

RESEARCH ARTICLE

Open Access

# Conversion events in gene clusters

Giltae Song<sup>1\*</sup>, Chih-Hao Hsu<sup>2</sup>, Cathy Riemer<sup>1</sup>, Yu Zhang<sup>1</sup>, Hie Lim Kim<sup>1</sup>, Federico Hoffmann<sup>3</sup>, Louxin Zhang<sup>4</sup>, Ross C Hardison<sup>1</sup>, for NISC Comparative Sequencing Program<sup>5</sup>, Eric D Green<sup>5</sup> and Webb Miller<sup>1</sup>

## Abstract

**Background:** Gene clusters containing multiple similar genomic regions in close proximity are of great interest for biomedical studies because of their associations with inherited diseases. However, such regions are difficult to analyze due to their structural complexity and their complicated evolutionary histories, reflecting a variety of large-scale mutational events. In particular, conversion events can mislead inferences about the relationships among these regions, as traced by traditional methods such as construction of phylogenetic trees or multi-species alignments.

**Results:** To correct the distorted information generated by such methods, we have developed an automated pipeline called CHAP (Cluster History Analysis Package) for detecting conversion events. We used this pipeline to analyze the conversion events that affected two well-studied gene clusters ( $\alpha$ -globin and  $\beta$ -globin) and three gene clusters for which comparative sequence data were generated from seven primate species: CCL (chemokine ligand), IFN (interferon), and CYP2abf (part of cytochrome P450 family 2). CHAP is freely available at [http://www.bx.psu.edu/miller\\_lab](http://www.bx.psu.edu/miller_lab).

**Conclusions:** These studies reveal the value of characterizing conversion events in the context of studying gene clusters in complex genomes.

## Background

Recent comparative genomics studies have revealed how the human genome has been shaped by various evolutionary forces. Some regions in the human genome are strongly conserved among many mammalian species [1,2], while others have seen accelerated change in the human lineage compared to other species [3,4]. Some regions, the so-called “gene clusters”, are composed of multiple similar copies of gene-containing segments in close proximity. Because such structurally complex regions have been implicated in human genetic diseases, their study has become of great interest. For example, the number of gene copies in the CCL cluster influences susceptibility to HIV [5], the IFN cluster is associated with sarcoidosis [6], and the CYP2abf cluster is implicated in lung cancer [7].

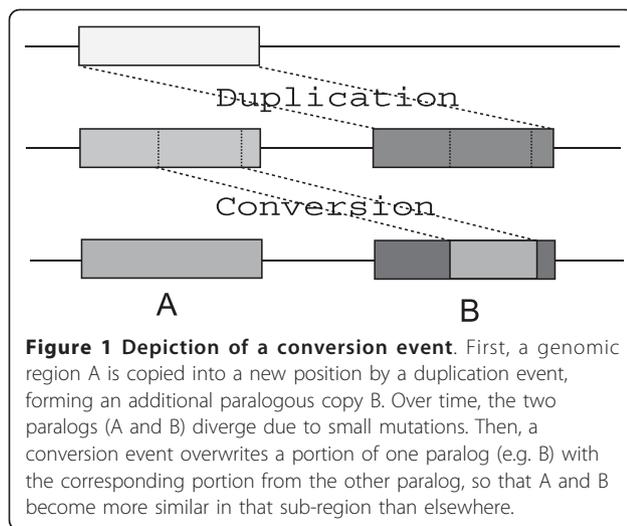
Analyzing the evolution of biomedically relevant gene clusters can inform studies aiming to discover the molecular mechanisms underlying their genetic disease

association (e.g. [8]). However, gene clusters are difficult to analyze because they contain multiple similar genomic regions, and tend to have complex evolutionary histories involving a variety of large-scale mutation events (such as duplications, deletions, inversions, and conversions). One of the key problems in analyzing gene clusters lies in distinguishing between orthologs, defined as genes that derive from speciation events, from paralogs, defined as genes that derive from duplication events. Traditional efforts to trace the relationships among these regions have focused on constructing a phylogenetic tree or a multiple alignment of homologous sequences, but these approaches suffer from uncertainty problems whereby different methods can produce substantially different results [9]. Even when using a single method, the tree topology of a sequence dataset can change depending on which region is selected [10]. One of the main factors confounding such evolutionary analyses is the phenomenon of conversion between paralogs (a.k.a. non-allelic or ectopic conversion), where sequence from one region overwrites part of a similar paralogous region via the recombinational machinery (Figure 1). This is sometimes called “gene conversion”, though it

\* Correspondence: [gsong@bx.psu.edu](mailto:gsong@bx.psu.edu)

<sup>1</sup>Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802 USA

Full list of author information is available at the end of the article



does not necessarily involve any genes. It is typically caused by DNA double-strand breaks or by a double Holliday junction dissolution mechanism [11]. Previous genome-wide studies of conversion revealed that such events have occurred quite often (e.g. 7.5% of all paralogous pairs in the mouse genome [12] and 13.5% in the human genome [13]). Moreover, some of these paralogs have undergone recurring conversion events, complicating matters even further.

Many computational methods have been devised for detecting gene conversion (e.g. [12,14-16]). According to our evaluation study [17], the method we developed [12] outperformed the others tested when applied to gene cluster data and both sensitivity and false discovery rate were taken into account. In addition, the other methods for detecting converted regions pose the inconvenience of requiring a multiple sequence alignment for each set of homologous sequences. Here, we report the development of an easy-to-use software package called CHAP (Cluster History Analysis Package) for detecting conversion events using gene cluster sequences from multiple species. CHAP uses the conversion-detection method from [12], which is based on a statistical test proposed by [18]. The CHAP package includes procedures for identifying all pairwise orthologs required by the conversion-detection method, utilizing an extension of our CAGE program [19]. Thus, users only need to prepare sequence datasets for their gene clusters of interest, and provide an overall phylogenetic tree for the species involved. The package also includes an extended version of our Gmaj viewer [20], which provides a convenient way to examine the detected conversion events. The entire CHAP package is available for free download from our website at [http://www.bx.psu.edu/miller\\_lab](http://www.bx.psu.edu/miller_lab). Users will also need the well-known RepeatMasker

program [21], which is available from <http://www.repeatmasker.org>.

CHAP includes the following differences from our previous program [12]. (1) CHAP focuses on analyzing gene clusters rather than entire genomes. In particular, it incorporates an ortholog identifier specifically designed for dealing with the clusters' complex evolutionary histories [19], whereas the program from [12] relies on external, generic ortholog assignments. (2) It detects conversion events in multiple related species at once, not just a single primary species as in [12]. (3) It leverages the multi-species conversion data to estimate the time of the events in the species tree. (4) It includes an additional, alternative test criterion that enables it to detect conversions covering entire paralogs, which were missed by [12]. (5) It provides a visualization tool to help users easily investigate the conversion results.

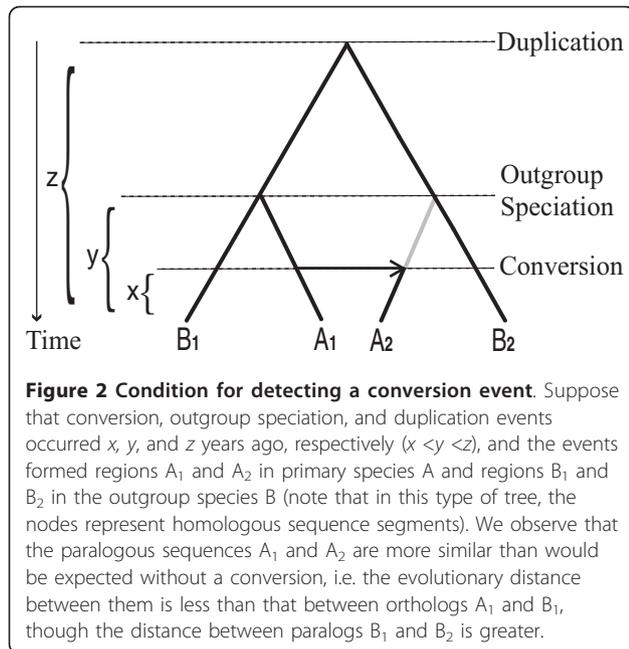
We used CHAP to analyze conversion events in five primate gene clusters. This involved using a combination of sequences available from public sources and newly generated ones (see Additional file 1, Table S1 in the supplement for the GenBank accession numbers of the new sequences). The clusters were chosen based on their relevance to human health, recent duplications in the human lineage, and small size:  $\beta$ -globin (hg19.chr11:5,224,419-5,314,419),  $\alpha$ -globin (hg19.chr16:190,000-235,000), CCL (chemokine ligand; hg19.chr17:34,310,693-34,812,885), IFN (interferon; hg19.chr9:21,058,760-21,481,698), and CYP2abf (part of cytochrome P450 family 2; hg19.chr19:41,324,635-41,712,359).

## Results and discussion

### Overview of detecting conversion events affecting gene clusters

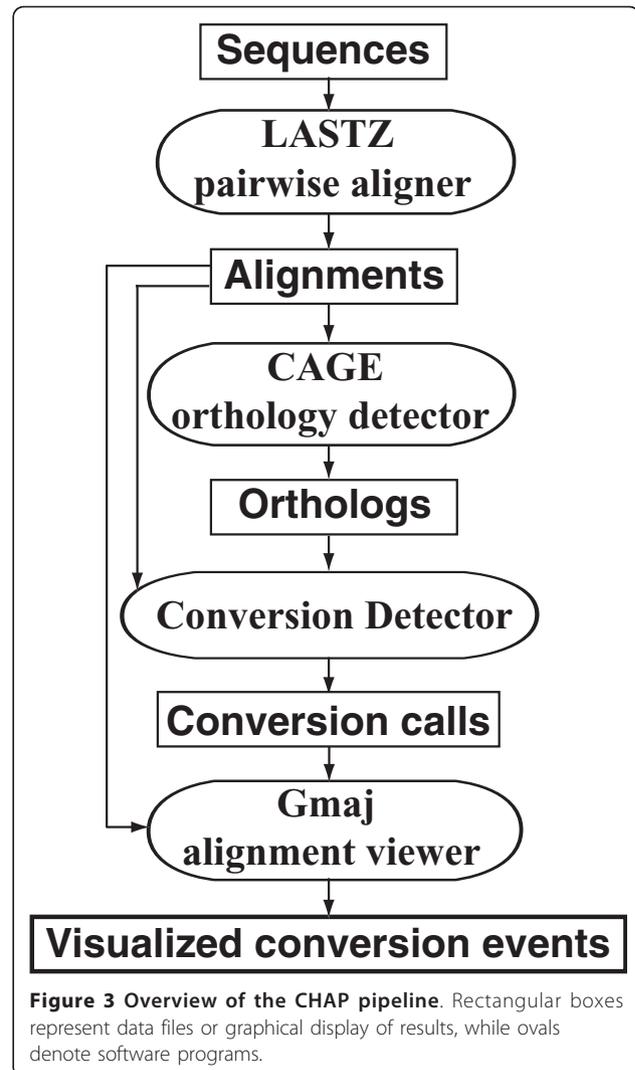
A conversion event produces a within-species alignment where one part is better conserved than other parts. To distinguish this from variation caused by natural selection, our method compares the alignment of paralogs with that of an ortholog in another species to factor out the effects of purifying selection. The use of sequence from an outgroup species that split from the species of interest after the duplication that gave rise to the paralogs, but before the conversion event, can help to illuminate the situation.

