VISTA - computational tools for comparative genomics

Kelly A. Frazer[1], Lior Pachter[2,3], Alexander Poliakov[3], Edward M. Rubin[3,4], and Inna Dubchak[3,4,*]

[1]Perlegen Sciences, Inc., 2021 Stierlin Ct., Mountain View, CA 94043
[2]Department of Mathematics, UC Berkeley, Berkeley, CA, 94720
[3]Genomics Division, Lawrence Berkeley National Laboratory, MS 84-171, Berkeley, CA 94720
[4]Department of Energy Joint Genome Institute, 2800 Mitchell Ave, Walnut Creek, CA 94598

*To whom correspondence should be addressed. Tel: (510) 495-2419; Fax: (510) 486-5614; Email: ildubchak@lbl.gov

Abstract

Comparison of DNA sequences from different species is a fundamental method for identifying functional elements in genomes. Here we describe the VISTA family of tools created to assist biologists in carrying out this task. Our first VISTA server at http://www-gsd.lbl.gov/VISTA / was launched in the summer of 2000 and was designed to align long genomic sequences and visualize these alignments with associated functional annotations. Currently the VISTA site includes multiple comparative genomics tools and provides users with rich capabilities to browse pre-computed whole-genome alignments of large vertebrate genomes and other groups of organisms with VISTA Browser, submit their own sequences of interest to several VISTA servers for various types of comparative analysis, and obtain detailed comparative analysis results for a set of cardiovascular genes. We illustrate capabilities of the VISTA site by the analysis of a 180 kilobase (kb) interval on human chromosome 5 that encodes for the kinesin family member 3A (*KIF3A*) protein.

Introduction

In light of the increasing amount of available genomic sequences from multiple species, the need for comparative genomics tools to functionally annotate this DNA sequence is also growing. These tools require efficient alignment algorithms as well as easy-to-interpret visualization strategies for investigating megabases of genomic intervals and whole genome assemblies. Although many individual programs have been developed separately for alignment and visualization, few services have attempted to integrate the two, and with the exception of the VISTA site and the PipMaker suite of tools (1-3) there have been few examples of continuous extensive development of web accessible software packages.

Our VISTA family of tools (4-6) is based on global alignment strategies and a curve-based visualization technique for rapid identification of conserved sequences in long alignments. Unlike other existing tools at the time of starting the VISTA project, the AVID alignment program allowed for real time global alignment of megabase-long sequences and the accompanying visualization program provided an easy method for the visual and computational analysis of conservation. This approach was extended to the pair wise and three-way alignment of whole genome assemblies by adding a mapping component as a first step before global alignment of putative orthologous regions of two species (7,8). This method is also used for aligning individual sequences against whole genome assemblies of several species. Improved prediction of functional signals, such as transcription factor binding sites, was introduced by taking into consideration conservation among species, and this feature also became available as a part of the VISTA family of tools (9). All VISTA tools use a standard platform of software for the analysis of conservation and visualization making is easy to compare results from different applications.

VISTA is a result of close collaboration of biologists, mathematicians, and computer scientists, and has been widely used by the biological community. A number of biological studies have utilized VISTA to answer various questions, from comparing genes from the same gene families (10,11), to discovering functional non-coding elements (12,13) and funding pattern of conservation on a whole-genome scale (14,15).

As we have mentioned, the VISTA system is fundamentally based on global alignments, and this should be contrasted with the PipMaker tools, which are based on local alignment strategies. A comparative review of the alignment and visualization features of PipMaker and VISTA has been recently puiblished (16). In addition, a recent paper (17) attempts to carefully analyze the benefits and drawbacks of different alignment methods and programs. It is important to note, that as alignment algorithms become more sophisticated, it is becoming harder to distinguish between local and global alignment tools. For example, a chaining option for BLASTZ (3) allows for the extraction of global alignments from BLASTZ local alignments, and similarly Shuffle-LAGAN (18) and MAVID (19) being global aligners explicitly deal with rearrangements between sequences.

VISTA Suite of Tools

The web page http://www-gsd.lbl.gov/vista/ serves as a portal for access to the suite of VISTA tools.

