Comparative gene finding in chicken indicates that we are closing in on the set of multi-exonic widely expressed human genes

Robert Castelo*, Alexandre Reymond[1,2], Carine Wyss[1], Francisco Câmara, Genís Parra, Stylianos E. Antonarakis[1], Roderic Guigó and Eduardo Eyras

Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra, Centre de Regulació Genòmica, E08003 Barcelona, Spain, [1]Department of Genetic Medicine and Development, University of Geneva, Medical School and University Hospital of Geneva, CMU, 1, rue Michel Servet, 1211 Geneva, Switzerland and [2]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

Received December 2, 2004; Revised December 13, 2004; Accepted March 8, 2005

ABSTRACT

The recent availability of the chicken genome sequence poses the question of whether there are human protein-coding genes conserved in chicken that are currently not included in the human gene catalog. Here, we show, using comparative gene finding followed by experimental verification of exon pairs by RT–PCR, that the addition to the multi-exonic subset of this catalog could be as little as 0.2%, suggesting that we may be closing in on the human gene set. Our protocol, however, has two shortcomings: (i) the bioinformatic screening of the predicted genes, applied to filter out false positives, cannot handle intronless genes; and (ii) the experimental verification could fail to identify expression at a specific developmental time. This highlights the importance of developing methods that could provide a reliable estimate of the number of these two types of genes.

INTRODUCTION

The analysis of the (almost) finished human genome sequence (1) concludes that the number of protein-coding genes should lie in the interval of 20 000–25 000. However, the models used by computational techniques to annotate genes and obtain gene-number estimates are oversimplified, and as the knowledge of the biology involved in this problem increases, this number can be updated. One such update has recently taken place with the release and analysis of the sequence of the Tetraodon genome (2). The authors found 904 putative novel human genes with a computational approach, of which 63% had expressed sequence tag (EST) evidence projecting an additional 2.6% of genes over the current version of the human gene catalog (Ensembl 34d) which contains 22 287 genes. Similarly, the recently obtained sequence of the chicken genome (3), which is currently the closest available vertebrate genome to mammals, can allow us to detect human genes that are not present, or were not detected, in rodents and are poorly represented in human cDNA sets, because of low-and highly-specific expression. Here, we show that the additional set of novel human genes found with a bioinformatic protocol followed by experimental verification on the sequence of the chicken genome does not add more than 0.2% to the current human gene catalog. We further analyzed the 904 putative novel human genes inferred with the Tetraodon genome and found that using our filtering protocol the additional set would add a percentage of genes similar to the one estimated in this paper, thus supporting the hypothesis that the current human protein-coding gene catalog is close to completion. It should, however, be noted that there are other types of coding genes, such as intronless and quickly evolving genes, that remain unseen by our method and for which the chicken sequence could still be exploited.

There are two main ways to automate gene annotation: homology-based approaches and *de novo* approaches. The former annotates genes by aligning transcript and protein evidence to the genome and the latter by using a statistical model trained beforehand on an available subset of annotated genes. Homology-based approaches, e.g. Ensembl (4), are highly accurate but require previous evidence (e.g. an EST, a known paralog or ortholog, etc.); on the other hand, *de novo* approaches, while less accurate, can reliably identify genes with no transcriptional evidence beforehand. The simultaneous use of two or more genomes by *de novo* gene predictors

*To whom correspondence should be addressed. Tel: +34 932 240 884; Fax: +34 32 240 875; Email: rcastelo@imim.es

(called comparative gene predictors) has proven extremely effective, not only in improving sensitivity over single-genome methods but also in finding novel genes (5–7), where novel specifies genes that fall outside a core set of annotated gene loci for a given genome. Here, we use SGP2 as a comparative gene predictor (8).

MATERIALS AND METHODS

We have applied a bioinformatic protocol analogous to the one used for searching novel human genes with the mouse genome (6), using the newly released chicken genome as reference. To further enhance the accuracy, we took advantage of the differentiated intron–exon conservation observed between human and chicken orthologous genes. Our pipeline includes three sequential phases: (i) prediction of multi-exon genes in the two organisms and generation of homologous chicken–human prediction pairs; (ii) filtering out the models not supported by exonic structure conservation and not meeting the novelty criteria; and (iii) experimental verification by RT–PCR.

