

A Novel Comparative Method for Locating Human Conserved DNA

1 Abstract

Functional elements in our genome are under purifying selection, and accumulate fewer mutation events than non-functional neutrally evolving material. This can be used to locate conserved and putatively functional DNA. Current methods are primarily based on models of nucleotide substitution, and have been quite successful at localising protein-coding genes. Many other types of functional elements however, such as transcription-factor binding sites and RNA genes, are less conserved at the sequence level, and are therefore harder to find. This proposal concerns a novel method that is orthogonal to current methods, and focuses exclusively on insertion and deletion events. Preliminary work showed that the method is highly sensitive, and is able to locate material that is weakly conserved at the sequence level, but is under purifying selection with respect to indels. We have e.g. shown a high sensitivity for RNA genes, despite a generally low sequence-level conservation of such genes. A preliminary analysis further revealed a novel class of high-period tandemly repeated sequences whose structure is strongly conserved, and which are highly overrepresented in sub-telomeric regions. The method is a first step toward more specific automatic annotation method for e.g. RNA genes of transcription factor binding sites. This pilot project proposes to set up a large, structured database of conserved elements in the human genome, to allow this data to be flexibly combined with downstream annotation methods, and to foster internet-enabled collaborations building further onto this novel method.

2 Background

The completion of the Human Genome Sequencing project in 2001 was heralded with much excitement, as it was expected to greatly enhance the understanding of our basic biology and the mechanisms of disease. The completion of the mouse, rat and dog genomes was another significant leap forward. Together, these advances made it possible for the first time to systematically compare mammalian genomes on a large scale and to detect biologically functional DNA, which tends to be evolutionary conserved and thus tends to stand out from neutrally evolving non-functional DNA.

The analysis of our genome has however proved to be surprisingly difficult. For example, even a basic statistic such as the estimated number of protein-coding genes is still contentious, with recent estimates ranging from an initial 100,000 to the current 20,000-25,000 [Stein 2004]. However, despite many remaining issues, protein coding genes are the functional class which is best understood and for which annotation procedures are most highly developed. Presently, 36.6 Megabase, or 1.2% of all human euchromatin, is annotated as protein-coding exon, forming the currently largest class of annotated elements by far. Still, it is well known that protein-coding genes comprise only a part of the functional fraction of our genome. Indeed, it has been suggested that many phenotypic differences among mammals may be due not to changes in gene products, but in their expression levels and patterns, highlighting the importance of elements that regulate this expression [Rodriguez-Trelles 2004, Khaitovich 2004], most of which will be non-genic functional elements such as transcription-factor binding sites and microRNAs [Gaffney and Keightley 2004]. A statistical analysis based on the distribution of nucleotide substitutions led to the estimate that as much as 5% of the human genome is under purifying selection and putatively functional [Chiaromonte et al. 2004], suggesting that the majority of functional elements do not consist of protein-coding genes, and remain to be annotated. The localization and characterisation of the non-genic functional fraction is one of the key challenges that is currently facing biology, as its resolution has a potentially major impact on our understanding of our basic biology and human disease.

Recent work has confirmed the existence of specific non-genic conserved regions of which the biological function is currently unknown [Antonarakis et al. 2002, Dermitzakis et al. 2004], some of which showing extremely strong conservation [Bejerano et al. 2004]. Functional elements that can accept some mutations are more informative on the internal structure of these elements (cf. the triple periodicity of protein-coding genes). The non-genic elements in the non-genic functional fraction whose structure is known, such as RNA genes and transcription factor binding sites (TFBSs), are much harder to localise than protein-coding genes because of a generally larger tolerance for nucleotide substitutions. These observations provide strong motivation for developing more sensitive computational methods, to find a larger set of such conserved elements. A detailed annotation of our genome is not only of fundamental scientific interest, but will be an essential tool for evolutionary biologists and bioinformaticians, and as functional annotations are added, for wet biologists and medical researchers, in the same way that the repository of annotated protein-coding genes has revolutionised biological research.

Comparative genomics, although considerably more difficult than originally thought, still promises to be the most effective approach to annotation. Many sequencing projects are underway or are nearing completion, including those for cow, chimp, dog, opossum and rhesus macaque, and many more are planned, such as for cat, pig, rabbit, wallaby and platypus [NHGRI 2004, GOLD 2004]. These genomes, spread across all mammalian clades, form a highly informative dataset that will further increase the power of comparative methods. It is however important to stress that

the information contained in this dataset increases sub-linearly with the number of species, and there is a parallel and urgent need for the development of new methodology to analyse this data.

