Size of gel slices or stabs excised (bp)	PCR Cycle #	Library concentration (ng/ul)
YHDASA	5	150－200,slice	18	20.0
YHDBSA	5	150－200,slice	18	10.0
YHCDSA*	2	200-225,slice	18	39.5
YHCDSB*	2	200-225,slice	12	32.9
YHCDSC**	2	200-225,slice	15	42.6
YHCDSD**	2	200-225,slice	10	26.2
YHCDSE**	2	200-225,slice	8	23.4
YHCDSF**	2	250,stab	18	17.0

(b)

Library #	Starting DNA amount (ug)	Size of gel slices or stabs excised (bp)	PCR Cycle #	Library concentration (ng/ul)
YHPEA	5	200	18	9.3
YHPEB	5	500	18	10.1

* YHCDSA and YHCDSB were from the same gel slice, but had different PCR cycles.

** These four libraries were from the same ligation products. YHCDSC, YHCDSD, and YHCDSE were from same gel slice but had different PCR cycles; YHCDSF was from the stab at 250bp.

Table S2: PCR validation of a subset of inconsistent SNPs and a subset of small indels between the assembled YH consensus and Illumina 1M BeadChip genotyping. In total, 50 SNPs and 45 small indels were PCR amplified and re-sequenced using traditional Sanger sequencing technology. The alleles that were identical between the assembled consensus and the PCR sequencing were taken as accurate.

		Total examined	Validated	Validation rate (%)
SNP (coding)		30	24	80
SNP (other)		20	17	85
Indel (coding)	1,2 bp	20	18	90
	3 bp	4	4	100
Indel (other)	1,2 bp	17	17	100
	3 bp	4	4	100

Table S3: Full list of the PCR validation of the SNP sites which are inconsistent between the assembled consensus and genotyping.

TableS3_ListPCRvalidation_SNPs.xls

Table S4: Full list of the PCR validation of the indel sites.

TableS4_ListPCRvalidation_Indels.xls

Table S5 Percent of small indels in the YH or the NCBI36 genome that have the same allele type as the chimpanzee genome. The identified 1–3 bp indels between YH and the NCBI reference genome were checked against the chimpanzee syntenic regions that were longer than 100-bp and greater than 95% identity. The alleles that are identical to those in the chimpanzee were taken as the ancestral allele types.

		# in Chimp	YH as ancestral type	NCBI Ref as ancestral type
Deletion in YH	HOM	34,743	34.00%	66.00%
	HET	24,342	15.70%	84.30%
Deletion in NCBI Ref	HOM	35,212	66.20%	33.80%
	HET	20,052	34.40%	65.60%

Table S6: List of genes overlapped with SVs.

TableS6_GeneLoseInSVs.xls

Table S7: Rate of SNPs and 1–3 bp indels in complete autosomes and in defined genetic regions. The Rate between YH and the NCBI reference ("vs Ref") was calculated using the population mutation parameter $\theta=K/aL$, $a=1+1/2+\ldots+1/(n-1)$, where K is the number of variant sites found by sequencing n chromosomes in a region of length L. YH is a diploid sequence, while the NCBI genome represents one set of human reference chromosomes, so n=3.

	# of SNPs	SNP rate (x1e-4)			# of 1-3bp indels	Indel rate (x1e-5)		
		HOM	HET	vs Ref		HOM	HET	vs Ref
Whole autosomes	3,004,206	5.16	6.94	8.06	132,781	3.06	2.08	3.42
CDS	15,686	2.56	3.35	3.94	65	0.16	0.09	0.17
5'-UTR	2,817	4.36	4.79	6.10	85	2.21	0.72	1.95
3'-UTR	15,885	3.83	5.27	6.07	978	3.20	2.36	3.70

Table S8: Full list of genes containing non-synonymous SNPs.

TableS8_ GenesWithNonsyn.xls

Table S9: Full list of genes containing frameshift indels.

TableS9_GenesWithIndels.xls

TableS10: Full list of genes involved in predicted selective sweeps.

TableS10_GenesInSelectiveSweeps.xls

Table S11: List of HGMD alleles which are positive in YH genome.

TableS11_HGMD_PositiveAlleles.xls

Supplementary Discussion

Data production and short read alignment

The use of paired-end reads as compared to single-end reads provided a higher percentage of unique alignments (90.2% vs 83.6%). In agreement with this, an estimation made by mapping simulated reads from the NCBI reference genome provided similar percentages of uniquely aligned reads: 95.1% paired-end and 86.0% single-end reads had a single best hit.