To illustrate the general idea, suppose these conversion, outgroup speciation, and duplication events occurred  $x$ ,  $y$ , and  $z$  years ago, respectively ( $x < y < z$ ), forming regions  $A_1$  and  $A_2$  in primary species A and regions  $B_1$  and  $B_2$  in outgroup species B (Figure 2). Thus before the conversion,  $A_1$  is orthologous to  $B_1$  and  $A_2$  to  $B_2$ , while  $A_1$  is paralogous to  $A_2$  and  $B_1$  to  $B_2$  by the definitions of orthology and paralogy [22,23]. Let  $\text{dist}(X, Y)$  denote the evolutionary distance between two regions X and Y. The distance between species A and B



is  $2y$  because A and B split  $y$  years ago, so in the absence of conversion, we expect  $\text{dist}(A_1, B_1) = 2y$  and  $\text{dist}(A_2, B_2) = 2y$ . Similarly, for the paralogs we expect  $\text{dist}(B_1, B_2) = 2z$  and  $\text{dist}(A_1, A_2) = 2z$ . However, the conversion event causes  $\text{dist}(A_1, A_2) = 2x$  and  $\text{dist}(A_2, B_2) = 2z$  instead. Whenever we observe two paralogous regions  $A_1$  and  $A_2$  and an outgroup species B such that  $\text{dist}(A_1, B_1) \leq \text{dist}(B_1, B_2)$  and  $\text{dist}(A_2, B_2) \leq \text{dist}(B_1, B_2)$ , but  $\text{dist}(A_1, B_1) > \text{dist}(A_1, A_2)$  and  $\text{dist}(A_2, B_2) > \text{dist}(A_1, A_2)$  when  $A_1$  is orthologous to  $B_1$  and  $A_2$  to  $B_2$ , then we infer a conversion event between  $A_1$  and  $A_2$ . Note that our statistical method was designed for unequal rates of evolution as well as a molecular clock, although we illustrate only the latter case here for simplicity.

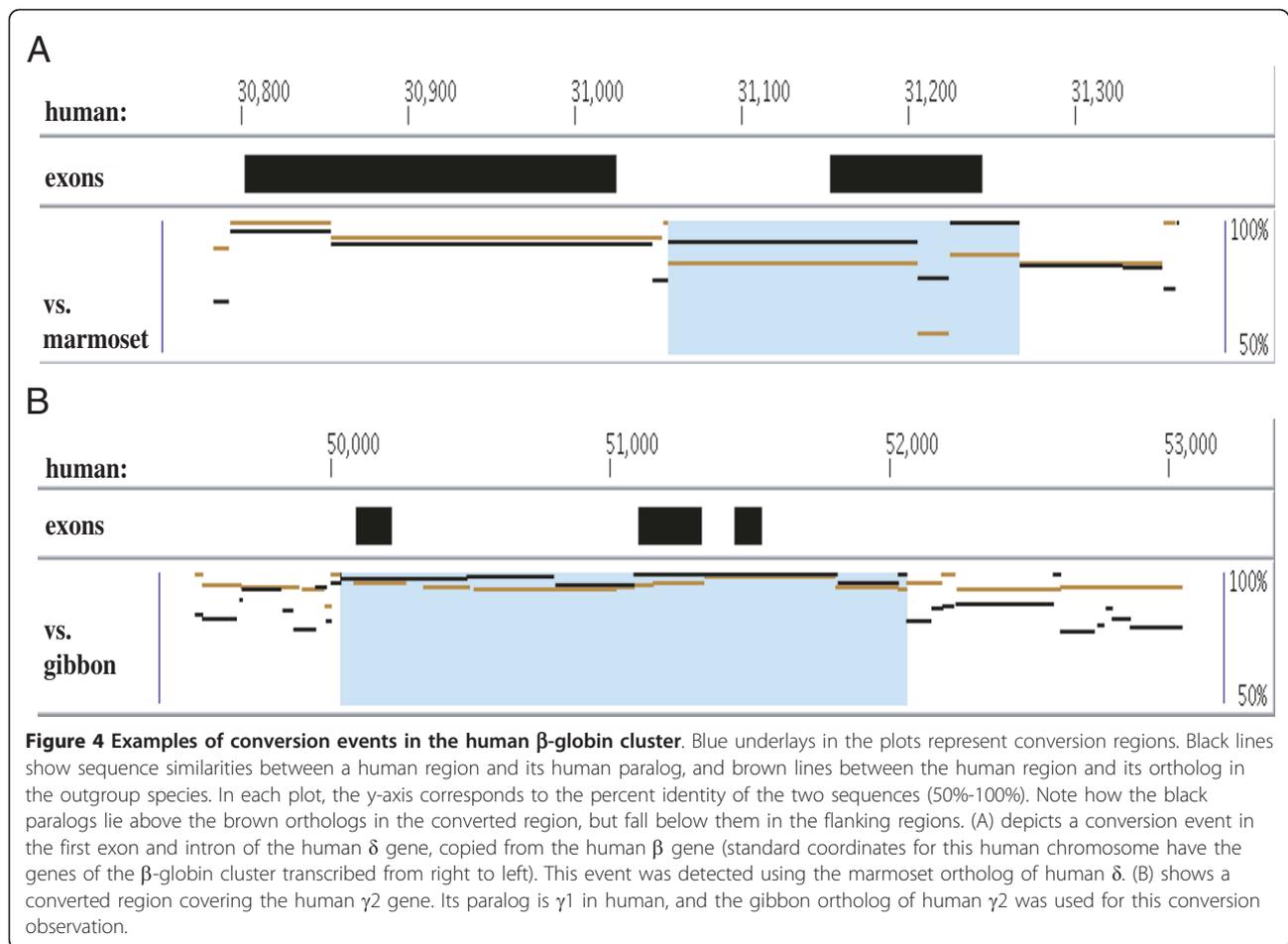
The CHAP pipeline takes as input a set of FASTA sequences from a number of species for a particular gene cluster, together with a phylogenetic tree describing the overall relationships among the species. With a single command, it detects conversion events in all of the species using the others as outgroups. Figure 3 illustrates the steps. First, all self alignments (aligning each sequence to itself) and pairwise inter-species alignments are generated (in MAF format) using the LASTZ aligner [24]. Next, the CAGE software [19] uses a combinatorial approach to identify all pairwise orthologs in the alignment data, according to the definition of orthology in [22,23]; its output is a subset of the inter-species alignments, also in MAF format. Then, our conversion detector [12] examines each pair of paralogous regions, together with their orthologs in each outgroup species, and performs several statistical tests to infer conversion



events (see Methods). The primary output from this step is a tab-separated text file listing the conversion observations for each combination of reference and outgroup species, which are grouped into unique events (since the same event may be observed in multiple species and/or by using multiple outgroups) and localized to a particular sub-lineage in the phylogenetic tree. Finally, a second command invokes an extended version of our Gmaj viewer [20] to visualize the detected conversion observations in a chosen species using an interactive, graphical display. Figures 4, 5, 6, 7, 8 and 12 show examples of conversion events detected and visualized by CHAP.

#### Conversion events affecting the primate $\beta$ -globin cluster

The human  $\beta$ -globin cluster is composed of five genes (listed 5'→3'): HBE ( $\epsilon$ ), HBG2 ( $\gamma_2$ ), HBG1 ( $\gamma_1$ ), HBD ( $\delta$ ), and HBB ( $\beta$ ). Each gene contains three exons separated by two introns. The human cluster also includes a well-

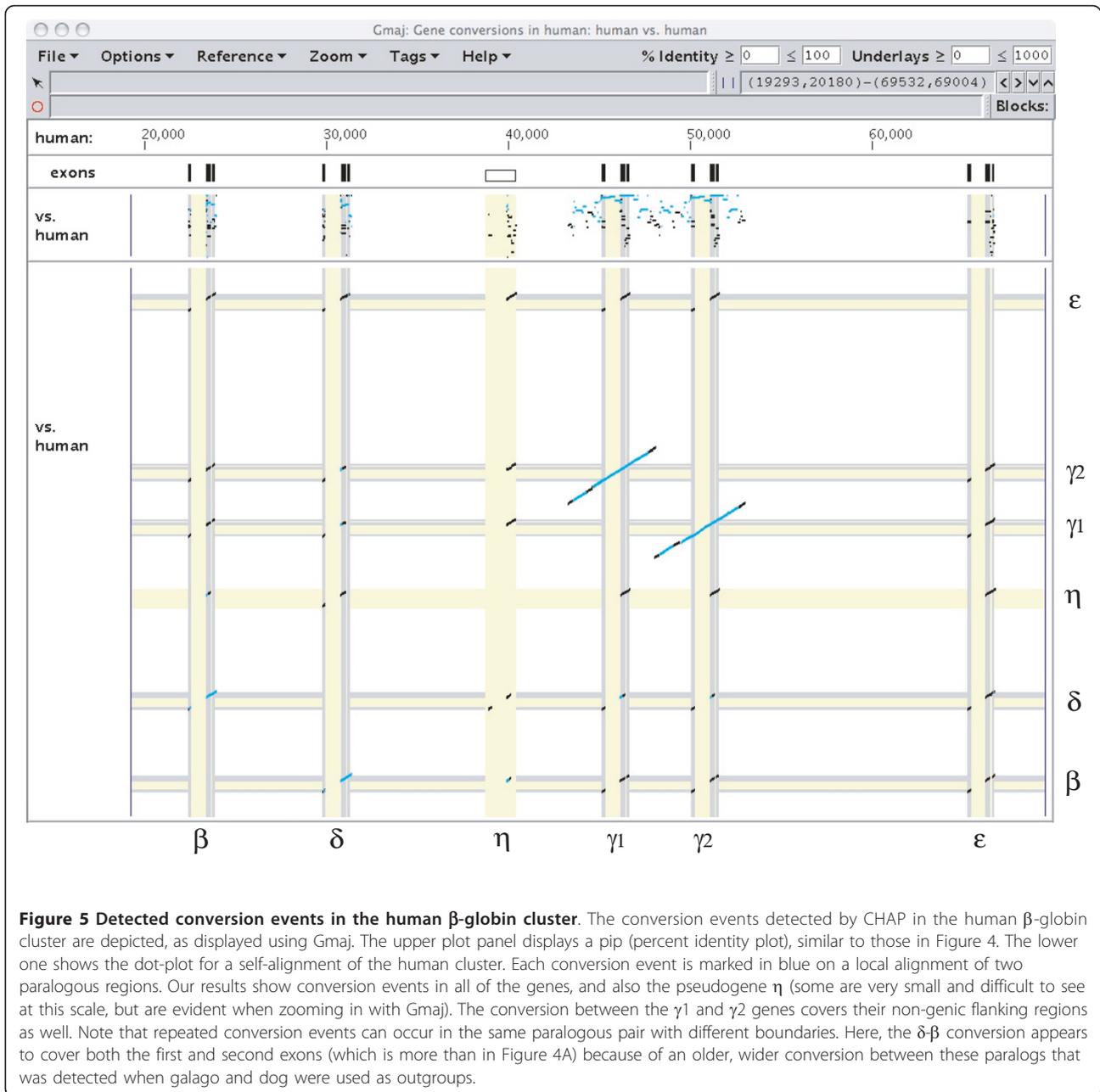


known pseudogene HBH ( $\eta$ ) located between the  $\gamma 1$  and  $\delta$  genes. The human sequence and annotation data for this cluster were obtained from the UCSC Genome Browser ([genome.ucsc.edu](http://genome.ucsc.edu)). We downloaded the sequences of 11 other primate species from the GenBank and Ensembl databases, and obtained gene annotation information for the non-human primate species using the GeneWise program [25]. The non-human primate species were used by the CHAP pipeline as outgroups to detect conversion events in the human lineage. We also used dog and elephant sequences from the UCSC Genome Browser as additional outgroup species.

