16. Threadgill, D. W., Hunter, K. R. & Williams, R. W. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome* **13,** 175–178 (2002).
17. Paigen, K. & Eppig, J. T. A mouse phenome project. *Mamm. Genome* **11,** 715–717 (2000).
18. Batzoglou, S. *et al.* ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
19. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. 2. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
20. Yokoyama, T. *et al.* Conserved cysteine to serine mutation in tyrosinase is responsible for the classical albino mutation in laboratory mice. *Nucleic Acids Res.* **18,** 7293–7298 (1990).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements We thank the Mouse Genome Sequencing Consortium for sequencing and SNP discovery efforts and the University of Queensland for providing study leave to C.M.W. We thank W. Frankel and K. Scott for providing DNA from seven B6 founder stocks from the Jackson Laboratories. We also thank R. Jaenisch, D. Page, A. Chess, W. Dietrich, J. Hirschhorn, D. Altshuler, J. Barrett and M.-P. Reeve for comments on the manuscript; B. Gilman, S. Schaffner and E. Karlsson for computational assistance; and L. Gaffney for assistance with figures. M.J.D. is supported through a computational biology fellowship funded by Pfizer, Inc.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.D. (e-mail: mjdaly@genome.wi.mit.edu) or K.L.T. (e-mail: kersli@genome.wi.mit.edu).

Numerous potentially functional but non-genic conserved sequences on human chromosome 21

Emmanouil T. Dermitzakis*, Alexandre Reymond*, Robert Lyle*, Nathalie Scamuffa*, Catherine Ucla*, Samuel Deutsch*, Brian J. Stevenson†‡, Volker Flegel†‡, Philipp Bucher†§, C. Victor Jongeneel†‡ & Stylianos E. Antonarakis*

** Division of Medical Genetics, 1 Rue Michel-Servet, University of Geneva Medical School and University Hospitals of Geneva, CH-1211 Geneva, Switzerland*
† Swiss Institute of Bioinformatics, § Swiss Institute for Experimental Cancer Research, and ‡ Office of Information Technology, Ludwig Institute for Cancer Research, 155 Chemin des Boveresses, CH-1066 Epalinges, Switzerland

The use of comparative genomics to infer genome function relies on the understanding of how different components of the genome change over evolutionary time[1–3]. The aim of such comparative analysis is to identify conserved, functionally transcribed sequences such as protein-coding genes and non-coding RNA genes, and other functional sequences such as regulatory regions[4,5], as well as other genomic features. Here, we have compared the entire human chromosome 21 with syntenic regions of the mouse genome, and have identified a large number of conserved blocks of unknown function. Although previous studies have made similar observations[6,7], it is unknown whether these conserved sequences are genes or not. Here we present an extensive experimental and computational analysis of human chromosome 21 in an effort to assign function to sequences conserved between human chromosome 21 (ref. 8) and the syntenic mouse regions. Our data support the presence of a large number of potentially functional non-genic sequences, probably regulatory and structural. The integration of the properties of the conserved components of human chromosome 21 to the rapidly accumulating functional data for this chromosome[9,10] will improve considerably our understanding of the role of sequence conservation in mammalian genomes.

The sequence of human chromosome 21 (ref. 8) was obtained from the National Center for Biotechnology Information (NCBI) and aligned with PipMaker[11] to the mouse orthologous sequences (both sequences were hard-masked with Repeatmasker). Details and parameters of the alignment are described in the Methods. Briefly, 33.5 megabases (Mb) of human chromosome 21 sequence was compared to approximately 21 Mb of mouse sequence from chromosomes 16, 17 and 10 (refs 7, 12). A large number of ungapped conserved sequences were identified and their