Generation of human–chicken homologous prediction pairs

Using the comparative gene predictor SGP2 (8), we have generated two sets of 40 005 human and 29 430 chicken genes. Each set of predictions was generated using the other genome as comparative reference and a large part of the difference in their size ($\sim$36%) is owing to a fragmentation of the human predictions since the number of predicted exons in human, 238 737, is only 17% larger than in chicken, where 203 834 exons were predicted. The compactness of the chicken genome, the chicken genome is $\sim$40% the size of the human genome (3), probably alleviates the tendency of gene predictors to fragmentate gene models. From the two prediction sets, we built a collection of 66 390 human–chicken putative homologous gene pairs using BLAST (9). For each chicken prediction, we drew a maximum of five human predictions whose protein alignment had an expected value $\leqslant 1 \times 10^{-10}$ in a BLASTP search.

Filtering by conservation and novelty

The set of 66 390 gene pairs were filtered by the conservation of the exonic structure in the following manner. We performed a pairwise global alignment of the corresponding proteins using T-Coffee (10), essentially carrying out the Needleman–Wunsch (11) alignment algorithm, and compared the intron locations in each alignment. The degree of conservation of the exonic structure is then defined as the proportion of the number of aligned intron positions to the number of introns that can be aligned (i.e. the minimum number of introns between the two genes).

There were 41 186 gene pairs (62%) that showed at least one intron position aligned (which is the minimum positive degree of conservation), and from these we kept 34 816 (52%) that had at least 50% of similarity along the 15 amino acids on either side of one or more of the conserved intron positions. This filter has already been shown to considerably reduce the false-positive rate of computational predictions (6).

Next, we filtered by novelty through the following two steps. First, we rejected pairs showing a high degree of homology in BLAST searches to the protein sets in Ensembl (release NCBI34d) and RefSeq (as of May 2004) and to the 41 118 full-length cDNAs from the H-Invitational database (12). More concretely, in the homology search for proteins we rejected those predictions with a match showing an expected value $<1 \times 10^{-50}$, while in the homology search for cDNAs we rejected those that showed at least 1 high-scoring segment pair (HSP) with an expected value $<1 \times 10^{-5}$ and a minimum length and identity of 60 bp and 95%, respectively.

Second, we considered as novel those human predictions that had no position overlap in the human genome with any of the annotations in Ensembl (release NCBI34d), RefSeq mRNAs (13) (UCSC mapping of May 2004), Vega manual annotations (http://vega.sanger.ac.uk) and full-open reading frame mRNAs from the Mammalian Gene Collection (14,15).

This definition of novelty includes all novel (i.e. found in new loci) transcripts even if they belong to known gene families and yielded a set of 332 pairs of chicken–human genes (0.5% of the initial set of 66 390 gene pairs) involving 311 human and 328 chicken genes. The BLAST searches in protein and cDNA sets make the approach highly conservative, but ensure that we discarded existing mRNAs that may be mis-annotated in the genome (16) due to, for instance, aligning an mRNA to the wrong locus in the genomic sequence.

Experimental verification by RT–PCR

The 311 human predicted genes form a set of putative novel human genes and can be downloaded from http://genome.imim.es/datasets/ggalhsapgenes2005. To estimate which fraction of these models are bona fide genes, we tried to experimentally verify by RT–PCR a subset of 50 promising candidates built in the following way. For a given predicted human gene, the experimental verification by RT–PCR was made on one pair of consecutive predicted exons; hence, the selection of candidate predictions not only involved choosing the most promising genes but also the most promising exon–exon junction. Thus, from the set of 311 putative novel human genes, we made a ranking of all the aligned human–chicken exon–exon junctions to select those most likely to correspond to real gene sections. The criterion to build this ranking was the abrupt change of intron–exon conservation between human and chicken. We captured this feature by first performing TBLASTX alignments (WashU-BLAST with parameters -nogap Z = 3000000000 B = 9000 V = 9000 -hspmax = 500 -topcomboN = 100 -filter = xnu+seg -matrix = BLOSUM62 -w = 5) between the genomic sequences of each gene pair, as well as between the pairs of aligned human–chicken exon–exon junctions. Second, we calculate the correlation coefficient at nucleotide level (17) (CCn) between the TBLASTX HSPs and the human gene predicted coordinates. The CCn ranges from -1 to 1, where the value 1 corresponds to the case when the conservation between the human and chicken sequences matches perfectly the coordinates of the predicted human gene. We made the ranking of the aligned human–chicken exon–exon junctions by decreasing CCn value, first for the entire gene coordinates (CCnG) and second for the exon–exon junction (CCnI). From the entire ranking,

