Human Molecular Genetics, 2006, Vol. 15, Review Issue 1 ***R51–R56***
doi:10.1093/hmg/ddl056

Evolution at the nucleotide level: the problem of multiple whole-genome alignment

Colin N. Dewey[1,*] and Lior Pachter[2]

[1]Department of Electrical Engineering and Computer Sciences and [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA

Received February 1, 2006; Revised and Accepted March 9, 2006

With the genome sequences of numerous species at hand, we have the opportunity to discover how evolution has acted at each and every nucleotide in our genome. To this end, we must identify sets of nucleotides that have descended from a common ancestral nucleotide. The problem of identifying evolutionary-related nucleotides is that of sequence alignment. When the sequences under consideration are entire genomes, we have the problem of multiple whole-genome alignment. In this paper, we first state a series of definitions for homology and its subrelations between single nucleotides. Within this framework, we review the current methods available for the alignment of multiple large genomes. We then describe a subset of tools that make biological inferences from multiple whole-genome alignments.

INTRODUCTION

Comparative genomics (1,2) is the use of molecular evolution as a tool in the investigation of biological processes. Nucleotide sequences common to the genomes of several diverged species are indicative of shared biology, whereas differences in genomic sequence and structure may shed light on what makes species distinct. The identification of genomic elements that have been conserved over time allows biologists to focus their experiments on those parts of the genome that are fundamental to much of life. Thus, methods for the comparison of genomes and prediction of elements constrained by evolution have been actively researched as of late.

Often implicit in the discussion of conserved or common sequences is the concept of 'homology'. Homology, famously defined by Richard Owen as 'the same organ in different animals under every variety of form and function', is accepted by most as 'common ancestry' (3). This important concept relies on the identification of evolutionary 'characters', distinct entities between which we may assign ancestral relationships. First used in reference to morphological characters, such as eye color or petal number, homology has since been used in reference to characters of all levels, from the molecular to the behavioral. Our recently acquired ability to identify single nucleotides in the genomes of different species allows us to specify homology at the smallest scale. The definition and identification of homology at this scale is the focus of our review.

The prediction of homology between nucleotides relies on the fact that genomic positions derived from a common ancestral position are more likely to have the same 'state': one of A, C, G or T. With only four states and, often, billions of genomic positions, we cannot simply use the coincidence of bases at two positions as a basis for assigning homology. Therefore, we must take advantage of context. Positions adjacent in an ancestral sequence are likely to be adjacent in the extant sequences. Predicting homology between genomic positions is thus the problem of identifying colinear segments having statistically significant numbers of similar states. Because we are faced with analyzing multiple large genomes, this task requires expertise from the fields of computer science, statistics, and mathematics. In these fields, the task of identifying related positions in sequences is the problem of alignment.

Although alignment is based on the identification of similar sequences, similarity is not equivalent to homology. Similar, but unrelated sequences may arise simply by chance or through convergent evolution. On the other hand, sequences may be homologous but not share a single similar character. In general, alignments may be used to specify relationships other than common ancestry, such as structural or functional similarities. Although identifying other classes of similarities between sequences is important, such similarities are best

*To whom correspondence should be addressed at: Department of Electrical Engineering and Computer Sciences, 207 Cory Hall No. 1772, University of California, Berkeley, CA 94720-1772, USA. Email: cdewey@eecs.berkeley.edu

understood in the light of evolution. Therefore, we focus on the problem of 'evolutionary alignment', which aims to identify only homologous relationships between nucleotide positions.

In this paper, we describe the various computational and statistical methods that have been developed for the evolutionary alignment of genomes. We restrict our review to methods that may be used to align multiple large genome sequences. Starting with the concept of nucleotide homology, we first provide a series of definitions by which to frame our discussion of alignment methods. After reviewing the latest in alignment technology, we discuss methods that utilize whole-genome alignments for biological discovery.