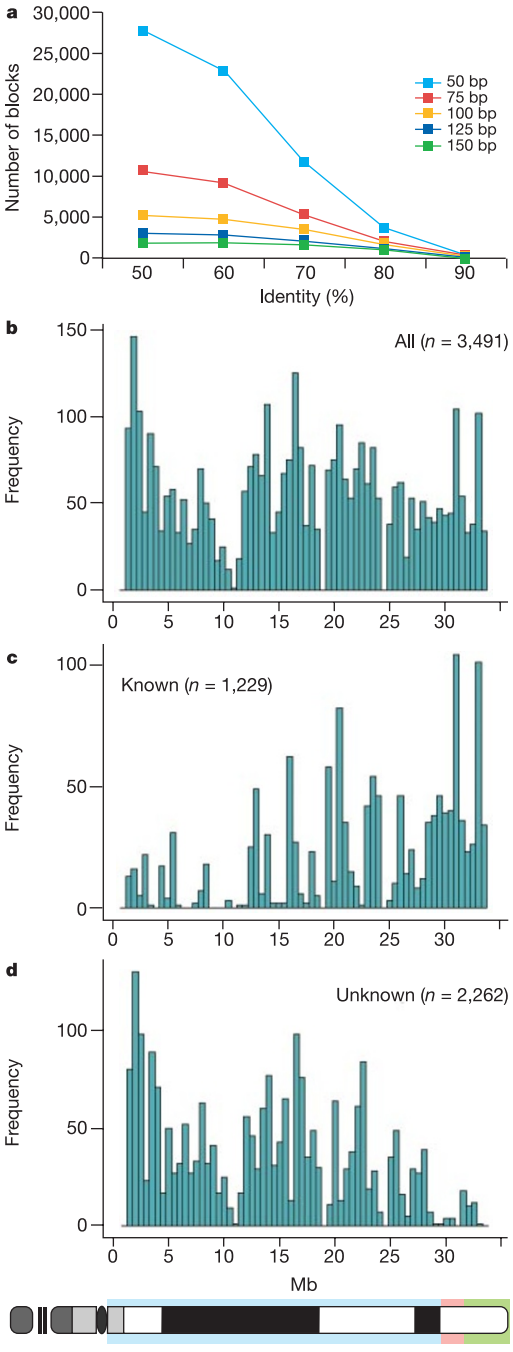

Figure 1 Distribution of conserved blocks on the 33.5 Mb of human chromosome 21 (long arm). **a**, Number of conserved sequence blocks identified (known and unknown) as a function of the threshold criteria for the size and percentage identity of the ungapped conserved blocks. **b**–**d**, The distribution of conserved blocks on the 33.5 Mb of human chromosome 21 at ≥100 bp and ≥70% identity is shown. Histograms showing all 3,491 conserved blocks (**b**), 1,229 conserved blocks corresponding to known exonic sequences (**c**), and 2,262 conserved blocks of unknown function (**d**). Below the histograms, human chromosome 21 with its banding pattern, and with the corresponding mouse syntenic regions is shown (mouse chromosomes 16, 17 and 10 are indicated in blue, pink and green, respectively).

distribution depending on the percentage identity and size thresholds is shown in Fig. 1a. We chose to analyse ungapped sequences of ≥100 base pairs (bp) with ≥70% identity, yielding a total of 3,491 blocks. These threshold criteria are informative for the identification of important genomic functional elements[13,14]. From those 3,491 blocks, 1,229 correspond partly or fully to known exonic sequences and annotated pseudogenes of human chromosome 21 (denoted as known)[8,10,15–18], and the remaining 2,262 blocks have unknown function (denoted as unknown). When compared with each other they did not have significant similarity, indicating that they are not repeats or widespread structural features. They also appear to be single copy in the human genome. Figure 1b–d shows the distribution of conserved blocks along the long arm of human chromosome 21. A large number of the conserved sequences are present in the gene-poor region of human chromosome 21 (ref. 8), suggesting that this region is not a 'functional desert'.

Unknown conserved blocks probably belong to the following six categories with respect to their function: (1) alternatively spliced, unknown exons of known genes; (2) exons of unknown genes; (3) non-coding RNAs; (4) *cis*-regulatory regions; (5) functional sequences of unknown significance; and (6) non-functional sequences with low divergence due to low substitution rate. In this study we investigated how the 2,262 unknown conserved blocks are distributed in the above six categories of conserved sequences.

To test experimentally the coding potential of the unknown 2,262 conserved blocks we applied four different methods to obtain candidate gene models in the human sequence: GrailEXP, Pro-Gen[19], human and mouse expressed sequence tag (EST) matches and adjacent conserved blocks. We also used NCBI annotations mostly on the basis of the gene prediction program GenomeScan, and performed BLAST analysis to identify which of the conserved blocks matched NCBI annotations. Out of 454 hypothetical loci in the NCBI human chromosome 21 annotations, most of which are GenomeScan predictions, only 18 matched 59 unknown conserved blocks. Furthermore, a total of 123 gene models (24 GrailEXP predictions, 10 Pro-Gen predictions, 26 EST matches and 63 adjacent conserved blocks) were tested experimentally. Oligonucleotides spanning putative introns of the gene models were used for polymerase chain reaction with reverse transcription (RT–PCR) amplification and sequencing on a panel of complementary DNA pools from 20 human adult and fetal tissues. Out of 123 models, only 2 spliced transcripts (1 Pro-Gen model and 1 GrailEXP model) were confirmed (1.7%). By using this RT–PCR methodology in other studies we documented transcripts of more than 98% of the human chromosome 21 genes, and discovered 19 additional transcripts[10,18]; we also found 139 additional transcripts in the whole mouse genome[12]. All 26 EST and 63 adjacent block models were also tested in the presence and absence of reverse transcriptase in hepatoma cell line and brain cDNA pools, to reveal possible unspliced non-coding RNAs. None of the 89 models showed evidence for the presence of an unspliced transcript in the tissues tested. In addition, using QRNA[20], which predicts coding sequences and non-coding RNAs on the basis of the pattern of substitution between two sequences, only 225 and 193 conserved blocks were predicted as coding or non-coding RNAs, respectively.