A very powerful and general approach to localising functional elements of a particular class is to use Bayesian or likelihood-based methods based on a probabilistic model for the evolution of such elements. This general approach has been very successfully applied in the case of protein-coding genes [Burge and Karlin 1998]. For TFBSs and RNA genes, this approach is to date much less successful. One reason is that the statistical signal, under state-of-the-art models, does not yield a sufficiently low false-positive rate of detection [Rivas and Eddy 2000, Rivas and Eddy 2001]. Because of the sheer size of our genome, this gives rise to unacceptably many false positive hits. One possible way of increasing the power of these methods, is to add an initial filtering step. This step would be based on a detailed model of neutral evolution, rejecting data that shows a close fit. The substitution-based approach referred to above, with which the 5% estimate was obtained, has however inadequate descriptive power to distinguish neutrally evolution sequence from sequence under purifying selection, at reasonable sensitivities and specificities.

3 Programme and Methodology

AIMS AND OBJECTIVES

This project aims to develop a novel comparative method for analysing the pattern of evolutionary conservation in the human genome, and to develop a large, structured database containing a substantial proportion of evolutionary conserved elements in the human genome, with a particular focus on non-genic elements, both serving as a basis for subsequent collaborative annotation projects. In more detail, this project comprises seven subsidiary objectives:

- To obtain an estimate of the proportion of conserved genome, improving upon the earlier estimate of 5%;
- To obtain a database (≈ 50 Mb) of conserved elements, independent of the type of conserved DNA;
- To publish a preliminary whole-genome analysis of conserved elements in the human genome;
- To develop a flexible and extensible XML database format to store and share evolutionary conserved segments;
- To develop a web server allowing easy (human and machine) access to this database;
- To develop statistical “re-alignment” procedures, optimised using an objective measure of alignment quality;
- To develop an XML language to describe hidden Markov models (HMMs), and a compiler to automatically generate efficient HMM parsers.

In contrast to previous approaches, the neutral model referred to above will be based on the neutral behaviour of *insertion/deletion (indel) events*, and is independent of nucleotide-level conservation due to selection acting on substitution events. Compared to substitutions, indels are often even more disruptive of function, and are heavily selected against. This is well known for protein-coding genes, but is true also for RNA genes and TFBSs, which are far less conserved at the nucleotide level and thus harder to find by substitution-based methods. In extreme cases, the sequence content may be arbitrary but its length conserved, for example when the spacing between TFBSs is critical [Bergman et al. 2002], or in the case of a conserved loop region in an RNA gene.

In a preliminary study [Lunter 2004] we have shown that the evolutionary signal left by indel events is surprisingly informative. We applied a novel model describing the evolution of genomic sequences under indel events on a whole-genome alignment of the human and mouse genomes. Using this we were able to identify a large subset of conserved elements (40 Mb) at a controlled and low false-positive rate (1%), over half of which was non-protein-coding (see section B for details). This project proposes to (i) further develop this model, (ii) to identify a large proportion of the genomic elements that are conserved between most mammals, and (iii) to perform a preliminary analysis of the resulting set of conserved elements.

The database is the main output of this project, and may serve as a pre-filtered dataset to improve the power of any subsequent annotation project aimed at particular classes of genetic elements. In order to facilitate such research, a key objective is to ensure rapid publication of the database of conserved elements. Particular attention will be given to the database format, to make sure that the annotation is easily usable, and that future developments and additions can be incorporated while ensuring backward compatibility.

We will use the database to perform a preliminary whole-genome analysis of non-genic conserved elements. Since the key strength of the proposed method is that it is independent of conservation with respect to nucleotide substitutions, we will focus our attention on *length-conserved elements*: those elements that are under strongly purifying selection with respect to indels, but evolve close to neutral with respect to substitutions. A preliminary scan has indicated that these elements are surprisingly abundant, and that they exhibit a strongly non-random distribution over the genome, with particular enrichment in sub-telomeric regions.

