Published in final edited form as:

Methods. 2009 September ; 49(1): 50–55. doi:10.1016/j.ymeth.2009.05.010.

Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships

Eric J. Vallender

Division of Neurosciences, New England Primate Research Center, Harvard Medical School, Southborough, MA 01772, USA

Abstract

Non-human primate genetic research defines itself through comparisons to humans; few other species require the implicit comparative genomics approaches. Because of this, errors in the identification of non-human primate orthologs can have profound effects. Gene prediction algorithms can and have produced false transcripts that have become incorporated into commonly used databases and genomics portals. These false transcripts can arise from deficiencies in the algorithms themselves as well as through gaps and other problems in the genome assembly. Putative genes generated can not only miss microexons, but improperly incorporate non-coding sequence resulting in pseudogenes or other transcripts without biological relevance. False transcripts then become identified as orthologs to established human genes and are too often taken as gospel by unwary researchers. Here the processes through which these errors propagate are isolated and methods are described for identifying false orthologs in databases with several representative errors illustrated. Through these steps any researcher seeking to make use of non-human primate genetic information will have the tools at their disposal to ascertain where errors exist and to remedy them once encountered.

1. Introduction

Implicit in all studies utilizing comparative genetics, non-human primate or otherwise, is the belief that the gene under study in one species is functionally and/or evolutionarily related to that in another species. This means successfully identifying and differentiating orthologs and paralogs. In mammals generally and non-human primates particularly, this has been a traditionally straightforward task. When gene sequences were being identified on a case-by-case basis, a successful cloning of a gene from another species using cross-species primers often gave all the information one needed to be fairly confident in the evolutionary relationships between the two. When problems arose they were generally immediately obvious and one could be confident that at least one set of eyes had looked at an alignment between the genes from the two or more species. With our entry into a genomics era, however, many of these old controls and safeguards are gone; no longer do human eyes necessarily sanity check putative orthologs and the errors of today are less easily recognized. As the importance of true species-typical sequences grows and our reliance upon genomically-derived gene sequences increases, we must take extra care.

Address correspondence to: Eric J. Vallender: eric_vallender@hms.harvard.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Identifying genes across species is generally accomplished through homology searches. Homologs are genes which share sequence identity, but the term homolog offers no comment on evolutionary history, though homology between genes generally connotes some evolutionary relationship [1]. Orthologs, genes which have originated from a single gene in a shared ancestor, are the most commonly sought homologs and are easily identified heuristically for single-copy genes. Rates of identity between orthologs correlate directly with time to the last common ancestor and usually orthologs are syntenic between species, sharing common flanking regions and context within chromosomes. Paralogs, on the other hand, are the result of duplication events and divergence between them is not necessarily correlated with speciation. Shared synteny is lost between paralogs.

While a common problem when dealing with distantly related species or prokaryotes, the differentiation between orthologs and paralogs within primates or mammals is generally not difficult, especially when complete genomes offer information on synteny. With the exception of certain gene families such as the olfactory receptors, duplication events are fairly uncommon and where they do occur their evolutionary history has not been mottled by the confounding multiple hits that plague distant branches. Simply counting the number of times a gene appears in one primate genome compared to another is often enough to presume orthology.

When considered in the abstract primate orthology is simple. Given a complete genome or transcriptome assigning evolutionary relationships is straightforward. Indeed, it is this basic conceit that has led to the identification of numerous orthologous groups among publicly accessible databases. There is a common assumption that because the theory behind their identification is simple, computer-aided identification of orthologous clusters of genes is infallible. This is not the case, however. Algorithms for identifying orthologs are dependent upon the sequence input and here there can be, and are, numerous failures. False negatives can occur when researchers fail to find an ortholog (or worse, only find an apparent pseudogene) and false positives are common due to alternative splicing or errors in gene prediction.