The per-base sequencing depth on the reference genome exhibited a Poisson-like distribution with a median depth of 34-fold on the autosomes and 19-fold on the sex chromosomes (Supplementary Fig. 4). The variance in the sequencing depth of the experimental data for the autosomes and sex chromosomes, respectively, was 2.5 and 2 times larger than that of their theoretical Poisson distribution. The variance in the sequencing depth can be due to numerous factors. Local chromosomal structure, for example, can influence the randomness of DNA fragmentation, especially when DNA is sheared into pieces between 135-500 bp in size. Additionally, the efficiency of PCR amplification is known to vary with GC content. In this regard, we did observe a negative linear correlation between sequencing depth and GC content (Supplementary Fig. 5): chromosomes with a higher GC content had a significantly lower sequencing depth. For example, chromosome 4, which has the lowest GC content (38.2%), had the greatest sequence depth (36-fold), while chromosome 19, with the highest GC content (48.4%), had the lowest sequence depth (28-fold). Having an understanding of the impact of such sequencing-bias patterns guided us in generating sufficient sequence data for different genomic regions to enable us to assemble a high-quality consensus sequence for these regions.

SNP and indel identification

Given that for genotype estimation we set a higher prior probability for SNP sites that have been deposited in dbSNP (see Methods for details), we may have missed many heterozygotes in novel sites. To estimate the rate of missed heterozygotes in novel sites, we calculated the consensus sequence of the YH genome without using any

probabilities based on dbSNP information, and checked the rate of missed heterozygotes in the identified 1M genotyped alleles. Our data indicate that about 7.3% of the heterozygotes have one allele missing after the quality filtering. Theoretically, given a 1.45% sequencing error rate and assuming errors were random and independent, requiring at least 4 occurrences to call a SNP allele would result in about a 0.076% false positive rate with only 0.501% heterozygotes missed if the SNP density were 1 in 1 kb. The difference we see can mainly be explained by our using a stringent strategy for discovering novel SNPs using prior probabilities as low as 2e-6 for heterozygotes and 1e-6 for homozygous SNPs. Our estimate, however, is still much lower than the undercall rate in the SNP analyses of the Watson[4] (24.2%) and Venter[3] (21.6%) genomes.

Mutation and selection

The rate of heterozygous SNPs, which is an indication of the sequence diversity in the YH genome, is 6.94×10^{-4} across the autosomes (Supplementary Table 7). As estimated above, about 7% of the novel heterozygotes may have one allele missing, thus the rate calculated here is a likely to be slightly low. The rate of small indels in the YH genome is 2.08×10^{-5}, which is 3 times lower than the estimate for the Venter genome[3]. This difference is likely because the sequencing of YH was done using very short reads, which makes it impossible to identify long indels and unlikely to detect indels in highly repetitive regions. In the coding regions, the YH genome SNP rate (3.35×10^{-4}) is 2.1 times lower, and the rate of small indels (0.09×10^{-5}) is 23 times lower than the average in the whole genome. This pattern is similar to what was observed in Venter's genome[3], but the numbers we report here show fewer number of SNPs in coding regions, possibly indicating stronger purifying selection as compared to Venter's genome; again, more data from additional individual genomes is required to make any firm conclusions.

The SNP rate in 5'-UTR (4.79×10^{-4}) and 3'-UTR (5.27×10^{-4}) is 31% and 24% lower than in the whole genome, while the rate of small indels in 5'-UTR (0.72×10^{-5}) is 2.3 times lower than that in 3'-UTR (2.36×10^{-5}). By adopting the population mutation parameter, which is a measure to correct for sample size or

number of chromosomes, the estimated rate of SNPs and small indels between YH and the NCBI reference is 8.06×10^{-4} and 3.42×10^{-5}, respectively. The rate of indels is 17.6% lower than the estimate in SeattleSNPs (4.1×10^{-5}), which was derived from gene region resequencing[21]. That the rate of indels is only slightly lower indicates that most of the small indels in gene regions have been identified in the YH genome.

The frequency of heterozygous and homozygous SNPs in our dataset that are in dbSNP (validated) was nearly equal (51.8% and 48.2%), but in the set of novel SNPs, the frequency of heterozygotes was 5.7 times higher. Such a difference in frequency may be because most common alleles have already been identified and placed in dbSNP, whereas novel alleles are likely to be rare and often exist as heterozygotes. Overall, the ratio of heterozygous to homozygous SNPs in the YH autosomes is 1.34, which is lower than expected from the Hardy-Weinberg principle. This might be due to the possible existence of rare alleles in the NCBI reference genome[3] and by the miss-identification of some heterozygotes in the YH genome. Transitions in the YH genome SNPs are four times more frequent than transversions, but there is little obvious bias among each type of transition or transversion combination; this is the same pattern seen in a previous study on NCBI dbSNPs[22].