Evidence suggests that the common ancestor of the primates had a structure of 5'- $\epsilon$ - $\gamma$ - $\eta$ - $\delta$ - $\beta$ -3' in this cluster [26-28]. Except for  $\gamma 1/\gamma 2$ , duplication events among these genes predate the split of primate species [29], i.e. the sequence similarities of two paralogs among those genes are usually lower than the similarities to their orthologs in other primate species [8]. However, Figure 4A shows a case where a region containing the first exon and part of the first intron of the human  $\delta$  gene

has a higher sequence similarity to its paralog  $\beta$  than to its marmoset ortholog, while the ortholog pair shows the expected higher similarity levels than the paralogs in the flanking regions and other parts of human  $\delta$ . CHAP detected that the blue-shaded region of human  $\delta$  was copied from human  $\beta$  via a conversion event. This conversion inference is strongly supported by a  $P$ -value of  $2.31 \times 10^{-5}$ ; it also agrees with other studies of evolutionary relationships in the  $\beta$ -globin cluster (e.g. [28,30]).

Previous studies on the duplication of the  $\gamma 1$  and  $\gamma 2$  genes have traced that tandem duplication to a time before the split of the catarrhine primates (humans, apes, and Old World monkeys) although sequence comparison of  $\gamma 1$  and  $\gamma 2$  shows a high similarity level at about 98% identity (e.g. [31]). This has been reported as an example of gene conversion [11,32]. CHAP detected a conversion event in the human  $\gamma 2$  gene copied from  $\gamma 1$ , e.g. when using gibbon as the outgroup (Figure 4B). Our result indicates that this conversion includes the flanking regions of the two  $\gamma$  genes as well as their exons and introns, in total involving approximately 2000 bp. In addition to these two main conversion events in



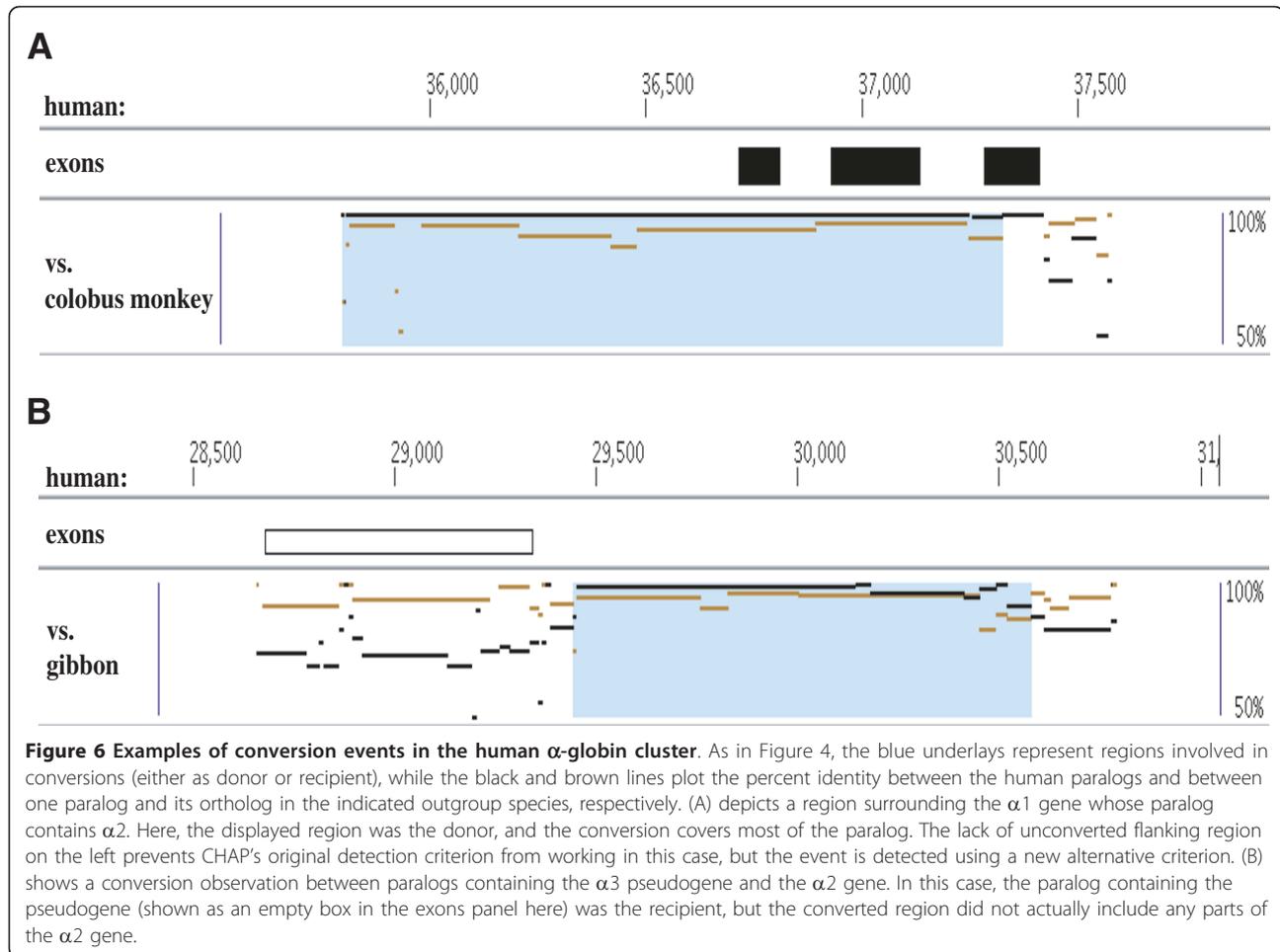
the human  $\beta$ -globin cluster, there are several minor ones as well (Figure 5).

#### Conversion in the $\alpha$ -globin cluster and events covering entire paralogs

The human  $\alpha$ -globin gene has the same origin as the  $\beta$ -globin gene, i.e., both were generated via duplication about 450 million years ago from a common ancestral globin gene [33]. Since then, they have formed their current clusters independently [28]. Similar to the  $\beta$ -globin cluster, the human  $\alpha$ -globin cluster comprises five genes (listed 5'→3'): HBZ-T1 ( $\zeta_1$ ), HBK ( $\mu$ ), HBA-T2 ( $\alpha_2$ ),

HBA-T3 ( $\alpha_1$ ), and HBQ ( $\theta$ ). Each gene contains three exons separated by two introns. In addition, there are two well-known pseudogenes: HBZ-T2 ( $\zeta_2$ ) and HBA-T1 ( $\alpha_3$ ). We obtained sequences and annotation information for human and 12 non-human primate species, cow, and elephant for the  $\alpha$ -globin cluster as above for the  $\beta$ -globin cluster (except that pseudogene locations were inferred from the human self-alignment). We then ran the CHAP pipeline to detect conversion events in this cluster.

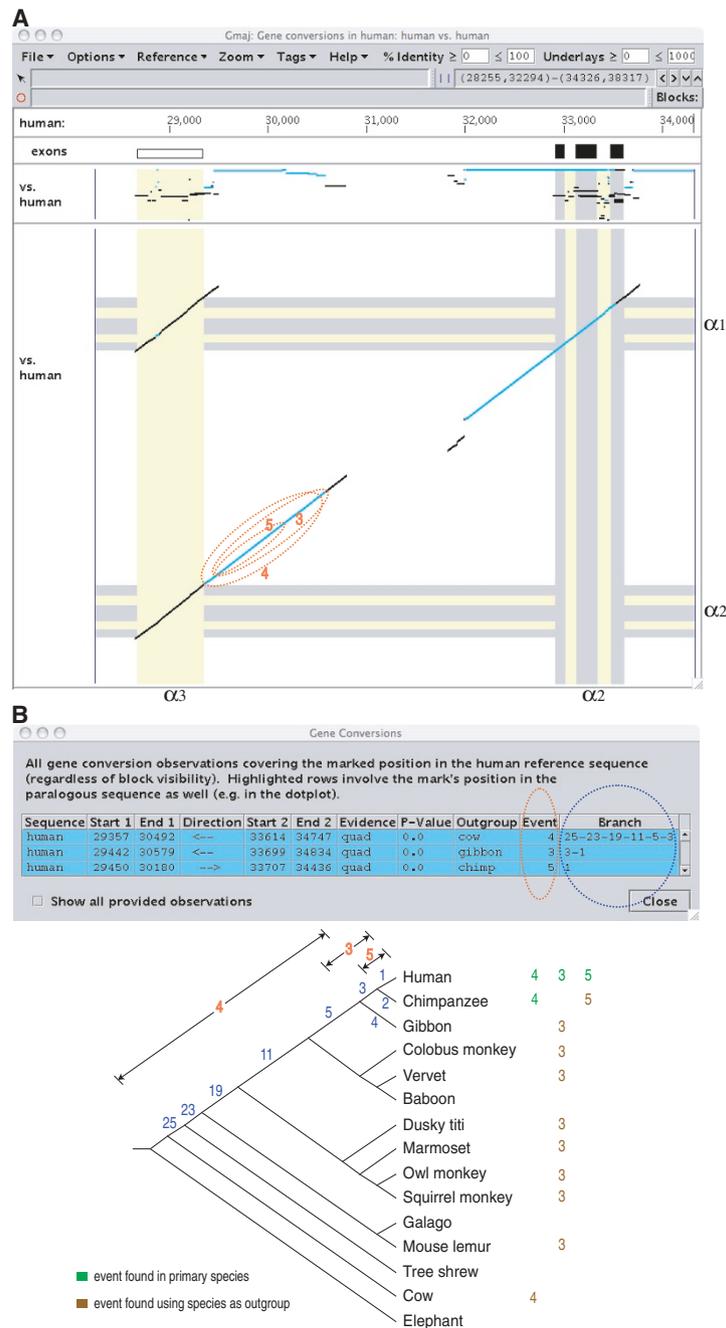
One of the interesting results in the human  $\alpha$ -globin cluster is a conversion involving the  $\alpha_1$  and  $\alpha_2$  globin



