

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

Robert Castelo^{1*}, Alexandre Reymond^{2,3}, Carine Wyss², Francisco Câmara¹, Genís Parra¹, Stylianos E. Antonarakis², Roderic Guigó¹, Eduardo Eyras¹

¹Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació Genòmica, E08003 Barcelona, Spain, ²Dept. of Genetic Medicine and Development, University of Geneva, Medical School and University Hospital of Geneva, CMU, 1, rue Michel Servet, 1211 Geneva, Switzerland, ³Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

*Correspondence should be addressed to: Psg. Marítim 37-49, E08003 Barcelona, Spain. E-mail: rcastelo@imim.es, Telephone: +34.932.240.884, Fax: +34.932.240.875.

The experimental verification by RT-PCR of the 50 out of the 311 putative novel human genes yielded 6 positive cases to which we shall refer by their identifier given by the comparative gene predictor SGP2 which includes the chromosome on which they map. Within these experimentally-verified predictions, three had hits with Pfam domains and four had matches to human ESTs verifying the exon-exon junctions with an HSP of at least 100bp and a percentage identity of at least 95%. All matched ESTs (82) are locus-specific, i.e., the locus corresponds to the best hit on the human

genome, and 98% of them (80) lack annotation of either cell type or developmental stage.

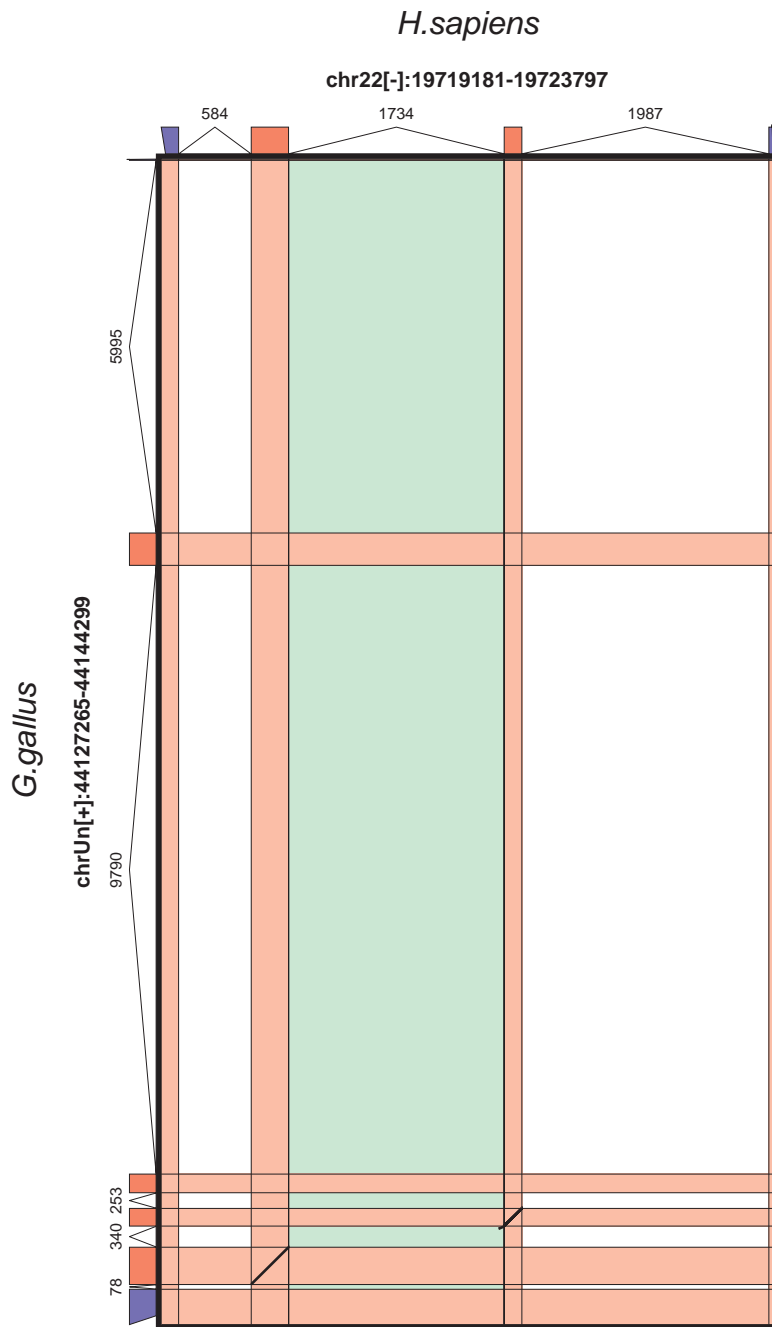
Two predictions are homologous to known human genes from different loci in the genome (see Table 3 in the main article). The first prediction (chr15_51, see Fig. S1) maps on HSA15 and encodes a protein similar to the one encoded by the HSA20 gene ELMO2 (NM_133171, NM_022086 and NM_182764), an activator of RAC1, a gene required for phagocytosis and cell migration(1). This prediction presents a hit to two Pfam domains (PF04006, PF02931), but no locus-specific ESTs. The second prediction, chr22_143 (see Fig. S2), is homologous to P2RXL1, a HSA22 gene encoding a P2X receptor(2). The prediction maps on the same chromosome, 100kb away of the P2RXL1 locus and on the opposite strand. The specificity of the primers designed to amplify this prediction, the uncovering of two locus-specific ESTs that further validate the predicted exon-intron structure, as well as a matching Pfam domain (PF00864) suggest that chr22_143 represents a genuine novel gene.