To illustrate broadly these difficulties, one needs look no further than the recent papers on the completion of the genomes themselves. The human genome, by far the most “finished” of the mammalian genomes to date, appears to harbor between 20,000 and 25,000 genes [2]. While it is certain that some portion of these will be either very recent, human-specific, duplicates, or otherwise of sufficient youth to make orthologs and paralogs difficult to dissociate, 1:1 orthologs for only two-thirds of these genes (13,454) were able to identified in the chimpanzee genome [3]. When the rhesus macaque genome was unveiled only half (10,376) of the human genes had 1:1:1 human-chimpanzee-rhesus orthogous trios [4]. Indeed, the rhesus genome assembly recognizes these failings estimating over 2,000 orthologous trios each being discarded as a result of flaws in the chimpanzee and rhesus assemblies. Conservatively, they report that finishing the genomes of chimpanzee and rhesus would increase the number of identified orthologous groups by 23% [4]. While, a recent study was able to update the number of 1:1:1 human-chimpanzee-rhesus orthologous trios to ~14,000 [5], one third of human genes still don't demonstrate chimpanzee and macaque orthologs. While the decrease as additional species are added is to be expected, it is perhaps noteworthy that the addition of the chimpanzee to established human-rhesus-mouse-rat quartets results in a loss of roughly 20% of orthologous clusters [4]. It seems biologically implausible that one in five genes conserved across humans, old-world monkeys and rodents would be lost in the chimpanzee.

These numbers are pointed out not to highlight inadequacies in the current available genomes. Rather, they suggest that the results of automated ortholog discovery approaches that manifest in commonly used databases must necessarily act as a starting point for researchers. Too often these orthologs are viewed as an endpoint, particularly among those uncomfortable or without grounding in genetics or bioinformatics. Below, some common problems are cataloged and

bioinformatic methods are offered which will help researchers more easily and confidently assess orthology for their genes of interest, ensuring that downstream studies are built upon sturdy premises and avoiding unneeded delays.

2. Gene Prediction

Many of the problems in ortholog identification arise, either directly or indirectly, from difficulties in gene prediction (also known as genome annotation). Because of this, it is important to understand how these methods work and why these errors arise. Three approaches are typically used, increasingly in combination, to identify and annotate genes in a genome: cis alignments, trans alignments, and de novo predictions. Cis alignments, which align cDNA sequences with genomic sequences, remain the gold standard. The primary difficulty with this method is the time and cost necessary to develop comprehensive cDNA libraries. This drawback is particularly exacerbated in non-human primates for whom fewer cDNAs are available than many model organisms, but if cDNAs are available for a gene, often this is sufficient to establish orthology without relying on genomic sequence. Further, the relatively close relationships between non-human primates and humans allow for better implementation of other methodologies, specifically trans alignments and de novo predictions.

2.1 De novo gene prediction

De novo gene prediction methods use only the genomic sequence and patterns regarding transcriptional initiation and termination and splice sites. These methods were pioneered early on in the genomic era when few genomes were available. Reviewed elsewhere [6], de novo gene prediction programs have become significantly more complex and powerful over time. The first of these programs, GENSCAN, debuted in 1997 and is still one of the two most widely used de novo prediction programs [7]. More recently, N-SCAN successfully incorporated multiple genomic sequences in order to improve these de novo predictions and has rapidly gained in use [8]. N-SCAN gene predictions are commonly available on the UCSC web browser (<http://genome.ucsc.edu>) [9] and have been used in the annotation of the rhesus genome. Though de novo gene prediction tools are becoming more reliable they are still far from perfect. While true de novo methods are still widely used among organisms for which no close relative is available, in species with annotated sister taxa such as mammals they have merged with trans alignment methods with corresponding increases in predictive power.