As noted, there are 7,062 non-synonymous SNPs in the YH genome and these are distributed throughout 4,293 genes. The ratio of the non-synonymous and synonymous mutation rate (dN/dS) over all these genes was 0.35. We examined the GO[23] classification of genes that had a dN/dS ratio greater than 0.35, and found that these genes primarily belong to functional categories that are known to be under relaxed selection or to have a high divergence rate, these include genes that encode zinc-finger proteins, and genes related to the immune system, to antioxidant activities, and to physiological responses to stress or stimuli[24]. The other functional categories that also have a higher fraction of non-synonymous SNPs are genes involved in the basement membrane, proteinaceous extracellular matrix, and enzyme inhibitors (Fisher exact test $p<0.001$). (See complete list in Supplementary Table 8.)

The distribution of the number of indels with sizes ranging from 1 to 3 bp shows an exponential decay in the whole genome. Of the 66 indels that are located in

coding regions, 47.0% of these were 3-bp indels, which was more frequent than the percentage of indels in the whole genome (p=1e-55) (Supplementary Fig. 6). This higher percentage of 3-bp indels in coding regions is likely because 1- or 2-bp insertions or deletions cause frameshifts, which are generally under higher purifying selection than are 3-bp indels, as these will result in an insertion or deletion of a single amino acid. Our analysis of the position of 1- or 2-bp indels indicate that 35 genes in the YH genome will contain frameshifts (Supplementary Table 9); 18 (51%) of these genes are homozygous, indicating that there are unlikely to be any functional copies of these genes in the YH genome. Most of these potential non-functional genes belong to multicopy gene families. Of these, 21 (60%) also have non-synonymous SNPs, indicating that they are under relaxed selection and have accumulated mutations.

To check whether some of the genes that are present in YH but are inactive in the NCBI reference genome, we aligned human cDNAs onto the reference genome to identify indels using BLAT[25]. Indels supported by both YH genome and human cDNAs were compared with other vertebrate genomes (rhesus, mouse, dog, opossum, chicken and stickleback). We found one case where there was the ancestral version of the gene in YH, but in the NCBI reference genome, there was an apparent frameshift in this gene (Supplementary Fig. 7). Future surveys of more personal genomes are required to distinguish population specific or individual specific gene loss.

Selective sweep is a process that causes a reduction or elimination of variations present in neighboring neutral nucleotides. It often occurs when a beneficial mutation appears that greatly increases the fitness relative to other alleles in the population. We used Tajima's D test[26] to examine candidate regions of selective sweep (regions with $p<0.05$) from the whole genome alignments of YH, Venter, Watson, and the NCBI reference genomes. A recent study[27] looked at regions >100 kb, but because we have a higher SNP density (1.88 SNPs/kb) than did this previous study (1.01 SNPs/kb, HapMap phase II), we could reduce size of the analyzed regions down to 50 kb, which allowed us to increase the sensitivity for finding candidate selective sweep regions. In total, we identified 813 candidate regions with lengths >50 kb, of which 323 genes appear to be involved (Supplementary Table 10). Of these

genes, 36 have been previously reported to be involved in the process of selective sweep[27-30]. For example, the gene *CYP3A*, which belongs to the CYP gene family, is thought to play a role in selective sweep as it exerts a strong influence on the bioavailability and clearance of numerous exogenous compounds, such as therapeutic agents and prescription drugs. In our study, we identified a gene belonging to this same family, *CYP2A6*, located at 19q13.2, in one of the candidate selective sweep areas.

It is not surprising that we did not find a greater overlap between genes identified in our study and those in other studies, as each study used different methods and different data sets— and these were often aimed to detect different groups of sweeps. Additionally, some of these methods and the data used could have provided false predictions for sweep regions especially if there have been population bottlenecks. Because of such compounding factors, the development of more robust methods for selective sweep detection will require the analysis of population data in conjunction with detailed demographic history of human ethnic populations and the recombination landscape[31].