More mammalian genomes will be sequenced in the next few years. Efficient use of this increasingly rich source of data hinges on the availability of reliable whole-genome multiple alignments, and of statistical models that describe

the simultaneous evolution of multiple genomes along their phylogeny. One key objective is therefore to develop practical statistical alignment procedures that can be applied to entire genomes. Probabilistic alignment procedures are formulated in terms of Hidden Markov models (HMMs). Implementation of the associated algorithms is time-consuming and error prone but essentially mechanical. To help facilitate this work, we will develop a compiler tool, and an associated XML description of such HMMs, which automates this process. Because of the widespread use of HMMs in biological (sequence) modelling, this tool will be of wider significance for the bioinformatics community.

This proposal comprises a small pilot project. It will be carried out by a single postdoctoral researcher with considerable research experience, who will manage the project himself. He will be responsible for organising weekly progress meetings with the PI. He will also be responsible for software development, and for scientific output in the form of papers and conference presentations for the duration of the project.

4 Research methodology

Below we describe in more detail the techniques that will be used to achieve the objectives outlined above.

A NEUTRAL MODEL OF EVOLUTION

An accurate understanding and modelling of the process of neutral evolution is fundamental to the proposed study. It is basic to the improved estimate of the amount of functional genome, and to the localisation of conserved elements. Besides these main applications, it also serves as a building block in alignment procedures, improvements to which directly impact every downstream analysis. From a modelling perspective, focusing on neutral evolution is advantageous as these models can benefit from the availability of large amounts of data for parameter inference (“training”) in the form of ancestral repeats (ARs, i.e. transposable elements that were inserted prior to speciation). Besides evolving neutrally, ARs generally occur in high copy numbers, allowing the ancestral sequence to be identified by consensus, further simplifying training.

We will focus on one aspect of sequence evolution: sequence insertions and deletions. Compared to substitutions, the modelling of the indel process has received comparatively little attention. Two aspects of the process in particular have not been included in models thus far. Indel rates vary with sequence content, with higher rates for more extreme C+G contents. This is possibly related to slippage, a major cause of small indels, being more likely in locally uniform sequences. Secondly, indel lengths have a “fat tailed” distribution, which is insufficiently accounted for by the usual affine-gap alignment penalties that essentially fit a geometric indel length distribution. These improvements will impact on alignment algorithms. Importantly, indel modelling forms the main ingredient for a conservation confidence measure, and moreover suggest a novel and objective way to measure the quality of alignment algorithms on neutrally evolving DNA. These applications are discussed in the next three sections.

B LARGE-SCALE ANALYSIS OF FUNCTIONAL GENOMIC ELEMENTS

The original estimate for the genomic proportion of functional elements (GPFE)¹ rested on the assumption that the raw observed substitution count on human-mouse ancestral repeats (ARs) follows that of general non-functional DNA [Mouse Genome Sequencing Consortium 2002]. This assumption can be disputed. Even assuming uniformity of the mutational process on these two genomic subsets, the observed substitution frequencies differ because ARs do have different nucleotide content from general non-conserved DNA. For instance, the ubiquitous Alu element is CG-

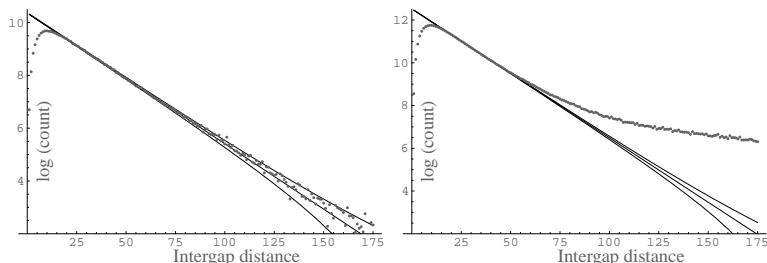


Figure 1: Histogram of observed intergaps in human-mouse alignments, on ancestral repeats (left), and the full genome (right). Lines represent best fit to a geometric distribution, and 95% confidence intervals (using a Bernoulli distribution per bin). The dip in observed short intergaps (≤ 20 nt) is caused by “gap attraction”. Purifying selection results in an overabundance of long genomic regions without gaps, showing up as a fattening of the distribution for long intergaps (from length ≈ 60), which is absent from the measurements on ancestral repeats. (Unpublished data)

rich and contains many CpG dinucleotides, both serving to increase the instantaneous substitution frequency. This leads to overestimates of the neutral substitution count, in turn leading to an overestimate of the GPFE. Moreover, this procedure is not well suited to locating conserved elements. Only for extremely low substitution counts can the hypothesis of neutral evolution be confidently rejected, corresponding to a small minority of the purported 5% conserved DNA.

