Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny

Nadav Ahituv[1,2*,] Shyam Prabhakar[1,2*,], Francis Poulin[1,3], Edward M. Rubin[1,2], Olivier Couronne[1,2]

1. Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
2. US DOE Joint Genome Institute, Walnut Creek, CA, USA
3. Present address: Department of Integrative Biology, University of California, Berkeley, CA, USA

* The first two authors contributed equally to the work.

To whom correspondence should be addressed:

Edward M. Rubin, Genomics Division, One Cyclotron Road, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. Email: emrubin@lbl.gov

Phone: (510) 486-5072, Fax: (510) 486-4229

ABSTRACT

Our inability to associate distant regulatory elements with the genes that they regulate has largely precluded their examination for sequence alterations contributing to human disease. One major obstacle is the large genomic space surrounding targeted genes in which such elements could potentially reside. In order to delineate gene regulatory boundaries we used whole-genome human-mouse-chicken (HMC) and human-mouse-frog (HMF) multiple alignments to compile conserved blocks of synteny (CBS), under the hypothesis that these blocks have been kept intact throughout evolution at least in part by the requirement of regulatory elements to stay linked to the genes that they regulate. A total of 2,116 and 1,942 CBS >200 kb were assembled for HMC and HMF respectively, encompassing 1.53 and 0.86 Gb of human sequence. To support the existence of complex long-range regulatory domains within these CBS we analyzed the prevalence and distribution of chromosomal aberrations leading to position effects (disruption of a gene's regulatory environment), observing a clear bias not only for mapping onto CBS but also for longer CBS size. Our results provide a genome wide dataset characterizing the regulatory domains of genes and the conserved regulatory elements within them.

INTRODUCTION

With the availability of the human and other vertebrate genomes, the annotation and analysis of distant regulatory elements using comparative genomics was greatly facilitated (1, 2). Studies using this approach suggest that cis-regulatory elements may lie in very distant regions from the genes that they regulate (3, 4), eluding to one of the difficulties in associating alterations within them to human disease. Primary insights into the ability of these alterations to cause human disease were obtained through chromosomal rearrangements causing position effects. Position effects lead to a phenotype similar to that resulting from mutations within the gene, thought to be brought about by removal of a gene's regulatory environment (5, 6), and thus provide evidence that disruption of distant regulatory architecture can lead to human disease. However, the association of human disease with nucleotide changes amongst these distant regulatory elements is hindered by the unavailability of a regulatory code and by the inability to link them to particular genes.

One mechanism by which evolutionary constraints against chromosomal breakage are thought to be maintained is the need for distant cis-regulatory elements to remain in the vicinity of the genes they act on. Based on this assumption, synteny blocks (chromosomal segments in which all sequences are in the same order and orientation in the species analyzed) can be used to delimit borders for distant cis-regulatory elements regulating a given gene, a strategy that has been minimally explored (3, 7, 8). To identify syntenic blocks on a whole-genome scale we generated multiple alignments of the human, mouse, and chicken genomes as well as alignments of the human, mouse, and frog genomes. We reasoned that these genomes would be the most suitable to carry out this analysis allowing adequate evolutionary divergence. Characterization of these conserved blocks of synteny revealed a decrease in gene density and increase in the density and evolutionary conservation of conserved non-coding sequences with block size. In order to validate the existence of distal regulatory networks within these blocks we assessed the prevalence and distribution of position effects within them.

RESULTS

To identify syntenic blocks on a whole-genome scale we generated multiple alignments of the human, mouse, and chicken genomes as well as alignments of the human, mouse, and frog genomes. Alignments were carried out by locally aligning the genomes to one another and then applying a computational algorithm to cluster all these alignments into an *n*-dimensional segmental map (materials and methods). HMC and HMF were found to have 2,116 and 1,942 conserved blocks of synteny (CBS; Supplementary Material, Table S1), the largest being 5.68 and 2.93 Mb respectively with a cumulative length in human of 1.53 and 0.86 Gb (Table 1).

Characterization of conserved non-coding sequences and genes within CBS

We then characterized these CBS in order to assess whether they have any unique sequence attributes. Analysis of their conserved non-coding sequences (CNSs) was carried out using Gumby (S. Prabhakar, manuscript in preparation), a tool for detecting statistically significant conserved regions in pairwise or multiple alignments of DNA sequences (materials and methods). Gumby identified a total of 37,191 CNSs for HMC and 9,884 CNSs for HMF, with a conserved Gumby *P*-value of < 0.01 [at this threshold Gumby assigns 3% of the human genome to human-mouse conserved regions compared to the 5% quoted in ref. (9)].