One of them is VISTA Browser that allows users to view pre-computed whole genome alignments of many species. There are three VISTA servers: GenomeVISTA, mVISTA, and rVISTA, which allow the user to submit DNA sequences for analysis. For GenomeVISTA (7) the user submits a single sequence (draft or finished) which is compared with publicly available completed whole genome assemblies. mVISTA (4,6) is the original program designed for comparison of orthologous sequences of different species. rVISTA (9) combines a transcription factor binding sites database search (20) with a comparative sequence analysis. The Phylo-VISTA program, a new member of the VISTA family of tools, allows a user to visualize submitted multiple sequence alignment data while taking the phylogenetic relationships between sequences into account (21). The VISTA web site also provides access to the comparative analyses of the set of cardiovascular genes, studied by the Berkeley Program for Genomic Applications (PGA).

VISTA pages provide extensive help on selecting a type of analysis, finding optimal parameters for a particular project, and navigating the web site.

VISTA Browser for pre-computed pairwise and multiple whole-genome alignments.

We have developed an automatic computational scheme for the alignment and analysis of conservation between large vertebrate genomes, which originally was applied to the comparative study of the human and mouse genomes (7,22). Our method uses the BLAT (23) local alignment program to find anchors on the base genome to identify regions of possible homology for a query sequence. These regions are post processed to find the best candidates that are then globally aligned using the AVID (6) or LAGAN (24) global alignment programs. When the rat genome assembly became available, the method was expanded to the comparison of three genomes, where the global alignment stage was accomplished with the MLAGAN multiple alignment program (8,15). Details on the strategy have been recently published elsewhere (7,8). The resulting whole-genome alignments exhibit a high degree of sensitivity, covering more than 90% of known coding exons in the human genome. More details about validation of our alignments and comparison to other methods are in (7,8). It is important to note that whole genome alignment is an ongoing area of active research (3, 19, 25, 26) and reference therein) and the alignment tools used in the VISTA servers are undergoing constant development, and the alignments continue to be analyzed accompanied by a methodical analysis. Although we focus on biological applications of VISTA in this paper, the VISTA browser has proved to be extremely useful as a tool for comparing and contrasting alignments.

When new genome assemblies become available they are aligned to previously available genomes in timely manner. Currently our site provides access to multiple human-mouse-rat alignment, pairwise alignments of the human genome with the chicken and chimpanzee assemblies, *D. melanogaster* with *D. pseudoobscura*, *C. elegans* with *C. briggsae*, and alignments of several plant genomes.

Visualization of aligned genome sequences

There are two schemes of visual data presentation on the whole genome scale available for the user – the VISTA Browser, the VISTA track on the mirrored UCSC genome browser.

VISTA Browser is a Java applet, very efficient for interactively visualizing results of comparative sequence analysis in the VISTA format on the scale of whole chromosomes along with annotations. The user may select any genome as the reference or base, and display the level of conservation between this reference and the sequences of another species in a particular interval. Conserved segments with percent identity X and length Y are defined to be regions of the alignment in which every contiguous subsegment of length Y in the base sequence was at least X% identical to its paired sequence. A user can use default values for conservation cutoffs (X% over Y bp), or specify them. These regions are highlighted under the curve, with different colors used for coding and noncoding sequences. The browser has a number of options, such as zoom, extraction of a region to be displayed, user-defined parameters for conservation level, and options for selecting sequence elements to study. VISTA track accessible through the VISTA Browser displays results of our comparative analysis in the context of the whole human genome annotation on the mirrored UCSC Human Genome Browser (27). VISTA track dynamically creates VISTA plots for each defined region and unlike VISTA Browser displays up to multiple individual plots if there is an overlap in alignments.

VISTA Browser and VISTA tracks are linked to the Text Browser that allows a user to examine detailed information about each sequence aligned to the selected region on the base genome. For each region, information such as exact locations of alignments on both genomes, the sequences, alignments, and coordinates of conserved regions are easily retrieved. Text Browser also gives access to rVISTA to obtain a prediction of potential transcription factor binding sites for any region of a base genome (see detailed description of rVISTA below).

In addition to alignments of wole genome assemblies, VISTA Browser provides access to multiple alignments of orthologous sequences of different species for genomic intervals containing cardioivascular genes currently under investigation in the Berkeley PGA program (28).