Structural variations

In addition to searching for SVs using paired-end methods (PEM), we also identified copy number variations (CNV) based on read depth. By modeling sequencing depth distribution on different levels of GC content, we found 1701 CNV regions that had a lower copy number (1–47 kb in length, median at 1 kb) and 1299 that had a higher copy number (1-105 kb in length, median at 1 kb) than NCBI36. Approximately 82% of the CNV regions that had a lower copy number and 61% CNVs that had a higher copy number had more than a 50% overlap with the annotated repeats, which was not surprising. However, these regions may be somewhat inaccurate because sequencing depth can be affected by many factors, including GC content and alignment difficulties in repetitive region.

Genetic ancestry insights

We examined the YH allele frequencies with those in the HapMap phase II data[32] (Supplementary Fig. 8). We found that most of the YH homozygous alleles are

common in all populations. (This included both YH SNPs that are identical to and that are different from those in the NCBI reference genome). The YH homozygous allele frequencies are, however, higher in the CHB (Han Chinese in Beijing, China)/JPT (Japanese in Tokyo, Japan) populations than in the CEU population (Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection) and the YRI population (Yoruba in Ibadan, Nigeria). For heterozygotes, the alleles that are in common between the YH and NCBI reference sequence are present at a much higher frequency than are the alleles that are unique to the YH genome. Additionally the alleles they have in common also appear at higher frequency in the CEU and YRI populations, in contrast it is only the unique YH alleles that have a higher frequency in CHB/JPT. Combining both homozygote and heterozygote alleles, only 0.8% YH alleles are not present in the CHB/JPT population; while 3.6% of the NCBI reference alleles are absent in the CEU population. The presence of a high rate of rare alleles in the NCBI reference is may be due to sequencing errors in the NCBI reference genome that were included in the HapMap SNP set.

Known phenotypic or disease risk variant screen

In addition to comparison with OMIM, we also compared YH genotypes with all available genotypes in the Human Gene Mutation Database (HGMD)[33] A total of 20,559 genotypes matched with 1,478 HGMD genes, of which 318 genotypes were associated with increased disease risk (Supplementary Table 11). Many of these specific variations that are potentially associated with disease risk have not yet been tested in sufficiently large population samples or have not been surveyed in different ethnic populations to provide a good assessment of the risk inherent in the presence of these genotypes in YH. A much more extensive survey of the frequencies of the diseases associated with variants in a broad range of populations and samples will be required to validate the level of risk of disease for an individual.

Supplementary Full Methods

Data availability

The data have been deposited in the EBI/NCBI Short Read Archive (Accession number: ERA000005). All Yanhuang data have been released for public use and can be freely accessed at http://yh.genomics.org.cn/download.jsp. The entire dataset includes all raw reads, alignment results, pseudo-chromosome consensus sequences, annotation of DNA variants including SNPs, small indels (1-3bp), and structural variations (≥100bp in size), newly assembled contigs (≥100bp in size), and relevant tools. SNPs and indels have been submitted to NCBI dbSNP and will be available in dbSNP version 130.

DNA library construction and sequencing

Genomic DNA was extracted from peripheral venous blood, and the blood sample was collected using the guidelines dictated by the institutional review board of the Beijing Genomics Institute (BGI).

Library preparation followed the manufacturer's instructions (Illumina). Briefly, 2–5 ug of genomic DNA in 50ul TE buffer were fragmented by nebulization with compressed nitrogen gas at 32psi for 9 minutes. Nebulization generated double-stranded DNA fragments with blunt ends or with 3' or 5' overhangs. The overhangs were converted to blunt ends using T4 DNA polymerase and Klenow polymerase, after which an "A" base was added to the ends of double-stranded DNA using Klenow exo- (3 'to 5' exo minus). Next, DNA adaptors (Illumina) with a single "T" base overhang at the 3' end were ligated to the above products. These products were then separated on a 2% agarose gel, excised from the gel at a position between 150 and 250 bp, and purified (Qiagen Gel Extraction Kit). The adapter-modified DNA fragments were enriched by PCR with PCR primers 1.1 and 2.1 (Illumina). Separate 8-, 10-, 12-, 15-, and 18-cycle reactions were used for sequencing. The concentration of the libraries was measured by absorbance at 260nm.

The template DNA fragments of the constructed libraries were hybridized to the surface of flow cells and amplified to form clusters. After dsDNA was denatured

to ssDNA and non-specific sites were blocked, genomic DNA sequencing primers were hybridized for DNA sequencing initiation. In brief, cluster generation was performed on the Illumina cluster station, and the basic workflow (based on the standard Illumina protocol) was as follows: Template hybridization, isothermal amplification, linearization, blocking, and denaturization and hybridization of the sequencing primers. The fluorescent images were converted to sequence using the Illumina base-calling pipeline (SolexaPipeline-0.2.2.6).