Analysis of the distribution of conserved non-coding sequences and genes within our CBS shows that CBS size is directly correlated with the median CNS density (8-fold difference between the shortest and longest CBS; Fig. 1A) and inversely correlated with gene density (decreasing from 0.9 genes per 100 kb in the shortest segments of both multiple alignments to 0.4 and 0.3 genes per 100 kb in the longest HMC and HMF CBS; Fig. 1A). We also found CBS size to be directly correlated with the degree of CNS conservation, with median Gumby *P*-value dropping from 10^{-7} and 10^{-6} for the smallest segments to 10^{-14} and 10^{-11} for the largest segments in HMC and HMF respectively (Fig. 1B). This indicates that longer CBS harbor fewer genes and denser evolutionarily conserved non-coding sequences. As an example of these trends, our analysis of chromosome 16 shows a long CBS covering 5.6 Mb with a gene and non-coding density per 100 kb of 0.5 and 7.5, compared to a shorter CBS covering 894 kb with a gene and non-coding density per 100 kb of 1.5 and 1.0 (Fig. 2).

To examine whether there is a functional uniqueness to the genes that are located near conserved non-coding sequences within the CBS we analyzed their biological processes and molecular functions using the GO database (10, 11). We identified the genes closest in distance to each CNS, and when many CNSs had the same closest gene, this gene was counted only once. Overall, we observed that the set of genes flanking these conserved non-coding sequences is enriched for transcription factors involved in developmental processes, with 40% (P-value = 10^{-17}) and 69% (P-value=10^{-19}) more such genes than expected for HMC and HMF respectively. This indicates that these clusters of highly conserved non-coding sequences tend to reside in the vicinity of developmental transcription factors and are likely to regulate them.

Mapping and distribution of position effects within CBS

In order to validate that our CBS harbor distant regulatory structure we analyzed the prevalence and distribution of position effects within them. We searched the literature for position effects leading to human disease where the regulatory elements were removed from the postulated regulated gene; for the 17 that fit these criteria (Table 2) we used the target gene as an anchor for our alignments. As control groups we chose the entire set of known genes in the human genome and the deleted subset of large-scale copy number polymorphisms (CNPs) (12), chromosomal deletions leading to no apparent phenotype, encompassing 44 alignable CNPs. Due to the incomplete nature of the frog genome, several of these regions were absent in the HMF alignments, so this analysis was performed using only HMC. In terms of prevalence we observed a skew where 88% (15/17) of position effects mapped to our CBS versus ~50% of CNPs and all known genes (Table 3). It is worth noting that the two genes that did not map to our >200 kb CBS are alpha-globin and beta-globin, most likely due to the extensive genomic variations between species in these regions (13). Interestingly, these are the only non-developmental genes in our position effect list and other than SOST they are the only genes that are not transcription factors.

We next analyzed the size of the CBS of only the ones that mapped within the HMC alignment. This examination revealed that while 80% (12/15) of the position effects fall within the top half of CBS in terms of length only 47.6% (10/21) of CNPs and 43.6% (4,598/10,546) of all known genes are in this upper 50th percentile of CBS (Fig. 3). Combined, these results confirm the existence of long-range regulatory elements within our CBS and suggest that disruption of long syntenic blocks is more likely to lead to a more striking and observable phenotype.

DISCUSSION

By examining multiple alignments between several vertebrates we characterized conserved blocks of synteny on a whole-genome scale. CBS size showed an inverse correlation with gene density and a direct correlation with the density and evolutionary conservation of non-coding sequences. These findings are in accordance with those from the chicken genome analysis (14) where it was observed that many of the human-chicken CNSs appear in clusters far away from genes. Combined, these observations suggest that longer regions of unbroken synteny are enriched for regulatory information.

Our analysis of the genes in the vicinity of these conserved non-coding sequences displayed enrichment for transcription factors involved in developmental processes. This is consistent with the characterization of human gene deserts (the 3% longest intergenic intervals in the human genome with the shortest one covering 640 kb), where stable gene deserts, deserts whose sequence harbors substantially greater human-chicken CNSs, were shown to be enriched for developmental transcription factors (15). In addition, a recent study using the *D. melanogaster* and *C. elegans* genomes (16) demonstrated that transcription factors and developmental genes are flanked by significantly more intergenic DNA than other genes with simpler functions. In our study, a similar theme was observed in vertebrates, in addition to the observation that these regions have more dense and evolutionary conserved non-coding sequences. This further supports the theory proposed in the metazoan study (16) that an expansion of intergenic regions occurred during evolution in order to accommodate numerous cis-regulatory elements. The drawback in this evolutionary scenario is the increased probability of a random chromosomal rearrangement event that would disrupt the cis-regulatory architecture and would lead to developmental defects. Thus, the complex architecture of regulatory domains may be one of the major constraints on chromosomal rearrangement during evolution.