VISTA Browser annotation of the *KIF3A* on human chromosome 5q31

Kinesin family member 3A (*KIF3A*), is expressed in the kidney and photoreceptor cells where it is required for the proper formation and maintenance of cilia. Tissue specific inactivation of *KIF3A* in the kidneys of mice causes polycystic kidney disease (29) and inactivation in photoreceptor cells leads to cell death as found in retinitis pigmentosa (30). Here we use the VISTA Browser to analyse the 180 kb interval on human chromosome 5 (5q31) surrounding *KIF3A* to identify conserved non-coding sequences which may potentially regulate its expression. In figure 1a the pre-computed alignments of the human, mouse, and rat sequences for the *KIF3A* interval are visualized in the VISTA Browser. In addition to encoding for *KIF3A*, the 180 kb interval contains the 3' end of *RAD50* (the protein product is required for repair of double stranded breaks) and the entire coding sequences of two cytokines; interleukin 4 (*Il-4*) and interleukin 13 (*Il-13*). Using the default parameters for defining a conserved element (70% identity over 100 bp length) there are 125 elements in the 180 kb interval that are evolutionarily conserved in all three species of which 36 are coding and 89 are non-coding sequences. The interval located immediately downstream of *KIF3A* contains several conserved non-coding elements, and thus is a reasonable candidate region for regulating the tissue specific expression of the gene. To allow for a biologist to easily design experiments for testing whether or not the elements in this interval are involved in regulating the expression of *KIF3A* the VISTA Browser has a function that generates a list of the details of the conserved sequences (fig. 1b). The list contains the positions, lengths and percent identities, and whether or not the element is coding or non-coding. Equally important for prioritizing conserved non-coding sequences for functional studies is the ability to determine how the boundaries of these elements change under different thresholds of conservation. As shown in figure 1c the number and location of elements considered evolutionarily conserved in the interval downstream of *KIF3A* changes dramatically as the percent identity and/or length thresholds are altered.

VISTA servers for comparative analysis of user submitted sequences.

GenomeVISTA

Genome VISTA is an automatic server that allows the user to find candidate orthologous regions for a draft or finished DNA sequence from one species on the base genome of a second species, and provides detailed comparative analysis. You can currently align your sequence to the following base genomes: human, mouse, rat, *D. melanogaster, C.elegans*, *A. thaliana,* rice. We are constantly adding new base genomes to the server when their assemblies become available. Genome VISTA uses the same computational strategy as was used for the alignment of whole genome assemblies, where query sequence contigs were anchored on the base genome by local alignment matches (23) and then globally aligned to candidate regions with the AVID program (6,7).

A sequence up to 300 kilobases long can be submitted by pasting it into a window in plain FASTA format, by uploading a FASTA file from your computer, or by providing a GenBank accession number to the server. After submitting the sequence you immediately receive a link to the computation results. The resulting alignments of the query sequence against the base genome and detailed comparative analysis of conservation can be viewed via VISTA Browser and Text Browser. When two or more high scoring alignments are obtained for the query sequences and the base genome sequence, the results for all alignments are provided to a user in the Text Browser. For each alignment a link to rVISTA is also provided.

Use of Genome VISTA to annotate the *KIF3A* interval

It is well established that the human and dog genomes have a higher level of sequence similarity with each other than either one has with the mouse genome (5,31). Thus, the landscape of conservation observed in the pair-wise comparison of orthologous human and dog DNA sequences can be quite different than that observed in the pair-wise comparison of orthologous human and mouse DNA sequences. Here, we used Genome VISTA to align the orthologous dog sequence to the human 5q31 interval containing *KIF3A* (Fig 2). In the 180 kb interval humans and dogs have 362 elements conserved at the VISTA default thresholds of conservation (70% identify over 100 bp), in contrast to 150 elements between humans and mice and 137 elements between humans and rats As it was shown earlier (5), more stringent thresholds of conservation are required for the dog/human comparison. VISTA analysis revealed that some of the conserved non-coding elements are uniquely present between humans and only one of the three species (dogs, mice, rats) whereas other elements are conserved in all four species. One hypothesis is that some of the non-coding sequences conserved in a limited number of mammals (in this case only humans and dogs) will be responsible for gene expression differences between species (32).

mVISTA

mVISTA is designed to perform pairwise alignments of DNA sequences from two or more species up to megabases long and to visualize these alignments together with annotations. **AVID is the alignment engine** behind mVISTA, it allows to globally align DNA sequences of arbitrary length (6). The key features of the algorithm are speed, accuracy, the ability to detect weak homologies and to align with one of the sequences in draft (by ordering and orienting the contigs automatically). The **mVISTA visualization module** is designed to display global sequence alignments of genomic sequences from different species (4).
To use mVISTA for comparative sequence analysis, two or more sequences in FASTA format (plain text only) or GenBank accession numbers together with a gene annotation file are submitted to the Web server. One of the two sequences is selected as the base or reference sequence. The server automatically uses RepeatMasker to mask repetitive elements in the reference sequence. The X-axis of the generated plot represents the base sequence and the Y-axis represents the percent identity in the predefined window of an alignment. If a user provides an annotation of the base sequence, the genes will be shown above the plot as dark gray arrows and the exons and UTRs will be marked by colored rectangles. mVISTA can also display positions and orientation of draft sequences, indicate gaps in the alignment, display locations and types of repeats and show SNPs on the base sequence.