Public data used

The human reference genome, together with genes and repeats annotation, was downloaded from UCSC database (http://genome.ucsc.edu/), which has the same sequence as the NCBI build 36.1. The NCBI reference genes with prefix "NM" were mapped to the reference genome using BLAT by UCSC. Hits with >90% identity were retained for further analysis, and only one transcript was retained for each gene. dbSNP v128 and HapMap release 23 were used. The SNP set from Venter's genome was downloaded from the public FTP site of JCVI (ftp://ftp.jcvi.org/pub/data/huref/), and the SNP set of Watson's genome was provided by Baylor College of Medicine (BCM).

Short reads alignment

We used SOAP to align each read or read-pair to a position on a chromosome of the NCBI36 human reference genome that had least number of nucleotide differences between the read and the reference genome, and called this a "best hit". If a read had only a single best hit, it was considered uniquely aligned. Reads that had more than one "best hit" (meaning they could be aligned to multiple positions that each had the same number of mismatches) were considered repeatedly aligned. For repeatedly aligned reads a random position was chosen from all of its best hits for placement on the reference genome for sequencing depth calculation.

In the specific alignment process, at most 2 mismatches were allowed between a read and the reference, and best hits were selected. Since errors can accumulate during sequencing, the quality of the last several base pairs at the end of reads can be relatively low. We thus set option –c 52 during our alignment procedure. Thus, if a

read could not be aligned, we discarded the first base, and iteratively trimmed 2 bp at the 3' end until the read could be aligned or the remaining sequence was shorter than 27 bp. For paired-end reads, two reads belonging to a pair were aligned with both being in the correct orientation and proper span size on the reference genome. If a pair could not be aligned without gaps but allowing at most 2 mismatches on each read, a gapped alignment was then performed with a maximum gap size of 3 bp. If the two reads could not be aligned as a pair, they were aligned independently.

Consensus assembly

We used a statistical model based on Bayesian theory and the Illumina quality system to calculate the probability of each possible genotype at every position from the alignment of short reads on the NCBI reference genome. A calibration matrix was built based on all uniquely mapped reads to estimate the probability for a given genotype *T* to have an observed base *X* located at a position *k* of its original read with quality score *S*. For a variety of reasons, similar sequencing errors are often repeated, thus, the *i*-th occurrence of base *X* covering a particular position would contribute less to denote an *X* in consensus by an adjustment formula. In brief, likelihood P(*X*|*T*) is a function of (*k, S, i, X, T*), not simply of F(*S*). The total likelihood of all observed bases (*O*) covering a site P(*O*|*T*) is the product of each one.

From observed prior probability, the SNP rate is expected to be about 0.1%, and the most common SNPs should already be present in dbSNP. Therefore, for positions without known polymorphisms, on one haploid, the reference bases will dominate the prior probability as 0.999; others will share the remaining 0.1% mutation rate. Because sequencing errors would look like HETs, a penalty factor of 0.001 is multiplied to the HET prior probability. At dbSNP sites, bases already observed dominate the prior probability equally and HET penalty factor is 0.01. As a result, the prior probabilities were as follows: a) 0.45 for a homozygote and 0.1 for a heterozygote at a SNP site that has been validated in dbSNP; b) 0.495 for a homozygote and 0.01 for a heterozygote at a SNP site that has not been validated in dbSNP; and c) 1×10^{-6} for a homozygote and 2×10^{-6} for a heterozygote at a potentially novel SNP site (one that is absent in dbSNP).

Using the information above, we calculated the posterior probability of each genotype using a Bayesian formula. The genotype of each position was assigned as the allele type that had the highest probability. A rank sum test was applied to adjust for the probability of heterozygotes. The final consensus probabilities were transformed to quality scores in Phred scale.

Calling SNPs

We used six steps to filter out unreliable portions of the consensus sequence: 1) we used a Q20 quality cutoff; 2) we required at least four reads; 3) the overall depth, including randomly placed repetitive hits, had to be less than 100; 4) the approximate copy number of flanking sequences had to be less than 2 (this was done in order to avoid misreading SNPs as heterozygotes caused by the alignment of similar reads from repeat units or by copy number variations (CNVs)); 5) there had to be at least one paired-end read; and 6) the SNPs had to be at least 5bp away from each other. For Chr X and Y, condition (2) was altered by requiring only 2 unique reads with at least 1 PE. In the SOAP algorithm, a gap-free alignment is done first then a gapped alignment. Thus, we required condition (6) because most of the discrepancies between YH and NCBI reference genome that are too close to each other are due to mismatches across indels. After filtering, we were confident in the calculated YH consensus sequence, and discrepancies between YH and NCBI reference genome were called as SNPs.