We will follow a different approach, and focus on the indel process instead. The neutral hypothesis predicts that “intergap distances” (i.e. the separation in nucleotides of two neighbouring indels) follow a geometric distribution. This prediction follows from

¹We equate “conserved” and “functional” here, which strictly speaking is unproven.

the key observation that under neutrality, individual indel events occur independently of each other. As a model prediction, it is borne out to surprisingly great accuracy over a wide range of intergap distances. Indeed, for distances in the range 20–50 basepairs, no deviations outside the tight sampling confidence intervals are observed (see Fig. 1). However, this near-exact model fit breaks down for short and very long intergaps. The short-range breakdown is due to a general phenomenon of alignment procedures, termed “gap attraction” (see section D), and is not due to evolutionary processes. The breakdown at longer distances however is due to either departures from uniformity of the neutral mutation process itself, or the effects of purifying selection. By comparing this statistic on ancestral repeats (i.e. in the absence of selection) and on the entire genome, we were able to determine that (i) indel rate non-uniformities do not contribute significantly to this observed overabundance, and (ii) that ancestral repeats indeed do evolve neutrally. This hitherto was a hypothesis which, although reasonable, had never been proved. These two observations imply that the GPFE can be quantified, avoiding any hypothesis on the substitution process. A preliminary analysis yielded a figure of 3.0%, significantly lower than the substitution-based estimate of 5%, but substantiating the claim that the majority of conserved genetic elements remain unannotated.

A second application of the neutral indel model is as the basis for a “conservation confidence” measure. When fewer indels are observed than expected under neutrality, the hypothesis of neutral evolution can be rejected at a well-quantifiable confidence level. Using this, genomic subsets of conserved elements can be identified. A first estimate indicated that conserved regions of medium size (\approx 150–200 bp in a two-sequence alignment) could be identified at a 1% level, making the method potentially relevant for finding clusters of regulatory elements, RNA genes, and small first exons in protein-coding genes. Indeed, the conserved subset identified in this way contained 205 of 222 currently annotated microRNA genes, suggesting a very good sensitivity within this structural class.

This way of identifying functional regions does not require evolutionary modelling of the functional element. While sacrificing some sensitivity, this makes the method relevant for identifying conserved regions of unknown function or structure. In addition, independence of nucleotide-level conservation implies that the information from substitution-based measures can further substantiate hypotheses of conservation. The method will be refined in several ways, e.g. by using local predictions of indel rates to account for non-uniformity, or by extending to multiple genomes. This, and combining indel and substitution modelling, will further greatly increase the sensitivity of this conservation confidence estimate.

C ANALYSIS OF CONSERVED ELEMENTS

The methodology detailed above will provide a set of conserved elements, at a controlled and low false-positive rate. Compared to other methods for measuring conservation, this set will be enriched with elements that show low conservation at the sequence level, and thus provides a unique opportunity to study these elements.

Several aspects of this dataset will be analysed. First, to ascertain that the majority of elements is indeed functional, we will look for independent confirmation, such as alignment to divergent species as pufferfish and chicken, and the presence of sequence-level conservation. Even though each is in itself not sufficient to establish conservation per element, on the entire dataset it can give conclusive support. Second, we will investigate the genomic distribution of the conserved elements found. In particular, the distribution with respect to annotated genes will be investigated. It is expected that protein-coding exons will be heavily represented, followed by UTRs, and regions proximal to 5' and 3' gene ends. The extent to which the density of established conserved elements decreases with increasing distance to genes will be a relevant indicator of the average extent of promoter regions.

We will also investigate the density of conserved elements in sub-telomeric regions. A preliminary scan revealed that elements with low sequence-level conservation are particularly enriched in these regions. Such elements have not yet been systematically characterised, and we will make a first attempt to do so. First observations seem to indicate a striking prevalence of tandem repeat structures. These structures align to chicken and pufferfish, with average sequence identity indistinguishable from neutral sequence, but with identical repetitive structure. Although the existence of such structures in sub-telomeric regions is widely known, their evolutionary conservation (with respect to indels, not substitutions) has not been noted before.