HOMOLOGY OF NUCLEOTIDES

When Watson and Crick (4) noted that 'the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material', they alluded to the most fundamental level of ancestry. Although homology is used at many levels of biology, it is most directly defined with respect to nucleotide sequences. It is not clear from the literature that people have agreed on a precise definition of nucleotide homology. Given that the molecular mechanisms of nucleic acid replication are well known, it is important from an evolutionary theory standpoint that such definitions are established. Moreover, if we design and compare methods that predict homology between nucleotides, we must have concrete definitions of the problem at hand. These definitions, however, must be based on biology and not on what is possibly identified by our methods. Adhering to this ideology, we propose definitions for nucleotide homology.

At the nucleic acid level, an evolutionary character is a position in single- or double-stranded DNA or RNA. The copying mechanism for nucleic acids is a single-stranded phenomenon and therefore we begin our definitions with the single-stranded case. For a single-stranded nucleic acid, a character x has two properties: its position, traditionally counted from the 5′ end of the polymer, and its state, which is one of A, C, G, T or U. A single-stranded character x is a 'copy' of a character y if x was initially base-paired with y at the time when x was added to its polymer. In such cases, the process by which x is added to its polymer is called 'template-dependent synthesis' (5) and y is called the 'template' for x. Positions added to a polymer without a template (e.g. adenines added during poly(A) extension of mRNAs) have no such relationships.

In the double-stranded case, a character x comprises two base-paired single-stranded characters, x^+ and x^-. Like a single-stranded character, double-stranded characters have a position (usually given as the position of x^+) and a state. The state of a double-stranded character depends on a third property, its orientation, which we indicate by one of $+$ or $-$. If x has an orientation of $+$, then its state is that of x^+ (the character on the forward strand), otherwise it is that of x^- (the character on the reverse strand). One of x^+ or x^- is usually a copy of the other, with exceptions occurring due to mechanisms such as replication slippage (5). Given a double-stranded character x and a single-stranded character y, x is a copy of y, if one of x^+ or x^- is a copy of y. Conversely, y is a copy of x if y is a copy of x^+ or x^-. If both x and y are double-stranded, then x is a copy of y if one of x^+ or x^- is a copy of y^+ or y^-.

We now address mutation, the second major mechanism in molecular evolution. Because characters are positions, point mutations of single-stranded characters do not change their copy relationships. For example, if x is a copy of y and a point mutation changes the state of x from A to G, then x is still a copy of y. However, in double-stranded DNA, repair mechanisms may use the template of an opposite strand or a homologous region to replace damaged positions. Whenever a position is excised and restored using a template, a new copy relationship is established.

Having discussed the concepts of copying and mutation, we now define homology. For both types of characters, we say that x is 'derived' from y if there is an ordered set of characters, $x_1, x_2, \ldots, x_T$, such that $y = x_1$, $x = x_T$ and x_{t+1} is a copy of x_t. The ordered set may include both single-stranded and double-stranded characters. A character x is homologous to a character y if there exists (or existed) a character z such that both x and y are derived from z.

Molecular homology has traditionally been divided into three subrelations: orthology, paralogy and xenology (6). Although these refinements have distinct biological implications (7), it is difficult to state unambiguous definitions for them in terms of biological mechanisms. Nevertheless, the distinctions made by orthology, paralogy and xenology are important and the alignment methods we discuss distinguish between them. We therefore describe how orthology, paralogy and xenology are applied at the nucleotide level.

Homology is first refined by the relation of xenology. Consider two homologous nucleic acid positions, x and y, whose last common ancestor is z. These characters are xenologous if at least one is derived from a position w, derived from z, that was horizontally transferred. That is, the species to which w belonged changed during w's existence (excluding changes from a parent to a child species).

If x and y are not xenologous, then they are either orthologous or paralogous, depending on the events undergone by z and its copies. Replication of genomic nucleic acids is a regular occurrence in cells, with copies of the same genetic material normally separating from each other during cell division. When cell divisions (either through mitosis, meiosis or binary fission) do not separate genomic copies, paralogous relationships are established. To make this more precise, suppose that z is copied, resulting in two derived characters z_1 and z_2 in the same cell, where x is derived from z_1 and y is derived from z_2. If z_1 and z_2 are not subsequently separated by cytokinesis, then x and y are paralogous. Otherwise, x and y are orthologous.