Identification of short indels

As number of SNPs is roughly one order of magnitude larger than that of indels, in the first stage of alignment we did not allow any gaps. Thus, some read pairs containing real indels could not be mapped when PE requirements were satisfied. After the first alignment stage, we mapped the unmapped read pairs by allowing up to 3bp insertion/deletion enable them to meet PE requirements. This limited the indels that could be detected in our study to gaps of 1-3bp in length. If different read pairs provided the same outer coordinates in mapping, they are likely to be duplicated products of a single fragment during PCR. We merged these redundant pairs prior to looking for indels. Gaps that were supported by at least 3 non-redundant paired-end

reads were extracted. If the number of ungapped reads that crossed a possible indel was no more than twice that of gapped reads, then an indel was called. In Chr X and Y, we required all indel sites to be covered by only gapped reads because valid indels on sex chromosomes are expected to be homozygous.

Annotation of SNPs and indels

SNPs and indels were compared with NCBI dbSNP v128 to distinguish known SNPs (those that had been deposited to dbSNP) and novel SNPs (those that were not in dbSNP). All the SNPs were annotated by comparing their position to other genomic features including gene regions, repeat elements, etc.

We carried out multi-alignments of cDNAs from different species (rhesus, mouse, dog, opossum, chicken, and stickleback) that covered YH indels that might cause a frameshift. If the YH sequence had the same sequence as the outgroups, we defined it as the ancestral type and considered the genes on NCBI reference to have the frameshift. In regions other than exons, we only used the chimpanzee genome as an outgroup to determine whether the YH or NCBI reference genome had the ancestral allele.

Experimental Validation of SNPs and indels

We use the genotyping platform Illumina HapMap 1M Beadchip to validate our consensus calling. Only those genotypes that provided consistent results between two replicates on the chip were selected for evaluation. We randomly selected 30 CDS and 20 non-CDS disagreements for PCR-Sanger dideoxy sequencing validation using the AB 3730XL. Additionally, 20 indels of 1–2 bp and 4 indels of 3 bp in the coding region and 21 indels in noncoding regions were selected for PCR-Sanger sequencing validation. All intensity trace files were checked manually.

Detection of structural variations

We defined a read pair as a diagnostic paired-end (PE) if the two ends of a read pair 1) could both be aligned but 2) could not meet the pair-end insert size and/or orientation requirement. We grouped abnormally mapped paired-end reads with coordinate distances smaller than the maximum insert size on both ends into diagnostic PE clusters. In order to avoid misalignment, PE clusters with <4 pairs were

discarded.

Common structural variations like deletions, translocations, duplications, inversions etc. were examined and summarized into alignment models. We checked the diagnostic clusters to fit models to detect all possible SVs. If an SV overlapped with another SV in a spanned region, and they could not be combined to form a larger SV, then we deemed such cases as being related to multiple segmental changes and defined them as complex SVs. With SVs that fit proper models, we recovered the YH genome in its own linear structure and searched for reads that crossed the breakpoints. The reads were then assembled to verify the specific coordinates of the SV elements. For those where sequencing depth was too low for us to specify boundaries, we first defined the region where the breakpoint should be using the alignment position of the abnormal paired-end reads. Conservatively, inner coordinates of two possible breakpoint regions were then defined as the boundaries of the SV elements.

This paired-end method (PEM) is substantially biased to deletion events, which is likely related to the fact that deletions can be identified by observing a cluster of paired-ends that have an unexpectedly long insert size. For insertion detection, however, it can be difficult to identify insertions that are longer than the span size of our paired-end libraries. Even with this limitation, we were able to detect candidate regions where large insertion events had possibly occurred. Insertions would interrupt normal paired-end alignment relationships. Thus regions where possible insertions had occurred were not likely to be spanned by normal paired-end reads. We checked all 500-bp–sliding windows on the genome and those that had single-end/paired-end ratios that were significantly ($p<0.001$) larger than the genome average were selected as regions where insertions might have occurred. We defined a "bridge pair" as two reads where one of the paired-end reads mapped to the candidate region and the other mapped to another region on the genome. If a candidate region had at least 20 bridge pairs that linked it to any other elements of same kind (SINE/MIR, LINE/L1, etc, but not necessarily at the same region on NCBI reference given that they are repeats), then a candidate insertion of this element was called.