D WHOLE-GENOME MULTIPLE STATISTICAL ALIGNMENTS

Any method in comparative genomics is critically dependent on an accurate alignment. Recent advances in alignment methods are therefore highly relevant for this project.

On a coarse scale, this goal has been mostly accomplished. For instance, human and mouse are relatively divergent mammalian species, but conservation of large-scale synteny means that anchors for local multiple alignments can be identified with confidence. Alignments are then built from these using a technique called *progressive alignment* [Brudno et al. 2003, Kalafus et al. 2004]. For the preliminary analyses referred to above, we used publicly available BlastZ multiple whole-genome alignments [Schwarz et al. 2003]. The BlastZ procedure takes great care to resolve the many large-scale ambiguities due to transposable element repeats and segmental duplications to obtain a reliable high-level alignment scaffold.

On the nucleotide scale however, the reliability of these alignments are unclear at best [Ellegren et al. 2003].

Uncertainties in gap placement mean that alignment confidence diminishes near gap boundaries. This has considerable impact on e.g. annotation of transcription factor binding sites, which usually reside in regions that are hard to align. Moreover, progressive multiple alignment algorithms have the weakness of “freezing in” early alignment decisions, biasing subsequent alignments higher up in the phylogenetic tree. Since the true homology can never be determined with certainty, it is imperative to understand and quantify the statistical aspects of alignments. We recently have made considerable progress in the development of models and practical algorithms that deal with this probabilistic aspect of alignment [Lunter et al. 2003, Lunter et al. 2004]. Incorporating the residual uncertainty by weighing alignment variants according to their posterior probability removes an important source of noise and bias in subsequent analyses, and increases sensitivity. Instead of a single alignment, the output of such procedures is a distribution over all possible alignments. In practise, one still wants a single “best” alignment, but such alignments can be annotated with per-column reliabilities, significantly improving the usefulness of such alignment “point estimates”. In more sophisticated analyses, several alignments sampled from the posterior distribution will be used, effectively mitigating alignment biases, while not significantly complicating the downstream analysis.

Besides accounting for uncertainties, improvements to the core homology model on which the alignment procedure is based remains a goal. A problem with this is that it is very difficult to assess improvements, as in particular for neutral alignments, no “gold standard” alignment is available. However, for large datasets, alignment quality can be measured in a statistical way. The gap attraction phenomenon mentioned above (Section B) is unavoidable, but the extent to which it occurs is related to the accuracy of the underlying homology model. The gap attraction phenomenon is caused by the fact that the maximum likelihood and true alignments do not generally coincide. On alignments, this results in nearby (true, i.e. indel-associated) gaps “attracting” each other and merging. The observed shortage of short intergap distances in the alignment, compared to the neutral model prediction, thus serves as a proxy for alignment accuracy, and gives an objective standard against which to evaluate different homology models.

E SOFTWARE DEVELOPMENT

Having identified the database of conserved elements a key output of this project, it is imperative that its is easily accessible. To this end, a web server will be developed to allow simple access to the database. The entire database will also be made available as a structured XML document, in order to combine a detailed and extensible annotation with the possibility of simple parsing using widely available tools.

Secondly, we will make available the re-alignment procedures mentioned above. These will include an optimisation method based on the gap-attraction phenomenon mentioned before.

Besides these end-user products, we will also develop software tools for more fundamental research purposes. The alignment procedures alluded to in section C, and indeed many of the models encountered in bioinformatics, are implemented as hidden Markov models (HMMs). HMMs are versatile and relatively straightforward tools, but implementing the various algorithms is often time-consuming and error-prone, especially for complicated models when efficient algorithms are required, making it hard to explore model space. We are currently developing a compiler for converting a high-level XML description of an HMM into efficient C++ code, which is currently being used by several students in the group. Although written with computational biology in mind, it is very general and not tied to any particular application. Models for RNA secondary structure use a formalism known as Stochastic Context Free Grammars (SCFGs), which are generalisations of HMMs. The complexity of SCFG-related algorithms makes a compiler-based approach particularly relevant for these models, and the extension of the compiler to SCFGs will be a particular priority.

