

.....

The Human Genome and Its Upcoming Dynamics

M. Platzer

Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute,
Jena, Germany

Abstract

The mapping, sequencing and analysis of the human genome is a milestone in biomedical research and a fundamental advance in self-knowledge. Because the sequence was intended to serve as a universal and permanent foundation of biomedical research, enormous international efforts were undertaken to reach the highest level of accuracy and completeness possible. The current assembly of the 24 DNA molecules covers ~99% of the euchromatic portion. Including gaps, the euchromatin is ~2.88 Gb and the overall size of the human genome ~3.08 Gb. Repeated sequences account for more than half of the human genome. Remarkable is the high proportion of segmental duplications. Until recently it was assumed that tiny variations in an otherwise universal reference sequence are the genetic bases of individual human traits. Now, with the nearly-complete reference in hands, it becomes increasingly evident, that our concept of genome plasticity has to be extended from seemingly fixed human segmental duplications to interindividual, large-scale structural polymorphisms.

Copyright © 2006 S. Karger AG, Basel

‘The human genome underlies the fundamental unity of all members of the human family, as well as the recognition of their inherent dignity and diversity. In a symbolic sense, it is the heritage of humanity.’

Universal Declaration on the Human
Genome and Human Rights
(http://www.unesco.org/human_rights/hrbc.htm)

Humans are much more than simply the product of a genome, but in a genetic sense we are, both collectively and individually, defined within our genome(s). The mapping, sequencing and analysis of the human genome is therefore not only a milestone in biomedical research but also a fundamental advance in self-knowledge. Hopefully, application of this knowledge will, in time, benefit almost everyone in the world.

Because the human genome sequence was intended to serve as a universal and permanent foundation of biomedical research, enormous international efforts were undertaken to reach the highest level of accuracy and completeness currently possible as well as to provide (and preserve!) free and unlimited access to the data. The current sequence (Build 35) represents the first near-complete assembly of the euchromatic portion of a vertebrate genome.

Historical Background of the Human Genome Sequence

The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a highly accurate sequence of the vast majority of the euchromatic portion of the human genome within a 15-year-time period. As our genome is the common heritage of all humanity, the HGP adopted two important principles: collaboration would be open to centres from any nation and, since 1996, rapid and unrestricted data release (referred to as ‘Bermuda rules’) [1]. Thus, the International Human Genome Sequencing Consortium (IHGSC) was formed as an open collaboration involving twenty centres in six countries [2]. In 1998 a biotechnology company, Celera Genomics, initiated its own effort to sequence the human genome [3], rising the danger that basic information would be withheld from the public domain.

In February 2001, the IHGSC [2] and Celera Genomics [4] each reported draft sequences providing an initial overall analysis of the human genome. Both draft sequences, however, had important shortcomings. The IHGSC sequence, e.g., omitted $\sim 10\%$ of the euchromatic genome; it was interrupted by $\sim 150,000$ gaps; and the order and orientation of many had not been established. The IHGSC thus turned to the challenge of completing the sequence of the euchromatic genome. Celera, as a market-based enterprise, did not intend to finish its working draft and, on the background of the public progress of the HGP, cancelled efforts in commercialising access to genome-sequence information and released its assembly to the public databases in 2005 [5].

Operationally, IHGSC defined a finished sequence as having an error rate of mostly one event per 10^4 bases, and the goal for completion was coverage in finished sequence of at least 95% of the euchromatic genome, with the only gaps being those refractory to all available techniques (see <http://www.genome.gov/10000923>). This goal was a challenge because the human genome is rich in dispersed repeats and segmental duplications. In fact, near-complete sequences had been obtained so far only for three multicellular organisms: a nematode [6], mustard weed [7] and the fruitfly [8]. These genomes are all ~ 30 -fold smaller than the human genome and have much simpler repeat structures.

The results of the multi-year finishing effort by the IHGSC were published in October 2004 [9], illustrating the analyses with far reaching biological impact made possible only by the high-quality near-complete sequence. In parallel, a series of papers is being written describing the organisation of the individual chromosomes in detail (<http://www.nature.com/nature/focus/humangenome>) such that the HGP reaches its formal completion by the end of 2005 – as initially envisaged in 1990.

Genomic Landscape

The human nuclear genome comprises 46 chromosomes (22 autosomes in pairs and the gonosomes X and Y). The current assembly of these 24 DNA molecules (Build 35) contains 2.85 billion nucleotides with an error rate of only about one event per 100,000 bases, covers ~99% of the euchromatic portion and contains only 341 gaps (table 1). Thirty-three gaps (~198 Mb) reflect heterochromatin, which was not targeted by the HGP, and 308 gaps (~28 Mb) are euchromatic. Gap sizes were estimated by interphase and fibre FISH as well as interphase nuclei mapping [10, 11]. Thus, the euchromatic portion is ~2.88 Gb and the overall size of the human genome ~3.08 Gb.