2.2 Trans alignments

Trans alignment methods use sequences derived from other species, generally cDNAs or annotated genomic sequences, to identify orthologous loci in the genome of another species. As ever greater numbers of genomes are being sequenced and transcriptomes, particularly in humans and mice, become more complete, these methods have gained in use as effectiveness of trans alignments is directly related to the level of homology between the bait cDNAs and the prey genomes [10]. Among mammals, human-mouse divergence for coding sequences is roughly 30% [11], while human-dog divergence in coding regions is in the range of 15-20% [12]. Within primates, human-chimpanzee cDNA divergence is expected to be less than 0.5% [3,13], human-rhesus (and all old world monkeys) cDNA divergence less than 2.5% [4], and even human-new world monkey cDNA divergence will likely fall in the range of 5%-10% when all is said and done. While it remains unclear to what level true gene gain and gene loss events (or more appropriately transcript gain and loss) occur, it seems likely that trans alignment methodologies will prove particularly successful in primates. In fact these approaches are integrated into many of the gene prediction pipelines currently in use [14,15].

2.3 Gene prediction in practice

As previously mentioned, the largely de novo prediction algorithm implemented by N-SCAN was used for the annotation of the rhesus genome. Also used were NCBI's Gnomon models [14] and the ENSEMBL annotation pipeline [15]. Both of these methods use human sequences and macaque ESTs to seed their algorithms, and represent the hybrid between de novo methods and trans alignment methods that are becoming more common. When these three programs were used to annotate the rhesus genome they identified roughly the same number of loci as were found in the human genome, though an additional 25% of genes thought to be false positives were also identified by at least one of the methods [4]. Beyond their use in the annotation of the rhesus genome, these programs also form common entry points into non-human primate genetic resources for the majority of researchers, whether under their own aegis or through other portals such as HomoloGene [16], the UCSC Genome Browser [9], or UniRef [17].

There are many gene prediction programs available, some more complicated and inaccessible to the casual user than others. A recent study, the ENCODE Genome Annotation Assessment Project, aimed at assessing the overall reliability and status of these efforts [18]. This project focused on using computational methods to annotate the ENCODE regions of the human genome and to compare the different methods to the known reference set of annotations. The results were mixed. While some programs achieved as high as 90% sensitivity and specificity for identification of coding nucleotides, there still remain significant deficits even in the most highly refined mammalian genome. Notably, there are significant difficulties in the identification of alternative splice forms and delineating non-coding regions (such as 5' and 3' untranslated regions) is extremely unreliable. It should be noted, however, that even in the four years since this study was concluded significant progress has, and continues to be made.

With regard to non-human primate researchers, these findings should send a clear message. While the programs and methodologies that are used to identify genes in non-human primates are largely reliable, and while non-human primate annotations are likely to be the best in genomes sequenced to comparable depths, there will still remain significant errors and misannotations. Databases remain strewn with incomplete data sets and false positives. It is important that end users of this information realize this whether they be focused on single genes or on large scale genomic comparisons.

3. Identifying Orthologs

Ortholog identification is accomplished primarily through homology and genic synteny. Given complete information, identification of orthologs is relatively straightforward, especially in species as closely related as primates and which harbor relatively few genomic rearrangements. Therein, however, lies the rub. Correct orthology, especially when determined through automated computational methods, requires complete information. In species that rely largely on predicted genes there is a greater likelihood that missing information exists, rendering ortholog identification incomplete or worse. These errors can take two forms. Misorthology, or the false identification of evolutionary relationships between homologs, is traditionally what bioinformaticians worry about when assessing orthology. While noteworthy problems exist, this actually represents a fairly small proportion in non-human primates. Issues of xenology caused by horizontal gene transfer are not present, and convergent evolution at the sequence level simply has not had the requisite time to occur. Gene gain and loss is uncommon enough outside of large gene families such as the olfactory receptors to make inference fairly straightforward when it does occur. Further, the close relationship between the species renders differences in methodologies largely moot.