Position effects are an ideal subset of chromosomal aberrations to study disruption of long-range regulatory domains. Using this subset we were able to validate that our CBS harbor distant regulatory architecture which when disrupted may lead to human disease. One such example is *SALL1*, a gene that maps to the longest CBS in our dataset, 5.6 Mb in size (Fig. 2). Mutations in *SALL1* lead to Townes-Brocks syndrome while a translocation in one patient ~180 kb telomeric to *SALL1* leads to the same syndrome (17) (Table 2). Analysis of this region in our CBS list (Supplementary Material, Table S1) suggests that removal of this 3.1 Mb region encompassing 241 CNSs (Fig. 2) is the cause for Townes-Brocks syndrome in this patient. In addition, transgenic mouse data indicate that a significant percentage of non-coding sequences conserved between human and fugu in this segment behave as enhancer elements with expression patterns recapitulating those of *SALL1* (Ahituv N., manuscript in preparation).

An important point regarding the disruption of regulatory architecture as a cause for human disease is that the phenotype caused by this disruption may only be a subset of the phenotype brought about by mutations in the coding region of the gene, thus making it difficult for clinicians to associate it to that disease/gene. One such example is the postulated *SHH* limb enhancer located 1Mb away from the gene. Mutations in the coding region of *SHH* lead to a large spectrum of phenotypes among which the most prominent is holoprosencephaly (18), while both a chromosomal breakpoint and single nucleotide changes within the limb enhancer are suggested to cause preaxial polydactyly (PPD) (19, 20). Analysis of both the gene and the limb enhancer in our data set show that they both map to the same CBS totaling 1.94 Mb in size (Table 2); moreover, additional position effects leading to holoprosencephaly (21) also map within this block. We could thus speculate that numerous genetically unaddressed phenotypes some even leading to embryonic lethality could be caused by disruption of distant regulatory elements of genes, which would likely also map to our CBS.

In general, using syntenic blocks to delineate boundaries of regulatory domains would seem an obvious approach when undertaking a comparative genomics endeavor, though it is not usually taken advantage of. The *SALL1* and *SHH* examples show that this approach provides one with the ability to obtain a better understanding of the boundaries of regulatory domains surrounding these genes and the CNSs within them. One could query a gene of interest for its CBS using our list (Supplementary Material, Table S1) and get a better sense of the domains and the conserved non-coding sequences within them. One major limitation to this method is when the gene of interest lies within blocks smaller than our 200 kb cutoff size, as we observed with the alpha- and beta-globin genes. A way around that is to use the UCSC browser chained BLASTZ alignments (11), with the limitations being that these are pairwise alignments that use less stringent filters, and consequently tolerate very large insertions and deletions. Another limitation in the usage of our CBS list is that HMC and HMF may not have enough evolutionary information to define a sufficiently precise cis-regulatory domain map, an obstacle which may be addressed in the future by the completion of several additional vertebrate genomes. In summary, our results provide a regulatory terrain for researchers around their gene of interest, highlighting evolutionary conserved non-coding sequences and delimiting their borders, in order to facilitate the search of regulatory mutations leading to human disease.

MATERIALS AND METHODS

Data source

Human (hg17, May 2004), mouse and chicken genomes and associated data (genes annotations, spliced EST, mRNA, repeats) were downloaded from the UCSC Genome website (11) (http://genome.ucsc.edu). The Xenopus Tropicalis v3.0 was downloaded from the JGI website (http://jgi.doe.gov). Gene Ontology (10) data was obtained from http://www.geneontology.org and http://genome.ucsc.edu. The human gene set and their GO data used in this work is from the 'Known Genes' set developed at UCSC and available at http://genome.ucsc.edu.

Segmental Map

The segmental homology map was computed using a clustering algorithm which takes BLASTZ (22) local alignment as an input. All pairwise alignments were merged into n-dimension anchors and then clustered by PARAGON (23). The n-dimensional segmental map problem was resolved in a graph-theoretic framework. Conserved BLASTZ anchors comprised the vertices of the graph and these were connected by edges if the distance between the anchors was less than 150 kb in all the aligned genomes. Each connected subgraph represents a conserved block of synteny. Human segments shorter than 200 kb were disregarded. We then realigned all the conserved blocks of synteny using the global aligner MLAGAN (24). We filtered out 3-way syntenic segments with a human-mouse non-coding nucleotide mismatch rate higher than two standard deviations above the whole-genome average, since these are likely to result from alignment artifacts and paralogy, rather than true orthology.

Identification of conserved regions.

MLAGAN alignments of synteny blocks were scanned for evolutionarily conserved regions using Gumby (v1.5; S. Prabhakar, manuscript in preparation). Gumby goes through the following three-step process to identify statistically significant conservation in the global alignment input: 1) non-coding regions in the alignment are used to estimate the local neutral mismatch rates between all pairs of aligned sequences (25). The rates are used to derive a log-likelihood scoring scheme for slow versus neutral evolution, where the slow rate is set to 2/3 times the neutral rate. 2) each alignment position is then assigned a conservation score using a phylogenetically weighted sum-of-pairs scheme; 3) Conserved regions of any length are identified as alignment blocks with a high cumulative conservation score, and assigned *P*-values using Karlin-Altschul statistics (26). We used a threshold *P*-value of 0.01 in a baseline human sequence length of 100 kb. Transcribed sequences in the conserved set were filtered out using known genes, spliced ESTs, and mRNA annotations obtained at the UCSC genome browser.