Advanced mVISTA options include: utilizing an algorithm that simultaneously compares two sequence alignments to determine percent identity and length thresholds for identification of non-coding sequences conserved in all three species (5); displaying a level of sequence difference rather than conservation (used for evolutionary close species). In the latter case the Y-scale is calculated automatically to allow for optimal visual analysis of a plot.

rVISTA

rVISTA (regulatory Vista) combines searching the major transcription factor binding site database TRANSFAC™ Professional from Biobase with a comparative sequence analysis. It can be used directly or through links in mVISTA, Genome VISTA, or VISTA Browser.

Identifying candidate transcriptional regulatory elements in non coding genomic sequences is a challenging problem. Analyzing non coding sequences for the presence of known transcription factor binding sites produces a huge number of false positive predictions that are randomly and uniformly distributed. Combining database searches with comparative sequence analysis reduces the number of predicted transcription factor binding sites by several orders of magnitude (9). rVISTA makes predictions by the Match™ program (33) based on TRANSFAC Professional library or user-submitted matrices to identify potential transcription factor binding sites in each of the two aligned sequences, and determines which of the predicted sites are aligned and conserved between the species in our alignment. Predictions can also be based on user-submitted position weight matrices or a consensus sequence. TRANSFAC searches are performed using the default core and matrix similarity values or parameters submitted by a user. The visualization program for rVISTA allows the user to look at binding sites for a single transcription factor and/or various combinations of transcription factor binding sites which allows one to easily examine the clustering of binding sites for factors that are believed to interact with one another. Both global (AVID), and local (Blastz) alignment algorithms are incorporated into rVISTA.

Use of rVISTA to annotate the candidate regulatory region of *KIF3A*

An immediate question usually asked about a candidate regulatory region is; can transcription factor binding sites be computationally identified in the interval? Here we use rVISTA to address this question about the candidate regulatory region which is located downstream of *KIF3A* and contains several conserved non-coding elements (Fig 1a). From VISTA Browser we submit this interval to TRANSFAC using default parameters (core similarity values of 0.7 and matrix similarity values of 0.75). VISTA Browser offers the option to the user of looking at all possible transcription factor binding sites or only those sites that are aligned and evolutionarily conserved between humans and mice. Examination of the list of transcription factors with evolutionarily conserved binding sites reveals one that is known to be involved in kidney development (*AP2REP*) and one that is expressed in the brain (*ZIC2*), two tissues in which *KIF3A* is functionally important. In Fig. 3 the location of the evolutionarily conserved binding sites for these transcription factors in the interval immediately downstream of *KIF3A* are shown.

Phylo-VISTA for visualization and analysis of multiple sequence alignments

Phylo-VISTA program with associated Web server presents a novel method for the visualization and analysis of conservation in multiple sequence alignments by providing several significant extensions to VISTA tools (21). It displays the similarity of DNA sequences from multiple species while considering an associated phylogenetic tree. Features include a broad spectrum of resolution parameters for examining the alignment and the ability to easily compare any sub-tree of sequences within a complete alignment dataset. Phylo-VISTA uses not an individual sequence, but the entire multiple alignment as a base in the x-axis, which is similar to the Synplot method for pairwise alignments (34). As a result, the tool is capable of displaying location and length of gaps in all sequences as well as providing annotations beyond a single base sequence.

The Phylo-VISTA server requires submission of a multiple alignment file in the multi-FASTA format, phylogenetic tree used in the alignment program or produced by it, and annotation files associated with individual sequences if available**.**

Future Directions

The VISTA family of tools has proven to be useful for biologists carrying out various comparative genomics studies. VISTA web site with all associated programs has been actively maintained and improved for the past 4 years. Since the introduction of our first online VISTA server mVISTA in 2000 this tool alone has processed more than 50,000 comparative analysis queries. In addition we distributed close to 2000 copies of the stand-alone mVISTA software in academic and commercial institutions of 53 countries.