Traditionally, classification of orthology is done through either a reciprocal best hit [19] or reciprocal smallest distance [20] method. The basic procedure entails collecting all the genes in two species and comparing them all to one another. If genes from two species identify each other as their closest partners then they are considered orthologs. This can be, and is, a major problem in highly divergent species [21,22]. Recently, studies have attempted to compare different databases of putative orthologs [23] including HomoloGene [16], ENSEMBL [15], and OMA [24]. Similar to studies of gene prediction methodologies, the findings are mixed. As expected, the different methodologies result in differing degrees of success. There is a tradeoff between specificity and coverage. They will both produce the same results in very closely related species.

The more common, and pernicious, problem of false orthology in primates arises as a result of incomplete or poor genome annotation. Two different alternative splice forms of the same gene can be, and often are, identified as orthologs when the corresponding alternative splice forms are not present in the other species. When this occurs, correctly orthologous genes are often paired as falsely orthologous transcripts and regions of the gene, usually specific exons, are incorrectly identified as orthologous when they may not even share homology let alone common descent. In non-human primate genomes this is particularly noteworthy in alternative first and last exons, which gene prediction programs struggle with at greater levels than internal exons. Internal micro exons that often evade gene prediction algorithms can also be a problem, but usually resulting in false negatives.

4. Sources of Errors

In the context of non-human primate genetics there are three primary sources of error that must be addressed: routine errors in gene identification, gap-introduced errors, and sequencing errors. Routine errors are not unique to non-human primates; indeed they plague many of the genome projects. Errors introduced by gaps in sequencing or assembly are again not limited to non-human primates but are more common in genomes with low coverage. While prices for genomic sequencing are coming down, current cost saving measures have increasingly favored low coverage genomes. It is likely that problems thus introduced will continue to affect non-human primate genome projects going forward. Lastly, it is important to consider the effects of sequencing error in close species. This is particularly true across the great apes. Divergence between human and chimpanzee in particular has been a major focus and one that may suffer the greatest from errors in any base calling.

4.1 False transcripts

Both routine gene identification errors and gap-introduced errors create orthology issues through the same basic mechanism. False transcripts are created which then are identified as orthologous to actual human transcripts. While the orthology of the gene is largely correct, non-orthologous exons may be introduced causing problems. In the case of true alternative transcripts the exons themselves are biologically meaningful but are simply not orthologous. Given that the exonic structure of only 50-60% of human genes is thought to be correct [6, 18] and that these human annotations form the basis of the other non-human primate gene predictions, numerous problems arise. This is particularly troublesome when incorrect human annotations are coupled with conservation based methods for gene detection. False exons in the humans are not likely to harbor above average conservation in other species, leading to a “correct” failure to predict a false exon but also failing to identify the orthologous region. These errors can be particularly misleading for genome-scale analyses where the logical and biological consistency of individual genes is not considered. Recent computational methods have begun to address the identification of these mispredicted proteins through the application of protein dogma to databases [25], but while these approaches will help clean up the databases, they will be insufficient to entirely perfect them.

Gap-introduced predicted alternative splice forms also introduce false orthology, but equally or perhaps more problematic, the transcripts are not biologically meaningful either. It is not difficult to find examples of these kinds of errors in non-human primate sequences; one particularly illustrative example is given in Figure 1. The exonic structure of human *SIX3* is fairly well known [26], yet the putative orthologs in both the chimpanzee and rhesus macaque are likely false. Focusing first on the 5' end of the gene, gaps in the chimpanzee and rhesus genomes each separately result in a missing initiation codon and early amino acids. Recognizing that genes exist here, prediction programs attempt to find alternative 5' exons to replace the missing regions. Without context the result is false putative orthology for the 5' end of the gene, though placed in a genomic framework the problems are obvious.