5 Statement of Timeliness

Comparative genomics is a rapidly developing field, and will continue to be so in coming years, with the increasing number of available mammalian genomes. Ad-hoc strategies were successful for initial analyses, but failed to reliably identify evolutionary conserved elements in these genomes, with the notable exception of protein-coding genes. The methodology outlines in this proposal will make a significant contribution to this effort, and will help to locate such diverse elements as microRNAs and transcription factor binding sites, which currently attract a great deal of interest in the field.

6 Justification of Resources

The research will be carried out by a postdoc, who will be fulltime dedicated to this project for one year. The intended researcher has considerable experience and will need minimal supervision apart from weekly progress meetings. The proposed research requires that whole-genome analyses be carried out routinely, which justifies the request for a large-memory two-processor compute server with high-capacity hard drive storage. The results of the research will be presented at UK (e.g. MASAMB) and international conferences (e.g. RECOMB, ECCB or ISMB).

7 References

- Stein LD (2004)**, Human genome: End of the beginning. *Nature* 431, 915 - 916
- Rodriguez-Trelles F (2004)**, Transcriptome evolution - much ado about nothing?, *Heredity* 93, 405-406
- Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansong W, Pääbo S (2004)**, A Neutral Model of Transcriptome Evolution, *PLoS Biol.* 2(5) 0682-0689.
- Gaffney DJ, Keightley PD (2004)**, Unexpected conserved non-coding DNA blocks in mammals *TRENDS in Gen.* 20(8) 332-337
- F. Chiaromonte, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D. (2004)**, The share of human genomic DNA under selection estimated from human-mouse genomic alignments., *The Genome of Homo Sapiens Vol. LXVIII*, 245–254, Cold Spring Harbor Press, Cold Spring Harbour, New York.
- Antonarakis SE (2003)**, CpG Dinucleotides and Human Disorders, *Enc. Human Gen.* vol. 1, Nat. Publ. Group
- Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE (2002)**, Numerous potentially functional but non-genic conserved sequences on human chromosome 21, *Nature* 420, 578-582
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004)**, Ultraconserved elements in the human genome. *Science* 304, 1321-1325
- NHGRI (2004)**: Genome Sequencing Proposals, <http://www.genome.gov/Research/>
- GOLD (2004)**: Genome On Line Database, <http://www.genomesonline.org/>
- Burge CB, Karlin S (1998)**, Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8, 346-354
- Rivas E, Eddy SR (2000)**, Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs, *Bioinformatics* 16:573-585
- Rivas E, Eddy SE (2001)**, Noncoding RNA gene detection using comparative sequence analysis, *BMC Bioinf.* 2:8
- Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, de Jong PJ, Botas J, Rubin GM, Celniker SE (2002)**, Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome, *Gen. Biol.* 2002, 3(12):research0086.
- Lunter GA (2004)**, Indels are Informative for Identifying and Quantifying Conserved Genomic Elements (Poster), Identification of Functional Elements in Mammalian Genomes meeting, Cold Spring Harbour.
- Mouse Genome Sequencing Consortium (2002)**, Initial sequencing and comparative analysis of the mouse genome, *Nature* 420, 520-562
- International Human Genome Sequencing Consortium (2004)**, Finishing the euchromatic sequence of the human genome, *Nature* 421, 921-945
- Brudno M, Do C, Cooper G, Kim MF, Davydov E, Reen ED, Sidow A, Batzoglou S (2003)**, LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA, *Genome Research* (4), 721-731
- Kalafus KJ, Jackson AR, Milosavljevic A (2004)**, Pash: Efficient Genome-scale Sequence Anchoring by Positional Hashing, *Gen. Res.* 14, 672-678
- Schwarz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003)**, Human-Mouse Alignments with BlastZ, *Gen. Res.* 13:103-107.
- Ellegren H, Smith NGC, Webster MT (2003)**, Mutation rate variation in the mammalian genome, *Curr Opin Gen Dev* 13:562-568.
- Lunter GA, Miklós I, Song YS, Hein J (2003)**, An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.*, 10(6):869-889
- Lunter GA, Drummond AJ, Miklós I, Hein J (2004)**, Statistical Alignment: Recent Progress, New Applications, and Challenges, in: Rasmus Nielsen (ed.), "Statistical methods in Molecular Evolution", Springer Verlag.