Two of the remaining four predictions have similarities with mouse RIKEN cDNAs(3) (see Table 3 in the main article). One of them (chr4_1746, see Fig. S3), amplified from human testis and fetal liver, retains only a very low sequence similarity to the mouse RIKEN sequence (see alignment in Fig. S4). The other prediction showing homology to a RIKEN cDNA, chr18_515 (Fig. S5), has a match with a Pfam domain-associated to the SKI/SNO proto-oncogene family. It was amplified only from human fetal brain tissue and has no EST matches (Tables 1, 2 and 3 in the main article). The RIKEN cDNA similar to this second prediction, however, shows its best alignment to the human genome on chromosome 15 and not 18, where chr18_525 is located. This together with the fact that the used primers were

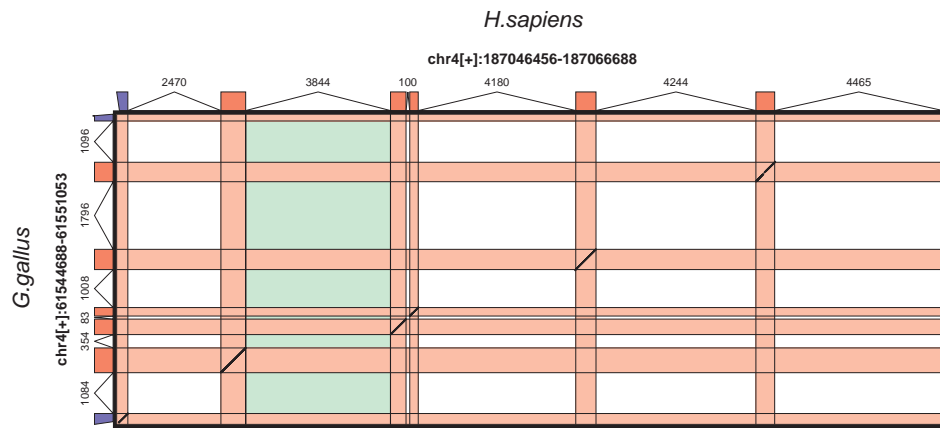
specific for chromosome 18 indicates that this prediction corresponds to a *bona fide* novel gene. One of the last two predictions (chr5_400, see Fig. S6) was only amplified from human stomach tissue (see Tables 1, 2 and 3 in the main article) and it maps in the locus of the SelP precursor mRNA that translates into a selenoprotein(4). This overlap with a known gene was overlooked due to the fact that the genomic locus of the SelP gene was not in the annotation databases used for filtering and the predicted gene had a frameshift (in order to skip the stop codon that codes for selenocysteine) that prevented the BLASTP search of matching the two proteins. We need to determine whether this prediction might correspond to a new transcript variant of the SelP gene. The last prediction (chr4_55, Fig. S7) is similar to a Tetraodon gene of unknown function, it has several specific EST matches and it amplified from multiple human tissues, strongly supporting its correspondence to a *bona fide* novel gene (Tables 1, 2 and 3 in the main article). This was the only positive common to the set of putative novel human genes obtained using the Tetraodon genome sequence.