VISTA web site with all associated programs has been actively maintained and improved for the past 4 years. We are planning to work on more efficient algorithms and software implementation to be able to efficiently compare the DNA sequences of a wide range of species at varying evolutionary distances. As more whole genome sequences become available we will incorporate those as base genomes in the VISTA Browser. Additionally, we plan to link VISTA Browser to a number of external databases of relevant genomic information.

References

1. Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C.,, Bouck J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res*., **10**, 577-86.

2. Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., Miller. W. and NISC Comparative Sequencing Program. (2003) MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.,* **31**, 3518-24.

3. 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.

4. Mayor, C., Brudno, M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer, K.A., Pachter, L.S. and Dubchak, I. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics,* **16**, 1046-1047.

5. Dubchak,I., Brudno, M., Pachter, L.S., Loots, G.G., Mayor, C., Rubin, E.M. and Frazer, K.A. (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons. *Genome Res.,* **10,** 1304-1306.

6. Bray, N., Dubchak, I. and Pachter, L. (2003) AVID: A Global Alignment Program. *Genome Res.,* 13, 97-102.

7. Couronne ,O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L. and Dubchak, I. (2002) Strategies and Tools for Whole Genome Alignments, *Genome Res.,* **13**, 73-80.

8. Brudno, M., Poliakov, A., Salamov, A., Cooper, G.M.,, Sidow, A., Rubin, E.M., Solovyev, V., Batzoglou, S. and Dubchak, I. (2004) Automated Whole-Genome Multiple Alignment of Rat, Mouse, and Human. *Genome Res.,* **14,** 685-92

9. Loots, G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin, E. (2002) rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. *Genome. Res.,* 12, 832-839.

10. Parent, S. A., Zhang, T., Chrebet, G., Clemas, J. A., Figueroa, D. J., Ky, B., Blevins, R. A., Austin, C. P. and Rosen, H. (2002) Molecular characterization of the murine SIGNR1 gene encoding a C-type lectin homologous to human DC-SIGN and DC-SIGNR. *Gene,* **293,** 33–46.

11. Chen, J., Kitchen, C. M., Streb, J. W. and Miano, J. M2002) Myocardin: A Component of a Molecular Switch for Smooth Muscle Differentiation. *J Mol. Cell Cardiol.,* **34,** 1345-1356.

12. Anguita, E., Sharpe, J. A., Sloane-Stanley, J. A., Tufarelli, C., Higgs, D. R. and Wood, W. G. (2002) Deletion of the mouse a -globin regulatory element (HS 26) has an unexpectedly mild phenotype. *Blood,* **100**, 3450-3456.

13. Touchman, J.W., Dehejia, A., Chiba-Falek, O., Cabin, D.E., Schwartz, J.R., Orrison, B.M. Polymeropoulos, M.H. and Nussbaum, R.L. (2001) Human and mouse alpha-synuclein genes: Comparative genomic sequence analysis and Identification of a Novel Gene Regulatory Element. *Genome Res*., **11,** 78-86.

14. 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

15. Rat Genome Sequencing Project Consortium (2004) Genome Sequence of the Brown Norway Rat Yields Insights into Mammalian Evolution. *Nature,* **428**, 493-521.

16. Frazer, K.A., Elnitski , L., Church, D.M., Dubchak, I., Hardison, R.C. (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res*., **13,** 1-12.

17. Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E, and Eisen, M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, Jan 21.

18. Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I. and Batzoglou, S. (2003) Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, **1**, I54-I62

19. Bray, N. and Pachter, L. (2004) MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.,* **14,** 693-9

20. Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F.A., Podkolodny, N.L.and Kolchanov NA. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.,* **26,** 362-7.

21. Shah, N., Couronne, O., Pennacchio, L.A., Brudno, M., Batzoglou, S., Bethel, E.W., Rubin, E.M., Hamann, B. and Inna Dubchak. (2004) Phylo-VISTA: An Interactive Visualization Tool for Multiple DNA Sequence Alignments. *Bioinformatics,* **20**, 636-43..

22. Waterston, R.H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature, **420,** 520-62.

23. Kent,W.J. (2002) BLAT—the BLAST-like Alignment Tool. *Genome Res.,* **12**, 656–664.

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

25. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., Miller, W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res., **14**, 708-15.

26. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.,* 5, R12

27. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.,* **12**, 996–1006.

28. Cheng, J,F,, Pennacchio, L.A. (2003) Comparative and functional analysis of cardiovascular-related genes. *Pharmacogenomics.* **4**, 571-82.