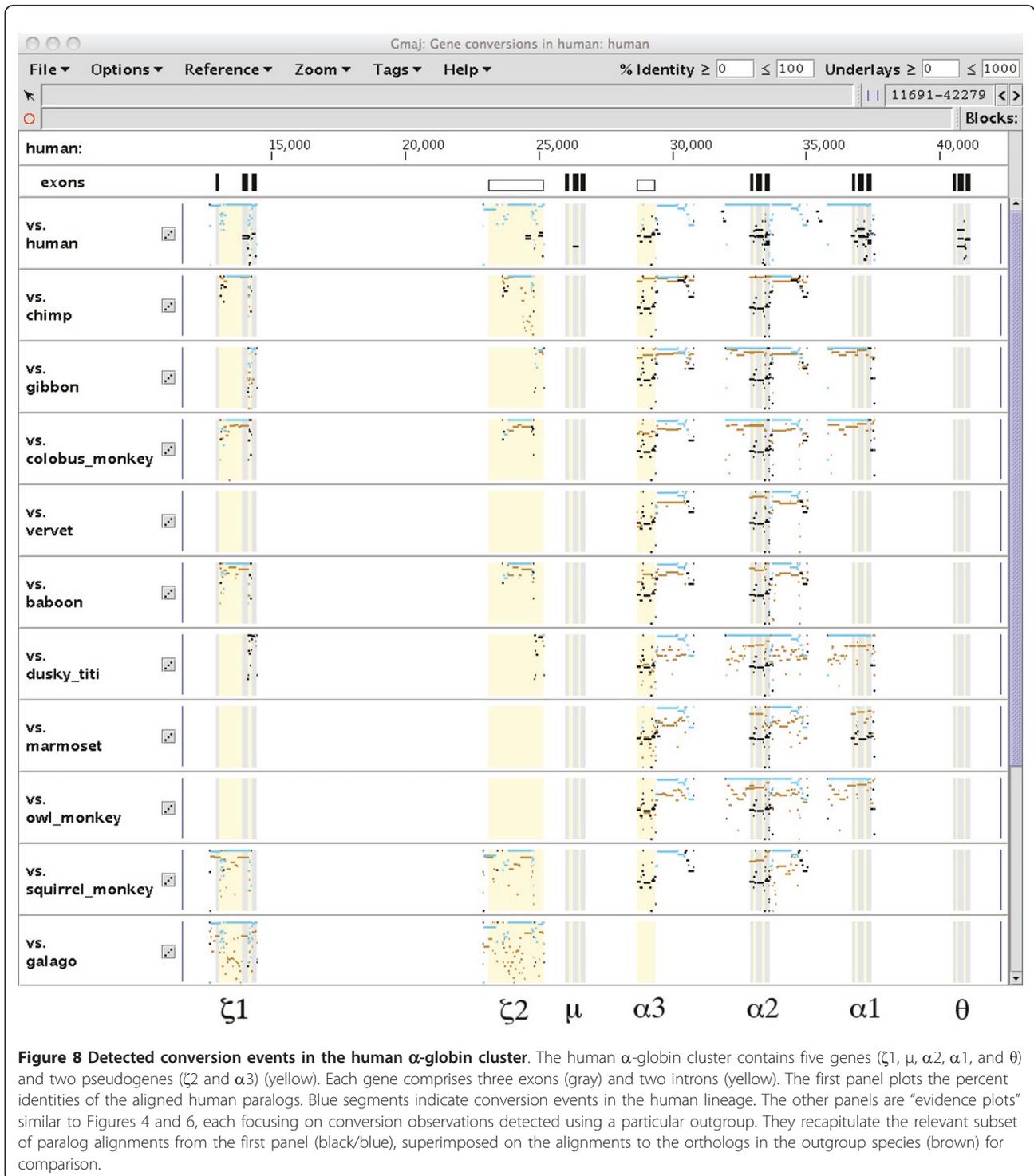
genes. The duplication giving rise to these paralogs predates the separation of the apes from New World monkeys [34]. However, the evolutionary distance between these paralogous sequences is less than that between  $\alpha 1$  and its ortholog in colobus monkey (Figure 6A). The ortholog situation is straightforward, since the colobus orthologs that the CAGE step identified for the human  $\alpha 1$  and  $\alpha 2$  genes are one-to-one; i.e. CAGE does not find any post-speciation duplications involving  $\alpha 1$  or  $\alpha 2$  in either the human or colobus lineages, so the human paralogs are mapped to single, distinct regions in colobus. Nevertheless, the original criterion used in [12] cannot detect the conversion event shown in Figure 6A because that criterion requires switch points of percentage identity differences between paralogs and orthologs, as shown in Figure 4 and Figure 6B. The reason that such pattern switches do not appear in Figure 6A is that nearly the entire sequence is involved in the conversion, so there are insufficient flanking regions for performing the test. To detect such events, we added a new criterion to the conversion detector in our CHAP pipeline that does not require flanking regions (see Methods). As

a result, CHAP was able to detect this event, where according to the colobus outgroup, 1526 bp of the human paralog containing  $\alpha 1$  overwrote the corresponding part of the  $\alpha 2$  paralog. This observation is supported by a  $P$ -value of  $5.29 \times 10^{-20}$ . The same event is also observed using other outgroups, though some of them show the opposite direction of conversion and different endpoints for the converted region.

Another interesting conversion occurred between paralogs containing the human  $\alpha 2$  gene and  $\alpha 3$  pseudogene (Figure 6B). The duplication forming  $\alpha 2$  and  $\alpha 3$  predates the separation of the simian primates from prosimians [34]. Since the conversion can be observed using gibbon as the outgroup sequence, we conclude that it occurred after the separation of hominids from other apes. Although the human  $\alpha 3$  region is a pseudogene, its 3'-flanking region still has a sequence similar to that of the protein-coding gene  $\alpha 2$  due to the recent conversion event. We speculate that this conversion may play a role in conserving some function associated with the 3'-flanking regions of  $\alpha 2$  and  $\alpha 3$ . The  $\alpha 3$  pseudogene is present as a pseudogene in several species,



**Figure 7 Timing of repeated conversion events in the human  $\alpha$ -globin cluster.** Three sequential events occurred between paralogs containing the  $\alpha 3$  pseudogene and the  $\alpha 2$  gene. (A) The regions involved in each conversion are marked here with orange ovals on the self-alignment dot-plot for the human cluster, and labeled with their corresponding event id numbers. (B) The table displayed by Gmaj shows a summary of these events. The Event column (highlighted here with an orange oval) shows the assigned event id, and the Branch column (blue oval) indicates the estimated conversion time as a sub-lineage of numbered tree edges. The tree and colored numbers added here illustrate how to interpret the table and what information was used to infer the conversion time. Numbers in green are the ids of events found in the particular species (i.e. as primary), while those in brown indicate events found using that species as outgroup. Primary species are used to estimate the lower bound for the conversion time; e.g. we inferred that event 4 occurred before the split of human and chimpanzee because we observed this event in both species (in paralogous pairs that are orthologous to each other). In contrast, an outgroup species used for detecting an event provides evidence for the upper bound of its conversion time, i.e. the event occurred after the split of the primary and outgroup species. For example, event 5 must have happened after the separation of human and chimpanzee because it was detected when using chimpanzee as the outgroup and human as the primary species. CHAP does not draw any inference from negative results (lack of detection), since that may be due to other factors such as missing sequence.



which suggests that it is somewhat important, given that it lost function several million years ago yet is still around in some of the species examined. Further, its ortholog in prosimians is an active gene.

The conversion events detected here include those originally described for the duplicated  $\alpha$ -globin genes in

humans [35]. Also, the “gradient of gene conversion” [36] is apparent in the declining percent identity of the converted DNA segments in the 3’ flanking regions of the  $\alpha 3$  and  $\alpha 2$  genes (Figure 7). This gradient was attributed to repeated gene conversions over evolutionary time. By examining the patterns of conversion

detected using outgroups over increasing phylogenetic distance, CHAP inferred three different conversion events in the 3' flanking regions of the  $\alpha 3$  and  $\alpha 2$  genes, and estimated the time of each event as illustrated in Figure 7B. For instance, event 5 in Figure 7 is a recent conversion that happened after the split of human and chimpanzee, while event 4 occurred before their separation. Note that the conversion time for event 4 is difficult to determine precisely because we do not have sufficient information from many of the species (mainly due to missing sequences), whereas event 3 has eight outgroup results supporting the inference that it followed the separation of human and gibbon.

All of the conversion events detected by the pipeline in the human lineage of the  $\alpha$ -globin cluster are shown in Figure 8, which displays them on a human self-alignment together with inter-species alignments between human and the outgroups used for the event detection. In total, 11 conversion events were detected, all involving paralogs containing  $\zeta 1$ ,  $\zeta 2$ ,  $\alpha 3$ ,  $\alpha 2$ , and/or  $\alpha 1$ .

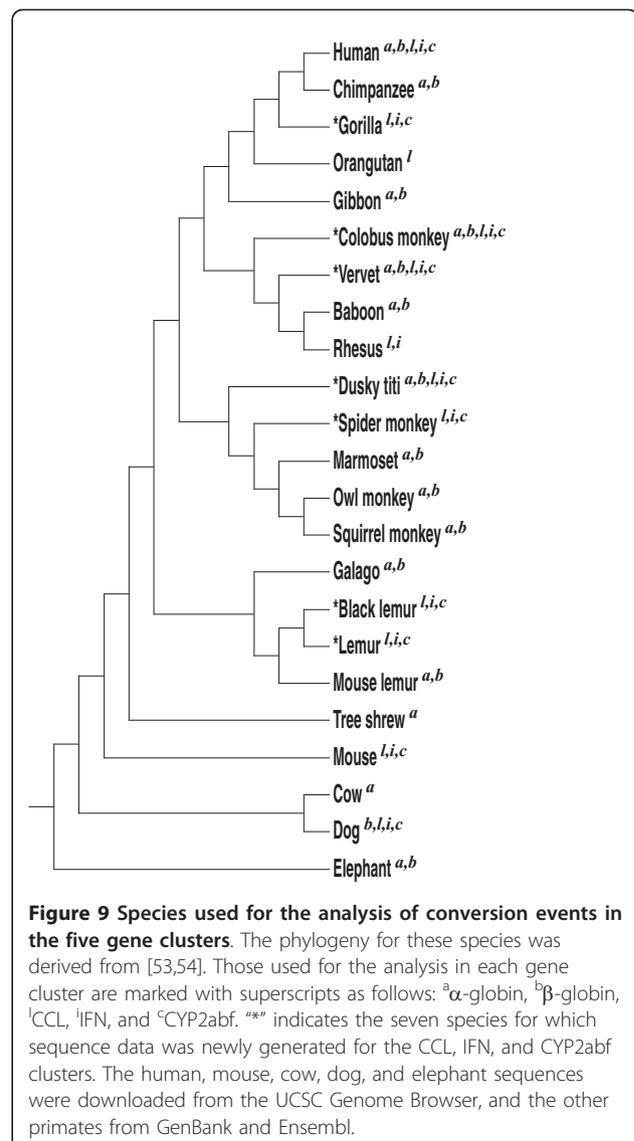
#### Summary of conversion events in five human gene clusters

In addition to the  $\beta$ -globin and  $\alpha$ -globin clusters, we obtained sequences for three more gene clusters: CCL, IFN, and CYP2abf. For these studies, comparative sequence data were generated from seven additional primate species in these clusters, indicated by "\*" in Figure 9.

We analyzed conversion events in the human lineage for these five clusters using the CHAP package. As shown in Figure 10, conversions are quite frequent in all of these clusters. Here, we classify the events into two categories: criterion 1 or 2. Criterion 1 means that the event was detected using the original triplet or quadruplet tests from [12], while criterion 2 is the new method for detecting events covering the entire paralogous sequence (or most of it), without sufficient unconverted flanking regions needed for criterion 1. Figure 10A shows the number of events detected by each criterion in the five clusters. Of the 256 total events, 38.3% were detected by criterion 2.