More specific subrelations of homology have been recently proposed and are often useful. Paralogy is divided into inparalogy and outparalogy depending on the collection of species being considered (7,8). To be fully described (C. Dewey and L. Pachter, manuscript in preparation) is the concept of 'topoorthology', a distinguished subrelation of orthology. Topoorthology is based on the classification of duplication events as either 'undirected' or 'directed'. Simply put, a duplication event is undirected if one cannot distinguish between the two copies of the duplicated material. Otherwise, the

duplication is directed, with one copy of the duplicated material called the 'target' if its removal would restore the genome to its original state. Once again, suppose that z is the last common ancestor of x and y. Characters x and y are topoorthologous if they are orthologous and neither is derived from a character w, derived from z, that was part of the target of a directed duplication. Characters x and y are 'monotopoorthologous' if they are topoorthologous and neither is derived from a character w, derived from z, that was part of an undirected duplication. Although orthology is generally a many-to-many relation, monotopoorthology is a one-to-one relation that is commonly identified between genomes. Figure 1 shows the division of homology into its subrelations. Figure 2 gives an example of homologous relationships between copied nucleotide positions.

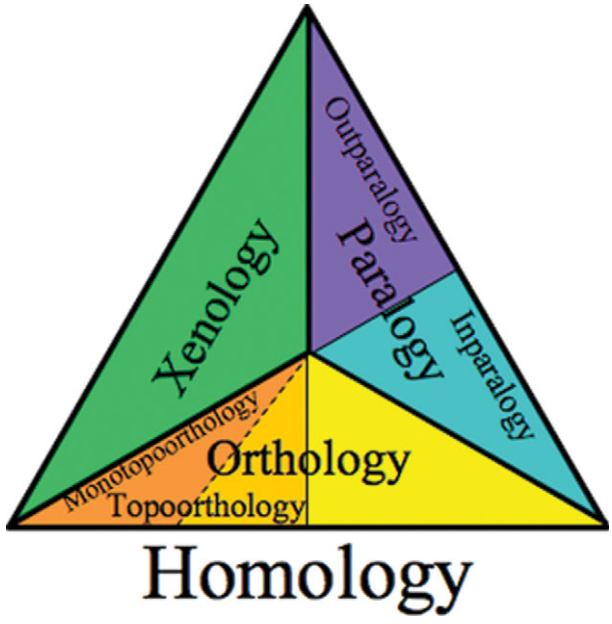

Figure 1. Refinements of homology.

METHODS FOR MULTIPLE WHOLE-GENOME ALIGNMENT

With the evolutionary relations that we wish to establish between genomic sequence defined, we review the tools available for this task. We focus on methods that take as input a set of three or more genomes and output alignments designating homology or its subrelations between individual genomic positions. There are two major strategies for aligning entire genomes: 'local' alignment and 'hierarchical' alignment. Figure 3 illustrates the main components of these strategies.

Local alignment

The local alignment strategy is first to find all similarities between pairs of genomes and then to combine these pairwise alignments into multiple alignments. Pairwise local aligners are unaffected by genome rearrangements, as they effectively compare every position in one genome to every position in another. Local alignments between two genomes represent both orthologous and outparalogous relations (xenology is rarely a concern, unless prokaryotes are involved). When the reference and query genomes are the same, local aligners can additionally find inparalogous relationships. However, as we will describe, pairwise local alignments are typically filtered for orthology before they are joined into multiple alignments.

Pairwise local alignment is a well-studied area (9). Most local aligners use a 'seed-and-extend' strategy in which short exact or inexact matches are used to initiate potentially larger alignments. Although BLAST (10) could be used as a local aligner for whole genomes, many other methods have been developed with large comparisons in mind (11–15).