Coincidentally, the 3' end of the rhesus gene also contains problems. The rhesus prediction was unable to identify the ortholog to the second human coding exon. Looking closely at the genome it is not simple to determine why this should be the case, though a gap or masked repeat may have eliminated a splice junction. Instead the rhesus prediction offers a continuation of the first coding exon into the intron. Interestingly, and perhaps informatively as to why this transcript was predicted, a similar run-on transcript is predicted in the mouse [27], though there is no evidence it exists in humans [26]. Regardless of whether this transcript represents a true alternative splice form in rhesus monkeys or not, it is the only one present in the database while the other is the only one present in the human. Again, alignments between the two putative orthologs will align the intronic run-on portion of the rhesus gene with the human second exon and suggest false orthology.

4.2 Sequencing errors

There remain also more fundamental sources of error. Sequencing errors, or more properly base calling errors, can be very difficult to detect. Indeed unless the error introduces a frame shift mutation it is unlikely to be identified. When it does introduce a frame shift mutation, however, it often will result in the prediction of a pseudogenization event, or what appears to be a highly divergent 3' end. Fortunately the raw trace files are available and with enough effort these errors can be rooted out. The rhesus macaque orthodenticle homeobox 2 gene (*OTX2*) is identified as a pseudogene and indeed perusing the genome will identify a single base insertion and corresponding frame shift. The function of this gene however is crucial to early brain development with knockout mice showing early embryonic lethality and a complete absence of forebrain and midbrain [28]; it seems unlikely that the gene should be absent from rhesus macaques. Indeed, when the trace files used to generate the rhesus genomic sequence are considered the miscall is obvious and a correct sequence, without frameshift, easily deduced (Figure 2).

Lest one believe that these are isolated incidents, it is worth noting that in a recent study 8% of chimpanzee genes in a fairly tightly controlled set of orthologs showed premature stop codons [29]. An earlier study using similar methodology found roughly three times the number of premature stop codons in chimpanzees as compared to humans [30]. The quality control undertaken in these studies to ensure the validity of the stop codons is rigorous, requiring high quality base calls around the stop codon itself, and many conclusions certainly are biologically plausible (for instance relatively higher rates of pseudogenization of olfactory genes and among genes on the Y chromosome). It is still worth noting, however, that the rhesus *OTX2* pseudogene would successfully pass through the controls as the miscall, and poor quality sequence, may be sufficiently far from the site of the stop codon to escape detection.

While discussing sequencing error it is important to note that the effects of sequencing error are particularly pronounced in close species comparisons such as those between humans and chimpanzees. Neutral divergence between humans and chimpanzee is roughly 1% or one mutation every hundred sites. Polymorphisms in chimpanzees occur at roughly 0.1%. A study

of the working draft of the chimpanzee genome suggested a sequencing error rate of 0.07% [31]. As the authors correctly note, that error rate will certainly be sufficient to ensure the validity of most gross comparisons, but when specific differences may not represent true biological importance. The authors suggest that in their comparison 25% of putative amino acid changing mutations between humans and chimpanzees are the result of false positives. As genomic studies comparing humans and chimpanzees proliferate, this must be kept in mind. More distant comparisons and higher quality sequence will be drowned out the noise of sequencing errors by the signal of fixed differences, but even a 5% false positive rate may corrupt biological meaning. The combination of short evolutionary distances and low quality sequence should make any researcher carefully consider the implications of false positives in their studies going forward.

5. Maximizing Utility

While the problems arising from failures during the genome assembly, gene annotation, and ortholog identification processes are not unique to primate genomes, their effects may be particularly pronounced in non-human primates given the close evolutionary relationships between the species under study. Yet while this relatively high degree of conservation makes the effects of errors greater it also makes it easier to identify the errors and eliminate them.

There are several steps that can be used in order to ensure that the sequence being used is appropriate and the best that it can be. First one needs to err on the side of parsimony. While it is possible that genes will become inactivated, even genes with apparently important functions, often the most parsimonious explanation is an error in the annotation. Similarly, while it is possible that exons have been lost over the course of evolutionary time and replaced with novel exons or previously non-coding intronic sequences, it is logical to first ensure that the annotation process was correct. Lastly, localized regions of high mutation may be the result of strong positive selective forces, but more often if they lack apparent homology they are in fact non-orthologous. This is particularly noteworthy when these regions of apparent non-homology also contain frameshifts or stop codons (Figure 3).