To investigate further the potential of functional transcription of the conserved blocks, we analysed a data set that reported the levels of transcriptional activity across the human chromosome 21 unique genomic sequence in 11 human cell lines, using Affymetrix oligonucleotide arrays[21]. We first created a non-redundant set of the oligonucleotide sequences reported as positive in at least one cell line[21], and then used BLAST to identify those sequences that matched conserved blocks. To avoid spurious positives we selected conserved blocks that corresponded to at least two positive oligonucleotides of the array. A total of 485 (21.4%) conserved blocks matched two or more positive oligonucleotides in the array.

We combined 6 pieces of computational and experimental evidence (GrailEXP prediction, GenomeScan prediction, human EST match, mouse EST match, QRNA prediction-coding, and RNA- and oligonucleotide array match) to obtain a signal of functional transcription for each of the 2,262 unknown blocks. Figure 2 shows that this signal is weak and that most of the blocks (63%) have no

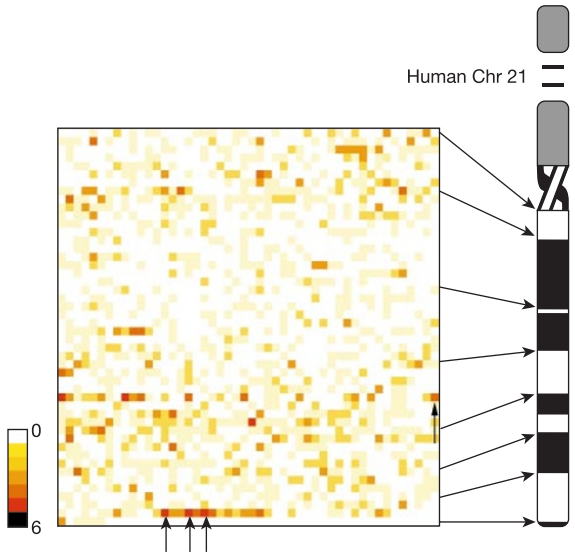

Figure 2 Colour-coded panel of the evidence of functional transcription (scale: 0 to 6) in 2,262 unknown conserved blocks. Arrows indicate the position of the last block of each row on human chromosome 21 (far right). The blocks are depicted as squares and the order along human chromosome 21 (centromere to telomere) is left to right and up to down. The 6 criteria are: GrailEXP, GenomeScan, human EST, mouse EST, QRNA (coding or RNA) and microarray. Vertical arrows indicate NCBI annotations with high support (bottom) and a gene identified during this analysis (right). Note the contrast of low support for most blocks compared with the blocks indicated by arrows.

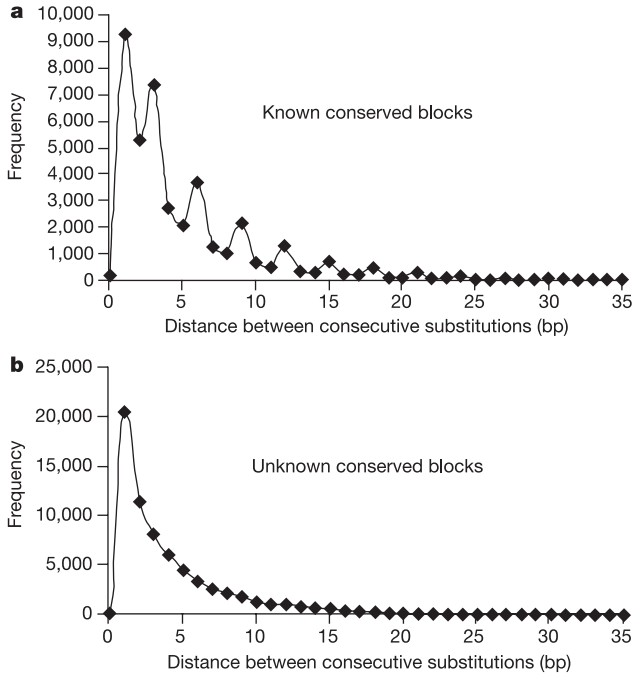

Figure 3 Frequency distribution of the distances between consecutive nucleotide substitutions in known (**a**) and unknown (**b**) conserved blocks. Note the periodicity (period = 3) in **a** due to the high frequency of synonymous substitutions at the third codon position, which is absent from **b**.