At the whole-genome scale, the only method currently available for combining pairwise local alignments into multiple alignments is MULTIZ (16). In the language of the authors of MULTIZ, a multiple whole-genome alignment is called a 'threaded blockset'. A threaded blockset is defined as a set of multiple alignments ('blocks') of colinear segments of the input sequences, where each position in the input sequences is included in exactly one block. Blocks are allowed to have just one sequence in cases where the sequence is not found to have any homologs. The purpose of MULTIZ is to join two threaded blocksets into one, given a local alignment of two of the input genomes. More precisely, given a threaded blockset containing species X and another containing species Y, the two threaded blocksets are joined by a pairwise alignment between X and Y.

The UCSC Genome Browser (17) currently provides MULTIZ genome alignments for vertebrates, insects and yeast. For each alignment, the pairwise BLASTZ (11) alignments given to MULTIZ as input are first filtered with a 'best-in-genome' criterion (18). Given a pairwise alignment between a reference and a query genome, this filter keeps only the best alignment for each position in the reference genome. The filtered alignments are assumed to specify only orthologous relationships. Unless applied in a reciprocal manner, these filters give many-to-one orthology relationships between a reference and a query genome. Although not capturing all orthologous relationships, the resulting reference-based multiple alignments have the convenient property that every column has at most one position from each genome.

Hierarchical alignment

A second strategy for multiple whole-genome alignment combines homology mapping with efficient global alignment. Homology maps identify sets of large colinear homologous segments between multiple genomes and are typically designed to find only monotopoorthologous relationships. For example, a homology map might specify that intervals 38,400,000–38,529,874 of human chromosome 17, 101,551,137–101,659,587 of mouse chromosome 11 and 90,483,833–90,585,675 of rat chromosome 10 (all intervals on the forward strand) contain monotopoorthologous and colinear positions (these intervals contain the *BRCA1* gene). Genomic global alignment programs, which require colinearity, are run on segments (such as those just mentioned as an

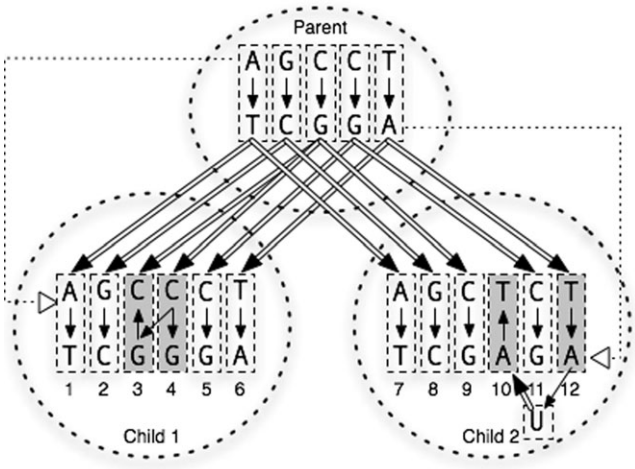

Figure 2. An example evolutionary scenario involving the replication of double-stranded DNA in a parent cell and division into two child cells. The two dotted arrows indicate the separation of the parent strands. Single- and double-stranded copy relationships are indicated by single- and double-edged arrows, respectively. Greyed double-stranded positions have participated in duplication events. Positions 3 and 4 are the result of an undirected duplication (due to replication slippage), whereas positions 10 and 12 are the result of a directed duplication (involving an RNA intermediate), with 12 as the source and 10 as the target. Position pairs (1, 7), (2, 8), (5, 11) and (6, 12) are monotopoorthologous, (3, 9) and (4, 9) are topoorthologous, (6,10) is only orthologous, and (3, 4) and (10, 12) are inparalogous.