There are numerous tools that can be used to identify regions of improper alignment and orthology, indeed any program that will alert the user to regions of non-conservation. Dot plots can be useful for detecting errors in orthology arising from either alternative splice forms, whether “real” or “fake”, incorrectly labeled as orthologous and for identifying missing exons. This can be particularly useful for large, multi-exonic genes, for which genome alignments are less intuitive. Programs such as PipMaker [32] and MultiPipMaker [33] allow the user to visualize alignments easily and identify regions of potential error for further study. In Figure 4 example output is shown for the human, chimpanzee, and rhesus macaque ribosomal protein S6 kinase 3 gene (*RPS6KA3*). Gaps along the diagonal, roughly one quarter in for the human-chimpanzee comparison and at roughly half in for the human-rhesus comparison, indicate regions of low homology. In this context they are regions of misorthology, specifically internal exons without relationship between the species. The important distinction to be made when using these programs comes in the identification of when low homology is beyond what is to be expected by chance and thus represents true misorthology rather than simply high divergence. While there are analytical methods for determining what constitutes a significant departure from null expectation [34], in reality two sequences sharing no common descent, even if selected for relative homology, will be immediately obvious in the context of primate high conservation levels.

For the vast majority of scientists using non-human primates, studies and research will be limited to a handful of genes of interest. In these situations the search for non-human primate orthologs must extend beyond simple database searches. The genes' exonic structure across

species should be compared and, where feasible, trace files examined. The latter is especially true of comparisons within the great apes where even low levels of sequencing error can have major effects. Further, while it may not always be possible, researchers are strongly encouraged to make use of cDNA sequences rather than predicted sequences from genomes. These sequences have likely had greater, if uneven, curation and are much more likely to represent real *in vivo* transcripts. They are also more likely to accurately represent some, if not all, of the untranslated regions.

For researchers focused on genome-wide comparisons, many of the labor intensive, manual approaches will be untenable. Nevertheless automated methods must be used to minimize errors. It is important that quality control checks be appropriately implemented and rigorously defined. While nearly all of the genomic studies on the human, chimpanzee, rhesus and other genomes produced have used quality control algorithms and methodologies, often these quality control assurances are hidden in supplementary material or glossed over in methods. Not only do researchers need to make quality control pipelines as transparent as possible, but readers need to actively seek out this information and assess not only its adequacies, but also its implications for the data presented. Both reader and author should consider the effects on the study if wrongly identified orthologs or orthologous regions are present. Culling sequences using base quality scores is common and necessary, especially for closely related species, yet quality control should not end there. Dot plot or sliding window analyses of orthologous pairs, at a minimum, appear to be required to identify regions of non-homology. While ideally these should then be reconstructed correctly, at a minimum they should be masked in future analyses. The implementation of programs designed to identify statistically improbable regions of clustered mutations can also be used to flag worrisome orthologs.

It is also important to consider the input source one is dealing with. Predicted RefSeq sequences (beginning with “XM” or “XP” designations) are less reliable than curated sequences (“NM” or “NP”) from the same source. HomoloGene, who uses these sequences to identify orthologous groups, also offers divergence values and BLAST alignments that can be used to identify potential sources for problems. Notably, significant changes in gene length or divergence values deviating significantly from the expectation (chimpanzee-human divergence greater than mouse-human divergence, for example), are identifiable. Generally speaking the gene prediction and ortholog identification in Ensembl is more refined and better. Again, alignments and divergence values are available for quick considerations while genomic sequences are also available. OMA, which makes use of annotations from several sources, also does a good job in avoiding many of the false positives that plague HomoloGene, but the presentation can make them more difficult to identify when present.

