

Review

Has the yo-yo stopped? An assessment of human protein-coding gene number

Christopher Southan

Oxford GlycoSciences, Abingdon, UK

Since the identification of ~ 25 000 proteins from the draft human genome assembly in 2001, estimates of the total have oscillated between 30 000 and 70 000. The recently announced genome closure has not generated a consensus gene count despite this being a key parameter for many areas of biology including drug target discovery and characterization of the human proteome. Contrary to earlier predictions of constitutive under-detection for eukaryotic genes, the latest model organism updates have produced minor increases in the worm but fly and yeast gene numbers have decreased. The postdraft, precompletion interval has produced large increases in human transcript coverage, continuous improvements in genome assembly and refinements in automated genomic annotation. Notably these enhancements have resulted in an Ensembl human protein-coding gene number of 22 184, a decrease of 1862 since the first release. Longitudinal database surveys indicate that redundancy-reduced human mRNA and protein collections are flattening out at ~ 28 000, although Ensembl maps ~ 20 000 known sequences. Observations suggest high-throughput cloning projects are predominantly extending known genes or sampling new splice forms and novel protein discovery has slowed to a trickle. The hypothesis that substantial numbers of short proteins remain experimentally and computationally undetected in mammalian genomes is neither supported by sequence data nor by the extensive homology between mouse and human proteins. Aggregating the independent annotations for complete transcripts from seven completed human chromosomes extrapolates to ~ 25 000 genes. The inclusion of partial putative genes would increase this to above 30 000 but recent data suggest these represent predominantly nonprotein-coding transcripts. Mass spectrometry-based proteomics has already verified more than 10% of human genes but has not identified significant numbers of unpredicted proteins. The available data are thus converging to a basal protein-coding gene number well below 30 000, which could even be as low as 25 000.

| | |
|----------|----------|
| Received | 8/7/03 |
| Revised | 18/11/03 |
| Accepted | 1/12/03 |

Keywords: Genes / Human genome / Proteome / Review / Transcriptome

Contents

| | | |
|---|--|------|
| 1 | Introduction | 1713 |
| 2 | Discovering eukaryotic genes | 1713 |

Correspondence: Dr. Chris Southan, AstraZeneca R&D, S-431-83, Mölndal, Sweden
E-mail: christopher.southan@astrazeneca.com
Fax: +46-31-776-3851

Abbreviations: **CDS**, coding sequences; **GP**, golden path; **IPI**, International Protein Index; **MPSS**, massively paralleled signature sequencing; **NCBI**, National Center Biotechnology Information; **smORF**, small open reading frame; **SPT**, SWISS-PROT/TrEMBL; **TARs**, transcriptionally active regions; **UTR**, untranslated region; **VEGA**, Vertebrate Genome Annotation

| | | |
|-----|--|------|
| 3 | High gene number arguments | 1714 |
| 4 | Gene numbers in model eukaryotes | 1714 |
| 4.1 | Genome annotation of model eukaryotes | 1714 |
| 4.2 | Post-completion gene number changes | 1714 |
| 4.3 | Proteomic sampling in model eukaryotes | 1715 |
| 5 | Annotation of the human genome | 1715 |
| 5.1 | The golden path genome assembly | 1715 |
| 5.2 | Ensembl gene totals | 1715 |
| 5.3 | Underlying trends | 1716 |
| 5.4 | Curation of completed chromosomes | 1716 |
| 5.5 | Comparing curated chromosomes with automated genome annotation | 1717 |
| 5.6 | Pseudogenes | 1718 |

| | | |
|-----|--|------|
| 6 | Post-genomic coverage of protein and transcript data | 1718 |
| 6.1 | Human protein databases | 1718 |
| 6.2 | Small proteins. | 1719 |
| 6.3 | Human mRNA databases | 1719 |
| 6.4 | EST coverage of mRNAs | 1720 |
| 6.5 | Mammalian protein increases | 1721 |
| 7 | Comparing human with other vertebrate genomes. | 1721 |
| 8 | Evidence of non-coding mRNA transcription | 1722 |
| 9 | Increasing the stringency of gene identification. | 1722 |
| 10 | Diminishing novel gene discovery. | 1723 |
| 11 | Proteomic sampling of human proteins . . . | 1723 |
| 12 | Conclusions | 1724 |
| 13 | References | 1724 |

1 Introduction

During the run up to the human genome in the year 2000, there was some surprise when the second, but arguably more complex, metazoan genome sequence from the fly turned out to have some 5000 less genes than the worm. This surprise was compounded when both the public and commercial versions of human sequence facilitated the annotation of only ~ 25 000 high confidence genes with an estimated maximum upper boundary of ~ 35 000 [1, 2]. However, a range of higher estimates with upper boundaries of 70 000 have continued to appear (see [3] for review). A selection of these is shown in Fig. 1.

This review will consider recent data supporting the number of human protein-coding genes, although there is evidence that higher mammals share similar gene numbers [6]. As was pointed out after the initial analysis of the human genome, gene number is a central issue for biological

complexity [7]. It is also a key parameter for other aspects of human biology such as defining the upper limits for the number of potential drug targets or therapeutic proteins and setting the baseline for initiatives to characterize the human proteome [8, 9].

The Guidelines for Human Gene Nomenclature define a gene as follows: “a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology” [10]. This review will be restricted to assessing numbers of protein-coding genes, defined as chromosome-derived transcripts giving rise to one or more protein forms with shared sequence identity that assign them as products of a single genomic locus and strand orientation. This must be considered as a baseline number because vertebrates produce many protein forms *via* multiple initiations, alternative splicing, post-translational modifications, constitutive proteolytic processing, and common genetic polymorphisms. The mammalian proteome has therefore been estimated to be at least an order of magnitude higher than the protein gene number [11].

2 Discovering eukaryotic genes

The delineation of proteins encoded in eukaryotic genomes utilizes the following types of bioinformatic and experimental evidence [12, 13]. (i) *Ab initio* prediction of potential ORFs from genomic DNA; (ii) Detection of known protein identity or homology in genomic DNA; (iii) Matches with ESTs that have coding potential and/or splice sites; (iv) Cross-species comparisons for homologous gene detection; (v) Gene anatomy features associated with ORFs such as CpG islands, core promoters, transcription start sites, splicing signals, polyadenylation signals and the absence of repeat elements; (vi) Cloning

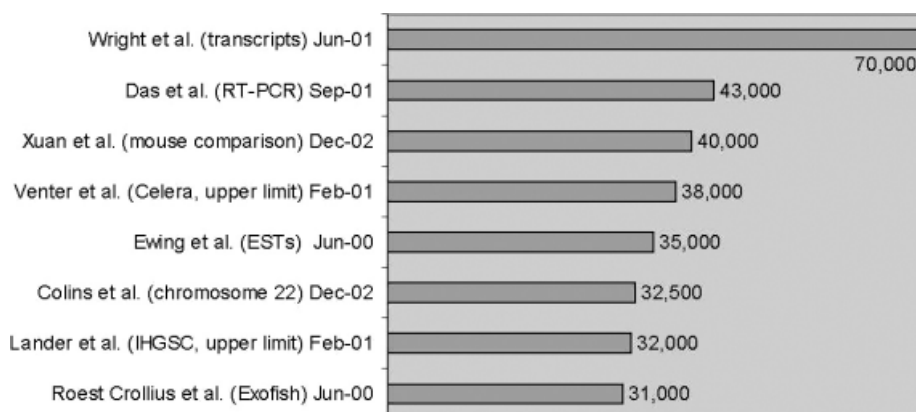


Figure 1. Estimates of human protein-coding gene number based on different lines of evidence. Taken from [3] with the addition of more recent papers [4, 5].

of predicted genes or extending partial transcripts; (vii) Characterization of deletion mutants or other loss-of-function approaches; (viii) Detection of active transcription by Northern blotting, or RT-PCR amplification; (ix) High-throughput transcript sampling by EST generation, SAGE tagging, massively parallel signature sequencing (MPSS) experiments or microarray profiling; and (x) Direct verification of a protein sequence by Edman sequencing, mass mapping or MS/MS sequencing.