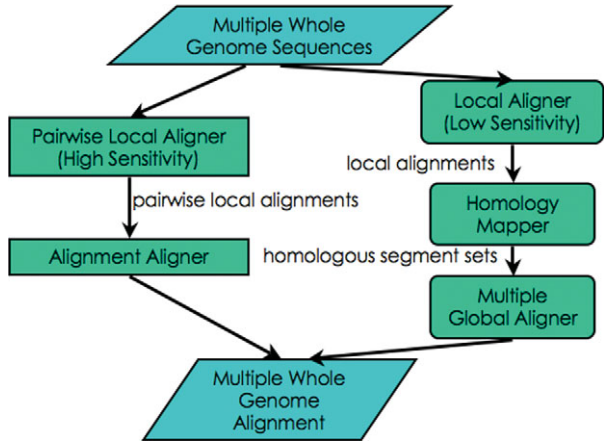

Figure 3. The local (left path) and hierarchical (right path) strategies for multiple whole-genome alignment.

example) specified by a homology map to produce nucleotide-level alignments.

Methods for homology mapping typically take as input sets of pairwise local alignments and output sets of genomic segments containing significant numbers of local alignments that occur in the same order and orientation. After the sequencing of the third large genome, that of the rat (19), several methods were developed for the construction of multiple genome homology maps. GRIMM-Synteny (20) combines the output of a sensitive local aligner, such as PatternHunter (12), between all pairs of k genomes to first produce k-way anchors. Nearby and consistent k-way anchors are joined to produce a k-way orthology map. Mauve (21), a related method that uses multiple maximal unique match (multi-MUM) local alignments (22) to construct orthology maps between multiple closely related species, has been demonstrated to create maps between the human, mouse and rat. Both Mauve and GRIMM-Synteny output one-to-one maps between genomes, which are indicative of monotopoorthology. PARAGON (23), another similar method that uses BLASTZ alignments as input, has been used to create orthology maps between more distant species.

Another method used to align the human, mouse and rat genomes (24) uses a progressive extension of a pairwise strategy engineered for aligning human to mouse (25). Using the BLAT (14) local aligner, a mouse–rat orthology map was first constructed. The orthologous segments were aligned using the LAGAN (26) global aligner, mapped to the human genome using BLAT and finally put into a multiple alignment using MLAGAN. The resulting maps represented all orthology relationships, although most genomic segments were found to be monotopoorthologous. A final method used for human, mouse and rat orthology mapping used bacterial artificial chromosome (BAC) end sequence comparisons as a basis for orthologous anchors (27).

Mercator, a monotopoorthology mapping method that we have designed (C. Dewey and L. Pachter, manuscript in preparation, http://bio.math.berkeley.edu/mercator/), takes as input a set of non-overlapping landmarks in each genome and pairwise similarity scores between all landmarks. A graph is constructed with landmarks as vertices and hits between them as edges. Within this graph, Mercator identifies high-scoring cliques, i.e. sets of landmarks where there is a significant hit between each pair. For example, if exons are used as landmarks, then the first exons of the human, mouse and rat *SHH* gene would be identified as a high-scoring clique. Such cliques indicate orthologous relationships. Starting with the largest cliques (those in which we are most confident), adjacent and consistent cliques (such as those formed from each exon of *SHH*) are joined into runs that represent orthologous segments. Edges not consistent with previously identified runs are discarded and smaller cliques are discovered in the graph and incorporated into runs. The algorithm iterates until cliques involving all possible combinations of genomes have been considered. Thus, unlike most other monotopoorthology mapping methods, Mercator produces maps comprising sets of segments that may be specific to any subset of the input genomes.

Once colinear homologous segments have been identified, multiple global alignment programs are used to assign homologous relationships between individual positions. Global aligners create a one-to-one mapping between the positions of two sequences. Thus, in the absence of recent tandem duplications, multiple global aligners will determine the monotopoorthologous positions in a set of colinear monotopoorthologous segments. The only methods that have been run on whole large genomes thus far are MAVID (28) and MLAGAN (26). Both rely on global 'chaining' of short matches between pairs of sequences. A 'chain' is simply an ordered set of locally aligned segments with the property that the coordinates of the segments of the ith local alignment in the chain are less than those of the segments of the jth local

Table 1. A comparison of the local and hierarchical multiple whole-genome alignment strategies