We compared the regions between two SV boundaries with variants present in

DGV to distinguish those that were known or novel. Those that had less than 10% of a segment that overlapped with known variants were identified as novel. We also compared SV elements identified in our study to genomic features, such as genes, repeat regions, etc. Genomic elements that had a greater than 10% overlap with SV elements were determined to be valid overlaps. SV elements that overlapped with exons (using the above cutoff) were defined as being rearrangements that might delete or substantially alter important gene structures and impact gene function.

Detection of copy number variations

GC% and averaged sequencing depth of every 1 kb sliding window of the YH genome were counted and the depth distribution was modeled to a normal distribution with an estimated mean depth and standard deviation for each level of GC content. Every region with 1) a depth that was significantly ($p<0.0001$) different from that of the whole genome average at the same level of GC content and 2) with flanking sequences that had a depth that was significantly ($p<0.0001$) different from that in the region was deemed potential CNVs.

Mapping and *de novo* assembly of novel sequence

We extracted unmapped reads that had no adapter contamination. Then we used SOAP to align these reads to unplaced human DNA fragments, novel sequences from HuRef (Venter genome), and novel sequences identified by Kidd, J. M. et al.[10]. Regions that were continuously covered were deemed novel sequences in the YH genome that were possibly been lost or not mapped in NCBI reference.

All 487 million unmapped reads went through an error correction process. Reads with N's or 15 mers that were present at <10 frequency were discarded, leaving 299 million reads for analysis. We then used Velvet[12] to assemble these reads. Assembled contigs were aligned against the NCBI core nucleotide database, including other mammal sequences (mouse or chimpanzee), using BLAT. Hits with >=90% identity were considered significant alignments and were categorized as belonging to human sequences or as belonging to other mammalian sequences (mouse and/or chimpanzee).

GO analysis

The gene ontology classification analysis was performed using WEGO[34] (http://wego.genomics.org.cn), a tool for visualizing, comparing, and plotting GO annotation results. We used the Fisher exact test and required p values to be smaller than 0.001 for a specific category of genes in GO classifications, in order for it to be considered significantly different.

Cluster analysis

We selected 87,216 completely genotyped loci that were shared by all 3 human populations that were included in the HapMap project (CHB (Han Chinese in Beijing, China)/JPT (Japanese in Tokyo, Japan), CEU (in Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection), and YRI (Yoruba in Ibadan, Nigeria)) to carry out a cluster analysis and to estimate the proportion of contributions from these populations to the YH donor's ancestry. All 270 HapMap samples together with the YH, Venter (HuRef), and Watson genotypes at these positions were collected to generate 273 multi-dimensional data vectors. The vectors were then clustered hierarchically, based on allele-sharing distance (DAS) using Ward's method[35]. DAS was defined to be:

$$D_{ij} = \frac{1}{L}\sum_{l=1}^{L} d_{ij}^{(l)},$$

where

$$d_{ij}^{(l)} = \begin{cases} 0, \text{ if 2 alleles are shared at the lth locus by individual i and j,} \\ 1, \text{ if 1 allele is shared at the lth locus by individual i and j,} \\ 2, \text{ if no allele is shared at the lth locus by individual i and j.} \end{cases}$$

After setting all 270 HapMap samples to be ancestral with respect to their populations, we used a frequentist program frappe[14] to estimate the donor's ancestry composition.

DNA divergence and Tajima's D test

SNP sets of the YH, Venter, and Watson genomes were aligned based on the NCBI reference coordinates. Watterson's θ_w was used to evaluate the DNA divergence in the population. θ_w was defined as:

$\theta_w = K/(L*a)$, where $a=1+1/2+\ldots+1/(n-1)$

where n is the sample size, K is the number of segregating sites, and L is the number

of total alignment sites. For the YH genome alone n=2, and for the NCBI, Watson, Venter, and YH genomes n=7, respectively.

The Tajima's D test was implemented according to Tajima's original method[26]. Due to the small sample size (the sequence number for the NCBI, Watson, Venter, and YH genome alignment was 7), we selected non-overlapping windows (50 kb) along each chromosome. Regions with p-value<0.05 were selected as candidates of selective sweeps.