3 High gene number arguments

The following arguments have been put forward in support of a final human gene number significantly above 30 000. (i) Model eukaryotes will show a postcompletion rise in gene number; (ii) The human genome assembly is incomplete; (iii) Gene prediction programs have a significant false-negative rate; (iv) Automated genome annotation pipelines are conservative; (v) Chromosome curation teams find genes missed by automated pipelines; (vi) Transcript coverage by mRNA and EST entries is incomplete; (vii) Novel proteins continue to be reported; (viii) Human/mouse/rat genomic comparison identifies many conserved sections; (ix) Sampling experiments, by proteomics, RT-PCR, SAGE, MPSS and microarrays, have revealed new genes; and (x) A fraction of rapidly evolving small proteins remain computationally and experimentally undetectable.

Although each of these propositions may be true for specific individual genes, or may have been true in the past, it will be argued in the sections below that only point (v) has the supporting data indicating a potentially significant contribution to the current gene total. The significance of argument (x) needs further explanation because it postulates a large undiscovered set, *i.e.* of the order of 3000 to 10 000 human proteins, that share the following characteristics; a propensity to be missed by *ab initio* gene prediction, not sampled in any mammalian mRNA or EST data and they cannot be detected by homology searching of known proteins against genomic data. In addition they are implied to be single-exon proteins with a low level and/or restricted pattern of tissue expression. The low sequence similarity to known proteins implies they have evolved rapidly and may be clade-specific. It is important to recognize that these characteristics have to be combined. This reduces the probability that large numbers of such genes have evaded detection. The sections below will consider the relationships between the types of evidence and arguments for high gene numbers in the context of current data.

4 Gene numbers in model eukaryotes

4.1 Genome annotation of model eukaryotes

Even before completion of their genomes the yeast, worm and fly were the focus of international efforts to identify and characterize all of their proteins. Essentially all the lines of evidence described in Section 2 have been utilized in this undertaking. The yeast alone has been the subject of hundreds of genome-wide analysis papers including large scale functional profiling [14]. The results, in addition to being published and submitted to the primary databases, have been aggregated in major web portals such as the *Saccharomyces* Genome Database (SGD), Flybase, and Wormbase [15–17].

However, the provision of updated and definitive eukaryotic gene sets remains problematic. Despite being completed some six years ago the yeast protein numbers vary according to source [12, 18]. Checking the three major portals for *Saccharomyces cerevisiae* (in October 2003) gave ORF totals of 6202 from the European Bioinformatics Institute complete proteome set, 5878 from SGD and 6723 from the Comprehensive Yeast Genome Database [19–21]. Additional discordance occurs in the literature. An approach focusing on small proteins reported 137 new genes and a total of 6000 proteins in 2002 [22]. An analysis later in the same year proposed a downward revision to 5400 [23]. The most recent analysis, in 2003, was based on comparisons with sequences of three related species, *Saccharomyces paradoxus*, *Saccharomyces mikatae* and *Saccharomyces bayanus* [24]. This resulted in revision of ~15% of all genes and a total reduced to 5726.

4.2 Postcompletion gene number changes

The changes in gene totals, between first releases and latest updates, along with more recently sequenced model eukaryotes, are shown in Fig. 1. By taking the latest published revised number *S. cerevisiae* has decreased by 9% over seven years [24, 25]. The 441 new genes in *Caenorhabditis elegans* accumulated over five years represent an increase of 2% [26, 27]. However, recent experimental evidence indicates the *C. elegans* ORF collection could still include a substantial proportion of pseudogenes [28]. The *Drosophila* gene number, initially determined as 13 601, has been revised down to 13 379, a drop of 2%, after a recent major re-annotation [29, 30].

The *Drosophila* re-annotation paper touches on several themes that are relevant to the human genome and individual chromosome reports discussed in Section 5.5. Firstly, community annotation and human curation have

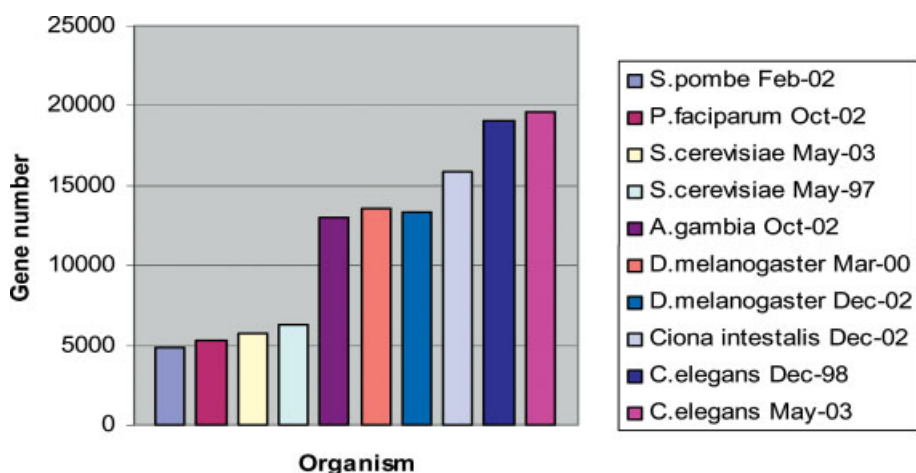


Figure 2. Gene numbers for completed model eukaryotic genomes as referenced in the text. The initial publication numbers are given with dates. These are compared with current numbers from websites or the most recent publications.

made key contributions to the quality of revised gene sets that include changes in approximately 45% of the predicted proteins. Secondly, as chromosome assemblies improve, the discovery of new genes is balanced by a reduction in gene number from the fusion of previously fragmented gene predictions. Thirdly, as more transcript data has accumulated, the average transcript length and total exon count have risen without increasing gene number. Lastly, the latest revision encompasses 1042 additional homology-based candidate genes detected immediately after the initial genome release [31]. Thus, what was initially proposed as a 7% increase has become a 2% decrease.

The yeast, worm and fly can now be compared with additional completed organisms that are included in Fig. 2. The second yeast, *Schizosaccharomyces pombe* turns out to have 23% fewer genes than *S. cerevisiae* [32]. The recent completion of a second insect, the *Anopheles* mosquito, gives a figure of 12 981 and the malarial protozoan *Plasmodium faciparum* 5268 [33, 34]. The first simple chordate *Ciona intestinalis* containing many vertebrate protein families gives a figure of 15 852 and even this is estimated to be a 5% over-prediction because of fragmentation in the draft assembly [35].

4.3 Proteomic sampling in model eukaryotes

The identification of 1484 yeast proteins, *i.e.* 23% of all predicted genes, by mass spectrometry, did not initially report any novel proteins [36]. However, a later survey, combining expression profiling with mass spectrometry, claimed the addition of 62 new yeast genes [37]. A recent analysis of the *P. faciparum* genome identified 1289 proteins from selected parasite stages, corresponding to

24% of the predicted proteins [38]. This report included 100 unmatched peptides but the numerical relationship between these orphan peptides and novel proteins remains to be elucidated. Although these reports demonstrate the potential of mass spectrometry-based proteomics to detect eukaryotic proteins that have eluded *in silico* annotation there is no evidence of a significant impact on gene totals.