Strategy	Local	Hierarchical
Programs	BLASTZ, PatternHunter, MUMmer, MULTIZ, CHAINNET	GRIMM-Synteny, Mauve, PARAGON, Mercator, MAVID, MLAGAN, TBA, MAP2
Relationships identified	Most commonly many-to-one orthology	Most commonly monotopoorthology
Sensitivity (genome-wide)	High	Moderate, depending on local alignments used for homology map construction
Sensitivity (within homologous segments)	Moderate	High
Speed	Slow, but often parallelizable	Fast and parallelizable
Short indels	Explicitly gapped	Explicitly gapped
Long indels	Implicitly gapped or interpreted as missing data	Explicitly gapped

alignment, when $i < j$. To create a multiple alignment, both methods use a progressive strategy. MAVID and MLAGAN differ in their identification of local alignment anchors (exact versus inexact) and the methods by which alignments are aligned at internal nodes of the phylogenetic tree (ancestral reconstruction versus via sum-of-pairs). Other multiple genomic global aligners that have not been run on whole genomes but are comparable are given in (13,16,29).

Comparison of alignment strategies

Currently, all local and hierarchical multiple alignment methods focus on orthology. They either identify many-to-many, many-to-one or one-to-one (monotopoorthologous) relationships. Hierarchical methods begin by using local alignments, but typically do not use local methods with their most sensitive parameter settings. This results in much faster running times at the expense of missing short and significantly diverged orthologous sequence. Although less sensitive at the genome-wide scale, the hierarchical strategy can afford to use more sensitive methods at a smaller scale, within the sets of orthologous segments identified by the map.

An important difference between the two strategies is the treatment of genomic segments that have been inserted or deleted during evolution. Given a set of orthologous segments, global aligners will gap all positions that are not found to have orthologous relations. With recent insertions of mobile elements, these gaps can often be very large. Local alignments, on the other hand, are not extended through longer insertions and deletions. Segments that are not part of any local alignment may be interpreted in two ways. One way is to treat orthologous relationships to such segments as missing data. A second interpretation is that segments not part of any alignment are implicitly gapped, i.e. they are believed to have been inserted or deleted. The choice of alignment strategy and the treatment of gaps are issues that researchers must be aware of when using multiple whole-genome alignments for biological inference. Table 1 summarizes the important differences between the two strategies.

FROM ALIGNMENTS TO BIOLOGICAL DISCOVERY

Multiple whole-genome alignments usually constitute only the first step of comparative genomic studies targeted at specific biological questions. We refer the reader to a number of excellent surveys on comparative genomics (1,2) for examples of how multiple whole-genome alignments have been utilized. However, we have selected for further discussion one key (unsolved) problem that is central to utilizing multiple alignments for functional genomics.

A multiple whole-genome alignment assigns homology between nucleotides, but it does not identify genomic positions that are under selection or evolving neutrally. The analysis of homologous nucleotides in a multiple alignment using an evolutionary model forms part of the emerging field of phylogenomics (30) and is essential for distinguishing functional elements from neutrally evolving regions in genomes.

The term ‘conserved nucleotide’ is used informally to describe nucleotides that appear to be mutating slower than suggested by a neutral model of evolution [usually based on a continuous time Markov model for point mutation (31)]. Groups of conserved nucleotides are called conserved elements. To our knowledge, there is no precise definition of conserved elements at this time. Software tools that have been developed for identifying conserved nucleotides and elements include GERP (32), PhastCons (33), BinCons (34) and Shadower (35). Conserved elements can also be identified by examining insertions and deletions within multiple alignments. This has been described in (36,37). There is a discussion of how the choice of alignment affects the determination of conserved nucleotides and estimation of evolutionary model parameters (C. Dewey, P. Huggins, K. Woods, B. Sturmfels and L. Pachter, manuscript in preparation).