support. Three conserved blocks that probably correspond to genes with high experimental support—according to the NCBI annotation—and a gene identified by our group during this analysis stand out as highly supported in Fig. 2 (vertical arrows). (Supplementary Table 1 summarizes the coding potential and expression characteristics of all 2,262 unknown conserved blocks as derived by the *in silico* and experimental analysis.) The pattern seen in Fig. 2 combined with the previously described experimental RT–PCR analysis suggests that the functional transcription potential of most of the 2,262 unknown blocks is low.

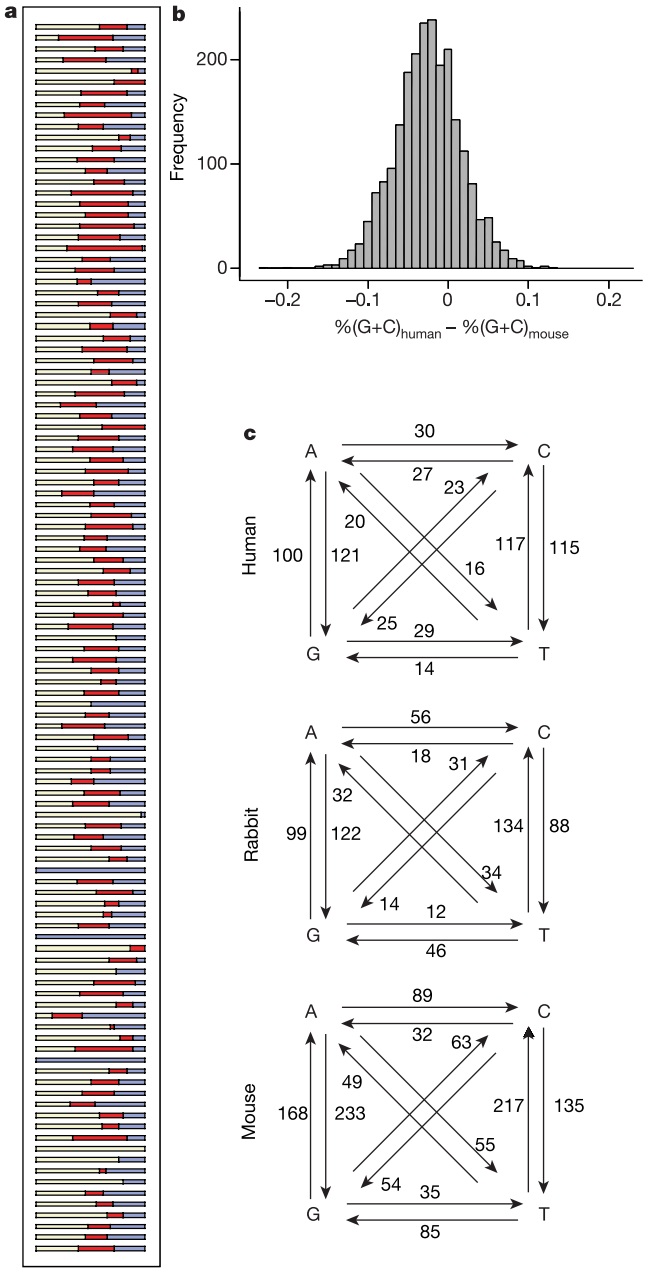

Figure 4 Analysis of sequence comparison of three species. **a**, Proportional representation of the species-specific nucleotide substitutions in human (blue), rabbit (red) and mouse (yellow). Each bar represents one of the 111 regions sequenced in rabbit, ordered by their position on human chromosome 21 (centromere to telomere). Note the apparent heterogeneity. **b**, Distribution of the difference in the percentage (G+C) content between human and mouse ($\%(G+C)_{human} - \%(G+C)_{mouse}$). The trend towards higher (G+C) content in mouse is obvious by the shift of the distribution to the left. **c**, Species-specific substitution counts in human, mouse and rabbit. The (A+T) to (G+C) bias in mouse and rabbit is evident.