For the 43 euchromatic chromosomal arms the proximal heterochromatic and the distal telomeric boundaries were identified in 30 and 21 cases, respectively. But even in these cases it cannot be excluded that there are additional unique sequences beyond the ends of the current reference assemblies, as e.g. a euchromatic 450-kb island recently identified within the pericentromeric repeats of the human Y chromosome [12]. More than half of the intra-euchromatic physical gaps are flanked by segmental duplications with ~90% sequence identity, although such duplications comprise only ~5% of the euchromatic genome [13]. The most extreme case occurs near the centromere of chromosome 9. The proximal 5 Mb on 9p and 4 Mb on 9q comprise 0.3% of the genome, but account for ~12% of the physical gaps in the euchromatic sequence. These two pericentric regions are unique in the genome with respect to density of segmental duplication and average degree of intrachromosomal sequence identity (98.7%; for more details see fig. 3 in [14]). The high sequence similarity between the two regions is likely to be the reason for a polymorphic inversion in the centric heterochromatin on chromosome 9, present with a 1%-frequency in the human population. Thus, targeted efforts and novel techniques have to be applied to resolve the remaining ~1% of the euchromatin residing in the gaps, although the current sequence has already reached a much higher degree of completion than initially anticipated.

The local GC content of the human genome undergoes substantial long-range deviations from its genome-wide average of 41%. There are huge regions

Table 1. Finished sequence and gaps (Build 35) [9]

Chromosome	Finished sequence, kb	Gaps	
		Euchromatic ^a	Heterochromatic
1	222,828	49	2
2	237,503	20	1
3	194,636	5	1
4	187,161	14	1
5	177,703	5	1
6	167,318	10	1
7	154,759	11	1
8	142,613	9	1
9	117,781	52	2
10	131,614	20	1
11	131,131	7	1
12	130,259	8	1
13	95,560	6	2
14	88,291	1	2
15	81,342	10	2
16	78,885	4	2
17	77,800	9	1
18	74,656	3	1
19	55,786	5	1
20	59,505	4	1
21	34,170	3	2
22	34,765	11	2
X	150,394	26	1
Y	24,872	16	2
Total	2,851,331	308	33

^aSubsumes physical gaps (for which no clone was available, totalling 215), sequence gaps (for which clones were found but reliable finished sequence could not be obtained, totalling 58) and gaps in the euchromatic boundary regions (totalling 35).

(>40 Mb) with a GC content far from the average, e.g. 47% GC on chromosome 1 or only 36% GC on chromosome 13. There are large shifts in GC content between adjacent multi-megabase regions in less than 300 kb with even wider swings in GC content, e.g., from 33% to 59%. Referring the 'isochore'-concept [15], which proposes that the long-range variation in GC content may reflect that the genome is composed of a mosaic of compositionally homogeneous regions of ~300 kb, substantial variation at many different scales can be observed and, in the

absence of a precise definition, the term ‘GC content domains’ was proposed [2]. The association between GC content domains and biological properties is of great interest. Strong correlations exist with repeat content, gene density and the cytogenetic Giemsa bands (darkest G-bands – low GC; lightest G-bands – high GC) [16].

‘CpG islands’, i.e. regions in which CpG dinucleotides are not methylated and occur at a frequency close to that predicted by the local GC content, are of particular interest because many are associated with the 5’ ends of genes [17]. Outside of these islands, the dinucleotide CpG is greatly ($\sim 5\times$) under-represented. This deficit occurs because these CpG dinucleotides are mostly methylated on the cytosine, and spontaneous deamination of methyl-C residues gives rise to thymine. As a result, methyl-CpG dinucleotides steadily mutate to TpG and CpA on the reverse strand, respectively. Using standard parameters (length >250 bp, ratio observed/expected CpGs >0.6) [18], $\sim 29,000$ CpG islands can be localized in the non-repetitive portion of the genome and $\sim 21,000$ within repeats (notably ‘Alu’-repeats which are GC-rich). More than 75% of the former consist of less than 850 bp. The longest CpG island (on chromosome 10) is >36 kb, and only about 300 are longer than 3 kb. The role of these large islands is uncertain, but many of the smaller islands are consistent with their previously hypothesized function as part of primary or alternative promoters. Nevertheless, it seems unlikely that most of the very small apparent CpG islands are functional. Own results of experimentally testing for the absence of cytosine methylation [unpublished] and *in silico* analyses of others [19] revealed more specific parameters for functional CpG island prediction in the human genome (>500 bp and $o/e >0.55$). The density of CpG islands varies substantially among chromosomes and correlates well with the gene density. The extreme outliers are chromosomes Y and 19 with ~ 3 and ~ 43 islands per Mb as well as ~ 3 and ~ 26 genes per Mb, respectively.