6. Concluding Remarks

Genomic science has come a long way in a very short time. Computational and bioinformatic tools are still struggling to keep up with the vast quantities of data being produced. While these tools continue to improve, they are not perfect nor can they adequately cope with errors in sequencing, low coverage, or poor assemblies. While several notable efforts have been made to improve the annotation of genomes and to ensure that genomic information is at its most efficacious it is clear that the recent trend toward bang-for-the-buck genomics has been low coverage genomes of many species. This emphasis has put the onus on researchers, bioinformaticists, geneticists, and all scientists relying on genomic information for even a single gene to implement better quality control methods and to ensure that the information being taken from the genome is correct, no more and no less.

It seems inevitable that as genome sequencing prices come down these problems will resolve themselves. When high quality, deep coverage of multiple individuals from a species is

available annotation will necessarily improve and with it ortholog identification. And while it seems that this day is approaching more rapidly than many believed, there will still be a significant time for which the *post hoc* analysis methods described here will be the only way to approach the questions that arise from the current state of genomics. Researchers focusing on non-human primate genetics in particular need to be aware of these issues. Not only are they more likely to arise in the field, where almost implicitly every gene will be compared directly or indirectly to human, but their impact is likely to be at its greatest.

Acknowledgments

This work is supported by grants from the National Institutes of Health (MH082507 to EJV) and the National Center for Research Resources (RR00168).

References

1. Koonin EV. Annu Rev Genet 2005;39:309–38. [PubMed: 16285863]
2. Finishing the euchromatic sequence of the human genome. 2004;431:931–45.
3. Consortium CG. Nature 2005;437:69–87. [PubMed: 16136131]
4. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, et al. Science 2007;316:222–34. [PubMed: 17431167]
5. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. PLoS Genet 2008;4:e1000144. [PubMed: 18670650]
6. Brent MR. Nat Rev Genet 2008;9:62–73. [PubMed: 18087260]
7. Burge C, Karlin S. J Mol Biol 1997;268:78–94. [PubMed: 9149143]
8. Gross SS, Brent MR. J Comput Biol 2006;13:379–93. [PubMed: 16597247]
9. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. Nucleic Acids Res 2009;37:D755–61. [PubMed: 18996895]
10. Mathe C, Sagot MF, Schiex T, Rouze P. Nucleic Acids Res 2002;30:4103–17. [PubMed: 12364589]
11. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. Nature 2002;420:520–62. [PubMed: 12466850]
12. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. Nature 2005;438:803–19. [PubMed: 16341006]
13. Chen FC, Vallender EJ, Wang H, Tzeng CS, Li WH. J Hered 2001;92:481–9. [PubMed: 11948215]
14. Pruitt KD, Tatusova T, Maglott DR. Nucleic Acids Res 2007;35:D61–5. [PubMed: 17130148]
15. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al. Nucleic Acids Res 2007;35:D610–7. [PubMed: 17148474]
16. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. Nucleic Acids Res 2009;37:D5–15. [PubMed: 18940862]
17. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. Bioinformatics 2007;23:1282–8. [PubMed: 17379688]
18. Guigo R, Flicke P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, et al. Genome Biol 2006;7:S2, 1–31. [PubMed: 16925836]
19. Moreno-Hagelsieb G, Latimer K. Bioinformatics 2008;24:319–24. [PubMed: 18042555]
20. Wall DP, Deluca T. Methods Mol Biol 2007;396:95–110. [PubMed: 18025688]
21. Koski LB, Golding GB. J Mol Evol 2001;52:540–2. [PubMed: 11443357]
22. Wall DP, Fraser HB, Hirsh AE. Bioinformatics 2003;19:1710–1. [PubMed: 15593400]
23. Altenhoff AM, Dessimoz C. PLoS Comput Biol 2009;5:e1000262. [PubMed: 19148271]
24. Schneider A, Dessimoz C, Gonnet GH. Bioinformatics 2007;23:2180–2. [PubMed: 17545180]
25. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L. BMC Bioinformatics 2008;9:353. [PubMed: 18752676]
26. Granadino B, Gallardo ME, Lopez-Rios J, Sanz R, Ramos C, Ayuso C, Bovolenta P, Rodriguez de Cordoba S. Genomics 1999;55:100–5. [PubMed: 9889003]