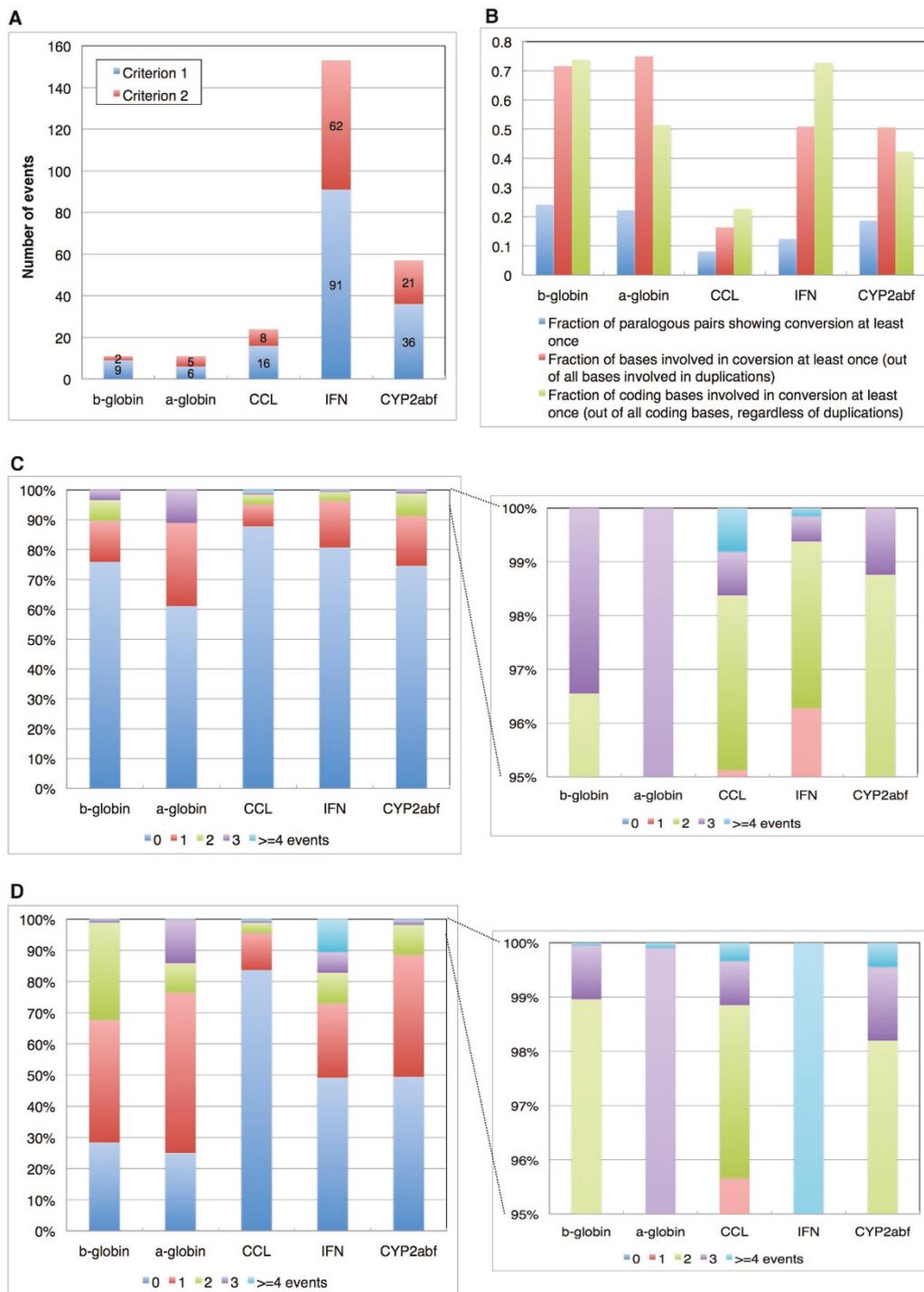
To examine the conversion frequency, we first computed the fraction of paralogous sequence pairs showing conversion, displayed in the blue bars of Figure 10B. There were a total of 975 paralogous sequence pairs in the five clusters, and of these, 194 (19.9%) experienced at least one conversion event. By this measure, the  $\alpha$ -globin cluster is the most active of the five, with 38.9% of its pairs showing evidence of conversion.

We also computed the conversion frequency using the number of bases instead of paralogous pairs, shown in the red bars of Figure 10B. A total of 725,694 bases in the five clusters lie in duplicated regions, and 37.7% of them have been involved in a conversion at least once. This



fraction is almost twice the one based on the number of pairs. In this comparison based on the number of bases, the  $\alpha$ -globin cluster again shows the highest conversion frequency, with 75.0% of its duplicated bases involved in conversions. According to both measures, the CCL cluster experienced the least conversion activity.

In addition to the conversion frequencies in the general paralogous sequences, we also analyzed the incidence of conversion in protein-coding exons (green bars in Figure 10B). On average, nearly 50% of the coding bases were involved in at least one conversion. In the  $\beta$ -globin and IFN clusters, this rose to more than 70%. The frequent conversion events between the coding regions serve to homogenize paralogs within species, which could contribute to the establishment of species-specific characteristics. Furthermore, conversions that



**Figure 10 Summary of detected conversions in the five gene clusters.** (A) Number of conversion events via criterion 1 and criterion 2. (B) Conversion frequencies in the general paralogous sequences and protein-coding exons. (C) Fraction of paralogous pairs by their number of conversion events, out of all paralogous sequence pairs. (D) Fraction of bases by their number of conversion events (involved as either source or target), out of all bases involved in duplications.

copy non-coding sequences into coding regions (e.g. as in the human CYP2A13 gene, discussed below) can introduce new functionality more quickly than point mutations, since many nucleotides are changed by a single event. IFN and CCL are involved in immune response to infection by pathogens such as viruses, bacteria, or tumor cells, and the CYP450 genes play crucial roles in the metabolism of exogenous substances such as drugs or carcinogens. Thus gene conversion may be an effective evolutionary mechanism facilitating adaptation to environmental changes.

Some paralogous pairs are involved in conversions more than once. Figure 10C shows how many pairs in each cluster experienced various numbers of conversion events. About 10% of the paralogous pairs in the  $\beta$ -globin and  $\alpha$ -globin clusters had at least two sequential events, and this figure was at least 4% for the other three clusters.

We also computed the proportion of bases involved in multiple conversions, either as source or target, out of all bases in duplicated regions (Figure 10D). More than 20% of the duplicated bases in the  $\beta$ -globin,  $\alpha$ -globin, and IFN clusters were involved in conversion events at least twice. In the IFN cluster, about 10% were involved in four or more events.

In addition, we identified the “hot-spot” protein-coding gene in each cluster according to the number of conversion events involving coding exons, either as donor or recipient (Table 1).

When conversions occur in protein-coding genes, the functions of the genes may be affected. For example, we detected recurrent conversions between the human  $\beta$ - and  $\delta$ -globin genes. Evidence for a similar but presumably much more recent event has been reported in a patient exhibiting mild microcytosis, whose  $\delta$ -globin gene contained part of the amino acid sequence of the  $\beta$ -globin gene [37]. The authors of that report hypothesized that this could have been due to a new gene conversion.

#### The human CYP2A13 gene as a hot-spot for conversions in the CYP2abf cluster

To examine the effect of conversion events on gene function, we focused on the CYP2A13 gene from Table

1. CHAP detected seven conversion events, all overwriting parts of this gene with donor sequences from elsewhere, and together they involved all nine of CYP2A13's coding exons (Figure 11). Interestingly, except for one event the donor paralogs are entirely non-coding, and according to annotation data from NCBI, none are considered to be pseudogenes. It is somewhat surprising that the gene is still functional, in spite of being overwritten with non-coding sequences so many times.

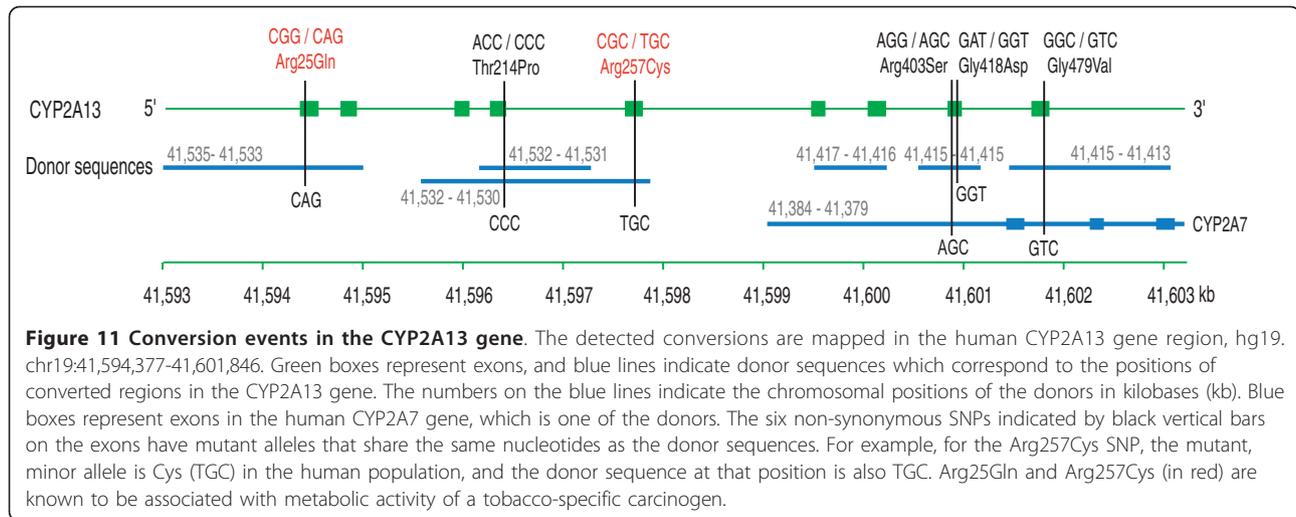
The CYP2A13 gene encodes an enzyme which metabolizes a tobacco-specific carcinogen, NNK (4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone) [38]. NNK is activated by the metabolism, leading to its carcinogenicity [39]. The gene is intensively expressed in the respiratory tract [38,40], and it has been studied as a candidate gene associated with tobacco-related cancers, such as lung cancers. In particular, the variations Arg25Gln and Arg257Cys have been shown to relate to metabolic activities involving NNK [41-43]. The minor alleles, Gln and Cys, decrease the activities of the enzyme as well as the gene expression, and could produce lesser toxicity of NNK in smokers [44].

To investigate the effects of conversions on the CYP2A13 coding sequences, we compared sites in these regions that are polymorphic in the human population against the nucleotide sequences of the donor regions. In particular, we used non-synonymous SNP sites from the NCBI dbSNP database in the converted CYP2A13 exon regions, and looked for cases where the mutant allele of the SNP matches the donor sequence of a conversion detected in this study (Figure 11). The mutant allele was determined under the assumption that the gorilla orthologous sequence is the ancestral type. Indeed, the donor sequences for the 25th and 257th amino acids of CYP2A13 do match the minor alleles, which are also the mutant alleles. We also found four other non-synonymous SNPs in this gene whose mutant alleles share the same nucleotides as the donor sequences. This frequent sharing could be explained by recurrent conversions from the donor sequences in the human population, as an alternative to point mutations alone. Thus we suggest that recent conversions, perhaps tending to recapitulate ancestral ones, may contribute to polymorphism and phenotypic variation in humans.