Figures

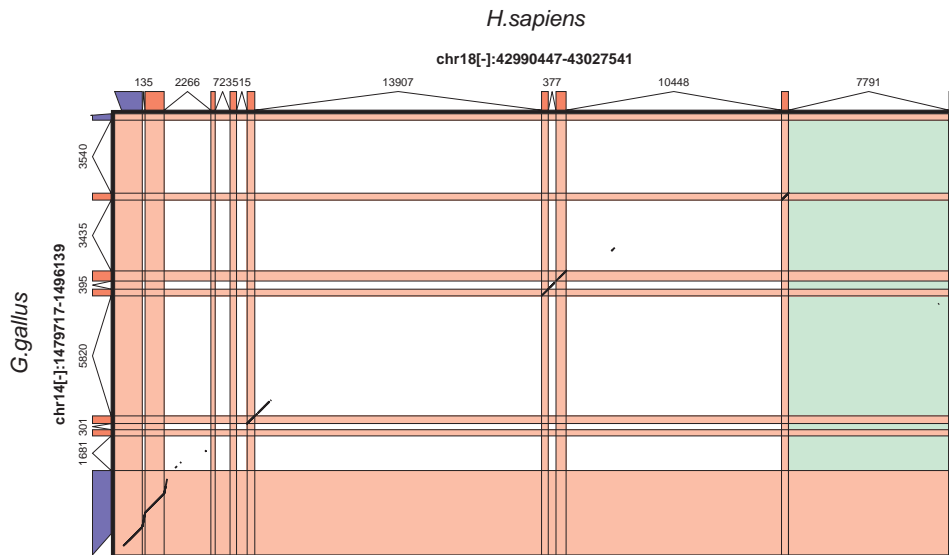
Each of the figures (except for Fig. S4) shows the TBLASTX alignments between the genomic sequence of the human predicted gene and a chicken homolog. We show the figures of the six experimentally verified cases. The exonic structures of the human and chicken genes are shown along the horizontal and vertical axes, respectively. Internal exons are in red and the first and terminal exons in blue. The exon coordinates are projected along the alignment space in light red, while the HSPs are plotted in black solid lines. The exon-exon junctions verified by RT-PCR are highlighted in light green. To produce a clearer picture, intron sizes have been scaled down proportionally. (their sizes in bp are indicated), whereas exons are drawn at their actual scale. These six figures have been produced using the `gff2aplot(5)` software.



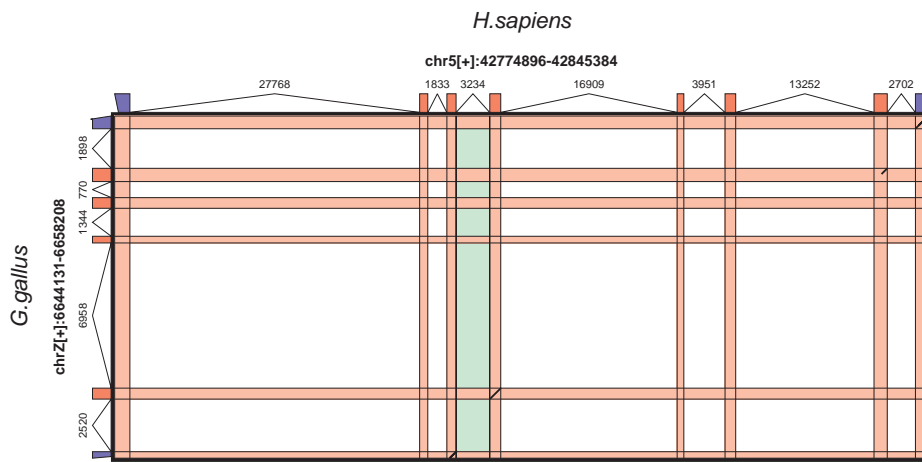
Supplementary Figure S2 Prediction chr22_143 is homologous to the P2RXL1 gene, that encodes a P2X receptor. It has a match to locus-specific ESTs and to a Pfam domain (PF00864).