5 Annotation of the human genome

5.1 The golden path genome assembly

The accuracy of gene prediction is crucially dependent on the quality of the underlying genome data. The early genome assemblies or reference sequences, colloquially known as the Golden Path (GP), were produced at the University of California at Santa Cruz [39]. Subsequently, the International Human Genome Sequencing Consortium has used assemblies generated by the National Center for Biotechnology Information (NCBI). The latest human reference sequence is based on NCBI Build 34 [40]. This covers about 99% of the gene-containing regions in the genome, and has been sequenced to an accuracy of 99.99%. The missing portions are comprised of several hundred defined gaps representing DNA regions with unusual structures that cannot be reliably sequenced using current technology.

5.2 Ensembl gene totals

Ensembl has become the *de facto* standard for automated annotation of eukaryotic genomes [41]. The end result is a set of transcripts grouped into genes by shared

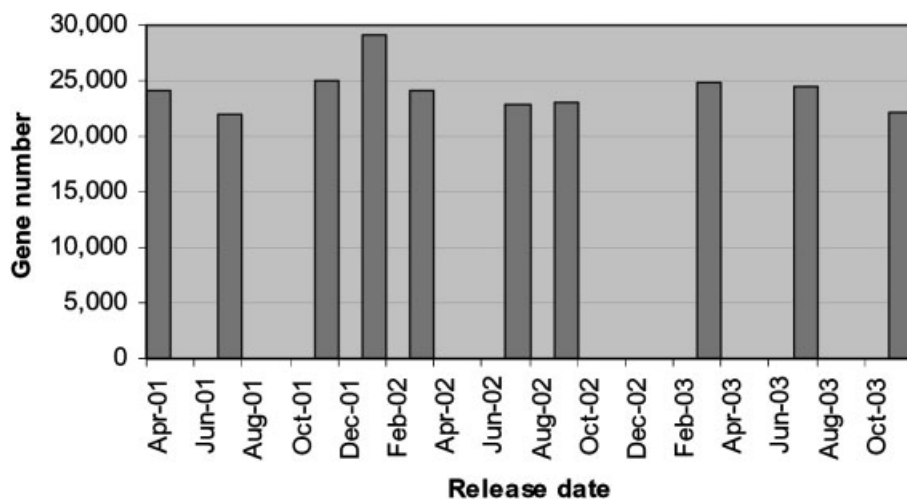


Figure 3. Gene numbers, excluding alternative transcripts, associated with major Ensembl releases [42]. The latest version, 18.34.1, November 2003, was built on GP34.

exons and supported by evidence of at least one form of sequence homology. Since the release based on the first draft of the human genome, Ensembl has gone through seven major releases of the GP. The gene numbers associated with these releases are shown below in Fig. 3.

The key feature in Fig. 3 is that the total, currently at 24 037, has decreased by nine genes since the first release at 24 046. If the pseudogene count is factored in, the gene number is reduced to 22 184, representing a decrease of 1862 since the first release. The maximum seen in the January 2002 release may be associated with clone orientation changes between the University of Santa Cruz hg8 genome assembly and NCBI GP26 [43].

5.3 Underlying trends

A number of important trends can be discerned from the Ensembl release statistics. The first is the detection of a higher proportion of known genes, rising from 90% to over 95%. The second is that the number of novel genes, defined as those less than 95% identical to known human proteins at build time, has fallen from the maximum of 12 398 in the November 2001 release to 2263 in the December 2003 release. The third trend is the increase in exons *per* gene, rising from 6.5 in January 2002 to 9.3 in July 2003 (although this has dropped back to 8.6 in the latest release). The fourth trend is an increase in the number of alternatively spliced transcripts from 3669 in November 2001 to 5827 in November 2003. The fifth trend, documented only for the last two releases, is an increase in annotated pseudogenes from 900 up to 1853.

These trends suggest that novel predicted genes are being converted to experimentally confirmed proteins against the background of an essentially static gene num-

ber. Genes have also been growing in average length and exon count. Similar effects of reduced genomic fragmentation and improved transcript coverage on gene annotation have been documented for *Drosophila* and for *Arabidopsis* [30, 44].

5.4 Curation of completed chromosomes

The completed human chromosomes, 6, 7, 13, 14, 20, 21 and 22, have been subject to major curation efforts [4, 45–51]. These analyses have been performed by large teams of authors from different groups. They include experimental confirmation of putative genes and, in some cases, the use of fish as well as mouse gene predictions to broaden the range of homology detection. They also include a large component of manual curation. Compared with automatic pipelines this provides a more reliable annotation of pseudogenes, splice variants, polyadenylation sites, and incomplete putative transcripts (reviewed in [52]). They therefore produce independent gene counts that can be compared to automated gene annotations. The continuing work of annotation groups for chromosomes 6, 13, 14, 20 and 22 is now included in the Vertebrate Genome Annotation (VEGA) database [52, 53]. This process includes a virtuous cycle whereby novel genes are cloned and submitted to public databases, thereby ensuring their incorporation into subsequent genome annotations.

Ensembl uses a similarity cut-off to classify gene products as known or novel. The annotation groups, including those contributing to VEGA, have introduced more graded definitions. Known genes are classified by their presence in the NCBI RefSeqNP collection rather than any identity match within the larger SWISS-PROT/

TrEMBL (SPT) human data set used by Ensembl. Novel transcripts are subdivided into three categories, novel coding sequences (CDSs) where an ORF can be determined, novel transcripts where none of the alternative potential ORFs can be frame-fixed by homology and putative transcripts, where spliced ESTs define intron/exon boundaries but are not sufficient to define an ORF. Annotated pseudogenes are defined as similar to known proteins but have frame shifts or premature stop codons which disrupt the ORF. An example comparison of VEGA and Ensembl annotation is shown below in Fig. 4.

5.5 Comparing curated chromosomes with automated genome annotation

The use of graded definitions and categories of supporting evidence by the chromosome curation teams clearly represents the reality of detailed gene annotation. However, comparisons with Ensembl numbers and extrapolations to total gene counts are difficult for the following reasons: (i) Groups reporting on 7, 14 and 21 have used their own assemblies rather than the public GP version; (ii) Publications from these groups have used different supporting evidence and definitions of gene categories. For example, not all publications include putative genes in the total count; (iii) Each chromosome report was made at different times and was therefore compared against different sets of public transcript data; (iv) The two groups reporting approximately simultaneously on chromosome 7 show different gene numbers and use different pseudogene definitions [46, 51]; (v) None of the groups have yet published a formal gene cross-mapping to Ensembl to determine the basis for any systematic discrepancies, al-

though at least for 6, 13, 14, 20 and 22, the result sets for Ensembl and VEGA can be visualized in the same browser (Fig. 4); (vi) Only one group has presented a longitudinal update of gene changes [4]. Nevertheless it is informative to attempt a comparison of the numbers between the curated and automated gene build sets. This is presented below in Table 1.

The data in Table 1 represents ~ 15% of all genes. A surprising observation is that Ensembl records ~ 20% more known genes. This is likely to be because knowns are defined by identity matches in SPT) in Ensembl rather than the smaller RefSeqNP set used by most of the chromosome reports. The second trend is the high ratio of pseudogenes: total genes of ~ 1:2.8. The third trend is the impact of putative genes on the total. If these are excluded then Ensembl finds, on average, ~ 3% less genes than the chromosome groups. If the putatives are included this drops to ~ 20% less.

The supporting evidence for the gene categories used by the chromosome annotation teams shows a relationship to the exon count and transcript size. Thus the reference (RefSeq) sequences on chromosome 14 have an average of 10.3 exons and 3 kbp transcripts, compared with 2.8 exons and 723 bp for putative genes with EST-only support [47]. The overall Ensembl figure is 9.3 exons *per* gene. This suggests either that Ensembl misses some short proteins or that the short putative genes catalogued by the chromosome annotation teams may include non-coding transcripts, expressed truncated pseudogenes or EST artefacts. Chromosome 22 provides the only longitudinal comparison where, between 1999 and 2003, the gene count only increased by one [4].