Table 1. Tissue distribution for the positive cases

Identifier	Br	He	Ki	Sp	Li	Co	SI	Mu	Lu	St	Te	Pl	Sk	PBL	BM	FB	FL	FK	FH	FU	Th	Pa	MG	Pr
chr18_515																+								
chr15_51				+		+		+	+	+						+								
chr4_1746											+						+							
chr5_400										+														
chr4_55								+	+	+			+		+	+					+			+
chr22_143	+							+																

we selected a final set for experimental verification based on the following criteria:

(i) CCnG ⩾ 30%.
(ii) CCnI ⩾ 70%.
(iii) Minimum intron length of 400 bp and maximum of 10 kb.
(iv) Minimum flanking exons lengths of 30 bp.

This left us with 87 exon–exon junctions involving 52 genes. We selected for each gene the one with the largest CCnI value. From the 52 exon–exon junctions, we rejected two, as they were identical at sequence level to a third one in the set, producing a final set of 50 exon–exon junctions from 50 different genes to test by RT–PCR.

cDNA preparation. Human cDNAs from 24 different tissues were synthesized using 12 poly$(A)^+$ RNAs from Origene, 8 from Clemente Associates/Quantum Magnetics and 4 from BD Biosciences as described previously (18,19). The relative amount of each cDNA was normalized by quantitative PCR using SyberGreen as intercalator and an ABI Prism 7700 Sequence Detection System.

Experimental verification. Predictions of human genes were assayed experimentally by RT–PCR as previously described and modified (6,19,20). Similar amounts of 24 *Homo sapiens* cDNAs (brain, heart, kidney, spleen, liver, colon, small intestine, muscle, lung, stomach, testis, placenta, skin, peripheral blood cells, bone marrow, thymus, pancreas, mammary gland, prostate, fetal brain, fetal liver, fetal kidney, fetal heart and fetal lung, final dilution 2500×) were mixed with JumpStart REDTaq ReadyMix (Sigma) and 4 ng/µl primers (Sigma–Genosys) with a BioMek 2000 robot (Beckman). The first 10 cycles of PCR amplification were performed with a touch-down annealing temperatures decreasing from 60 to 50°C; annealing temperature of the next 30 cycles was carried out at 50°C. Amplimers were separated on 'Ready to Run' precast gels (Pharmacia) and sequenced. This procedure was used to experimentally assay the 50 exon–exon junctions of putative novel human genes. The sequences of the amplified human exon–exon junctions are deposited in GenBank under accession nos. AY947523–AY947528.

RESULTS

From the 50 exon–exon junctions of the set of promising candidates, 6 (12%) yielded positive results with expression in an average of 3.3 tissues out of the 24 tested [confidence interval (0.2–6.4) at 95% significance level; SD 2.9] and with 67% of cases showing expression in only 1 or 2 tissues (see Tables 1 and 2 and Supplementary Figures 1–7). As expected considering previous studies (6,21), these novel genes show a far more restricted tissue distribution than the previously known vertebrate genes which show expression in an average of 7–8 tissues out of the 12 tested (18). Further analyses of these six novel genes are given in Table 3 and in the Supplementary Material.