**Table 1 Hot-spot genes for conversion events**

Cluster name	Gene name	Chromosome	Coding start	Coding end	Number of events
$\beta$ -globin	HBB	chr11	5,246,828	5,248,251	6
$\alpha$ -globin	HBZ	chr16	202,909	204,399	5
CCL	CCL15	chr17	34,324,803	34,328,531	5
IFN	IFNA14	chr9	21,239,365	21,239,934	13
CYP2abf	CYP2A13	chr19	41,594,377	41,601,846	7

The protein-coding gene in each cluster showing the most conversion events involving exons. Positions are given with respect to the UCSC hg19 assembly.



(Note that CHAP would not detect such recent conversions here, as it was only given the reference assembly for each species.)

#### Conversion in *Drosophila melanogaster* paralogs

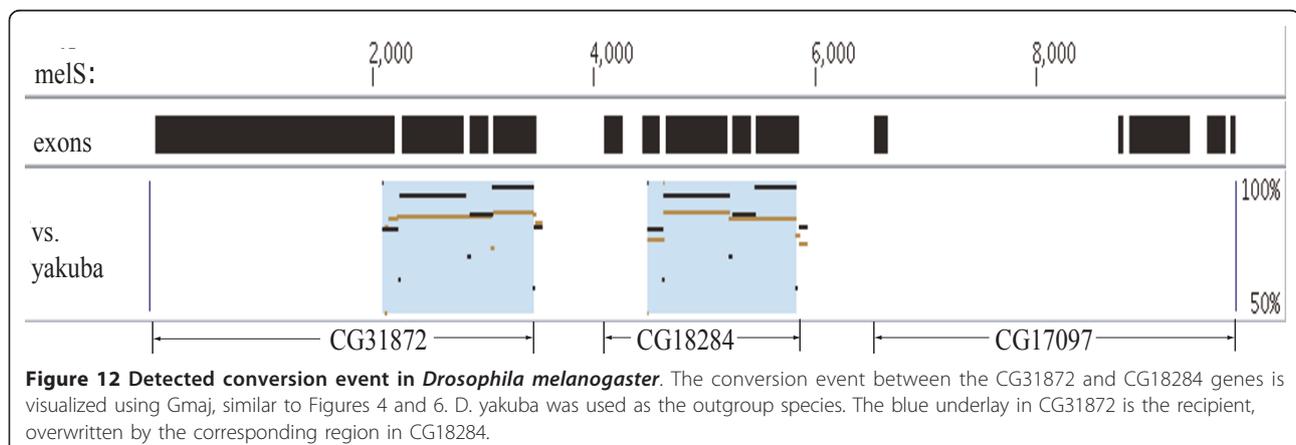
Conversion in the *D. melanogaster* genome has been studied by many researchers (e.g. [45-47]). Approximately 7.5% of its paralogous genes have experienced conversion [46]. To explore the applicability of CHAP for studying non-mammalian sequences, we used it to analyze a gene cluster where we suspected the occurrence of conversion.

The genomic region dm3.chr2L:10,639,900-10,649,700 (FlyBase R5.33) [48] contains three paralogous genes (CG31872, CG18284, and CG17097), which are associated with lipase activity and lipid metabolic processes [49]. The CG18284 and CG17097 genes comprise five exons each, while CG31872 has four. We located the corresponding genomic region in *D. yakuba* (droYak2.chr2L:7,042,800-7,053,500) using inter-species LASTZ alignments for use as an outgroup. CHAP detected a

1354-bp conversion event between the CG31872 and CG18284 genes that involves their last four exons and three introns (Figure 12). It is interesting to note that the conversion event maintained the exon-intron structure of the 3' portion of these two genes while the 5' structure of the genes diversified.

#### Conclusions

In this study, we developed the CHAP package for detecting conversion events in gene clusters, and used this computational tool for analyzing five human gene clusters. We found that 20.0% of the paralogous sequence pairs in those clusters have undergone at least one conversion event. This is somewhat higher than the genome-wide frequency (13.5%) for all human paralog pairs [13], and confirms that conversion is a common phenomenon that must be accounted for when studying gene cluster evolution. Interestingly, some paralogs have experienced conversion events repeatedly in the same intervals; this can be detected if the later ones change narrower regions in the target.



To detect the occurrence of conversion events and their boundaries, we used a statistical test based on two criteria: the original triplet and quadruplet tests implemented in [12], and a new alternative test for events covering most or all of their respective paralogous sequences, which was implemented by extending the original detector. Both criteria achieve comparable statistical rigor by calculating a  $P$ -value for each event. We have compared our conversion detector pipeline (including the CAGE program) to other existing methods available for detecting conversion events, and the overall accuracy of our method (considering both the sensitivity and false discovery rate) was superior to the others when applied to gene cluster data [17]. The quality of the non-parametric statistical test used in our conversion detector was also evaluated in [18].

CHAP works best when sequences are available from many species, because even if a conversion cannot be detected using one particular primary or outgroup species (e.g. due to difficulty assigning orthologs), it may still be detected using others. In addition, using more species in the analysis enables more precise estimation of conversion times. However, as illustrated in the *Drosophila* example above, even one outgroup sequence is often sufficient to find conversions.

Accurate detection of conversion events can contribute substantially to other studies of gene cluster evolution by correcting duplication ages distorted by conversion and by improving the identification of orthology relationships in complex gene clusters. For example, the orthology assigned by CAGE is based on the positions where duplications occurred and the species in which they are seen (which we call “orthology by position”). The orthology mappings based on this definition are suitable for our purpose of conversion detection. However, other studies may require orthology mappings based on a different definition: if we trace the orthology based on the origin of the sequence contents rather than the positions of duplication, conversion events can change the mappings (for example, in Figure 2 the ortholog of  $A_2$  becomes  $B_1$  instead of  $B_2$ ). This alternative concept of “orthology by content” corresponds to the mosaic structures of phylogenetic trees [50], and is necessary for certain types of evolutionary analysis, such as determining the rate of synonymous substitutions in coding regions. In fact, it may not even be appropriate to analyze synonymous substitution rates in genes subject to frequent conversion events, because their nucleotide changes are not introduced according to a Poisson process, but rather as segments of changes. For both concepts of orthology, CHAP can lead to more accurate ortholog identification that is essential for constructing multiple sequence alignments and phylogenetic trees. Finally, our ongoing efforts in this area can help to

more accurately model the full complexity of evolutionary processes, and to improve software for the analysis of gene cluster evolution.

## Methods

### Statistical triplet and quadruplet tests for conversion detection

The basic triplet test for conversion detection involves three homologous sequences: two paralogs from one species ( $A_1$  and  $A_2$  in Figure 2), and the (positional) ortholog of one of them in an outgroup species (say,  $B_1$ ). The alignments  $A_1$ - $A_2$  and  $A_1$ - $B_1$  are examined base-by-base using the hypergeometric random walk method of [18], which utilizes a cumulative score representing the difference in distances between these sequence pairs. This score is computed as follows: for each informative site in  $A_1$ , the score increases if the paralog has the same nucleotide, but the ortholog has a different one, while the score decreases if the ortholog has the same nucleotide, but the paralog does not. (Note that the tree topology in Figure 2 is rooted, so sites can be informative with only three sequences involved. This enables us to detect conversion even when only one ortholog in the outgroup species is available, e.g. due to deletion of the other ortholog.) At the same time, we count the numbers of sites that increase and decrease the score, denoted as  $m$  and  $n$  respectively. When a maximum descent of  $k$  is observed for the hypergeometric random walk with  $m$  up and  $n$  down steps ( $H_{m,n}$ ), the  $P$ -value to determine the significance of the maximum descent is computed as the probability that the maximum descent of  $H_{m,n}$  is greater than or equal to  $k$  by chance. When the  $P$ -value is below the cutoff threshold value, which is determined based on the multiple-comparison-corrected method proposed in [12], the interval of the maximum descent is taken as the converted region. See [12] for details, e.g. how to compute the  $P$ -value in a space- and time- efficient way in order to overcome the length limitation of  $\sim 400$  informative sites encountered by [18].

Our detector from [12] also improves the method in another way, by using a quadruplet test where possible. If both paralogs ( $A_1$  and  $A_2$ ) have positional orthologs ( $B_1$  and  $B_2$ , respectively), then the two corresponding triplet tests [ $A_1$ ,  $A_2$ ,  $B_1$ ] and [ $A_2$ ,  $A_1$ ,  $B_2$ ] are combined. At an informative site, if only one triplet shows an up step or down step, it is treated as usual. Otherwise, if both triplets show an up step for a column, we assign two up steps, and similarly if both show down steps (the combination of one up and one down step does not occur, because the two triplets have the same paralogous nucleotides). The maximum descent is determined based on the combined hypergeometric random walk. Then we use the same  $P$ -value formula as in the triplet

test. This quadruplet test can improve the specificity of detection, because if one triplet falsely indicates a conversion event due to variation in evolutionary rate, it may be neutralized by the other triplet. Similarly, the quadruplet test can also enhance the sensitivity, because if a true conversion would have been missed due to weak support from one triplet, it may be corrected by strong support from the other one. In addition, the quadruplet test is often able to determine the direction of a conversion event by computing the probabilities of going down within the common maximum descent regions of the two triplets and determining the significance of the difference between the two probabilities based on a binomial distribution [12].

#### Alternative criterion for conversions covering entire paralogous sequences

The paralogs for the triplet and quadruplet tests are obtained from intraspecies local alignments generated by LASTZ. Such alignments are highly sensitive to the parameter settings of the alignment program [51]. In addition, an intraspecies alignment tends to be less accurate than an interspecies one, due to the difficulty of computing alignment scores corresponding to the actual divergence time [52]. This is because each pair of paralogous regions can have its own unique divergence level according to their duplication time, whereas all of the orthologous pairs have been diverging over the same amount of evolutionary time. (Note that the LASTZ parameters and alignment scores used for this study are included as defaults in the CHAP package. They are recommended for conversion detection, but users can change these settings if desired.)

Alignment quality is especially problematic in the end regions of the intraspecies alignments, because (1) the ends of local alignments are generally defined by falling scores, and (2) the paralogs used by CHAP are entire local alignments, though their corresponding orthologs are usually excerpts from longer alignments. This increases the likelihood of calling false-positive conversions near the ends of the paralogs when using the original triplet and quadruplet tests. To avoid these spurious calls caused by erroneous end parts of intraspecies alignments, we skip these tests when the maximum descent covers almost the entire length of the paralogs (over 80%). We use 80% as the threshold because we observed that most conversion events are very short compared to the entire paralogous length, but the frequency of conversions called by the triplet/quadruplet tests increases drastically when the maximum descent covers more than 80% of the paralogous region. Note that the tests can handle the erroneous ends of intraspecies alignments better when the

conversions are smaller, because informative sites in the non-converted regions reduce the effect of the erroneous ends of the paralogs. Small or medium-sized conversions near the end of a paralog pair may sometimes be reported as extending overly far into the poorly-aligning end region, but are much less likely to be completely spurious.