None of the above methods (except for QRNA) make use of the pattern of substitution between the human and mouse sequences. In coding sequences most of the nucleotide changes are synonymous and occur at the third position of a codon. To exploit that characteristic[22], for each of the human–mouse alignments in the 2,262 unknown blocks and the 1,229 known (exonic) blocks we obtained measures of the distances (in bp) between consecutive nucleotide substitutions. Figure 3 shows the distribution of these distances in known and unknown conserved blocks. Within the known blocks there is a pronounced periodicity (period = 3) owing to the high frequency of synonymous changes in the third position, which is not at all present in the unknown set. We performed the same analysis in the unknown blocks identified by GrailEXP, GenomeScan, oligonucleotide array data, and QRNA, and none showed periodicity (see Supplementary Information). These data further demonstrate that most of the unknown conserved blocks are unlikely to be coding sequences.

To test for active (functional) conservation we attempted to amplify by PCR a total of 220 (about 10%) of the unknown conserved blocks in a third species, the rabbit. This species was chosen because, on the basis of mammalian phylogeny, it is almost equally distant to mouse and human (slightly closer to mouse). Thus there has been enough time since its separation from the two other species to allow efficient phylogenetic footprinting[23–25]. The distribution of conserved blocks amplified in rabbit is very similar to the distribution of the unknown conserved blocks on human chromosome 21. Out of the 220 unknown conserved blocks, we obtained 111 distinct rabbit sequences without any bias in their position on the chromosome (remaining sequences did not amplify or had multiple PCR products), and a total of 18,288 nucleotides were aligned in all three species with MultiPipMaker. Average measures of divergence of the rabbit sequences were within the expected level (Supplementary Information), which supports the hypothesis that these sequences are orthologous. Three-way species sequence analysis[26] allows for a detailed description of the pattern of substitution within these regions. Although in ref. 6 some three-way analysis was performed, the methodology did not allow for either reliable quantification of conservation or analysis of the pattern of substitution, which can reveal significant properties of these sequences.

To perform a fine analysis of substitution patterns, we counted the species-specific substitutions in each of the 111 conserved blocks (see Supplementary Information for methods) and constructed a 3×111 contingency table to test whether there is heterogeneity in the substitution pattern. A Fisher's exact test shows that the pattern is significantly heterogeneous ($P < 10^{-4}$). All three species seem to contribute to this heterogeneity (Fig. 4a) and thus it is not an artefact of the presence of paralogues in rabbit data or differentially accelerated substitution rate in mouse. Performing the test in a sliding window of blocks within 1 Mb or 500 kb of genomic sequence, most tests show $P < 0.001$; this heterogeneous pattern cannot be explained by regional variation of species-specific mutation or substitution rate. Therefore, despite the fact that these sequences are conserved as a whole between species, the heterogeneous species-specific substitution patterns suggest differential selective pressure. Fine scale species-specific changes in a conserved sequence could indicate differences in putative functional regions (for example, regulatory) that contribute to species differences[5,27].

A recent study suggested that variation in substitution rate among genomic regions[28] is probably due to different nucleotide composition bias between species in some genomic regions[29], which would increase the mutational bias of a sequence within these regions. This allows for a prediction of the effect of the genomic context to the neutral substitution rate. We explored the substitution pattern of the conserved regions by focusing on the type of nucleotide changes that have occurred. We calculated the difference in the (G+C) content between human and mouse in the

2,262 conserved blocks (Fig. 4b). There is a pronounced skew towards higher (G+C) content in the mouse sequence, which implies different mutational pressure in this genomic region in the two species. With the three-way species alignment we can determine the direction of these substitutions (Fig. 4c) and decide which of the species drive the (G+C) content differential in the conserved regions. Out of the 637 human-specific substitutions 293 (46%) were A+T to G+C and 271 (42.5%) were G+C to A+T. For mouse these numbers were 624 A+T to G+C out of 1,215 (51.3%) and 370 G+C to A+T out of 1,215 (30.4%); for rabbit these were 358 A+T to G+C out of 686 (52.1%) and 217 G+C to A+T out of 686 (31.6%). This demonstrates that there is no mutational bias on human chromosome 21, whereas there is an A+T to G+C mutational bias in the mouse and rabbit syntenic regions. The difference in the mutational bias between human and mouse/rabbit in these regions increases the pressure for substitutions so as to reach the species-specific nucleotide composition equilibrium. This observation suggests the presence of an extra force of nucleotide change (A+T to G+C bias) that increases the substitution rate relative to a non-bias model. High sequence conservation with the presence of this bias is an additional support for their functional significance.