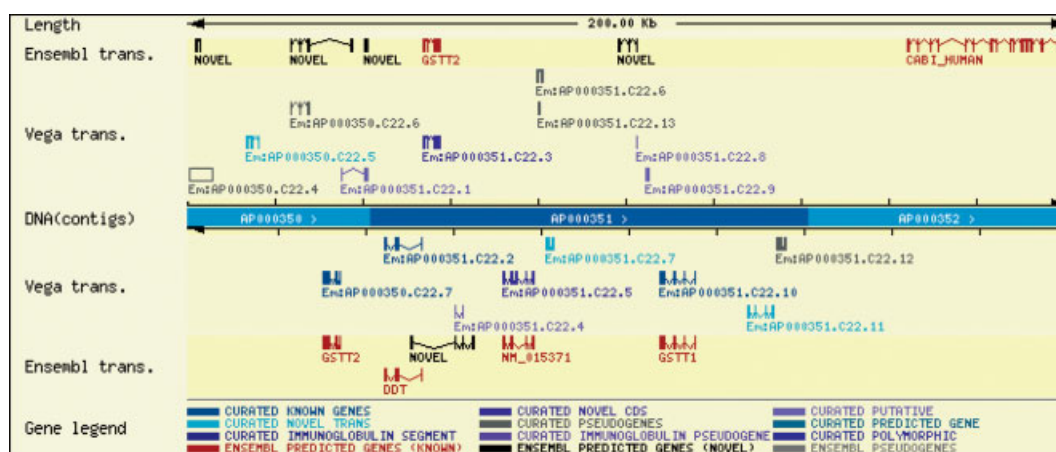


Figure 4. Comparison of VEGA and Ensembl annotation for release 15.33. This view shows the annotations centered on position 22593134, from 22.q12.1.

Table 1. Gene numbers for individual completed human chromosomes. These were extracted from the first publication for chromosome 7, the Ensembl March 2003 (GP31) release, the chromosome 21 website and the VEGA websites for 6,13,14, 20 and 22 [53, 54]

| Chr | Known | Nov cds | Nov trans | Put | Psgs | Genes | Ens total | Ens known | Ens nov |
|-------|-------|---------|-----------|------|------|-------|-----------|-----------|---------|
| 6 | 772 | 287 | 213 | 285 | 633 | 1272 | 1296 | 1037 | 259 |
| 7 | 863 | 71 | 521 | 213 | 144 | 1455 | 1269 | 925 | 344 |
| 13 | 228 | 101 | 132 | 161 | 281 | 461 | 455 | 322 | 133 |
| 14 | 510 | 119 | 23 | 207 | 296 | 652 | 736 | 572 | 164 |
| 20 | 448 | 99 | 68 | 258 | 173 | 727 | 704 | 614 | 90 |
| 21 | 176 | 29 | 28 | n.a. | 80 | 233 | 192 | 138 | 54 |
| 22 | 223 | 172 | 67 | 115 | 238 | 462 | 469 | 356 | 113 |
| Total | 3220 | 878 | 1052 | 1239 | 1845 | 5262 | 5121 | 3964 | 1044 |

Abbreviations: Chr, chromosome; Known, known genes; Nov cds, ORFs identical to spliced ESTs or similar to other proteins; Nov trans, similar to Nov cds but ORF cannot be determined; Put, identical to human ESTs that splice but do not contain an ORF; Psgs, pseudogenes; Genes, sum of known, novel cds and novel transcripts, *i.e.* excluding putatives; Ens total, all Ensembl genes on that chromosome; Ens known, Ensemble known genes only; Ens nov, Ensembl novel genes.

5.6 Pseudogenes

A potential source of false-positives in homology-based genome annotation is pseudogenes [55]. Estimates of the total number of these in the human genome have risen to 20 000 [56]. It seems certain that a proportion of these may have been counted as genes in automated annotation pipelines, especially those pseudogenes that have only minor disablements that distinguish them from functional paralogues. The mouse genome paper includes a formal estimate of 4000 pseudogenes that could be present in their gene total (see Section 7). The latest human Ensembl release has provided automatic annotation of 1853 pseudogenes and the latest rat release 1592. The situation is made more complex by the existence of transcribed pseudogenes because these would have both homology and supporting evidence of transcription. According to the current LocusLink statistics 115 of 2613 annotated human pseudogenes (4.4%) fall into this category [57].

6 Postgenomic coverage of protein and transcript data

Locating transcript identities and detecting protein homology in genome data are directly related to coverage in nongenomic data. The exponential growth of the primary nucleotide databases since the first draft human genome has included a massive increase in human mRNA and protein data. The EST data are also used for support-

ing evidence for predicted genes. Data from other mammals or vertebrates can be used for homology detection. The growth of these sources will be considered below.

6.1 Human protein databases

The International Protein Index (IPI) merges a set of experimental and predicted sequences derived from SWISS-PROT, TrEMBL, RefSeqNPs, RefSeqXPs and Ensembl [58]. This provides a resource of complete sets of human, mouse and rat proteins represented by one sequence *per* transcript with statistics on the overlap between sequences from the different sources. The process also includes a redundancy-reduced set of sequences from the larger SPTTr set. The growth of human protein numbers in SPTTr is shown in Fig. 5.

The human IPI total shows a surprising rise and fall since its first release of 33 013. The peak of 67 105 in mid-2002 was due to the deposition by the NCBI of large numbers of *ab initio* predictions (RefSeq XPs) into their protein databases [59]. Continual revision of the NCBI genomic annotation pipeline has reduced this set and consequently the IPI number has now fallen to 39 440.

The main feature in Fig. 5 is that the number of new proteins in the redundancy-reduced version is small (there are technical issues associated with the promotion of sequences from TrEMBL to SWISS-PROT and subfragment merging that cause intermittent drops in these numbers). The figures indicate that worldwide post-genomic

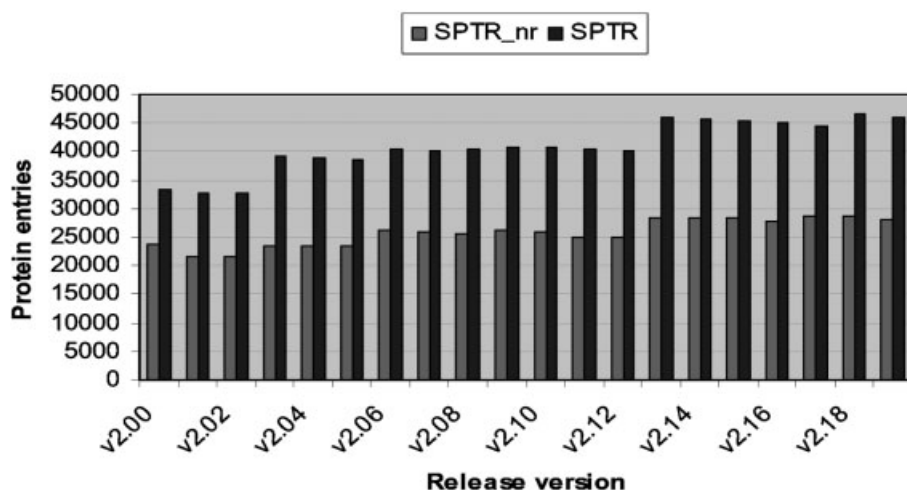


Figure 5. The growth of human sequence numbers in SPTr (upper bars) and a redundancy-reduced subset SPTr_nr (lower bars) (courtesy of Dr. P. Kersey, European Bioinformatics Institute). The X-axis represents monthly release versions from October 2001 to June 2003.

activity for cloning human genes over the last 18 months has produced an increase of ~ 4800 new proteins. In addition this has perceptibly flattened out over the last five releases at $\sim 28\,000$. This suggests that, although they may be covering more splice forms, extending previously partial sequences or confirming predicted entries, new protein submissions are predominantly resampling previously identified genes.