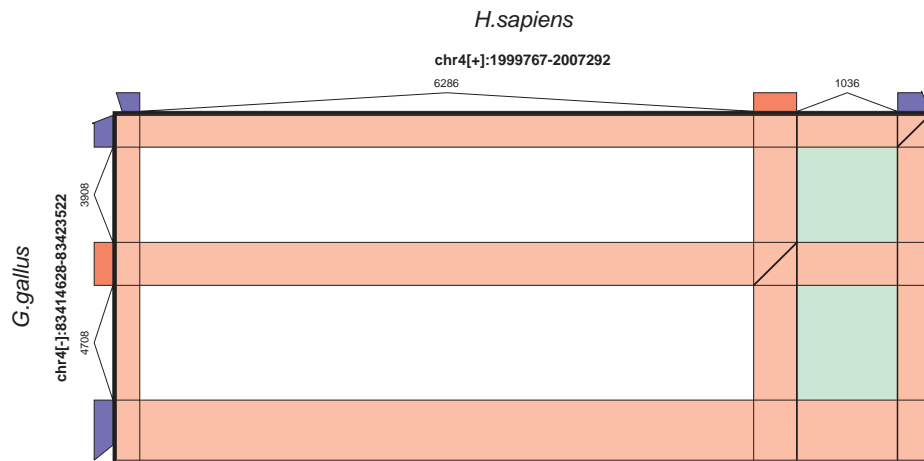
Supplementary Figure S3 Prediction chr4_1746 aligns with an homologous chicken gene mapping on GGA4. We amplified an exon-exon junction corresponding to this prediction from human testis and fetal liver. This prediction has 77.05% identity at the amino acid level with a mouse RIKEN cDNA (AK006501) of unknown function (see Supplementary Figure S7).



Supplementary Figure S5 The tested exon-exon junctions from predicton chr18_515 were amplified uniquely from human fetal brain tissue. This prediction has a match to a Pfam domain associated to the SKI/SNO proto-oncogene family, but no EST matches.



Supplementary Figure S6 Prediction chr5_400: It amplified only from human stomach tissue and maps in the locus of the SeIP precursor mRNA.



Supplementary Figure S7 Prediction chr4_55 was amplified from eight human tissues (see Tables S1 and S2). It present homologies to several specific EST matches and encodes a peptide similar to a Tetraodon predicted protein of unknown function.

References

1. Gumienny, T.L., Brugnera, E., Tosello-Trampont, A.C., Kinchen, J.M., Haney, L.B., Nishiwaki, K., Walk, S.F., Nemergut, M.E., Macara, I.G., Francis, R. *et al.* (2001) *Cell*, **107**, 27-41.
2. Urano, T., Nishimori, H., Han, H., Furuhata, T., Kimura, Y., Nakamura, Y. and Tokino, T. (1997) *Cancer Res*, **57**, 3281-3287.
3. Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. *et al.* (2002) *Nature*, **420**, 563-573.
4. Kryukov, G.V., Castellano, S., Novoselov, S.V., Lobanov, A.V., Zehtab, O., Guigo, R. and Gladyshev, V.N. (2003) *Science*, **300**, 1439-1443.
5. Abril, J.F., Guigo, R. and Wiehe, T. (2003) *Bioinformatics*, **19**, 2477-2479.