Repeat Content

Repeated sequences account for more than half of the human genome. Repeats fall into five classes: (1) transposon-derived repeats, referred to as interspersed repeats; (2) inactive retroposed copies of cellular genes, referred to as processed pseudogenes; (3) simple sequence repeats (SSR), consisting of direct repetitions of short oligomers; (4) segmental duplications, consisting of blocks of around 10–300 kb that have been copied from one region of the genome into another region; and (5) blocks of tandemly repeated sequences (centromeres, telomeres, the short arms of acrocentric chromosomes and ribosomal gene clusters that were beyond the euchromatic focus of the HGP and thus are underrepresented in the current assembly).

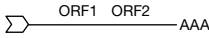
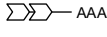
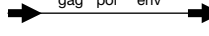
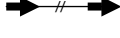


			Length kb	Number × 1,000	Genome fraction
LINES	Autonomous		6–8	850	21%
	Non-autonomous		0.1–0.3	1,500	13%
LTR retro-transposons	Autonomous		6–11	450	8%
	Non-autonomous		1.5–3		
DNA transposons	Autonomous		2–3	300	3%
	Non-autonomous		0.1–3		

Fig. 1. Classes of interspersed repeats in the human genome. Empty arrow: promoter; filled arrow: LTR (long terminal repeat); triangle: short terminal repeat.

Currently, ~45% of the euchromatic genome can be recognized as transposon-derived repeats. A considerable part of the remaining ‘unique’ DNA is also very likely derived from ancient transposon copies that have diverged too far to be recognized as such. In mammals, almost all transposable elements fall into one of four classes: (1) long interspersed elements (LINES), (2) short interspersed elements (SINEs), (3) LTR retrotransposons, and (4) DNA transposons (fig. 1). The first three transpose through RNA intermediates and the last directly as DNA by cut-and-paste mechanism.

LINES are one of the most ancient and successful transposons in eukaryotic genomes. In humans, three LINE families are found: LINE1, LINE2 and LINE3. Only LINE1 is still active and transposition is associated with comobilisation, deletion and inversion of flanking sequences [20]. The LINE machinery is believed to be responsible for most reverse transcription in the genome, including the creation of processed pseudogenes [21] and the retrotransposition of the non-autonomous SINEs [22]. SINEs do not encode proteins. The internal polymerase III promoters of all known SINEs are derived from tRNA sequences, with the exception of a single monophyletic family derived from the signal recognition particle component 7SL [23]. The human genome contains three distinct SINEs: the active 7SL-derived Alu, and the inactive MIR and Ther2/MIR3. LTR retrotransposons are autonomous elements falling into three classes each comprising many families with independent origins. 85% of the human LTR retrotransposon-derived ‘fossils’ consist only of an isolated LTR, with the internal sequence lost by homologous recombination between the flanking LTRs. DNA transposons in the human genome resemble bacterial transposons and fall into at least seven major groups, which can be subdivided into many families with independent origins.

The repeat content varies strikingly across the genome on the hundreds-kb scale between 89% (Xp11) and 2% (HOX gene clusters). The absence of repeats may be an indication of large-scale cis-regulatory elements that cannot tolerate being interrupted by insertions. There also exists a correlation between GC content and repeat distribution [24]. LINE sequences occur at much higher density in AT-rich regions ($4\times$), whereas SINEs (MIR, Alu) are enriched in regions with high GC content ($5\times$ for Alu).

SSRs, called ‘microsatellites’ with a repeat unit of 1–13 bases and ‘minisatellites’ with 14–500 base elements, comprise about 3% of the human genome, with the greatest single contribution coming from dinucleotide repeats (0.5%) [2]. On average, there is one SSR per 2 kb. With the exception of poly(A) tails from reverse transcribed messages, SSRs are thought to arise by slippage during DNA replication [25].

The human genome is remarkable for its high proportion of recent segmental duplications [26]. These are defined regions that are not transposable element copies, with ≥ 1 kb in length and sequence identity $\geq 90\%$ (which corresponds to duplication within the past ~ 40 million years (Myr)). Accurate analysis of segmental duplications became feasible only with the near-complete sequence, where artefacts (collapsed assemblies, artefactual duplications) are largely eliminated [9]. On this basis, segmental duplications cover $\sim 5.3\%$ of the euchromatic genome. The most extreme case is chromosome Y, which carries segmental duplication along $\sim 25\%$ of its total length and includes blocks as large as ~ 1.45 Mb with sequence identity of $\sim 99.97\%$ [27]. In addition, many pericentromeric and subtelomeric regions are particularly rich in dispersed segmental duplications [28].