Over the same time period covered in Fig. 5 the Ensembl gene total (Fig. 3) actually fell. Although there are caveats to comparing Ensembl protein sequences with SPTr sequences, for example the persistence of some early incorrect *ab initio* gene predictions in TrEMBL, the number of novel genes in Ensembl fell by ~ 8000 over this period. This exceeds the ~ 5000 new genes appearing in SPTr over the same period but some of the novel genes are likely to have been removed by the correction of previously fragmented genes in earlier assemblies. These trends suggest that new sequences appearing in the protein databases are making a diminishing contribution to the gene total.

6.2 Small proteins

As described in Section 3 a key argument for high gene numbers, in quantitative terms, is the hypothetical existence of large numbers of hitherto undetected small open reading frames of less than 100 residues, termed small open reading frames (smORFs). These are difficult to detect both experimentally and computationally [60]. The likelihood of this category contributing large numbers of new proteins was assessed by looking for smORF increases in recent and/or novel sequences submitted to the protein databases. A submission date was found

(October 2001) that divides SPTr roughly in two. The protein length distribution of the old and new halves of the databases was then compared. The most recent data showed, 5.5% of proteins below 100 residues, compared to 6.3% in the older data. Selecting entries that were classified by authors as “novel” at the time of submission showed only 3.4% of these sequences were below 100 residues. Thus, the size distribution of experimentally determined human proteins indicates the proportion of smORFs is falling.

6.3 Human mRNA databases

One way of addressing gene sampling in transcript data is to look at the longitudinal statistics of the UniGene system [61]. This includes a count of mRNAs along with EST cluster counts on the basis of shared sequence identity. Plots of the growth of these sets of data are shown in Fig. 6.

Figure 6 shows the contrast between the rapid growth of total submissions and the slow increase in the clustered set (the sporadic falls in mRNA occur when bulk preliminary submissions are replaced with revised sets). On average each unique mRNA has been sequenced 3.8 times. The slow increase suggests new entries are predominantly sampling the same set of transcripts.

The data above suggests the redundancy-reduced human mRNA and protein collections are converging to $\sim 28\,000$ genes. This presents a paradox in that they significantly exceed the 19921 known protein entries in Ensembl 18.34. However, there are technical caveats associated with the process of redundancy reduction such as increasing numbers of splice variants with significant length differences. The lower numbers in Ensembl

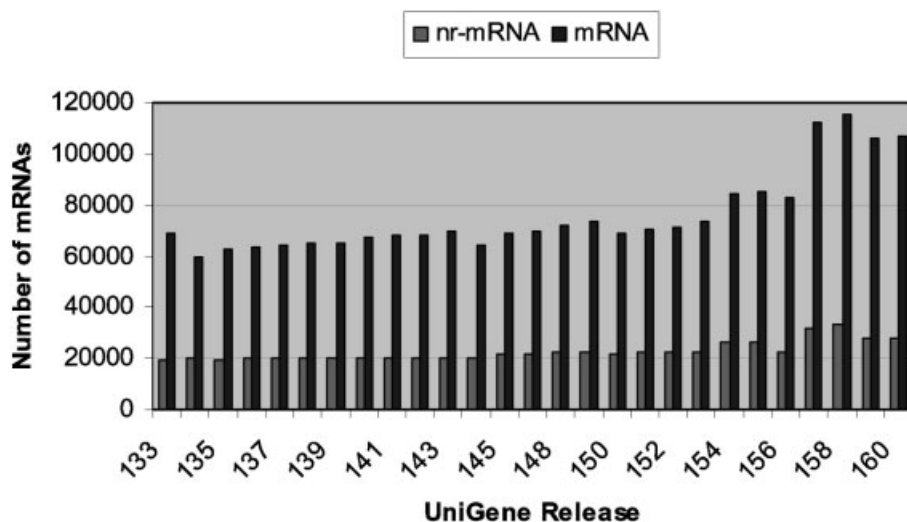


Figure 6. Plots of clustered (lower bars) and total (upper bars) human mRNA entries extracted from GenBank represented in UniGene releases from March 2001 to March 2003 (data courtesy of Dr. Lucas Wagner, National Center for Biotechnology Information).

Table 2. Transcript figures for the five mammals indexed in UniGene [61]. Fold coverage is the ratio of total mRNAs to the clustered mRNA number. EST coverage (recording of at least one extended identity match to an mRNA, even in the UTR) is calculated as mRNA clusters that include ESTs divided by all mRNA clusters

| Species | ESTs (millions) | Total mRNAs | mRNA clusters | mRNA coverage (fold) | EST coverage of mRNA |
|---------|-----------------|-------------|---------------|----------------------|----------------------|
| Human | 4.02 | 106,900 | 27,887 | 3.83 | 95.2% |
| Mouse | 3.34 | 47,518 | 16,177 | 2.94 | 93.1% |
| Rat | 0.33 | 11,611 | 4,882 | 2.38 | 84.2% |
| Cow | 0.14 | 3,661 | 1,650 | 2.21 | 81.4% |
| Pig | 0.06 | 1,942 | 1,262 | 1.53 | 59.0% |

are therefore not only due to effective redundancy removal by explicit mapping onto the genome but also, according to the latest release notes, that a significant proportion of apparently novel mRNA protein-coding genes are artefacts generated from chimeras and spurious ORFs translated from noncoding transcripts or pre-mRNA. Supporting evidence for this is that the current nonredundant RefSeqNP collection from all human mRNAs produces only 18 112 uniquely mapping proteins.

6.4 EST coverage of mRNAs

The current dbEST is approaching 10 million mammalian ESTs of which 5.3 million are human [62]. Tissue sampling has expanded to include more specialized sources such as human tumor cell lines, cow foetal libraries and mouse embryo stages. In addition, many subtractive strategies have been employed to enhance the detection of low-abundance transcripts. This has been accompanied by

major initiatives in the USA, Japan, Germany and other countries to convert as many human and mouse ESTs as possible to full-length mRNA sequences.

The argument (from Section 3) that a significant fraction of the expressed genome remains unsampled becomes less tenable as the tissue breadth and sequencing depth in dbEST expands. A predicted consequence of incomplete EST coverage would be that targeted cloning, including the confirming of genes predicted from genomic data, should increase the proportion of mRNAs that are not represented in EST data. In fact the UniGene data suggest the opposite trend *i.e.* EST numbers and mRNA coverage have increased in parallel. The comparison between the five mammals with the most sequence data are shown in Table 2.

Table 2 shows that ESTs, mRNA and EST coverage of mRNAs move in the same direction *i.e.* the ranking of these is the same across the species, with humans being the most highly sampled at ~ 50-fold more mRNA entries

than pig. The clusters, which can be considered as a redundancy-reduced set, also rank in the same order. The last column indicates that over 95% of unique human mRNAs have been sampled by human ESTs. A longitudinal assessment derived from the data sets used to prepare Fig. 6 shows that this figure has climbed slowly from ~ 92% over the last two years.

Does this indicate that dbEST is approaching saturation for human gene coverage? The fact that 5% of mRNA clusters are still not sampled by ESTs suggests saturation has not been reached but the longitudinal statistics do show a flattening out. Non-sampled genes include classes of proteins, such as seven-transmembrane receptors, that are known to have very low EST coverage. However, they are not lost in the human gene count because they are relatively easy to find by homology searching in genome data [63]. In addition, even if a human gene had no matches to human ESTs there is an increasing likelihood of an orthologous EST match *i.e.* the aggregate of 10 million mammalian ESTs may well be approaching saturation sampling of mammalian protein-coding transcripts.