29. Lin, F., Hiesberger, T., Cordes, K., Sinclair, A. M., Goldstein, L. S. B., Somlo, S. and Igarashi, P. (2003) Kidney-specific inactivation of the KIF3A subunit of kinesin-II inhibits renal ciliogenesis and produces polycystic kidney disease. *Proc. Nat. Acad. Sci.,* **100,** 5286-5291.

30. Marszalek, J. R., Liu, X., Roberts, E. A., Chui, D., Marth, J. D., Williams, D. S. and Goldstein, L. S. B. (2000) Genetic evidence for selective transport of opsin and arrestin by kinesin-II in mammalian photoreceptors. *Cell,* **102**, 175-187.

31. Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M. and Venter, J.C. (2003) The dog genome: survey sequencing and comparative analysis. *Science,* **301,** 1898-903.

32. Frazer, K.A., Tao, H., Osoegawa, K., Pieter J. de Jong, P.J., Chen, X., Doherty, M.F. and Cox, D.R (2004). Non-coding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res*., 14, 367-72.

33. Kel, A.E., Gossling, E., Reuter, I., Cheremushkin. E., Kel-Margoulis, O.V. and Wingender, E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.,* **31,** 3576-9.

34. Göttgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R. and Green, A.R. (2001) Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences, *Genome Res.,*. **11**, 87-97.

Acknowledgements

We are grateful to Simon Minovitsky, Shyam Prabhakar, Michael Teplitsky, Dmitriy Ryaboy for ongoing development, maintenance and support of all our tools. The VISTA project is a collaborative effort of a large group of scientists and engineers. Among contributors are Nicolas Bray (the author of AVID and contributor to the design of whole-genome alignment strategies), Michael Brudno (contributor to MLAGAN and multiple whole-genome alignment), Nameeta Shah (the author of Phylo-VISTA), Olivier Couronne (contributor to whole-genome alignment), Gabriela Loots (rVISTA), Ivan Ovcharenko (contributor to rVISTA and the whole-genome alignment), Chris Mayor (contributor to mVISTA), and Brian Klock and Lila Tretikov (VISTA Browsers). Our special thanks to the biologists of Genomics Division at LBNL (Dario Boffelli, Jim Bristow, Jan-Fang, Cheng, Marcelo Nobrega, Len Pennacchio, James Priest, and many others) for their help, support and critical comments.

This project was partially supported by the Programs for Genomic Applications grant from the NHLBI/NIH and Office of Biological and Environmental Research, Office of Science, U.S. Department of Energy.

VISTA has a clean output, allowing for easy identification of sequence similarities and differences, and is easily configurable, enabling the visualization of alignments of various lengths at different levels of resolution.

Figure legends.

Fig 1a. VISTA Browser (VGB2.0) plot of the 180 kb interval (chr5:131949456-132139102) containing *KIF3A*, accessible through the gateway at http://www-gsd.lbl.gov/vista or http://pipeline.lbl.gov. Visualization plots shows conserved sequences between humans and mice (top panel) and humans and rats (bottom panel) based on the multiple three-genome alignment by MLAGAN. The level of conservation (vertical axis) is displayed in the coordinates of the human sequence (horizontal axis). Conserved regions above the level of 70%/100bp are highlighted under the curve, with red indicating a conserved non-coding region, blue - a conserved exon, and turquoise – a UTR. Details of the display are given in the legend on the left hand side of the plot. 'UCSC' button opens another window with the mirrored UCSC browser of the same interval with integrated VISTA tracks. The browser is provided with extensive on-line help.

Fig 1b. VISTA Browser generated list of conserved human/mouse elements in the *KIF3A* region with their coordinates in the human (unbracketed numbers) and mouse (bracketed numbers) sequence, lengths and percent identities, and functional annotation. Elements from the beginning of the 180 kb interval in *RAD50* are shown.