A preliminary description of sequence conservation along human chromosome 21 has been presented in previous studies[6,7]. These studies have detected patterns of conservation similar to our study, but the question of whether these unknown conserved sequences are unannotated genes or are non-genic remained unanswered. This is an important question because genic and non-genic sequences require different experimental approaches for their functional analysis. The fact that many conserved sequences are not present in the current gene annotation does not necessarily mean that they are non-coding, as discussed in the accompanying manuscript, where a significant number of new genes have been verified experimentally[12]. In our study, we tested specific hypotheses about the importance of the unknown conserved sequences. With this analysis we rejected the hypothesis that the unknown conserved sequences on human chromosome 21 are unannotated genes, therefore emphasizing the need for specific experimental strategies to unravel the function of such non-coding sequences.

Another major aspect of the present study is the use of direct nucleotide alignment to perform the analysis. One study used high-density microarrays[6], which provide qualitative but not reliable quantitative data in terms of numbers and types of substitutions at levels of divergence such as those between human and mouse. Another study performed nucleotide comparisons[7], but the data were restricted to the degree of overlap of the conserved sequences with the known genes, and only involved two species. In our study we describe certain characteristics of the pattern of substitutions between human, mouse and rabbit that allowed us to test the coding potential of the conserved sequences, as well as the degree of active conservation. Our analysis demonstrates the power of combining computational, experimental and evolutionary analysis for the understanding of genome function.

Despite our extensive computational and experimental efforts to assign gene identity to the 2,262 unknown conserved blocks, the evidence we present indicates that most of these conserved blocks are not protein-coding or RNA genes. The selection of ungapped alignments for $\geq$100 bp is conservative, as ungapped sequences are more likely to be coding to maintain the open reading frame. Therefore the number of non-genic conserved sequences we identified here is an underestimate (see also Fig. 1a). Sequence conservation of at least 50% of a sample of the conserved blocks was confirmed in rabbit. This level of conservation, in light of the substitution biases in the sequences described here, shows that the selective constraint is high. If at least 50% of the 2,262 unknown blocks are selectively constrained and 63% of them have no support of being genes, as we have shown, then we estimate that there is a lower bound of 712 functional sequence blocks of $\geq$100 bp with no gene characteristics on human chromosome 21. This large number of functionally constrained sequences that are not genes probably consists of regulatory regions and unknown functional features of the genome that we are now beginning to discover. Further study of non-genic sequences will improve significantly our understanding of features shared between species as well as species differentiation. Trisomy for some of these non-genic sequences may also contribute to Down's syndrome phenotypes, and their functional characterization provides an attractive challenge for the understanding of genome function in health and disease. Although the present analysis is performed in human chromosome 21, the conclusions can be projected to the whole human and mouse genomes. As discussed in the accompanying manuscript[12] there is a high abundance of conserved regions not corresponding to known genes and a large fraction of those are not genic. The identification of the role of such sequences will probably reveal important clues for genome function and regulation. □

Methods

Alignment

Human sequence was obtained from NCBI (human genome version: build 28) and corresponded to the following contigs: NT_011512 (gi: 161170824), NT_030187 (sequence version gi: 16166615), NT_030188 (gi: 16166749), NT_011515 (gi: 16166537). Mouse genomic sequence was originally obtained from Celera genomics (ref. 7; see also http://www.celera.com) through subscription available at the Ludwig Institute for Cancer Research in Lausanne (CVJ). When the publicly available mouse genome was released all 3,491 conserved blocks were subjected to BLAST analysis to verify their presence in the public domain[12]. The accession numbers for the mouse sequences syntenic to human chromosome 21 are available as Supplementary Information. Both human and mouse sequences were masked for human and rodent repeats respectively with RepeatMasker (A. F. A. Smit and P. Green, unpublished data (http://ftp.genome.washington.edu/RM/RepeatMasker.html).)

Human and mouse sequences were searched with BLAST to identify the genomic boundaries of known genes, and using these boundaries the contigs were separated in slightly overlapping segments of approximately 2–3 Mb. Subsequently they were individually aligned using PipMaker. In all cases where we used BLAST for the identification of the characteristics of the conserved blocks, we required stringent matching criteria with an *e*-value $<10^{-20}$ and similarity >98%.