6.5 Mammalian protein increases

The discovery of novel proteins uses protein similarity matches as supporting evidence for putative exons in genomic DNA. One of the arguments for high gene number listed in Section 3 is the speculation that the human genome (and by extrapolation other mammals) encodes significant numbers of proteins with similarity matches below the thresholds used for exon detection. The increase in the mammalian protein numbers over a three year postgenomic period is shown in Fig. 7.

The data in Fig. 7 show that mammalian sequence coverage exhibited a ~ fourfold increase in numbers available for homology searching over the two years since the first

human draft. While this could arguably still be biased against very rare transcripts it includes the worldwide output of high-throughput cloning projects and the results of over 30 years of targeted gene discovery. The entirety of mammalian proteins therefore represents ~ 4.5-fold averaged genome coverage. The recent analysis of the *Fugu* genome shows that ~ 75% of fish ORFs have significant similarity scores with human sequences [64]. This suggests at least some of the ~ 40 000 non-mammalian vertebrate sequences could also contribute to mammalian annotations.

7 Comparing human with other vertebrate genomes

The human genome has now been joined by substantially complete assemblies for mouse, rat and fish [64–66]. The total gene numbers, transcripts and exons *per* gene are shown below in Table 3.

The mouse gene catalogue was compiled with the Ensembl pipeline and the use of ESTs to support gene predictions. An important conclusion from the mouse paper is that approximately 80% of mouse genes have 1:1 orthologues and the number of genes without human homologues is only ~ 1% [65]. By aligning the results from human and mouse a separate annotation effort was able to increase the specificity and sensitivity of *ab initio* gene prediction [67]. This detected approximately 12 000 additional exons which the authors, after selected PCR validation, suggest could add ~ 1000 new genes. There could also be in the order of 1000 genes in the missing 4% of the mouse assembly. However, these potential false-negatives would more than be balanced by the estimate of ~ 4000 pseudogenes. Given the prediction of slightly higher gene numbers because of the ~ 500 rodent odor receptors that have human pseudogene orthologues, the

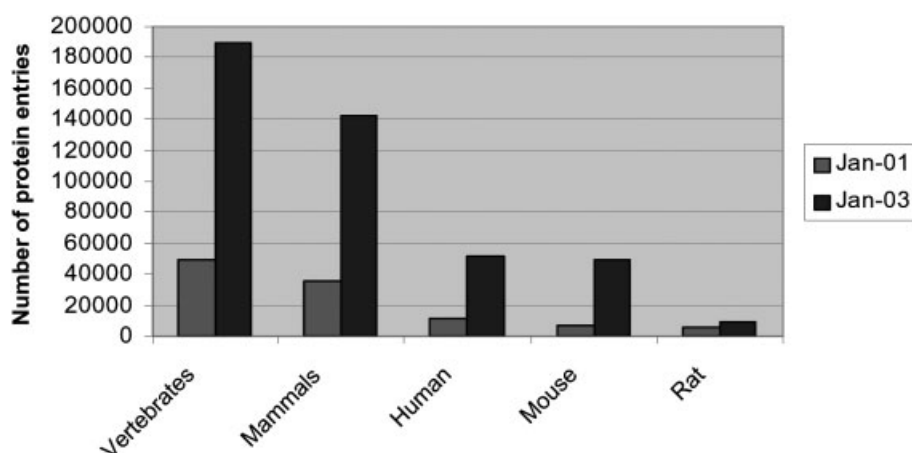


Figure 7. Changes in the species distribution in the SWISS-PROT/TrEMBL protein databases, at 1024626 sequences at the beginning of 2003, compared with the numbers available at the time of the first draft of the human genome, January 2001, at 430915 sequences.

Ensembl gene numbers for both rodents seem low (Table 3). However, this may result from the relative transcript coverage, shown in Fig. 7, that shows the same order of declining gene number *i.e.* human > mouse > rat.

The mouse genome issue of *Nature* included a comprehensive analysis of all available mouse mRNAs although these were not directly mapped to genome data [68]. The number breakdown is complex but the team collapsed 60 770 highquality cDNA sequences, 39 694 of which were new, with 40 106 public mouse mRNAs to produce 37 409 representative transcript units. Of these only 20 489 could be determined as protein-coding.

The *Fugu* fish has come out higher than both humans and rodents with just over 35 000 genes. However, it is known that teleost (ray-finned) fish contain larger numbers of duplicated genes compared with lobe-finned fish and tetrapods such as mammals [69]. In addition, it has a lower number of exons *per* gene that suggests there may still be some fragmentation within the draft sequence that has artificially elevated the gene number.

Table 3. Comparative data for assembled vertebrate genomes

| Organism | Genes | Transcripts | Exons/gene |
|----------------------|----------------------|-------------|------------|
| Human (18.34.1) | 22,184 ^{a)} | 29,864 | 8.6 |
| Mouse (MGSC) | 22,011 | 29,201 | 8.4 |
| Mouse (13.30.1) | 24,948 | 32,911 | 8.7 |
| Rat (18.3.1) | 22,159 ^{a)} | 30,232 | 7.9 |
| <i>Fugu</i> (18.2.1) | 35,180 | 38,510 | 4.7 |

a) Numbers exclude the Ensembl pseudogene counts.

8 Evidence of noncoding mRNA transcription

In addition to the three classical types of RNA, ribosomal, transfer and messenger, a number of additional categories of noncoding RNA have been identified [70]. Evidence has accumulated that substantial sections of the human genome can give rise to poly-A⁺ mRNA, presumably derived from RNA polymerase II activity, that does not encode protein. Examples of this are given in reports that have used DNA microarrays to reveal widespread transcription activity on chromosomes 21 and 22 outside the boundaries of what were known exons at the time [71, 72]. Such transcripts, termed transcriptionally active regions (TARs) could represent novel genes, novel exons

in known genes, extensions of previously undetected 5' or 3' untranslated regions (UTRs), transcribed pseudogenes, or noncoding transcripts of unknown function. The fact that neither publication presented any complete novel ORFs derived from their data argues in favor of the latter category. The combined data from these two publications also suggest that these TARs are expressed at a low level, they include antisense transcripts and many show sequence conservation between human and mouse.

The existence of TARs outside the boundaries of known protein-coding genes is supported by data from other sources. A recent mouse transcript analysis paper reported 4280 transcripts that lacked protein-coding potential but some had poly-A tracts and matches in both mouse and human ESTs [73]. Another paper has reported 2431 pairs of overlapping sense-antisense pairs from mouse transcript data [74]. A similar number of approximately 1600 human antisense transcriptional units have been verified by stand-specific microarrays, some of which were also represented in EST data [75]. These antisense transcripts are suggested to have some kind of regulatory function rather than coding for proteins [76].

9 Increasing the stringency of gene identification

Nearly all known proteins conform to the universally confirmed features of gene anatomy. A diagram of these is shown in Fig. 8.

In the past, merely the detection and/or determination of short sequences of transcribed fragments in cytoplasmic mRNA fractions has been used as sufficient evidence to infer the existence of new protein-coding genes. However, as described above in Section 8 it now seems that a significant proportion of TARs (in coverage rather than quantitative terms) are noncoding. This may explain the previously reported high gene numbers from transcript sampling experiments performed without corroborative full-length cloning [78]. Even the presence of protein homology in transcribed fragments can no longer be relied upon, not only because of expressed pseudogenes but also because of the reported existence of ancient protein "fossil" fragments in intergenic regions of the human genome [79]. This means that claims of novel protein discovery now require not only the determination of a transcript with an ORF that conforms to the universal features of gene anatomy but also submission of that complete sequence to the primary databases.

matches to genomic DNA as opposed to scoring spectral matches against translated and/or predicted protein sequences [88, 89]. However, an alternative explanation is that there are simply too few undiscovered genes being sampled by these experiments. Although mass spectrometry-based peptide identification will make a major impact on identifying post-translational modifications and correcting exon assignments there are no data to suggest this will have a significant impact on gene number.