The problem of identifying conservation within multiple alignments is inherently a statistics problem, but one that requires further advances by biologists in experimentally validating functional elements. Such advances are crucial for defining appropriate choices of evolutionary models and will subsequently inform computational biologists on the best ways to predict new functional elements.

ACKNOWLEDGEMENTS

C.N.D. was supported by the NIH (HG003150). L.P. was supported by the NIH (R01-HG2362-3 and HG003150) and an NSF CAREER award (CCF-0347992).

Conflict of Interest statement. None declared.

REFERENCES

1. Miller, W., Makova, K.D., Nekrutenko, A. and Hardison, R.C. (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet*., **5**, 15–56.
2. Hardison, R.C. (2003) Comparative genomics. *PLoS Biol*., **1**, E58.
3. Hall, B.K. (ed.) (1994) *Homology: The Hierarchical Basis of Comparative Biology*. Academic Press, San Diego, CA.
4. Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
5. Brown, T.A. (1999) *Genomes*. Wiley, New York.
6. Fitch, W.M. (2000) Homology: a personal view on some of the problems. *Trends Genet*., **16**, 227–231.
7. Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet*., **39**, 309–338.
8. Sonnhammer, E.L.L. and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*., **18**, 619–620.
9. Batzoglou, S. (2005) The many faces of sequence alignment. *Brief. Bioinform*., **6**, 6–22.
10. 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.
11. 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**, 103–107.
12. Ma, B., Tromp, J. and Li, M. (2002) Patternhunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
13. Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S. and Morgenstern, B. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.
14. Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
15. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol*., **5,** R12.
16. Blanchette, M., Kent, W., Riemer, C., Elnitski, L., Smit, A., Roskin, K., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. *et al*. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*., **14**, 708–715.
17. 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**, 51–54.
18. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
19. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E. *et al*. (2004) Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
20. Bourque, G., Pevzner, P. and Tesler, G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*., **14**, 507–516.
21. Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*., **14**, 1394–1403.
22. Darling, A.E., Mau, B., Blattner, F.R. and Perna, N.T. (2004) GRIL: genome rearrangement and inversion locator. *Bioinformatics*, **20**, 122–124.
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 DNA sequence and comparative analysis of human chromosome 5. *Nature*, **431**, 268–274.
24. Brudno, M., Poliakov, A., Salamov, A., Cooper, G., Sidow, A., Rubin, E., Solovyev, V., Batzoglou, S. and Dubchak, I. (2004) Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res*., **14**, 685–692.
25. Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L. and Dubchak, I. (2003) Strategies and tools for whole-genome alignments. *Genome Res*., **13**, 73–80.
26. Brudno, M., Do, C., Cooper, G., Kim, M., Davydov, E., Green, E., Sidow, A. and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*., **13**, 721–731.
27. Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., de Jong, P., Nierman, W., Strausberg, R. and Fraser, C. (2004) Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res*., **14**, 1851–1860.
28. Bray, N. and Pachter, L. (2004) MAVID: constrained ancestral alignment of multiple sequences. *Genome Res*., **14**, 693–699.
29. Ye, L. and Huang, X. (2005) MAP2: multiple alignment of syntenic genomic sequences. *Nucleic Acids Res*., **33**, 162–170.
30. Eisen, J. and Fraser, C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science*, **300**, 1706–1707.
31. Felsenstein, J. (2003) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, USA.
32. Cooper, G., Stone, E., Asimenos, G., Green, E., Batzoglou, S. and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*., **15**, 901–913.
33. Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., Richards, S. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.
34. Margulies, E.H., Blanchette, M., Haussler, D. and Green, E.D. (2003) Identification and characterization of multi-species conserved sequences. *Genome Res*., **13**, 2507–2518.
35. McAuliffe, J., Pachter, L. and Jordan, M. (2004) Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics*, **20**, 1850–1860.
36. Lunter, G., Ponting, C.P. and Hein, J. (2006) Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol*., **2,** e5.
37. Snir, S. and Pachter, L. (2006) Phylogenetic profiling of insertions and deletions in vertebrate genomes. In *Proceedings of RECOMB*.