Analysis of gene functionality

Each CNS was coupled to the closest gene. The distribution of GO terms of these genes was then compared to the overall GO term distribution of all the genes in the human genome.

Analysis of chromosomal aberrations

We searched the literature using PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=File&DB=pubmed) and OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=omim) for position effects leading to human disease where the regulatory elements were removed from the postulated regulated gene and that gene was used as an anchor for our alignments. As control groups we used all ‘Known Genes’ from the UCSC Genome website (11) and large-scale copy number polymorphisms (CNP) corresponding to deletions (loss) (12), under the assumption that deletions have a greater potential to disrupt regulation networks. We were only able to map 44 of these CNP deletions to the human May 2004 freeze using the coordinate conversion webtool at http://genome.ucsc.edu. and used those as our dataset.

ACKNOWLEDGMENTS

We thank members of the Rubin laboratory for helpful comments and suggestions. F.P. was supported by a fellowship from the Canadian Institutes for Health Research. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36 and was supported by NIH-NHLBI U1HL66681B.

CONFLICT OF INTEREST STATEMENT

None declared.

Table 1. Characteristics of the human-mouse-chicken and the human-mouse-frog CBS.

	Human-Mouse-Chicken	**Human-Mouse-Frog**
number of segments	2,116	1,942
length on human (Gb)	1.53	0.86
N50 length (Mb)	1.02	0.48
max length (Mb)	5.68	2.93

Table 2. Characterization of position effects within human-mouse-chicken CBS.

Gene	Disease	Synteny size	Human-mouse-chicken positions on human hg17	Reference
FOXC1	Glaucoma/autosomal dominant iridogoniodysgenesis	1,848,302	chr6:249245-2097547	(27)
FOXC2	Lymphedema-distichiasis	982,755	chr16:84885116-85867871	(28)
FOXL2	Blepharophimosis/ptosis/epicanthus inversus syndrome	414,470	chr3:140135506-140549976	(29)
GLI3	Greig cephalopolysyndactyly	2,775,157	chr7:39760342-42535499	(30)
HBA	α-thalassemia	_	_	(31)
HBB	γβ-thalassemia	_	_	(32)
HOXD	mesomelic dysplasia and vertebral defects	2,444,711	chr2:175489049-177933760	(33)
MAF	Cataract, ocular anterior segment dysgenesis, coloboma	1,529,488	chr16:76704419-78233907	(34)
PAX6	Aniridia	2,624,266	chr11:30308118-32932384	(35)
PITX2	Rieger syndrome	1,976,829	chr4:111754180-113731009	(36)
POU3F4	X-linked deafness	536,809	chrX:82130470-82667279	(37)
SALL1	Townes-Brocks syndrome	5,681,989	chr16:48615828-54297817	(17)
SHH	Holoprosencephaly	1,946,088	chr7:154763593-156709681	(21)
SIX3	Holoprosencephaly	2,363,759	chr2:43430275-45794034	(38)
SOST	Van Buchem disease	339,143	chr17:38973638-39312781	(39)
SOX9	Campomelic displasia	3,898,059	chr17:65045536-68943595	(40)
TWIST	Saethre-Chotzen syndrome	2,529,763	chr7:16989836-19519599	(41)

Table 3. Prevalence of known genes, copy-number polymorphism (CNP) deletions, and position effects in human-mouse-chicken CBSs.

	Known genes	CNP deletions	Position effects
Total	20,399	44	17
Map on HMC*	10,546	21	15
Ratio	51.7%	47.7%	88.2%

* Human-Mouse-Chicken

FIGURE LEGENDS

Figure 1: CBS size is directly correlated with CNS density and evolutionary conservation and inversely correlated with gene density in human-mouse-chicken and human-mouse-frog multiple alignments. (A) Median CNS and gene density compared to CBS size (B) Median Gumby *P*-value of the evolutionary conservation of CNS compared to CBS size.

Figure 2: Human chromosome 16 as an example of CBS trends. (A) Human-mouse-chicken CBSs (colors of blocks indicate the different chicken and mouse chromosomes that the sequence is derived from). (B) Normalized density of human-mouse-chicken conserved non-coding sequences, densities are normalized so that the darkest shade in each track denotes 3.5 times the genomic average. (C) Conservation plot of two human-mouse-chicken syntenic segments. Conserved regions with a Gumby *P*-value < 0.01 are depicted as blue (exonic) and magenta (non-exonic) bars (bar height is directly correlated with evolutionary conservation), with the gene structure shown bellow them. The longer CBS is 5.6Mb in human, containing 428 human-mouse-chicken CNSs and 27 genes including *SALL1*, with a non-coding and gene density per 100 kb of 7.5 and 0.5 respectively. The shorter CBS is 894kb long in human, contains 9 human-mouse-chicken CNSs and 13 genes, giving a non-coding and genes density per 100 kb of 1.0 and 1.5 respectively. The red arrow depicts the approximate region of the chromosomal translocation leading to Townes-Brocks syndrome in one patient [16].