12 Conclusions

Significantly, and perhaps unexpectedly, postcompletion revisions have produced gene number decreases in yeast and fly and only a minor increase in the worm. This has occurred despite the massive worldwide experimental and computational focus on these organisms. This is neither to suggest that gene discovery in the model eukaryotes is at an end nor that detailed revision of gene structures will not continue. However, it argues strongly against constitutive underdiscovery for eukaryotic genes.

The interval between the first draft assembly and the closure of the human genome announced on the 14 April 2003 has seen big increases in human mRNA coverage, EST production and continual refinement of automated genome annotation [90]. These advances, also perhaps unexpectedly, have produced a fall in the Ensembl gene count to just above 22 000. The initial mouse and rat annotations also indicate pseudogene-adjusted numbers close to this figure. In parallel, the public protein collections are showing a diminished rate of novel human gene discovery suggestive of saturation.

So is there any evidence for higher gene numbers? Certainly the total numbers for all categories extrapolated from the aggregated individual chromosome reports could be above 30 000. However, excluding the putative gene category reduces this to ~ 25 000. The total therefore depends on the future status of the novel and putative transcripts reported by the chromosome annotation groups. Although some of these will be found to have the gene anatomy features necessary for protein expression it seems likely that the majority will not be verified as protein-coding genes.

The postulate that anywhere between 10–30% of human genes consist of experimentally and computationally undiscovered small proteins seems untenable for the following reasons. Firstly, the mouse genome paper estimated that only ~ 1% of mouse genes had no detectable human homology [65]. Secondly, as shown in Section 5.8, there is no proportional increase in small proteins

amongst recently discovered genes. Thirdly, proteins shorter than 100 residues may fall below the threshold necessary to fold into functional domains [91]. Fourthly, although small proteins can evolve more rapidly, there is no precedent in the literature on protein evolution for the existence of large numbers of clade-specific mammalian proteins that have nonsynonymous mutation rates (K_a/K_s values) so high that they cannot be detected by cross-species protein sequence similarity [92].

So what would constitute gene number closure? Bioinformatic approaches will increasingly use comparative genomics to refine mammalian gene sets [93]. Experimentally, it could be possible to use MS to verify at least one unique peptide from all proteins isolated from *in vivo* sources. Although this goal may be too ambitious for current technology it is at least implicit in the aims of the Human Proteome Organisation (HUPO) and there is already a commercial initiative underway that has verified over 14 000 proteins from human cell lines and tissues [94, 95]. However, in the short term we are more likely to see “closure-by-expression-*in vitro*” and some academic centers are already building up the requisite human clone collections [96].

Given the lingering uncertainties for yeast protein number any short term expectation of human closure with four times as many genes seems unrealistic. However, we might hope to reach a point beyond which significant numerical changes would be unlikely. The current evidence points to this being well below 30 000 protein-coding genes, possibly as low as 25 000.

I would like to thank Paul Kersey of the EBI for the protein database statistics, Lucas Wagner of the NCBI for the retrospective UniGene data, and numerous other people at the NCBI, European Bioinformatics Institute, and Sanger Institute who have graciously answered many queries on their data collections. I would also like to thank Duncan Campbell and John Hancock for reviewing this manuscript, members of the OGS Proteome Discovery team for helpful discussion and my wife for perceptive proof-reading.

13 References

- [1] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W. *et al.*, *Science* 2001, 291, 1304–1351.
- [2] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C. *et al.*, *Nature* 2001, 409, 860–921.
- [3] Harrison, P. M., Kumar, A., Lang, N., Snyder, M., Gerstein, M., *Nucleic Acids Res.* 2002, 30, 1083–1090.
- [4] Collins, J. E., Goward, M. E., Cole, C. G., Smink, L. J. *et al.*, *Genome Res.* 2003, 13, 27–36.
- [5] Xuan, Z., Wang, J., Zhang, M. Q., *Genome Biol.* 2003, 4, R1.