Gene Content, Birth and Death

One of the central goals of the HGP was the compilation of an as complete as possible list of all human genes to serve as a universal foundation of biomedical research. Although this task remains challenging and will not be finished in the near future, the probably most surprising result of the HGP is certain: the human genome contains only little more genes than the nematode [6] and the fruitfly [8], and even less than mustard weed [7]. However, the genes are more complex, with more alternative splicing generating a larger number of protein products. More than 70% of human multi-exon genes show evidence for alternative splicing [29]. Moreover, the full set of proteins (proteome) is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs ($\sim 7\%$ of the total), but more to the fact that

vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures [2].

The current version of the human gene catalogue (Ensembl 34.35 g, Nov 2004), greatly aided by the near-complete sequence together with other improved resources (such as expanded cDNA collections, genome sequences from other organisms and better computational methods), contains 24,194 gene loci (with a total of 35,845 transcripts, corresponding to 1.48 transcripts per locus), including 1,976 pseudogenes. These gene loci have a total of 245,231 exons, with ~ 10.1 exons per locus. The total length covered by the coding exons is ~ 34 Mb or $\sim 1.2\%$ of the euchromatic genome; the untranslated regions of the processed transcripts are estimated to cover another ~ 21 Mb or $\sim 0.7\%$ of the euchromatic genome [9].

In addition to protein-coding genes the human genome contains thousands of genes with so called ‘non-coding’ RNAs (ncRNAs, non-protein-coding) as their ultimate product [30]. Their overall number is currently much more obscure than that of the protein-coding genes. As they are often small and not polyadenylated, established cDNA sequencing efforts fail and effective computational gene-finding techniques are still missing. High resolution mapping of polyadenylated and nonpolyadenylated RNAs revealed, that unannotated, nonpolyadenylated transcripts comprise the major proportion of the transcriptional output of the human genome [31] and was supported by a recent comprehensive analysis of the transcriptional landscape of the mouse genome [32].

About 500 tRNA genes are dispersed throughout the human genome, showing a striking clustering on a genome-wide scale. More than 25% of the tRNA genes are found in a region of only about 4 Mb on chromosome 6. More than half of the tRNA genes reside on either chromosome 1 or chromosome 6. Chromosomes 3, 4, 8, 9, 10, 12, 18, 20, 21 and X appear to have fewer than ten tRNA genes each; and chromosomes 22 and Y have none. Two-thirds of tRNA clusters colocalize with transcriptional hotspots (e.g. MHC on chromosome 6) and it was postulated that selection-mediated recruitment and/or hitchhiking are responsible for the observed clustering and localization [33]. The genes for rRNAs occur in the human genome as a 44-kb tandem repeat. It is assumed, that ~ 150 – 200 copies of this repeat unit are arrayed on the short arms of acrocentric chromosomes 13, 14, 15, 21 and 22 [34]. In addition, hundreds of rDNA-derived sequence fragments are dispersed throughout the complete genome, including one ‘full-length’ copy of an individual 5.8S rRNA gene which is not associated with a true tandem repeat unit. Small nucleolar RNAs (snoRNAs), responsible for rRNA modification, are almost all expressed from single copy genes. In contrast, there are multiple copies for several spliceosomal snRNAs; e.g., 44 dispersed genes for U6 snRNA, and 16 for U1 snRNA. The U2 RNA genes are

located in a tandem array of 10–20 copies of nearly identical 6.1-kb units [35]. Similarly, the U3 snoRNA genes are clustered, not in a tandem array, but in a complex inverted repeat structure of about 5–10 copies per haploid genome [36]. To date, 222 human microRNAs have been identified and recent findings suggest that the total number is at least 800, many of them clustered [37]. In addition, there is a striking proliferation of ncRNA-derived pseudogenes. There are thousands of sequences in the genome related to some of the ncRNA genes. The most prolific pseudogene counts come from RNA genes transcribed by RNA polymerase III promoters, including U6, the hY RNAs and SRP-RNA.

In general, a comprehensive search for pseudogenes requires a high quality genome sequence. Recent sensitive methods to detect even small and old pseudogenes have already identified ~20,000 processed (originated by reverse transcription of mRNAs) and unprocessed (originated by segmental duplications) pseudogenes [38]. This is certainly still an underestimate, because such analysis will miss pseudogenes that are extremely old or that contain primarily untranslated regions. Therefore, the total number of pseudogenes is likely to exceed the total number of functional genes.