Figure 3: Over-representation of position effect genes compared to deleted copy number polymorphisms (CNPs) and known genes in human-mouse-chicken CBS (blue = all genes, purple = deleted CNPs, red = position effects). The Y-axis represents the percentage of blocks proportional to CBS size, which is the X-axis.

References

1. Boffelli, D., Nobrega, M.A. and Rubin, E.M. (2004) Comparative genomics at the vertebrate extremes. *Nat Rev Genet,* **5,** 456-65.

2. Dermitzakis, E.T., Reymond, A. and Antonarakis, S.E. (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet,* **6,** 151-7.

3. Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science,* **302,** 413.

4. Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. and Shiroishi, T. (2005) Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development,* **132,** 797-803.

5. Ahituv, N., Rubin, E.M. and Nobrega, M.A. (2004) Exploiting human--fish genome comparisons for deciphering gene regulation. *Hum Mol Genet,* **13,** R261-6.

6. Kleinjan, D.A. and van Heyningen, V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet,* **76,** 8-32. Epub 2004 Nov 17.

7. Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R.J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K.C., McMorrow, T. *et al.* (2001) Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum Mol Genet,* **10,** 371-82.

8. Goode, D.K., Snell, P., Smith, S.F., Cooke, J.E. and Elgar, G. (2005) Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics,* **3,** 3.

9. Waterston, R.H. and Lindblad-Toh, K. and Birney, E. and Rogers, J. and Abril, J.F. and Agarwal, P. and Agarwala, R. and Ainscough, R. and Alexandersson, M. and An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature,* **420,** 520-62.

10. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* **25,** 25-9.

11. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res,* **31,** 51-4.

12. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science,* **305,** 525-8.

13. Hardison, R. (1998) Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *J Exp Biol,* **201,** 1099-117.

14. Hillier, L.W. and Miller, W. and Birney, E. and Warren, W. and Hardison, R.C. and Ponting, C.P. and Bork, P. and Burt, D.W. and Groenen, M.A. and Delany, M.E. *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature,* **432,** 695-716.

15. Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W. and Stubbs, L. (2005) Evolution and functional classification of vertebrate gene deserts. *Genome Res,* **15,** 137-45. Epub 2004 Dec 8.

16. Nelson, C.E., Hersh, B.M. and Carroll, S.B. (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol,* **5,** R25. Epub 2004 Mar 15.

17. Marlin, S., Blanchard, S., Slim, R., Lacombe, D., Denoyelle, F., Alessandri, J.L., Calzolari, E., Drouin-Garraud, V., Ferraz, F.G., Fourmaintraux, A. *et al.* (1999) Townes-Brocks syndrome: detection of a SALL1 mutation hot spot and evidence for a position effect in one patient. *Hum Mutat,* **14,** 377-86.

18. Wallis, D. and Muenke, M. (2000) Mutations in holoprosencephaly. *Hum Mutat,* **16,** 99-108.

19. Lettice, L.A., Horikoshi, T., Heaney, S.J., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N. *et al.* (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci U S A,* **99,** 7548-53.

20. Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. and de Graaff, E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet,* **12,** 1725-35.

21. Roessler, E., Ward, D.E., Gaudenz, K., Belloni, E., Scherer, S.W., Donnai, D., Siegel-Bartelt, J., Tsui, L.C. and Muenke, M. (1997) Cytogenetic rearrangements involving the loss of the Sonic Hedgehog gene at 7q36 cause holoprosencephaly. *Hum Genet,* **100,** 172-81.

22. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res,* **13,** 103-7.

23. Schmutz, J., Martin, J., Terry, A., Couronne, O., Grimwood, J., Lowry, S., Gordon, L.A., Scott, D., Xie, G., Huang, W. *et al.* (2004) The complete sequence of human chromosome 5. *Nature,* **In press**.

24. Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A. and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res,* **13,** 721-31. Epub 2003 Mar 12.

25. Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S. and Sidow, A. (2004) Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res,* **14,** 539-48.

26. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A,* **87,** 2264-8.

27. Davies, A.F., Mirza, G., Flinter, F. and Ragoussis, J. (1999) An interstitial deletion of 6p24-p25 proximal to the FKHL7 locus and including AP-2alpha that affects anterior eye chamber development. *J Med Genet,* **36,** 708-10.