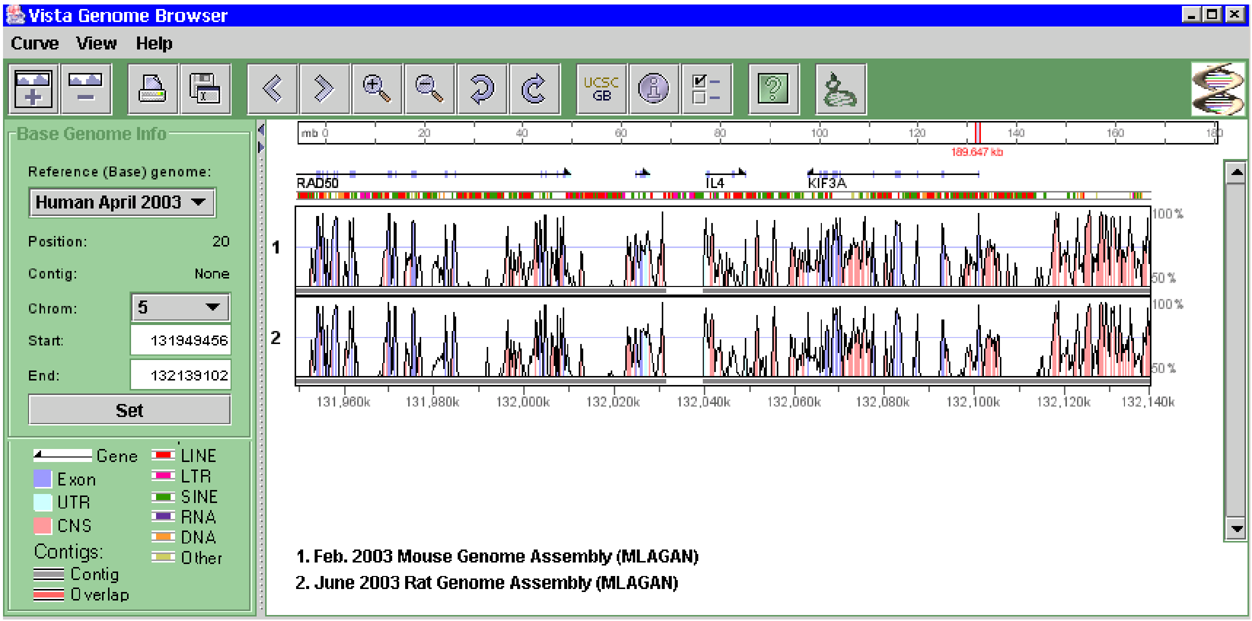
Fig 1c. Genomic fragment upstream of *KIF3A* gene containing multiple conserved non-coding elements. The number of conserved elements (colored in red) depends on the selected by a user percent identity and length cutoffs shown above each plot.

Fig 2. VISTA Browser plot generated by the submission of the draft dog genomic sequence (GenBank accession number AF276990) to the GenomeVISTA server. The dog sequence is automatically aligned against the orthologous human region. The bar at the bottom of the plot shows the locations of draft fragments in the aligned sequence, grey indicates that the sequence is present and white indicates that it is missing

Fig. 3. rVISTA visualization of predicted binding sites for the *AP2REP* and *ZIC2* transcription factors in the interval downstream of *KIF3A.* Only the predicted binding sites that are evolutionarily conserved are displayed.

.