In particular, recently arisen non-processed pseudogenes (nearly intact human genes that appear to have acquired an inactivating mutation) shed light on the phenomenon of gene death in the human lineage. A careful analysis identified 32 respective pseudogenes fixed in the human population, including ten that are derived from olfactory receptor genes [9].

The birth of new genes is of interest because extra copies of genes are able to undergo functional divergence in response to positive selection. Such analysis would have been unreliable with a draft sequence, because the artefactual local duplication and collapsed assemblies would have given rise to many false positives and negatives, respectively. Mining the near-complete human euchromatin, recent duplications are enriched in genes with immune and olfactory function, as well as those likely to be involved in reproductive functions. For example the family of cancer/testis antigen genes, which are normally expressed in the testis and show high expression in carcinomas [39] are, surprisingly, enriched on ‘female’ chromosome X [40]. The distribution of duplications shows a striking excess of genes corresponding to recent events occurring ~3–4 Myr ago [9]. Possible explanations are: (1) an explosion in the rate of gene duplication in the human lineage, (2) an ongoing process of gene conversion of older gene duplications, or (3) a transient state of duplicated genes that are too young relative to the characteristic time of deletion.

Olfactory genes occur prominently in both birth and death analyses, indicating a dynamic extension and contraction of this gene family. In accordance, also ‘resurrection’ of pseudogenes by gene conversion may generate diversity at the odorant binding sites [41].

Evolutionary Genome Dynamics

Mobile elements are drivers of genome evolution [42] and their age distribution in the human genome provides a rich ‘fossil record’ stretching over several hundred Myr. Determining the phylogeny of all three million interspersed repeats, several conclusions can be drawn [2]. First, most interspersed repeats in the human genome predate the eutherian radiation. This indicates the extremely slow rate with which nonfunctional sequences are cleared from vertebrate genomes (in contrast e.g. to the fly genome). Second, LINE and SINE elements are active over extremely long periods. The monophyletic LINE1 and Alu lineages are at least 150 and 80 Myr old, respectively. In earlier times, the prevailing transposons were LINE2 and MIR [23] until LINE2 became extinct 80–100 Myr ago. Third, there were two major peaks of DNA transposon activity. The first occurred long before the eutherian radiation; the second after this period. Because DNA transposons can cause large-scale chromosome rearrangements [42, 43], it is possible that this widespread activity could be involved in speciation events. Fourth, there is no evidence for DNA transposon activity in the past 50 Myr in the human genome. Finally, LTR retroposon activity appears to be extinct. Only a single LTR retroposon family (HERVK10) is known to have transposed since our divergence from the chimpanzee 7 Myr ago, with only one known copy (in the HLA region) that is not shared between all humans [44].

More generally, the overall activity of all transposons has declined markedly over the past 35–50 Myr. Indeed, apart from an exceptional burst of activity of Alus peaking since the mammalian radiation around 40 Myr ago a steady decline in activity can be observed in the hominid lineage. A possible explanation refers to hominid populations that tend to be small and underwent frequent bottlenecks. Evolutionary forces affected by such factors (inbreeding, genetic drift) might restrict the persistence of active transposable elements, as it is observed during long-term laboratory inbreeding of *C. elegans* [45].

Strikingly, recently transposed Alus show a target site preference for AT-rich DNA resembling that of LINEs, whereas progressively older Alus show a progressively stronger bias towards GC-rich DNA. These observations may be interpreted as the result of positive selection in favour of Alus in GC-rich regions within the last 30 Myr [2]. It was supposed that the beneficial effect of Alus in GC-rich DNA may result from SINE-RNA-mediated promotion of translation under stress [46].

There is strong evidence that the nucleotide substitution pattern varies as a function of local GC content [2]. GC base pairs are more likely to mutate towards AT in AT-rich regions than in GC-rich ones. This bias could be due to the tendency for GC-rich regions to replicate earlier in the cell cycle than AT-rich regions together with the fact that guanine pools become limiting during DNA

replication. Late in the S-phase, the depleted guanine may result in a small but significant shift in substitution towards AT base pairs [47]. Another theory proposes differences in DNA repair mechanisms, possibly related to transcriptional activity and thereby to gene density and GC content [48]. Moreover, there is an absolute bias in substitution patterns resulting in directional pressure towards lower GC content throughout the human genome. The human genome is not at equilibrium with respect to the pattern of nucleotide substitution. On the basis of nucleotide substitution patterns, the GC content would be expected to be about 7% lower throughout the genome [2]. But selection on coding regions and regulatory CpG islands may maintain the higher-than-predicted GC content.