28. Fang, J., Dagenais, S.L., Erickson, R.P., Arlt, M.F., Glynn, M.W., Gorski, J.L., Seaver, L.H. and Glover, T.W. (2000) Mutations in FOXC2 (MFH-1), a forkhead family transcription factor, are responsible for the hereditary lymphedema-distichiasis syndrome. *Am J Hum Genet,* **67,** 1382-8. Epub 2000 Nov 8.

29. Crisponi, L., Uda, M., Deiana, M., Loi, A., Nagaraja, R., Chiappe, F., Schlessinger, D., Cao, A. and Pilia, G. (2004) FOXL2 inactivation by a translocation 171 kb away: analysis of 500 kb of chromosome 3 for candidate long-range regulatory sequences. *Genomics,* **83,** 757-64.

30. Vortkamp, A., Gessler, M. and Grzeschik, K.H. (1991) GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. *Nature,* **352,** 539-40.

31. Barbour, V.M., Tufarelli, C., Sharpe, J.A., Smith, Z.E., Ayyub, H., Heinlein, C.A., Sloane-Stanley, J., Indrak, K., Wood, W.G. and Higgs, D.R. (2000) alpha-thalassemia resulting from a negative chromosomal position effect. *Blood,* **96,** 800-7.

32. Kioussis, D., Vanin, E., deLange, T., Flavell, R.A. and Grosveld, F.G. (1983) Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature,* **306,** 662-6.

33. Spitz, F., Montavon, T., Monso-Hinard, C., Morris, M., Ventruto, M.L., Antonarakis, S., Ventruto, V. and Duboule, D. (2002) A t(2;8) balanced translocation with breakpoints near the human HOXD complex causes mesomelic dysplasia and vertebral defects. *Genomics,* **79,** 493-8.

34. Jamieson, R.V., Perveen, R., Kerr, B., Carette, M., Yardley, J., Heon, E., Wirth, M.G., van Heyningen, V., Donnai, D., Munier, F. *et al.* (2002) Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. *Hum Mol Genet,* **11,** 33-42.

35. Fantes, J., Redeker, B., Breen, M., Boyle, S., Brown, J., Fletcher, J., Jones, S., Bickmore, W., Fukushima, Y., Mannens, M. *et al.* (1995) Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Hum Mol Genet,* **4,** 415-22.

36. Flomen, R.H., Vatcheva, R., Gorman, P.A., Baptista, P.R., Groet, J., Barisic, I., Ligutic, I. and Nizetic, D. (1998) Construction and analysis of a sequence-ready map in 4q25: Rieger syndrome can be caused by haploinsufficiency of RIEG, but also by chromosome breaks approximately 90 kb upstream of this gene. *Genomics,* **47,** 409-13.

37. de Kok, Y.J., Vossenaar, E.R., Cremers, C.W., Dahl, N., Laporte, J., Hu, L.J., Lacombe, D., Fischel-Ghodsian, N., Friedman, R.A., Parnes, L.S. *et al.* (1996) Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. *Hum Mol Genet,* **5,** 1229-35.

38. Wallis, D.E., Roessler, E., Hehr, U., Nanni, L., Wiltshire, T., Richieri-Costa, A., Gillessen-Kaesbach, G., Zackai, E.H., Rommens, J. and Muenke, M. (1999) Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. *Nat Genet,* **22,** 196-8.

39. Balemans, W., Patel, N., Ebeling, M., Van Hul, E., Wuyts, W., Lacza, C., Dioszegi, M., Dikkers, F.G., Hildering, P., Willems, P.J. *et al.* (2002) Identification of a 52 kb deletion downstream of the SOST gene in patients with van Buchem disease. *J Med Genet,* **39,** 91-7.

40. Wirth, J., Wagner, T., Meyer, J., Pfeiffer, R.A., Tietze, H.U., Schempp, W. and Scherer, G. (1996) Translocation breakpoints in three patients with campomelic dysplasia and autosomal sex reversal map more than 130 kb from SOX9. *Hum Genet,* **97,** 186-93.

41. Rose, C.S., Patel, P., Reardon, W., Malcolm, S. and Winter, R.M. (1997) The TWIST gene, although not disrupted in Saethre-Chotzen patients with apparently balanced translocations of 7p21, is mutated in familial and sporadic cases. *Hum Mol Genet,* **6,** 1369-73.