However, conversion events may sometimes genuinely cover more than 80% of their respective paralogs. In addition, the original triplet/quadruplet test causes false negatives for conversion events covering the entire paralogous alignments with no erroneous end parts, as well as false positives for erroneous alignments. When the maximum descent covers the entire paralogous length, the *P*-value computed by the original formula is equal to one due to lack of non-converted regions, undermining the significance of the event detection. Thus the original test is inadequate for detecting conversions covering entire or nearly entire paralogs. In order to avoid missing such events, we apply an alternative criterion utilizing the orthology mappings. When two paralogs show the descent pattern across more than 80% of their length, we distinguish conversion events from false positives according to whether or not the paralogs map to distinct ortholog regions in the outgroup species. If the two paralogs have distinct orthologs, we infer that they were formed by a duplication before the speciation. In this case a maximum descent covering nearly the entire paralog pair represents evidence for conversion, so we treat the descent region as a potential conversion event and calculate a *P*-value for its significance based on a binomial distribution. This test is inferior to the triplet/quadruplet test for typical conversions, but does not break down for the long ones. Other than this replacement of the *P*-value formula, all procedures for this alternative criterion are exactly the same as for the original triplet and quadruplet tests.

Suppose we know that the duplication occurred before the speciation event; then the probability of seeing an up step should be greater than the probability of seeing a down step, i.e.,  $P(-1) \leq 0.5$ . We further assume that the observations are independent. The probability of the number of down steps being  $> n$  by chance (i.e. without conversion) can be calculated as

$$P - value = \sum_{x=n}^{m+n} \left[ \binom{m+n}{x} p^x (1-p)^{m+n-x} \right]$$

where *m* up and *n* down steps are observed and  $p = \min(0.5, n/(n+m))$ . The parameter *p* is estimated from the observed number of up and down steps, and is at most 0.5.

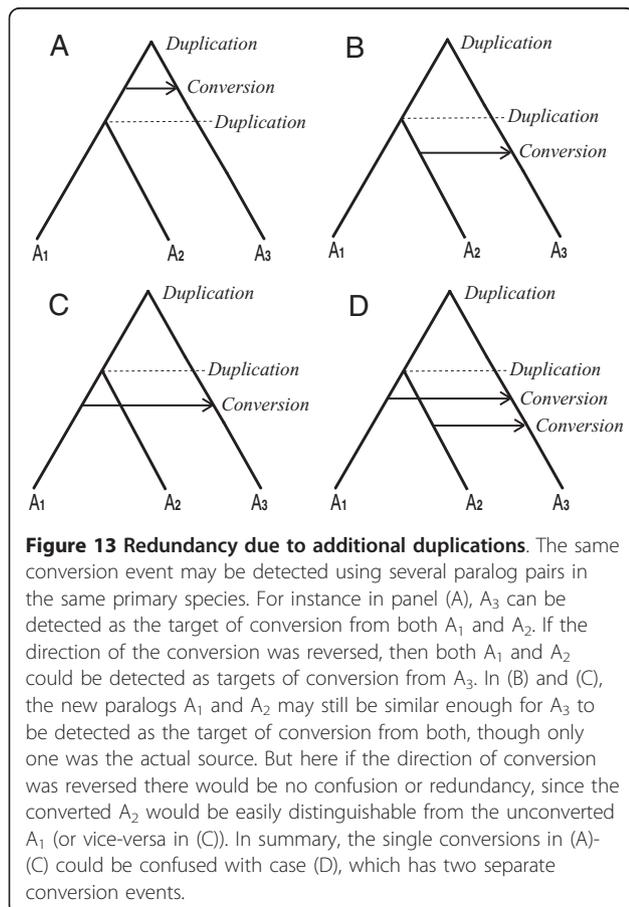
### Redundant evidence for conversions

CHAP runs its tests for every combination of primary species, paralog pair in the primary species, and outgroup species. Because of this, it is common for a single conversion event to be detected more than once:

- If additional speciation events occurred between the duplication and the conversion, then the conversion may be detected using all of those outgroup species.
- If additional speciation events occurred after the conversion, then the conversion may be detected in all of the resulting primary species.
- If additional duplication events occurred after (or perhaps even slightly before) the conversion, then the conversion may be detected in all of the resulting paralogs (Figure 13).

CHAP's conversion detector watches for observations with similar boundaries that match these scenarios and assigns them the same event identifier in the output, as follows.

Type (a): For conversions detected using several outgroup species for a particular paralogous pair, the case



of multiple observations for a single conversion event is distinguished from that of repeated conversion events by comparing the boundaries of the conversion regions. If their boundaries are equal or quite similar, we regard these multiple observations as representing the same conversion event. In the non-redundant version of the output, the observation with the smallest  $P$ -value is kept and the others are removed. Otherwise, they are considered to be separate, repeated conversion events.

Type (b): To remove redundancies in multiple primary species due to subsequent speciation, orthology information is needed. The CHAP pipeline obtains the orthologous relationships among all of the species using CAGE (Figure 3). In our procedure, if one conversion observation is detected between paralogs  $A_1$  and  $A_2$  in species A using outgroup species C, then all of the orthologous sequences of  $A_1$  and  $A_2$  in another primary species B, e.g.  $\text{Orth}(A_1) = \{B_{11}, \dots, B_{1n}\}$ , and  $\text{Orth}(A_2) = \{B_{21}, \dots, B_{2m}\}$ , are identified, and any conversion observations with similar endpoints between  $\text{Orth}(A_1)$  and  $\text{Orth}(A_2)$  which are detected using C as the outgroup species are assigned the same event id. This is repeated for all other species B, and in the non-redundant version of the output only one observation (the one with the lowest  $P$ -value) is kept for each event id.

Type (c): In Figure 13, two conversion observations, e.g.  $A_1 \rightarrow A_3$  and  $A_2 \rightarrow A_3$ , are detected for four different cases. However, in the first three of these there is only one actual conversion event. In order to distinguish these four cases, we first check the boundaries of the conversion regions. If the boundaries are equal or quite similar, we regard them as one of the three cases in Figure 13A-C. Otherwise, the events are inferred to be separate, as in Figure 13D. In order to discriminate the first three cases in Figure 13A-C, the similarities between three pairs of sequences,  $S_1 = \text{Sim}(A_1, A_2)$ ,  $S_2 = \text{Sim}(A_1, A_3)$ , and  $S_3 = \text{Sim}(A_2, A_3)$ , in the conversion regions are examined, where  $\text{Sim}(X, Y)$  is the similarity level (percentage identity) of regions X and Y based on their LASTZ alignment. The applicable case is determined as follows:

- If  $\max(S_1, S_2, S_3) = S_1$ , then the conversion occurred before the duplication of  $A_1$  and  $A_2$ , as shown in Figure 13A.
- If  $\max(S_1, S_2, S_3) = S_2$ , then the conversion occurred between  $A_1$  and  $A_3$ , as shown in Figure 13C.
- If  $\max(S_1, S_2, S_3) = S_3$ , then the conversion occurred between  $A_2$  and  $A_3$ , as shown in Figure 13B.

This is used to decide which observation is kept in the non-redundant version of the output; in case (1) the

observation with the lowest  $P$ -value is used to represent the event, while in cases (2) and (3) the actual conversion is kept and the incorrect one removed.

The redundancies are removed in the order of Type (a), followed by (b), then (c). The order does not matter in principle, but may affect the results if some redundant observations are missing.

#### Estimating the age of a conversion event

Redundancy information can be used as a guide to estimate the time of a conversion event, in terms of its placement on the species tree. Timing an event is composed of two parts. One is inferring an upper bound for the conversion time, i.e. its earliest possible branch edge in the tree, and the other is determining a lower bound.

An outgroup species used for detecting a conversion in a primary species helps to establish the upper bound, indicating that the conversion occurred after the separation of the primary and outgroup species. If an event in a primary species was detected using multiple outgroup species, the closest outgroup species to the primary one in the species tree is chosen to estimate the upper bound. When a conversion event was detected in multiple primary species, an upper bound for each primary species case is computed and then the most recent of these edges that is a common ancestor of all the primary species is assigned as the earliest possible conversion time.

In order to estimate the lower bound of the conversion time, a list of all primary species having the same conversion event is used. If we observed the same event in multiple primary species, that conversion predates the split of those primary species. So we infer the lower bound of the conversion time as the least common ancestral edge of those primary species in the species tree.

The conversion time is summarized as a path of possible edge numbers (sub-lineage) in the Branch column of Gmaj's table (Figure 7B). If a conflict is encountered when determining the conversion time due to an inconsistency in the data, then a question mark will appear in the output. However, this did not arise in our analysis of the five clusters.

#### Simulation studies

CHAP's goal is to detect actual historical conversion events as accurately as possible. However it is also useful to have a more computational formulation of the problem, so that the general quality of the results can be assessed even though little is known about the true history of most clusters. Since we are developing the CHAP analysis package in conjunction with our gene cluster evolution simulator [17], the success of CHAP

can be defined as the extent to which it can correctly recreate the known history of a simulated input sequence. However, we note that this framing of the problem shifts the burden of determining biological relevance onto the quality of the simulator, which currently emulates some, but not all, known evolutionary processes. Measuring the extent to which the simulated sequence reflects the characteristics of real-world gene clusters is a statistical problem that we leave for future work.

We conducted a study using our gene cluster evolution simulator from [17], in order to (1) more rigorously explore the effects of the main tuning parameter (the 80% paralog coverage threshold for criterion selection) on CHAP's sensitivity and false discovery rate, and to (2) measure CHAP's runtime performance on gene clusters of varying complexity. These results are available in the supplement (Additional file 1, Figure S1). We found that as expected, there is some trade-off between sensitivity and specificity of detection, but that the results are not greatly affected by the exact setting of this parameter. Our choice of 80% was initially selected by examining CHAP's output on real gene clusters for which some guidance is available in the literature regarding the actual converted regions. However, we found that 80% also looks reasonable, and arguably even optimal, on our simulated datasets where the true answers are known. We also observed that the runtime performance of our package is fairly fast (e.g. it finished the more complicated datasets within 15 minutes on average), despite using a probabilistic model.

#### Additional material

**Additional file 1: Supplement.** This includes Table S1 - GenBank accession numbers of the new sequences; Table S2 - Summary of detected conversions in the five human gene clusters; Table S3 - Fraction of paralogous pairs by their number of conversion events, out of all paralogous sequence pairs; Table S4 - Fraction of bases by their number of conversion events (involved as either source or target), out of all bases involved in duplications; Table S5 - Hot-spot segments of conversion events; and Figure S1 - Performance of our CHAP pipeline based on a simulation study.

#### Acknowledgements

This work was supported by grant HG02238 to WM from the National Human Genome Research Institute and grant DK065806 to RCH from the National Institute of Diabetes, Digestive and Kidney Diseases.

#### Author details

<sup>1</sup>Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802 USA. <sup>2</sup>Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD, 20892 USA. <sup>3</sup>Department of Biochemistry and Molecular Biology, Mississippi State University, Mississippi State, MS 39760, USA. <sup>4</sup>Department of Mathematics, National University of Singapore, 117543, Singapore. <sup>5</sup>NIH Intramural Sequencing

Center (NISC) and Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Bethesda, MD 20892 USA.

#### Authors' contributions

GS performed the experiments using the CHAP package. GS, CH, and CR implemented the pipeline. GS and CR wrote and finalized the manuscript. CH and YZ designed the statistical method for conversion detection and helped to write the Methods section. GS and HK analyzed the results. CR, FH, LZ, RH, EG, and WM helped with the analysis and writing. NISC Comparative Sequencing Program sequenced three new clusters in seven primates for this study. WM initiated, supervised, and coordinated the work. All authors read and approved the final manuscript.