Table 2. Two-letter code for the tissues in Table 1

Code	Tissue
Br	Brain
He	Heart
Ki	Kidney
Sp	Spleen
Li	Liver
Co	Colon
SI	Small intestine
Mu	Muscle
Lu	Lung
St	Stomach
Te	Testis
Pl	Placenta
Sk	Skin
PBL	Peripheral blood leukocyte
BM	Bone marrow
FB	Fetal brain
FL	Fetal liver
FK	Fetal kidney
FH	Fetal heart
FU	Fetal lung
Th	Thymus
Pa	Pancreas
MG	Mammary gland
Pr	Prostate

Since the tested subset was formed by our best selection of candidates, the 12% success rate can be considered as an upper bound success rate within the set of 311 putative novel genes. Applying this rate, we obtain an upper estimate of 37 additional human genes which add up to a 0.2% of the current estimate of the number of protein-coding genes given by Ensembl. In order to reconcile this estimate with 2.6% (within the 904 genes) obtained with the Tetraodon sequence, we applied the described filters for novelty and found that, in fact, 466 out of the 904 passed the filters; of these, 371 were intronless genes and 95 were multi-exon genes. Since our protocol only applies to multi-exon genes, we intersected the subset of 95 Tetraodon-based predictions with our 311 chicken-based predictions obtaining 6 common putative novel human genes. From these, 4 were among the 50 promising candidates we submitted for RT–PCR verification and 1 was among the 6 RT–PCR positives (chr4_55 in Table 3), yielding a success rate of 25%. Applying this rate to the 95 putative novel human genes obtained from Tetraodon, we obtain an estimate of 24 novel human genes, an additional 0.1% of the current human

Table 3. Analysis of the six experimentally verified novel human genes

Identifier	Pfam	ESTs	Homology	Identity (%)	Coverage (%)	Description
Chr15_51	Mpp10 (PF04006) Neur_chan_LBD (PF02931)		NP_997588.1	74% (amino acid)	56% (amino acid)	ELMO2 (*H.sapiens*): engulfment and cell motility 2, ced-12 homolog
Chr18_515	Ski_Sno (PF02437)		AK049035	80% (amino acid)	89% (amino acid)	Riken mouse cDNA
Chr22_143	P2X_receptor (PF00864)	BX096265 BE876713	P2RXL1	94% (amino acid)	94% (amino acid)	Purinergic receptor P2X-like 1, orphan receptor
Chr4_1746		BQ429300 CN281994 CN281995	AK006501	77% (amino acid)	98% (amino acid)	Riken mouse cDNA of unknown function
Chr5_400		BQ428697 BF218453 CN265566 CD248366 BG496466	BC039102.1	99% (nucleotide)	73% (nucleotide)	IMAGE Clone—SelP precursor
Chr4_55		BU075833 AW163448 BE047596 AL583585 BM688172	CAG12806.1	83% (amino acid)	71% (amino acid)	Tetraodon protein of unknown function

The identifiers of the predictions include the chromosome name, on which they were predicted and a label that differentiates different genes predicted in the same chromosome. Three of the genes match Pfam domains and four match with genome specific human ESTs (not all are given). The percentage identity of the alignments with the homologous sequence (Identity) and the proportion of the gene prediction covered in this alignment (Coverage) are also given. We distinguish whether the alignment is at the amino acid or at the nucleotide level.

gene catalog and in agreement with our estimate using the chicken genome.

Additionally, we were also interested in searching for distinctive features of the RT–PCR-positive exon pairs in the computational predictions to possibly enhance the sensitivity and specificity of our predictions. We found that the length of their 3′ exon was on average much shorter (91 bp) than that of the negative cases (167 bp), an observation which suggested a relationship between the reliability of longer predicted exons and the selection of primers for RT–PCR. The average distance of the 5′ start of the reverse primers selected from the downstream exons in the negative cases is 132 bp, while this distance in the positive cases is 60 bp, which in turn makes more likely an overlap with the region of high conservation around the intron, required by our protocol. This region includes 15 amino acids on either side of the intron and the average conservation found was 85% for the negative cases and 91% for the positive ones. Thus, the primers selected in the positive cases overlap with stretches of a higher conservation level. This observation suggests that the described bioinformatic protocol is more reliable in its prediction of short rather than long exons.