fig. 1

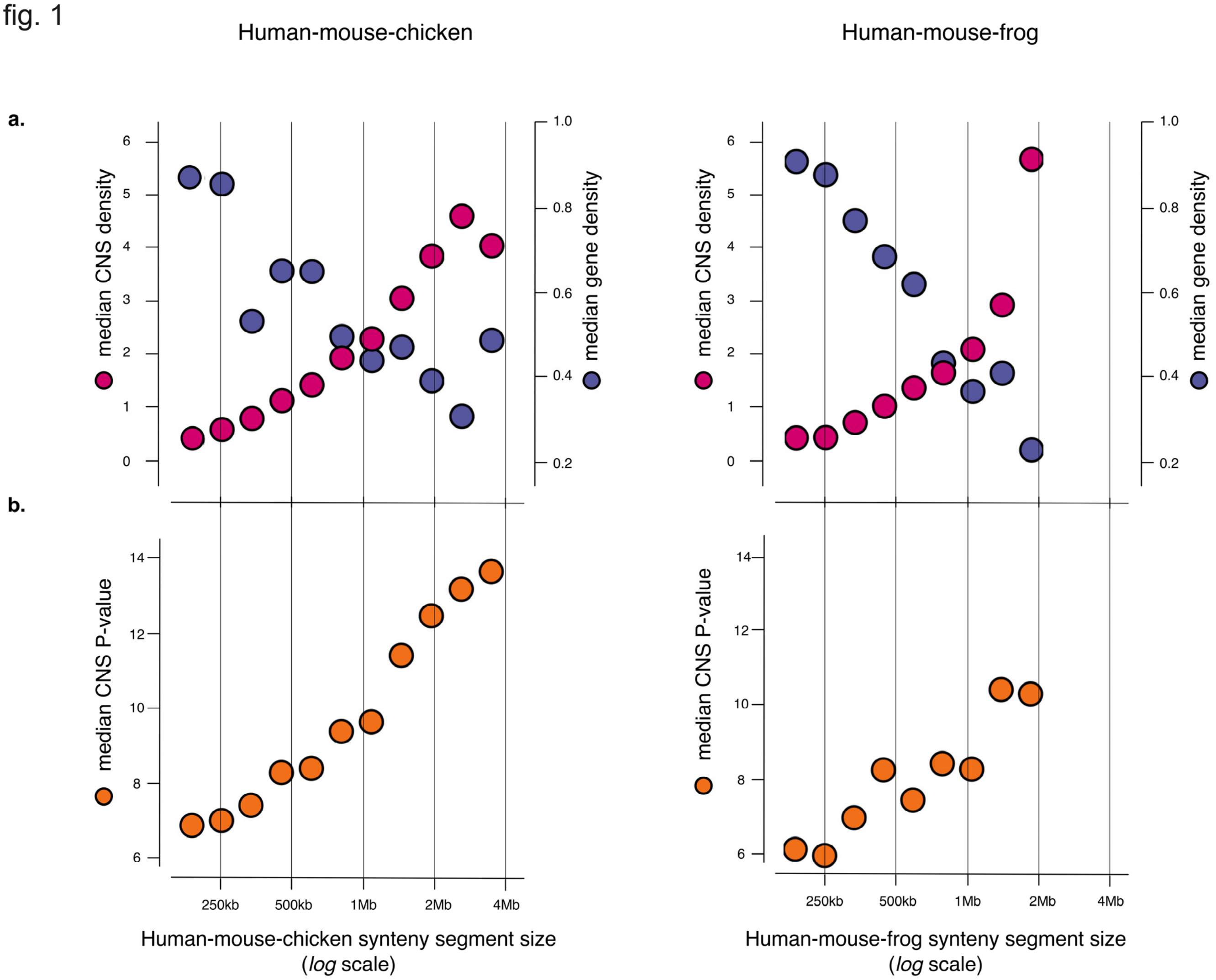

Human-mouse-chicken
Human-mouse-frog
a.
b.
median CNS density
median gene density
0
1
2
3
4
5
6
0.2
0.4
0.6
0.8
1.0
median CNS P-value
6
8
10
12
14
250kb
500kb
1Mb
2Mb
4Mb
Human-mouse-chicken synteny segment size
(*log* scale)
Human-mouse-frog synteny segment size
(*log* scale)