Fig.1a



Reference (Base) genome: Human April 2003

Position: 20

Contig: None

Chrom: 5

Start: 131949456

End: 132139102

Set

1. Feb. 2003 Mouse Genome Assembly (MLAGAN)
2. June 2003 Rat Genome Assembly (MLAGAN)

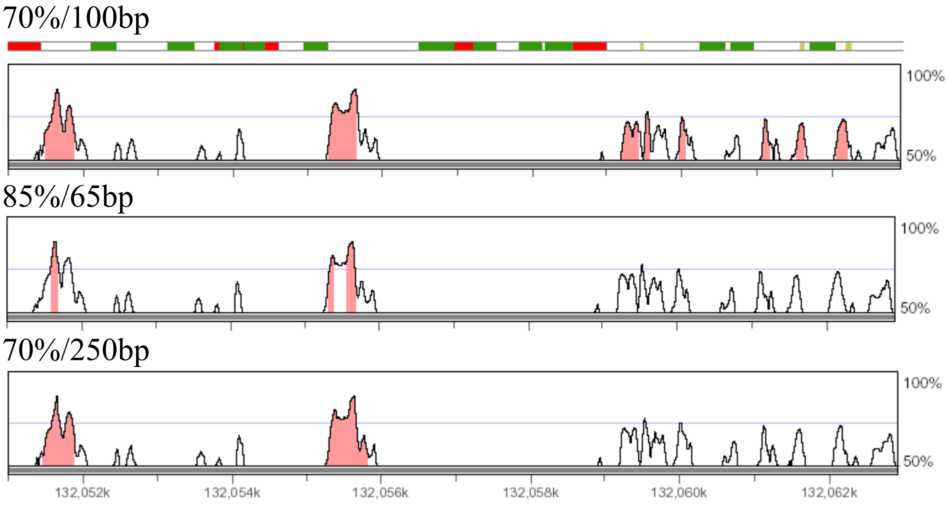
Fig 1b

```
Criteria: 70% identity over 100 bp

*************** Conserved Regions - Human (Mouse) ***************

 131952851  (54292441)  to 131953108  (54292210)  =      258bp  at  69.4%  noncoding
 131954117  (54291314)  to 131954245  (54291186)  =      129bp  at  89.9%  exon
 131954246  (54291185)  to 131954339  (54291091)  =       98bp  at  71.4%  noncoding
 131954479  (54290969)  to 131954644  (54290804)  =      166bp  at  87.3%  exon
 131954759  (54289473)  to 131954891  (54289341)  =      135bp  at  71.1%  noncoding
 131955242  (54288804)  to 131955435  (54288611)  =      194bp  at  89.7%  exon
 131956186  (54288222)  to 131956392  (54288016)  =      207bp  at  73.4%  exon
 131957525  (54284506)  to 131957654  (54284379)  =      130bp  at  70.0%  noncoding
 131957779  (54284180)  to 131957961  (54283998)  =      183bp  at  85.2%  exon
```

Fig 1c



70%/100bp

85%/65bp

70%/250bp

Fig 2

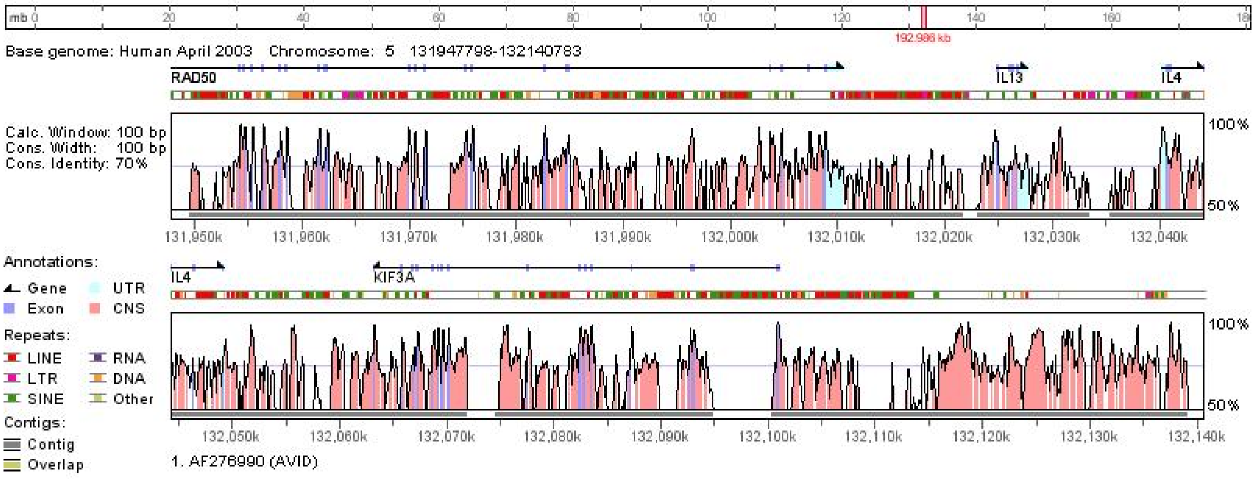
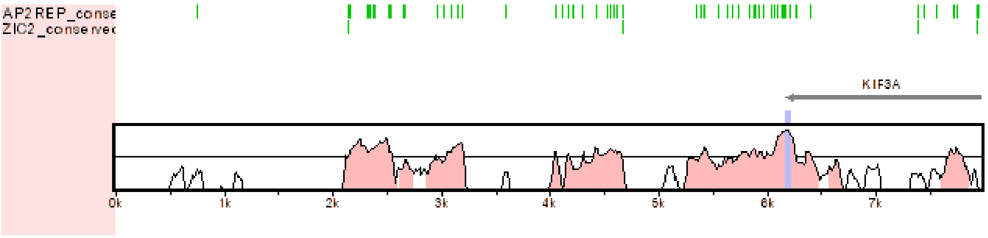


Fig 3



AP2REP_conse
ZIC2_conserved
KIF3A
0k
1k
2k
3k
4k
5k
6k
7k