- [6] O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J., Nash, W. G. *et al.*, *Science* 1999, 286, 458–462, 479–481.
- [7] Claverie, J. M., *Science* 2001, 291, 1255–1257.
- [8] Hopkins, A. L., Groom, C. R., *Nat. Rev. Drug Discov.* 2002, 1, 727–730.
- [9] Aebersold, R., Mann, M., *Nature* 2003, 422, 198–207.
- [10] Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J. *et al.*, *Genomics* 2002, 79, 464–470.
- [11] Rappsilber, J., Mann, M., *Trends Biochem. Sci.* 2002, 27, 74–78.
- [12] Snyder, M., Gerstein, M., *Science* 2003, 300, 258–260.
- [13] Southan, C., Barnes, M. R., Gray, I. C. (Eds.) in: *Genetic Bioinformatics: A Bioinformatics Guide for Geneticists*, John Wiley & Sons, Chichester 2003, pp. 71–79.
- [14] Giaever, G., Chu, A. M., Ni, L., Connelly, C. *et al.*, *Nature* 2002, 418, 387–391.
- [15] Weng, S., Dong, Q., Balakrishnan, R., Christie, K. *et al.*, *Nucleic Acids Res.* 2003, 31, 216–218.
- [16] FlyBase-Consortium *Nucleic Acids Res.* 1999, 27, 85–88.
- [17] Harris, T. W., Lee, R., Schwarz, E., Bradnam, K. *et al.*, *Nucleic Acids Res.* 2003, 31, 133–137.
- [18] Garrels, J. I., *Funct. Integr. Genomics* 2002, 2, 212–237.
- [19] <http://www.ebi.ac.uk/proteome/>.
- [20] <http://www.yeastgenome.org/>.
- [21] <http://mips.gsf.de/proj/yeast/CYGD/db/index.html>.
- [22] Kumar, A., Harrison, P. M., Cheung, K. H., Lan, N. *et al.*, *Nat. Biotechnol.* 2002, 20, 58–63.
- [23] Mackiewicz, P., Kowalczyk, M., Mackiewicz, D., Nowicka, A. *et al.*, *Yeast* 2002, 19, 619–629.
- [24] Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E. S., *Nature* 2003, 423, 241–254.
- [25] Mewes, H. W., Albermann, K., Bahr, M., Frishman, D. *et al.*, *Nature* 1997, 387, 7–65.
- [26] Consortium, C. e. S., *Science* 1998, 282, 2012–2018.
- [27] <http://www.wormbase.org/>.
- [28] Mounsey, A., Bauer, P., Hope, I. A., *Genome Res.* 2002, 12, 770–775.
- [29] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A. *et al.*, *Science* 2000, 287, 2185–2195.
- [30] Misra, S., Crosby, M. A., Mungall, C. J., Matthews, B. B. *et al.*, *Genome Biol.* 2002, 3, 8301–8321.
- [31] Gopal, S., Schroeder, M., Pieper, U., Sczyrba, A. *et al.*, *Nat. Genet.* 2001, 27, 337–340.
- [32] Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M. *et al.*, *Nature* 2002, 415, 871–880.
- [33] Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D. *et al.*, *Science* 2002, 298, 149–159.
- [34] Gardner, M. J., Hall, N., Fung, E., White, O. *et al.*, *Nature* 2002, 419, 498–511.
- [35] Dehal, P., Satou, Y., Campbell, R. K., Chapman, J. *et al.*, *Science* 2002, 298, 2157–2167.
- [36] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, *Nat. Biotechnol.* 2001, 19, 242–247.
- [37] Oshiro, G., Wodicka, L. M., Washburn, M. P., Yates, J. R., 3rd *et al.*, *Genome Res.* 2002, 12, 1210–1220.
- [38] Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M. *et al.*, *Nature* 2002, 419, 537–542.
- [39] <http://genome.cse.ucsc.edu/>
- [40] http://www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html.
- [41] Hubbard, T., Barker, D., Birney, E., Cameron, G. *et al.*, *Nucleic Acids Res.* 2002, 30, 38–41.
- [42] <http://www.ensembl.org/>.
- [43] Rouchka, E. C., Gish, W., States, D. J., *Nucleic Acids Res.* 2002, 30, 5004–5014.
- [44] Haas, B. J., Volfovsky, N., Town, C. D., Troukhan, M. *et al.*, *Genome Biol.* 2002, 3, RESEARCH0029.
- [45] http://vega.sanger.ac.uk/Homo_sapiens/mapview?chr=6.
- [46] Scherer, S. W., Cheung, J., MacDonald, J. R., Osborne, L. R. *et al.*, *Science* 2003, 300, 767–772.
- [47] Heilig, R., Eckenberg, R., Petit, J. L., Fonknechten, N. *et al.*, *Nature* 2003, 421, 601–607.
- [48] Deloukas, P., Matthews, L. H., Ashurst, J., Burton, J. *et al.*, *Nature* 2001, 414, 865–871.
- [49] http://vega.sanger.ac.uk/Homo_sapiens/mapview?chr=13.
- [50] Reymond, A., Camargo, A. A., Deutsch, S., Stevenson, B. J. *et al.*, *Genomics* 2002, 79, 824–832.
- [51] Hillier, L. W., Fulton, R. S., Fulton, L. A., Graves, T. A. *et al.*, *Nature* 2003, 424, 157–164.
- [52] Ashurst, J. L., Collins, J. E., *Annu. Rev. Genomics Hum. Genet.* 2003, 4, 69–88.
- [53] <http://vega.sanger.ac.uk/>.
- [54] <http://chr21.molgen.mpg.de/index.html>.
- [55] Mighell, A. J., Smith, N. R., Robinson, P. A., Markham, A. F., *FEBS Lett.* 2000, 468, 109–114.
- [56] Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M. *et al.*, *Genome Res.* 2002, 12, 272–280.
- [57] <http://www.ncbi.nlm.nih.gov/LocusLink/statistics.html>.
- [58] <http://www.ebi.ac.uk/IPI/IPIhelp.html>.
- [59] Pruitt, K. D., Maglott, D. R., *Nucleic Acids Res.* 2001, 29, 137–140.
- [60] Basrai, M. A., Hieter, P., Boeke, J. D., *Genome Res.* 1997, 7, 768–771.
- [61] <http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs>.
- [62] http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html.
- [63] Lee, D. K., George, S. R., O'Dowd, B. F., *Expert Opin. Ther. Targets* 2002, 6, 185–202.
- [64] Aparicio, S., Chapman, J., Stupka, E., Putnam, N. *et al.*, *Science* 2002, 297, 1301–1310.
- [65] Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J. *et al.*, *Nature* 2002, 420, 520–562.
- [66] http://www.ensembl.org/Rattus_norvegicus/.
- [67] Guigo, R., Dermitzakis, E. T., Agarwal, P., Ponting, C. P. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, 100, 1140–1145.
- [68] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J. *et al.*, *Nature* 2002, 420, 563–573.
- [69] Taylor, J. S., Braasch, I., Frickey, T., Meyer, A., Van de Peer, Y., *Genome Res.* 2003, 13, 382–390.
- [70] Eddy, S. R., *Nat. Rev. Genet.* 2001, 2, 919–929.
- [71] Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S. *et al.*, *Science* 2002, 296, 916–919.
- [72] Rinn, J. L., Euskirchen, G., Bertone, P., Martone, R. *et al.*, *Genes Dev.* 2003, 17, 529–540.
- [73] Numata, K., Kanai, A., Saito, R., Kondo, S. *et al.*, *Genome Res.* 2003, 13, 1301–1306.
- [74] Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., Hayashizaki, Y., *Genome Res.* 2003, 13, 1324–1334.
- [75] Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y. *et al.*, *Nat. Biotechnol.* 2003, 21, 379–386.
- [76] Carmichael, G. G., *Nat. Biotechnol.* 2003, 21, 371–372.
- [77] Barnes, M. R., in: Barnes, M., Gray, I. (Eds.) *Genetic Bioinformatics: A Bioinformatics Guide for Geneticists*, John Wiley & Sons, Chichester 2003, pp. 249–287.

- [78] Das, M., Burge, C. B., Park, E., Colinas, J., Pelletier, J., *Genomics* 2001, 77, 71–78.
- [79] Zhang, Z. L., Harrison, P. M., Gerstein, M., *J. Mol. Biol.* 2002, 323, 811–822.
- [80] Vitale, L., Casadei, R., Canaider, S., Lenzi, L. *et al.*, *Gene* 2002, 290, 141–151.
- [81] Wistow, G., Bernstein, S. L., Wyatt, M. K., Fariss, R. N. *et al.*, *Mol. Vis.* 2002, 8, 205–220.
- [82] Taylor, S. W., Fahy, E., Zhang, B., Glenn, G. M. *et al.*, *Nat. Biotechnol.* 2003, 21, 281–286.
- [83] Adam, P. J., Boyd, R., Tyson, K. L., Fletcher, G. C. *et al.*, *J. Biol. Chem.* 2003, 278, 6482–6489.
- [84] Han, D. K., Eng, J., Zhou, H., Aebersold, R., *Nat. Biotechnol.* 2001, 19, 946–951.
- [85] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J. *et al.*, *Mol. Cell. Proteomics* 2002, 1, 947–955.
- [86] Rappsilber, J., Ryder, U., Lamond, A. I., Mann, M., *Genome Res.* 2002, 12, 1231–1245.
- [87] Fountoulakis, M., Juranville, J. F., Dierssen, M., Lubec, G., *Proteomics* 2002, 2, 1547–1576.
- [88] Choudhary, J. S., Blackstock, W. P., Creasy, D. M., Cottrell, J. S., *Proteomics* 2001, 1, 651–667.
- [89] Kuster, B., Mortensen, P., Andersen, J. S., Mann, M., *Proteomics* 2001, 1, 641–650.
- [90] <http://www.sanger.ac.uk/Info/Press/2003/030414.shtml>.
- [91] Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., Tatusova, T. A., *BMC Evol. Biol.* 2002, 2, 20.
- [92] Liberles, D. A., Wayne, M. L., *Genome Biol.* 2002, 3, REVIEWS1018.



Dr. Chris Southan is Proteome Discovery Bioinformatician at Oxford GlycoSciences. Previously he was Head of Computational Biology at Gemini Genomics (2000–2001), Senior Investigator, Bioinformatics Target Discovery, SmithKline Beecham (1987–1999), Cancer Research Campaign

Protein Sequencing Facility (1986–1987) and Research Fellow, Charing Cross Hospital Medical School (1984–1986). He has a BSc in Biochemistry (Dundee University, 1974), an MSc in Virology (Reading University, 1975) and PhD from the Ludwig Maximilian University of Munich in 1983.

- [93] Ureta-Vidal, A., Ettwiller, L., Birney, E., *Nat. Rev. Genet.* 2003, 4, 251–262.
- [94] <http://www.hupo.org/>.
- [95] McGowan, S. G., Terrett, J., Brown, C. J., Adam, P. J. *et al.*, *Current Proteomics* 2004, 1, 41–48.
- [96] Braun, P., Hu, Y., Shen, B., Halleck, A. *et al.*, *Proc. Natl. Acad. Sci. USA* 2002, 99, 2654–2659.