Gene prediction

We used four different methods to obtain gene models. For the GrailEXP method (http://compbio.ornl.gov/grailexp/) we performed exon prediction on extended versions of the human sequence of the conserved blocks by adding 50 bp upstream and downstream to provide the appropriate genomic context, and we chose pairs of predicted exons that were less than 100 kb apart and that had the same orientation as putative exons of a gene. For the Pro-Gen method we carried out homology-based gene prediction using the genomic sequences of human and mouse corresponding to regions with high density of conserved blocks ($\geq$5 blocks within 40 kb). For the EST matches pairs of adjacent conserved blocks that matched the same human or mouse EST were considered putative pairs of exons of the same gene. The EST entries used were from the Human Genome Mapping Project (HGMP) database. The fourth method was adjacent conserved blocks. Pairs of adjacent conserved blocks that had a low ratio of replacement (K_a) to silent substitutions (K_s) in one of the six putative coding frames ($K_a/K_s < 0.5$) and similarity between human–mouse >85% were chosen.

RT–PCR analysis

The expression of the above predictions was tested by RT–PCR. Total RNA derived from 20 different normal human adult tissues (brain, heart, kidney, spleen, liver, stomach, colon, small intestine, muscle, lung, testis, placenta, skin, peripheral blood lymphocytes, bone marrow, fetal brain, fetal liver, fetal kidney, fetal heart and fetal lung) was extracted, reverse-transcribed and normalized. The quality of total RNA was tested by PCR using MLH1 primers located at intronic sequences flanking exon 12 (forward, 5′-TGGTGTCTCTAGTTCTGG-3′; reverse, 5′-CATTGTTGTAGTAGCTCTGC-3′), as an indicator of possible genomic DNA contamination. Primers for RT–PCR were designed using the Primer3 program (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Primers were designed from the sequence of distinct putative exons so that the possible amplification of genomic DNA could be distinguished from cDNA amplification. We chose a single PCR rather than a nested PCR approach to avoid false-positive results due to illegitimate transcription. Similar amounts of the 20 cDNAs (final dilution × 250) were mixed with JumpStart REDTaq ReadyMix (Sigma) and 4 ng ml^{-1} of each of the primers (Sigma-Genosys) with a BioMek 2000 robot (Beckman). The ten first cycles of PCR amplification were performed with a touchdown annealing temperature decreasing from 60 °C to 50 °C; annealing temperature of the next 35 cycles was 50 °C. Amplimers were separated on Ready to Run precast gels (Pharmacia), and positives were sequenced directly.

Microarray data analysis

We identified all of the positive oligonucleotides with the threshold values $R = 13$ and $D = 12Q$ (ref. 21). R and D are threshold values for the ratio and the difference between perfect match intensity and mismatch intensity, respectively. Thus varying these values gives different measures of sensitivity and specificity. We then used BLAST to identify conserved blocks that corresponded to at least two positive oligonucleotides so as to reduce the number of false positives.

Received 16 September; accepted 30 October 2002; doi:10.1038/nature01251.

1. Hardison, R. C., Oeltjen, J. & Miller, W. Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7,** 959–966 (1997).
2. O'Brien, S. J. *et al.* The promise of comparative genomics in mammals. *Science* **286,** 458–462 (1999) 479–481.
3. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17,** 373–376 (2001).
4. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16,** 369–372 (2000).
5. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19,** 1114–1121 (2002).
6. Frazer, K. A. *et al.* Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11,** 1651–1659 (2001).
7. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296,** 1661–1671 (2002).
8. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
9. Antonarakis, S. E., Lyle, R., Deutsch, S. & Reymond, A. Chromosome 21: a small land of fascinating disorders with unknown pathophysiology. *Int. J. Dev. Biol.* **46,** 89–96 (2002).
10. Reymond, A. *et al.* Human chromosome 21 gene expression atlas in the mouse. *Nature* **420,** 582–586 (2002).
11. Schwartz, S. *et al.* PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10,** 577–586 (2000).
12. Waterston, R. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
13. DeSilva, U. *et al.* Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res.* **12,** 3–15 (2002).
14. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288,** 136–140 (2000).
15. Davisson, M. T. *et al.* Evolutionary breakpoints on human chromosome 21. *Genomics* **78,** 99–106 (2001).
16. Gardiner, K., Slavov, D., Bechtel, L. & Davisson, M. Annotation of human chromosome 21 for relevance to down syndrome: gene structure and expression analysis. *Genomics* **79,** 833–843 (2002).
17. Reymond, A. *et al.* From PREDs and open reading frames to cDNA isolation: revisiting the human chromosome 21 transcription map. *Genomics* **78,** 46–54 (2001).
18. Reymond, A. *et al.* Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics* **79,** 824–832 (2002).
19. Novichkov, P. S., Gelfand, M. S. & Mironov, A. A. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* **17,** 1011–1018 (2001).
20. Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BioMed Central Bioinformatics* **2,** 8 (2001).
21. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919 (2002).
22. Nekrutenko, A., Makova, K. D. & Li, W. H. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* **12,** 198–202 (2002).
23. Madsen, O. *et al.* Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409,** 610–614 (2001).
24. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294,** 2348–2351 (2001).
25. Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409,** 614–618 (2001).
26. Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10,** 1304–1306 (2000).
27. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296,** 340–343 (2002).
28. Eyre-Walker, A. & Hurst, L. D. The evolution of isochores. *Nature Rev. Genet.* **2,** 549–555 (2001).
29. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99,** 803–808 (2002).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements This project was supported by grants from the Swiss National Science Foundation, National Center for Competence in Research 'Frontiers in Genetics', the European Union/Federal office of Education and Health 'Child Care' foundation (to S.E.A.), and a Swiss National Science Foundation grant (to P.B.). We thank E. Lander for advice and support, C. Rossier for core sequencing support and J. Yang for providing programs.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.E.A. (e-mail: stylianos.antonarakis@medecine.unige.ch).