Received: 9 March 2011 Accepted: 28 July 2011 Published: 28 July 2011

#### References

1. Miller W, Rosenbloom K, Hardison R, Hou M, Taylor J, Raney B, Burhans R, King D, Baertsch R, Blankenberg D, Kosakovsky P, Nekrutenko A, Giardine B, Harris R, Tyekucheva S, Diekhans M, Pringle T, Murphy W, Lesk A, Weinstock G, Lindblad-Toh K, Gibbs R, Lander E, Siepel A, Haussler D, Kent WJ: **28-way vertebrate alignment and conservation track in the UCSC genome browser.** *Genome Res* 2007, **17**:1797-1808.
2. Siepel A, Bejerano G, Pedersen J, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier L, Richards S, Weinstock G, Wilson R, Gibbs R, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15**:1034-50.
3. Pollard KS, Salama S, King B, Kern A, Dreszer T, Katzman S, Siepel A, Pedersen J, Bejerano G, Baertsch R, Rosenbloom K, Kent WJ, Haussler D: **Forces shaping the fastest evolving regions in the human genome.** *PLoS Genet* 2006, **2**(10):e168.
4. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies.** *Genome Res* 2010, **20**:110-121.
5. Degenhardt JD, de Candia P, Chobot A, Schwartz S, Henderson L, Ling B, Hunter M, Jiang Z, Palermo RE, Katze M, Eichler EE, Ventura M, Rogers J, Marx P, Gilad Y, Bustamante CD: **Copy number variation of CCL3-like genes affects rate of progression to simian-AIDS in Rhesus Macaques (Macaca mulatta).** *PLoS Genet* 2009, **5**:e1000346.
6. Akahoshi M, Ishihara M, Remus N, Uno K, Miyake K, Hirota T, Nakashima K, Matsuda A, Kanda M, Enomoto T, Ohno S, Nakashima H, Casanova JL, Hopkin JM, Tamari M, Mao XQ, Shirakawa T: **Association between IFNA genotype and the risk of sarcoidosis.** *Hum Genet* 2004, **114**(5):503-509.
7. Wang H, Tan W, Hao B, Miao X, Zhou G, He F, Lin D: **Substantial reduction in risk of lung adenocarcinoma associated with genetic polymorphism in CYP2A13, the most active cytochrome P450 for the metabolic activation of tobacco-specific carcinogen NNK.** *Cancer Res* 2003, **63**(22):8057-8061.
8. Bailey J, Gu Z, Clark R, Reinert K, Samonte R, Schwartz S, Adams M, Myers E, Li P, Eichler E: **Recent segmental duplications in the human genome.** *Science* 2002, **297**:1003-1007.
9. Wong KM, Suchard MA, Huelsenbeck JP: **Alignment uncertainty and genomic analysis.** *Science* 2008, **319**:473-476.
10. Minin VN, Dorman KS, Fang F, Suchard MA: **Phylogenetic mapping of recombination hotspots in human immunodeficiency virus via spatially smoothed change-point processes.** *Genetics* 2007, **175**:1773-1785.
11. Chen J, Cooper D, Chuzhanova N, Ferenc C, Patrinos G: **Gene conversion: mechanisms, evolution and human disease.** *Nat Rev Genet* 2007, **8**:762-775.
12. Hsu C, Zhang Y, Hardison R, NISC Comparative Sequencing Program, Green E, Miller W: **An effective method for detecting gene conversion events in whole genomes.** *J Comput Biol* 2010, **17**:1281-1297.
13. Hsu C, Bedoya O, Zhang Y, Song G, Riemer C, Hardison R, NISC Comparative Sequencing Program, Green E, Miller W: **Non-allelic gene-conversion events for the entire human genome.** *in preparation* 2011.
14. Westesson O, Holmes I: **Accurate detection of recombinant breakpoints in whole-genome alignments.** *PLoS Comput Biol* 2009, **5**:e1000318.
15. Martin D, Williamson C, Posada D: **Rdp2: recombination detection and analysis from sequence alignments.** *Bioinformatics* 2005, **21**(2):260-262.
16. Sawyer S: **Statistical tests for detecting gene conversion.** *Mol Biol Evol* 1989, **6**:526-538.
17. Song G, Hsu C, Riemer C, Miller W: **Evaluation of methods for detecting conversion events in gene clusters.** *BMC Bioinformatics* 2011, **12**(Suppl 1): S45.
18. Boni M, Posada D, Feldman M: **An exact nonparametric method for inferring mosaic structure in sequence triplets.** *Genetics* 2007, **176**:1035-1047.
19. Song G, Zhang L, Vinar T, Miller W: **CAGE: Combinatorial Analysis of Gene-cluster Evolution.** *J Comput Biol* 2010, **17**:1227-1242.
20. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit A, Roskin K, Baertsch R, Rosenbloom K, Clawson H, Green E, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14**(4):708-715.
21. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0.** [http://www.repeatmasker.org], Unpublished 1996-2010.
22. Fitch W: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**:99-113.
23. Sonnhammer E, Koonin E: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genet* 2002, **18**:619-620.
24. Harris R: **Improved pairwise alignment of genomic DNA.** *Ph.D. thesis. Pennsylvania State University* 2007.
25. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14**:988-995.
26. Hardison RC: **Comparison of the beta-like globin gene families of rabbits and humans indicates that the gene cluster 5'-epsilon-gamma-delta-beta-3' predates the mammalian radiation.** *Mol Biol Evol* 1984, **1**:390-410.
27. Hardies SC, Edcell MH, Hutchison CA: **Evolution of the mammalian beta-globin gene cluster.** *J Biol Chem* 1984, **259**:3748-3756.
28. Hardison R, Miller W: **Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters.** *Mol Biol Evol* 1993, **10**(1):73-102.
29. Opazo JC, Hoffmann FG, Storz JF: **Genomic evidence for independent origins of beta-like globin genes in monotremes and therian mammals.** *Proc Natl Acad Sci USA* 2008, **105**(5):1590-1595.
30. Opazo JC, Hoffmann FG, Storz JF: **Differential loss of embryonic globin genes during the radiation of placental mammals.** *Proc Natl Acad Sci USA* 2008, **105**(35):12950-12955.
31. Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL: **Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates.** *Proc Natl Acad Sci USA* 1991, **88**(16):7396-7400.
32. Slightom JL, Blechl AE, Smithies O: **Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes.** *Cell* 1980, **21**(3):627-38.
33. Czelusniak J, Goodman M, Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE: **Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes.** *Nature* 1982, **298**:297-300.
34. Hoffmann FG, Opazo JC, Storz JF: **Rapid rates of lineage-specific gene duplication and deletion in the alpha-globin gene family.** *Mol Biol Evol* 2008, **25**(3):591-602.
35. Michelson AM, Orkin SH: **Boundaries of gene conversion within the duplicated human alpha-globin genes. Concerted evolution by segmental recombination.** *J Biol Chem* 1983, **258**(24):15245-15254.
36. Hess J, Schmid C, Shen C: **A gradient of sequence divergence in the human adult alpha-globin duplication units.** *Science* 1984, **226**:67-70.
37. Adams JG, Morrison WT, Steinberg MH: **Hemoglobin Parichman: double crossover within a single human gene.** *Science* 1982, **218**:291-293.
38. Su T, Bao Z, Zhang Q-Y, Smith TJ, Hong J-Y, Ding X: **Human cytochrome P450 CYP2A13: predominant expression in the respiratory tract and its high efficiency metabolic activation of a tobacco-specific carcinogen, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone.** *Cancer Res* 2000, **60**:5074-5079.
39. Hecht SS: **Biochemistry, biology and carcinogenicity of tobacco-specific N-nitrosamines.** *Chem Res Toxicol* 1998, **11**:559-603.
40. Koskela S, Hakkola J, Hukkanen J, Pelkonen O, Sorri M, Saranen A, Anttila S, Fernandez-Salguero P, Gonzalez F, Raunio H: **Expression of CYP2A genes in human liver and extrahepatic tissues.** *Biochem Pharmacol* 1999, **57**:1407-1413.

41. Zhang X, Su T, Zhang QY, Gu J, Caggana M, Li H, Ding X: **Genetic polymorphisms of the human CYP2A13 gene: identification of single-nucleotide polymorphisms and functional characterization of an Arg257Cys variant.** *J Pharmacol Exp Ther* 2002, **302**:416-423.
42. Zhang X, Chen Y, Liu Y, Ren X, Zhang Q-Y, Caggana M, Ding X: **Single nucleotide polymorphisms of the human CYP2A13 gene: evidence for a null allele.** *Drug Metab Dispos* 2003, **31**:1081-1085.
43. Schlicht KE, Michno N, Smith BD, Scott EE, Murphy SE: **Functional characterization of CYP2A13 polymorphisms.** *Xenobiotica* 2007, **37**:1439-1449.
44. D'Agostino J, Zhang X, Wu H, Ling G, Wang S, Zhang QY, Liu F, Ding X: **Characterization of CYP2A13\*2, a variant cytochrome P450 allele previously found to be associated with decreased incidences of lung adenocarcinoma in smokers.** *Drug Metab Dispos* 2008, **36**:2316-2323.
45. Osada N, Innan H: **Duplication and gene conversion in the *Drosophila melanogaster* genome.** *PLoS Genet* 2008, **4**:e1000305.
46. Casola C, Ganote C, Hahn M: **Nonallelic Ggene Cconversion in the Genus *Drosophila*.** *Genetics* 2010, **185**:95-103.
47. Lazzaro BP, Clark AG: **Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the *Attacin* genes of *Drosophila melanogaster*.** *Genetics* 2001, **159**:659-671.
48. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, Zhang H, The FlyBase Consortium: **FlyBase: enhancing *Drosophila* Gene Ontology annotations.** *Nucleic Acids Res* 2009, **37**:D555-D559.
49. Mueller JL, Ripoll DR, Aquadro CF, Wolfner MF: **Comparative structural modeling and inference of conserved protein classes in *Drosophila* seminal fluid.** *Proc Natl Acad Sci USA* 2004, **101**(37):13542-13547.
50. Allaby RG, Woodwark M: **Phylogenomic analysis reveals extensive phylogenetic mosaicism in the human GPCR superfamily.** *Evol Bioinform* 2007, **26**(3):357-70.
51. Miller W: **Comparison of genomic DNA sequences: solved and unsolved problems.** *Bioinformatics* 2001, **17**:391-397.
52. Chiaromonte F, Yap V, Miller W: **Scoring pairwise genomic sequence alignments.** *Pac Symp Biocomput* 2002, 115-126.
53. Opazo JC, Wildman DE, Prychitko T, Johnson RM, Goodman M: **Phylogenetic relationships and divergence times among New World monkeys (Platyrrhini, Primates).** *Mol Phylogenet Evol* 2006, **40**(1):274-80.
54. Goodman M, Grossman LI, Wildman DE: **Moving primate genomics beyond the chimpanzee genome.** *Trends Genet* 2005, **21**(9):511-7.

doi:10.1186/1471-2148-11-226

Cite this article as: Song et al.: Conversion events in gene clusters. *BMC Evolutionary Biology* 2011 **11**:226.

Submit your next manuscript to BioMed Central  
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