Segmental duplications have recently been shown to be evolutionary nurseries in which coding sequences are undergoing strong positive selection [49] and have created much more of human differences from chimpanzee than single base-pair differences [50]. Moreover, the evolutionary plasticity of the human genome is illustrated by the fact that some regions of our genome are more closely related to gorilla than to chimpanzee [51]. By whole genome sample sequencing of gorilla, orangutan and rhesus macaque we could show that for about one fifth of the human genome the chimpanzee is not our closest relative [unpublished].

Polymorphisms and Interindividual Genome Dynamics

SSRs show a high degree of length polymorphism in the human population owing to frequent slippage by DNA polymerase during replication. Thus, genetic markers based on SSRs – particularly $(CA)_n$ repeats – have been extremely important in most human disease-mapping studies until recently. A longstanding mystery of vertebrate genetic maps – the observed deficiency of polymorphic $(CA)_n$ repeats on chromosome X – is not due to a smaller fraction of these SSRs there [2, 40]. It results either from population genetic forces (the smaller effective population size, owing to its hemizyosity in males, may lead to more frequent selective sweeps reducing diversity), or a lower mutation rate (owing to its more frequent passage through the less mutagenic female germ line).

Single-nucleotide polymorphisms (SNPs) are the most common variants in the human genome and have replaced SSRs in recent association studies [52]. Any two copies of the human genome differ from one another by approximately 0.1% of nucleotide sites (one variant per 1,000 bases on average) [53]. It has been estimated that, in the world's human population, about ten million sites (one variant per 300 bases on average) vary such that both alleles are observed at a frequency of $\geq 1\%$, and that these ten million common SNPs constitute 90% of the

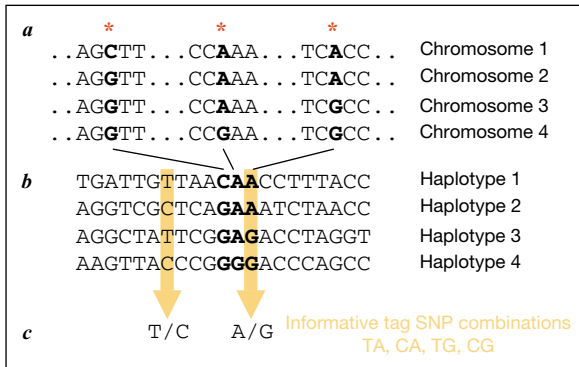


Fig. 2. SNPs, haplotypes and tag SNPs. **a** Three SNPs (red asterisks) within a short segment of four alleles of the same chromosome. **b** Four haplotypes comprised of 21 consecutive SNP alleles from larger regions of the same four chromosomes. **c** Two ‘tag SNPs’ are sufficient to distinguish the four haplotypes in genotyping efforts.

variation in the population [54] (fig. 2a). dbSNP (www.ncbi.nlm.nih.gov/projects/SNP), the most comprehensive public SNP database, currently contains already more than ten million human SNPs (version 124). But it has to be taken into account, that a recent metaanalysis of four confirmation studies estimated a false positive rate of ~15–17% [55].

The specific set of SNP alleles observed on a single chromosome, or part of a chromosome, is called a haplotype (fig. 2b). The coinheritance of SNP alleles on haplotypes leads to associations between these alleles in the population (known as linkage disequilibrium, LD). The strong associations between SNPs in a region have a practical value: genotyping only a few, carefully chosen ‘tag SNPs’ in the region will provide enough information to predict much of the information about the remainder of the common SNPs in that region (fig. 2c) [56].

The recently finished International HapMap Project [57] provides a public database of more than one million single SNPs for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations (<http://www.hapmap.org>). The analyses confirm the generality of hotspots of recombination, long segments of strong LD, and limited haplotype diversity. Most important is the extensive redundancy among nearby SNPs, providing (1) the potential to extract extensive information about genomic variation without complete resequencing, and (2) efficiencies through selection of tag SNPs and optimized association analyses.

Until recently it was assumed that tiny variations like SNPs and haplotypes thereof in an otherwise universal reference sequence are the genetic bases of individual human traits. Now, with this nearly-complete reference in hands, it

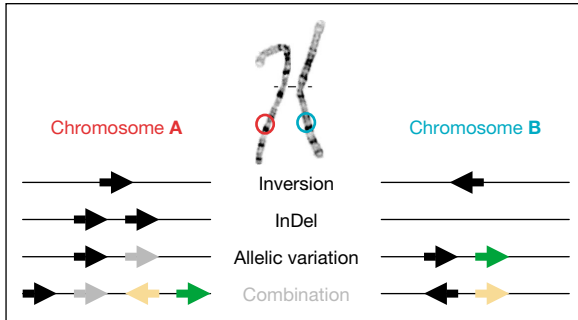


Fig. 3. Structural variants in the human genome. Arrows represent a particular locus, e.g. a gene or a group of genes. Different colours represent allelic variations of the locus. ‘InDel’ stands for insertion/deletion and ‘Combination’ for any possible combination of inversion, InDel and/or allelic variation.