Human chromosome 21 gene expression atlas in the mouse

Alexandre Reymond*†, Valeria Marigo†‡, Murat B. Yaylaoglu†§, Antonio Leoni‡, Catherine Ucla*, Nathalie Scamuffa*, Cristina Caccioppoli‡, Emmanouil T. Dermitzakis*, Robert Lyle*, Sandro Banfi‡, Gregor Eichele§, Stylianos E. Antonarakis* & Andrea Ballabio‡||

** Division of Medical Genetics, University of Geneva Medical School and University Hospital of Geneva, CMU, 1, rue Michel Servet, 1211 Geneva, Switzerland*
‡ Telethon Institute of Genetics and Medicine, Via Pietro Castellino 111, 80131 Naples, Italy
§ Max Planck Institute of Experimental Endocrinology, Feodor-Lynen-Str. 7, D-30625 Hannover, Germany
|| Medical Genetics, Second University of Naples, Naples, Italy
† These authors contributed equally to this work

Genome-wide expression analyses have a crucial role in functional genomics. High resolution methods, such as RNA *in situ* hybridization provide an accurate description of the spatiotemporal distribution of transcripts as well as a three-dimensional '*in vivo*' gene expression overview[1–5]. We set out to analyse systematically the expression patterns of genes from an entire chromosome. We chose human chromosome 21 because of the medical relevance of trisomy 21 (Down's syndrome)[6]. Here we show the expression analysis of all identifiable murine orthologues of human chromosome 21 genes (161 out of 178 confirmed human genes) by RNA *in situ* hybridization on whole mounts and tissue sections, and by polymerase chain reaction with reverse transcription on adult tissues. We observed patterned expression in several tissues including those affected in trisomy 21 phenotypes (that is, central nervous system, heart, gastrointestinal tract, and limbs). Furthermore, statistical analysis suggests the presence of some regions of the chromosome with genes showing either lack of expression or, to a lesser extent, co-expression in

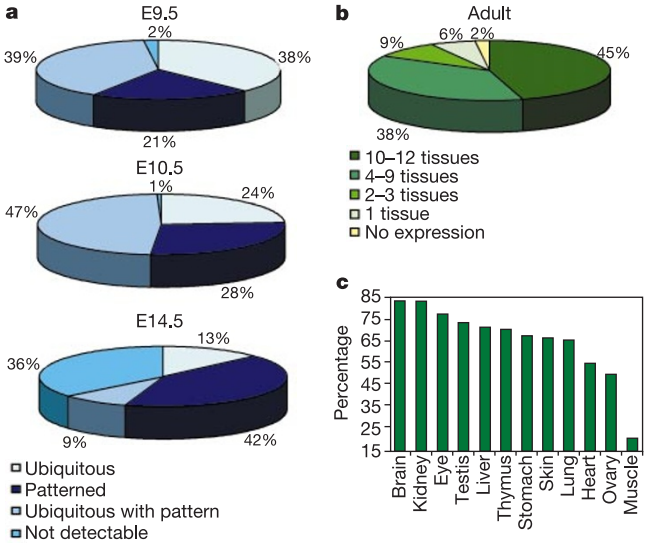

Figure 1 Distribution of expression patterns and transcriptome complexity. **a**, Each slice corresponds to the percentage of genes belonging to the four categories of expression pattern observed by ISH at E9.5 (whole mount), E10.5 (whole mount) and E14.5 (sections). **b**, Each slice represents the percentage of genes expressed in 0, 1, 2–3, 4–9 and 10–12 adult tissues by RT–PCR. **c**, Percentage of the analysed 161 human chromosome 21 murine orthologues identified in each murine adult tissue.