DISCUSSION

These results confirm the previous observation (6) that there does not appear to be a large population of conserved protein-coding genes in mammals outside of the currently existing cDNA sets. They also may reflect that our current methods for protein-coding gene annotation fail to identify the missing genes in the human gene catalog. For example, we cannot exclude that some predictions were not positively verified because they are expressed in only a few cells at a specific developmental time. Moreover, our bioinformatic screening to filter out false-positive predictions is based on the differentiated intron–exon conservation, thus directly discarding intronless and quickly evolving genes. The development of experimental and computational methods addressed to find missing genes under such specific circumstances is, therefore, crucial to clear up whether our current methods are heavily underestimating the size of the missing human gene catalog. As a byproduct, such an investigation could enlarge our knowledge on the biology involving gene evolution and expression.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Josep F. Abril and Oscar González for technical assistance and Charles Chapple for his useful comments on a first draft of the manuscript. This research is supported by grants BIO2000-1358-C02-02 from ‘Plan Nacional de I+D’ (Spain), Jérôme Lejeune, Childcare and Désirée and Niels Yde Foundations (Switzerland), the European Union Biosapiens Network of Excellence, the Swiss National Science Foundation and the NCCR Frontiers in Genetics. Funding to pay the Open Access publication charges for this article was provided by grant BIO2000-1358-C02-02 from ‘Plan Nacional de I+D’ (Spain).

Conflict of interest statement. None declared.

REFERENCES

1. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
2. Jaillon,O., Aury,J.M., Brunet,F., Petit,J.L., Stange-Thomann,N., Mauceli,E., Bouneau,L., Fischer,C., Ozouf-Costaz,C., Bernot,A. *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.
3. International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

4. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
5. Wu,J.Q., Shteynberg,D., Arumugam,M., Gibbs,R.A. and Brent,M.R. (2004) Identification of rat genes by TWINSCAN gene prediction, RT–PCR, and direct sequencing. *Genome Res.*, **14**, 665–671.
6. Guigó,R., Dermitzakis,E.T., Agarwal,P., Ponting,C.P., Parra,G., Reymond,A., Abril,J.F., Keibler,E., Lyle,R., Ucla,C. *et al.* (2003) Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci. USA*, **100**, 1140–1145.
7. Dewey,C., Wu,J.Q., Cawley,S., Alexandersson,M., Gibbs,R. and Pachter,L. (2004) Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res.*, **14**, 661–664.
8. Parra,G., Agarwal,P., Abril,J.F., Wiehe,T., Fickett,J.W. and Guigo,R. (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
11. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
12. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, 856–875.
13. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene centered resources. *Nucleic Acids Res.*, **29**, 137–140.
14. Strausberg,R.L., Feingold,E.A., Klausner,R.D. and Collins,F.S. (1999) The Mammalian Gene Collection. *Science*, **286**, 455–457.
15. Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
16. Furey,T.S., Diekhans,M., Lu,Y., Graves,T.A., Oddy,L., Randall-Maher,J., Hillier,L.W., Wilson,R.K. and Haussler,D. (2004) Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. *Genome Res.*, **14**, 2034–2040.
17. Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
18. Reymond,A., Marigo,V., Yaylaoglu,M.B., Leoni,A., Ucla,C., Scamuffa,N., Caccioppoli,C., Dermitzakis,E.T., Lyle,R., Banfi,S. *et al.* (2002) Human chromosome 21 gene expression atlas in the mouse. *Nature*, **420**, 582–586.
19. Reymond,A., Camargo,A.A., Deutsch,S., Stevenson,B.J., Parmigiani,R.B., Ucla,C., Bettoni,F., Rossier,C., Lyle,R., Guipponi,M. *et al.* (2002) Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics*, **79**, 824–832.
20. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
21. Slavov,D., Hattori,M., Sakaki,Y., Rosenthal,A., Shimizu,N., Minoshima,S., Kudoh,J., Yaspo,M.L., Ramser,J., Reinhardt,R. *et al.* (2000) Criteria for gene identification and features of genome organization: analysis of 6.5 Mb of DNA sequence from human chromosome 21. *Gene*, **247**, 215–232.