becomes increasingly evident, that our concept of genome plasticity has to be extended from seemingly fixed human segmental duplications [26, 58] to interindividual, large-scale structural polymorphisms (copy-number variations owing to insertions/deletions and/or inversions; fig. 3). Few examples of large duplication polymorphisms have been reported in the past [59]. However, the findings of hundreds of these polymorphisms with a size >100 kb [60, 61] and with a size >8 kb [62] suggest a widespread phenomenon. Owing to their size and gene content, these polymorphisms are unlikely to be selectively neutral. In particular, segmental duplications are of great medical interest because their structure often predisposes them to large-scale copy number polymorphisms with consequent phenotypic effects [63, 64]. This is consistent with the growing body of findings showing how inversions [65, 66] and copy-number variations [67] influence the predisposition to chromosomal rearrangements, higher fertility rates and susceptibility to viral infections or cancer [unpublished]. An attempt to consolidate information on structural genomic variants is currently being made by the Database of Genomic Variants project (<http://projects.tcag.ca/variation>). In this respect it has to be mentioned, that multisite variations, a new type of polymorphism representing the sum of the signals from many individual duplicon copies that vary in sequence content due to duplication, deletion or gene conversion, can masquerade as normal SNPs when genotyped [68]. Thus, new effective strategies must be established and deployed to identify multisite variation.

The recently detected large-scale heterogeneity argues for a more dynamic structure of the human genome than previously anticipated. The estimate of 99.9% total sequence identity between any two human beings will need to be reassessed over the coming years. In comparison to the human genome, the

mouse is thought to differ in having less recent segmental duplications [69]. As this and other vertebrate genome projects reach a comparable level of completeness and quality, it may be possible to determine whether the human genome is exceptional in respect to this dynamics.

Acknowledgements

I thank Ulrike Gausmann and Gernot Glöckner for critical reading of the manuscript.

References

- 1 Guyer M: Statement on the rapid release of genomic DNA sequence. *Genome Res* 1998;8:413.
- 2 International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
- 3 Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M: Shotgun sequencing of the human genome. *Science* 1998;280:1540–1542.
- 4 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al: The sequence of the human genome. *Science* 2001;291:1304–1351.
- 5 Kaiser J: Genomics. Celera to end subscriptions and give data to public GenBank. *Science* 2005;308:775.
- 6 *C. elegans* Sequencing Consortium: Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998;282:2012–2018.
- 7 Arabidopsis Genome Initiative: Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;408:796–815.
- 8 Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, et al: Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol* 2002;3:0079.0071–0079.0014.
- 9 International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004;431:931–945.
- 10 Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, et al: The DNA sequence of human chromosome 22. *Nature* 1999;402:489–495.
- 11 Hattori M, Fujiiyama A, Taylor TD, Watanabe H, Yada T, et al: The DNA sequence of human chromosome 21. *Nature* 2000;405:311–319.
- 12 Kirsch S, Weiss B, Miner TL, Waterston RH, Clark RA, et al: Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Res* 2005;15:195–204.
- 13 Eichler EE, Clark RA, She X: An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet* 2004;5:345–354.
- 14 Humphray SJ, Oliver K, Hunt AR, Plumb RW, Loveland JE, et al: DNA sequence and analysis of human chromosome 9. *Nature* 2004;429:369–374.
- 15 Clay O, Bernardi G: How not to search for isochores: A reply to Cohen et al. *Mol Biol Evol* 2005;22:2315–2317.
- 16 Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, et al: Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* 2001;409:953–958.
- 17 Bird AP: CpG islands as gene markers in the vertebrate nucleus. *Trends Genet* 1987;3:342–347.
- 18 Gardiner-Garden M, Frommer M: CpG islands in vertebrate genomes. *J Mol Biol* 1987;196: 261–282.
- 19 Takai D, Jones PA: Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci USA* 2002;99:3740–3745.
- 20 Kazazian HH Jr, Goodier JL: LINE drive: retrotransposition and genome instability. *Cell* 2002;110:277–280.