Haplotype construction

Haplotypes were constructed by PHASE[13] for all 700,300 known autosomal heterozygous SNPs. Each chromosome was divided into 100 kb windows. The haplotype was predicted on these fragments independently based on known phased genotypes in the Asian population of HapMap phase II. We then merged the neighbouring fragments to carry out another round of prediction. Final haplotypes were assembled to be as long as possible with the merged fragments that had fewer than 2 inconsistent phased SNPs as compared to that of separate fragments.

For evaluation of haplotypes by paired-end reads, we extracted paired-end reads covering two heterozygous SNPs that had been used in phasing, and then compared the nucleotides on these reads to the YH haplotype to check whether the two alleles on read pairs agreed with the phasing results.

References

21. Bhangale, T. R., Rieder, M. J., Livingston, R. J. & Nickerson, D. A. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14,** 59–69 (2005).

22. Zhao, Z. & Boerwinkle, E. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res.* **12,** 1679–1686 (2002).

23. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).

24. Vallender, E. J. & Lahn, B. T. Positive selection on the human genome. *Hum. Mol. Genet.* **13** (special issue no. 2), R245–R254 (2004).

25. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).

26. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).

27. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).

28. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449,** 913–918 (2007).

29. Carlson, C. S. *et al.* Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* **15,** 1553–1565 (2005).

30. Wang, E. T., Kodama, G., Baldi, P. & Moyzis, R. K. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl Acad. Sci. USA* **103,** 135–140 (2006).

31. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nature Rev. Genet.* **8,** 857–868 (2007).

32. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

33. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21,** 577–581 (2003).

34. Ye, J. *et al.* WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* **34,** W293–W297 (2006).

35. Ward, J. H. hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58,** 236–244 (1963).

The Yanhuang Project- Phase I

CONSENT TO PARTICIPATE

Whom can I talk to if I have questions or problems?

If you have questions about this sample collection, contact:
(PI)____Jun Wang_____ (phone) _0755-2527-3796____ _.

If you have questions about your rights as part of this research project, contact:
(IRB)___Songgang Li____(phone) 0755-2527-4287_____.

Consent and Signature

Please read the information below, think about your choice, and sign if you agree:
I agree:

- to be a potential donor for the Yanghuang project;
- that if I will be chosen as a donor for the Yanhuang Project, I agree to give a blood sample and to have a cell line made from the sample that will make an unlimited amount of my DNA for a long time;
- that if I will consent to donate my sample and to have a cell line made from the sample, I agree to have the cell line and DNA used in both the Yanhuang Project and in other approved studies of the type descried in the form;
- that if I will donate my sample, I agree to have the entire genetic code from the sample deposited in a scientific databases on the Internet ;
- that if I will consent to donate my sample, the sample or the data from my sample may be studied by researchers from various organizations, including companies, and that if any commercially valuable products result from these studies, I will not receive any profits; and
- that if I will consent to donate my sample, once the sample has been collected, I cannot withdraw my sample from the Repository in Beijing Genomics Institute in Shenzhen nor take the information about the sample out of the scientific databases.

I have read or listened to the reading materials for Phase I of the Yanghuang Project, I have asked any questions I had, and all my questions were answered. I know that giving a sample is my choice.

Your Signature _____ **Date** _____

Copy given to participant: _____**Yes**

The Yanhuang Project- Phase I

CONSENT TO DONATE

Whom can I talk to if I have questions or problems?

If you have questions about this sample collection, contact:

(PI)____Jun Wang_____ (phone) _0755-2527-3796_____ _.

If you have questions about your rights as part of this research project, contact:

(IRB)___Songgang Li____(phone) 0755-2527-4287_____.

Consent and Signature

Please read the information below, think about your choice, and sign if you agree.

I agree:

- to give a blood sample;
- to have a cell line made from the sample that will make unlimited amount of my DNA for a long time;
- to have the cell line and DNA used in both the Yanhuang Genomes Project and in other approved studies of the type descried in the form;;
- to have the entire genetic code from the sample deposited in a scientific databases on the Internet;
- that the sample or the data from my sample may be studied by various organizations, including companies, and that if any commercially valuable products result from these studies, I will not receive any profits; and
- that once the sample has been studied, I cannot withdraw my sample from the Repository in Beijing Genomics Institute in Shenzhen nor take the information about the sample out of the scientific databases

I have read or listened to the reading materials for Phase I of the Yanghuang Project, I have asked any questions I had, and all my questions were answered. I know that giving a sample is my choice.

Your Signature _____ Date _____

Copy given to participant: _____Yes