27. Oliver G, Mailhos A, Wehr R, Copeland NG, Jenkins NA, Gruss P. Development 1995;121:4045–55. [PubMed: 8575305]
28. Acampora D, Mazan S, Lallemand Y, Avantiaggiato V, Maury M, Simeone A, Brulet P. Development 1995;121:3279–90. [PubMed: 7588062]
29. Wetterbom A, Gyllensten U, Cavelier L, Bergstrom TF. BMC Genomics 2009;10:56. [PubMed: 19178713]
30. Wetterbom A, Sevov M, Cavelier L, Bergstrom TF. J Mol Evol 2006;63:682–90. [PubMed: 17075697]
31. Taudien S, Ebersberger I, Glockner G, Platzer M. Trends Genet 2006;22:122–5. [PubMed: 16406850]
32. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W. Genome Res 2000;10:577–86. [PubMed: 10779500]
33. Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, Green ED, Hardison RC, Miller W. Nucleic Acids Res 2003;31:3518–24. [PubMed: 12824357]
34. Wagner A. Genetics 2007;176:2451–63. [PubMed: 17603100]

Figure 1.

Schematic genome alignment of the putative *SIX3* orthologs in human, chimpanzee, and rhesus macaques. Coding sequences are shown in dark grey with untranslated regions in light gray. Gaps in the genomic sequence in chimpanzee and rhesus are illustrated by black bars. Sequences from RefSeq [14], genomic depiction adapted from UCSC browser [9].

Orthodenticle homeobox 2 (*OTX2*)

Figure 2.

Relevant partial trace files for human, chimpanzee, and rhesus macaque *OTX2* orthologs. RefSeq identifiers for the sequence are given as are trace identifier (TI) numbers. A one base pair insertion and missense mutation are introduced in the rhesus sequence resulting eventually in an incorrect call of a pseudogene. Sequences from RefSeq [14].

Asparagine-linked glycosylation 3 (ALG3)

Human	...CACAGGTTGGCAT—ACACAGAGATTGACTGGAAGGCCTACATGGCCGAGGTAGAAGGCGTCATCAATG—GTACCTAT...	NM_005787
Chimp	...CACAGGTTGGCAT—ACACAGAGATTGACTGGAAGGCCTACATGGCCGAGGTAGAAGGCGTCATCAATG—GTACCTAT...	XM_516912
	H R V A Y T E I D W K A Y M A E V E G V I N G T Y	
Rhesus	...CACAGGTTGGCAT—GTGAAGAGGTACAGTGGTGAATAAGAGATGCATTTCTAACTCCGATTGAAGGAGTCAGCAAAG—GTACCTAT...	XR_012354
	H R V A C E E V Q W * I R D A F L T P I E G V S K G T Y	

Figure 3.

Nucleotide alignment with translation of a wrongly included exon in the rhesus macaque *ALG3* gene. Solid black lines in the sequence indicate exonic boundaries. Sequences from RefSeq [14], ortholog identification through HomoloGene [17].

Figure 4.

Dot plots showing alignments between the coding sequences of human (X-axis) and chimpanzee (Y-axis top) or rhesus macaque (Y-axis bottom) *RPS6KA3* orthologs. Gaps in the diagonal indicate regions of low homology between the species. A simplified one dimensional depiction of the alignment is shown below the dot plot. Sequences from ENSEMBL [15], alignment and figure through PipMaker [32].