- 21 Esnault C, Maestre J, Heidmann T: Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* 2000;24:363–367.
- 22 Okada N, Hamada M, Ogiwara I, Ohshima K: SINEs and LINEs share common 3' sequences: a review. *Gene* 1997;205:229–243.
- 23 Smit AF: The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* 1996;6:743–748.
- 24 Smit AF: Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 1999;9:657–663.
- 25 Toth G, Gaspari Z, Jurka J: Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 2000;10:967–981.
- 26 Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al: Recent segmental duplications in the human genome. *Science* 2002;297:1003–1007.
- 27 Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al: The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 2003;423:825–837.
- 28 Horvath JE, Bailey JA, Locke DP, Eichler EE: Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum Mol Genet* 2001;10:2215–2223.
- 29 Johnson JM, Castle J, Garrett-Engel P, Kan Z, Loerch PM, et al: Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 2003;302:2141–2144.
- 30 Eddy SR: Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* 2001;2:919–929.
- 31 Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al: Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 2005;308:1149–1154.
- 32 Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al: The transcriptional landscape of the mammalian genome. *Science* 2005;309:1559–1563.
- 33 Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, et al: The DNA sequence and analysis of human chromosome 6. *Nature* 2003;425:805–811.
- 34 Sylvester JE, Whiteman DA, Podolsky R, Pozsgay JM, Respass J, Schmickel RD: The human ribosomal RNA genes: structure and organization of the complete repeating unit. *Hum Genet* 1986;73:193–198.
- 35 Pavelitz T, Liao D, Weiner AM: Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. *EMBO J* 1999;18:3783–3792.
- 36 Gao L, Frey MR, Matera AG: Human genes encoding U3 snRNA associate with coiled bodies in interphase cells and are clustered on chromosome 17p11.2 in a complex inverted repeat structure. *Nucleic Acids Res* 1997;25:4740–4747.
- 37 Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, et al: Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* 2005;37:766–770.
- 38 Torrents D, Suyama M, Zdobnov E, Bork P: A genome-wide survey of human pseudogenes. *Genome Res* 2003;13:2559–2567.
- 39 Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen YT: Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol Rev* 2002;188:22–32.
- 40 Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al: The DNA sequence of the human X chromosome. *Nature* 2005;434:325–337.
- 41 Glusman G, Yanai I, Rubin I, Lancet D: The complete human olfactory subgenome. *Genome Res* 2001;11:685–702.
- 42 Kazazian HH Jr: Mobile elements: drivers of genome evolution. *Science* 2004;303:1626–1632.
- 43 Gray YH: It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* 2000;16:461–468.
- 44 Medstrand P, Mager DL: Human-specific integrations of the HERV-K endogenous retrovirus family. *J Virol* 1998;72:9782–9787.
- 45 Malik HS, Eickbush TH: NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics* 2000;154:193–203.
- 46 Liu WM, Chu WM, Choudary PV, Schmid CW: Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* 1995;23:1758–1765.

- 47 Wolfe KH, Sharp PM, Li WH: Mutation rates differ among regions of the mammalian genome. *Nature* 1989;337:283–285.
- 48 Eyre-Walker A: Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* 1999;152:675–683.
- 49 Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, et al: Positive selection of a gene family during the emergence of humans and African apes. *Nature* 2001;413:514–519.
- 50 Cheng Z, Ventura M, She X, Khaitovich P, Graves T, et al: A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 2005;437:88–93.
- 51 Paabo S: The mosaic that is our genome. *Nature* 2003;421:409–412.
- 52 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996;273:1516–1517.
- 53 Wang DG, Fan JB, Siao CJ, Berno A, Young P, et al: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077–1082.
- 54 Reich DE, Gabriel SB, Altshuler D: Quality and completeness of SNP databases. *Nat Genet* 2003;33:457–458.
- 55 Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ: Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics* 2004;20:1022–1032.
- 56 The International HapMap Consortium: The International HapMap Project. *Nature* 2003;426:789–796.
- 57 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005;437:1299–1320.
- 58 She X, Jiang Z, Clark RA, Liu G, Cheng Z, et al: Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 2004;431:927–930.
- 59 Buckland PR: Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann Med* 2003;35:308–315.
- 60 Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al: Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–951.
- 61 Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al: Large-scale copy number polymorphism in the human genome. *Science* 2004;305:525–528.
- 62 Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al: Fine-scale structural variation of the human genome. *Nat Genet* 2005;37:727–732.
- 63 Stankiewicz P, Lupski JR: Genome architecture, rearrangements and genomic disorders. *Trends Genet* 2002;18:74–82.
- 64 Taudien S, Galgoczy P, Huse K, Reichwald K, Schilhabel M, et al: Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. *BMC Genomics* 2004;5:92.
- 65 Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, et al: A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet* 2001;29:321–325.
- 66 Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, et al: A common inversion under selection in Europeans. *Nat Genet* 2005;37:129–137.
- 67 Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, et al: The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 2005;307:1434–1440.
- 68 Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ: Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004;36:861–866.
- 69 Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al: Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;420:520–562.

Matthias Platzer

Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute
Beutenbergstr. 11, 07745 Jena (Germany)
Tel. +49 3641 656241, Fax +49 3641 656255, E-Mail mplatzer@fli-leibniz.de