fig.2

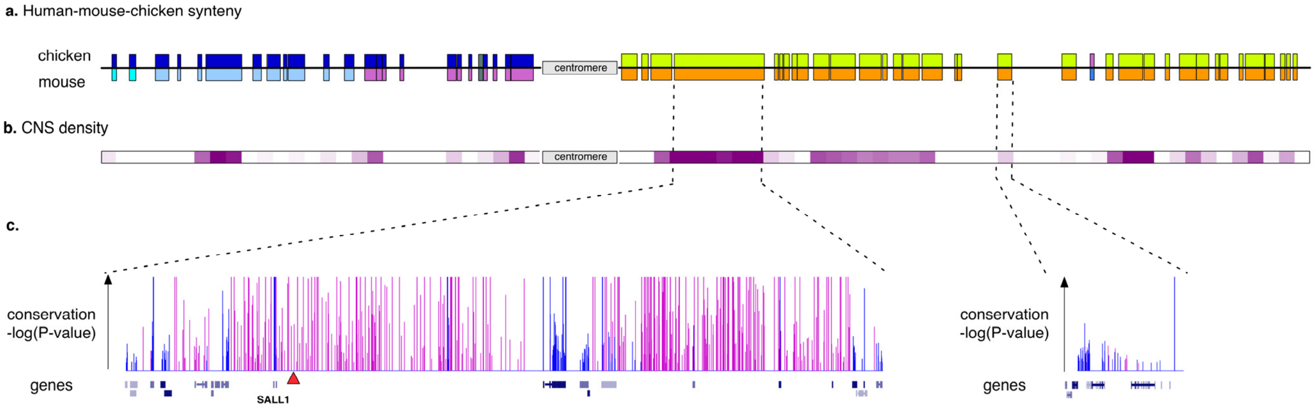
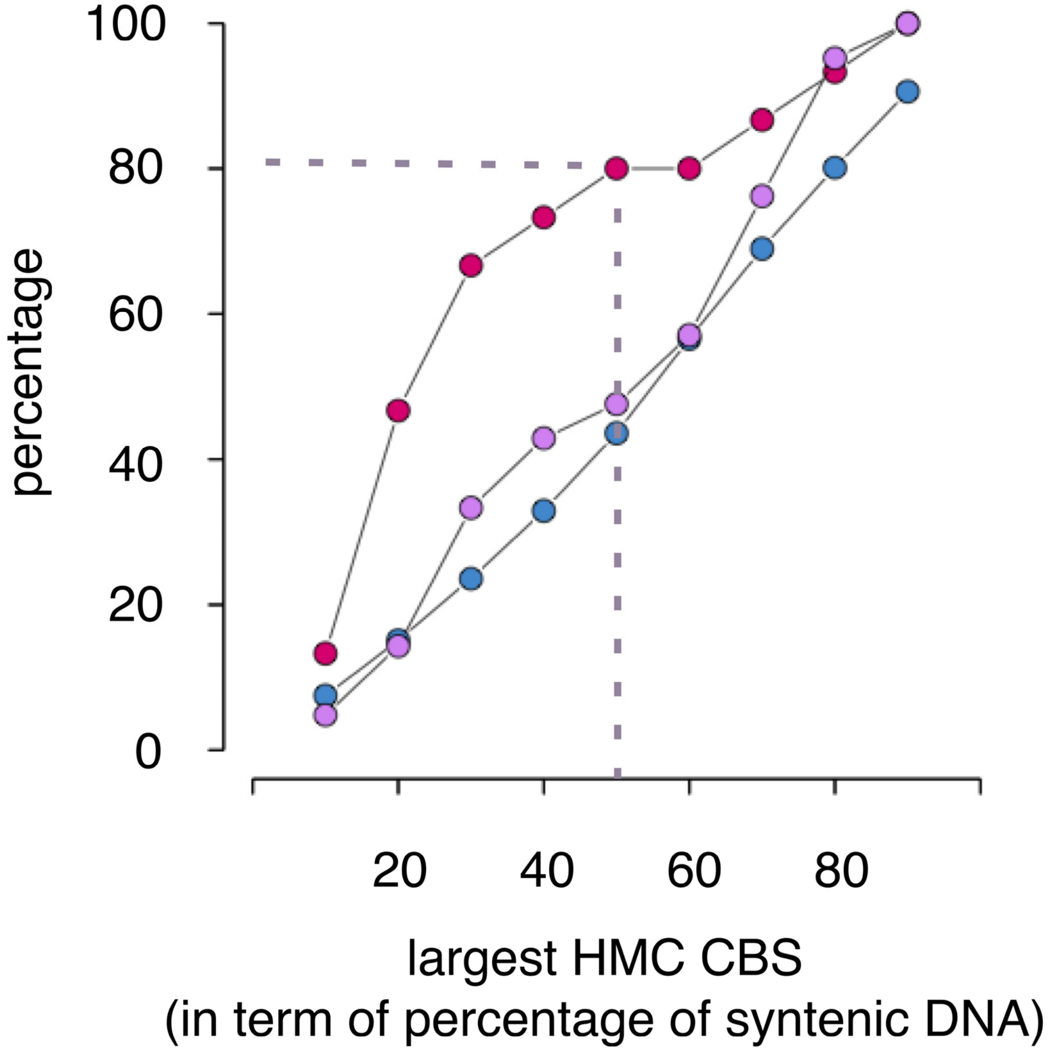

a. Human-mouse-chicken synteny
chicken
mouse
centromere
b. CNS density
centromere
c.
conservation
-log(P-value)
genes
SALL1
conservation
-log(P-value)
genes

fig. 3


100
80
60
40
20
0
percentage
20
40
60
80
largest HMC CBS
(in term of percentage of syntenic DNA)

ABBREVIATIONS

HMC= human-mouse-chicken HMC

HMF= human-mouse-frog

CBS= conserved blocks of synteny

CNSs= conserved non-coding sequences

Mb= Mega base

kb= kilo base

CNPs= copy number polymorphisms

PPD= preaxial polydactyly

ESTs= expressed sequence tags