# CTDGFinder: A Novel Homology-Based Algorithm for Identifying Closely Spaced Clusters of Tandemly Duplicated Genes

Juan F. Ortiz and Antonis Rokas*
Department of Biological Sciences, Vanderbilt University, Nashville, TN
*Corresponding author: E-mail: antonis.rokas@vanderbilt.edu
Associate editor: Claus Wilke

## Abstract

**Closely spaced clusters of tandemly duplicated genes (CTDGs) contribute to the diversity of many phenotypes, including chemosensation, snake venom, and animal body plans. CTDGs have traditionally been identified subjectively as genomic neighborhoods containing several gene duplicates in close proximity; however, CTDGs are often highly variable with respect to gene number, intergenic distance, and synteny. This lack of formal definition hampers the study of CTDG evolutionary dynamics and the discovery of novel CTDGs in the exponentially growing body of genomic data. To address this gap, we developed a novel homology-based algorithm, CTDGFinder, which formalizes and automates the identification of CTDGs by examining the physical distribution of individual members of families of duplicated genes across chromosomes. Application of CTDGFinder accurately identified CTDGs for many well-known gene clusters (e.g., Hox and beta-globin gene clusters) in the human, mouse and 20 other mammalian genomes. Differences between previously annotated gene clusters and our inferred CTDGs were due to the exclusion of nonhomologs that have historically been considered parts of specific gene clusters, the inclusion or absence of genes between the CTDGs and their corresponding gene clusters, and the splitting of certain gene clusters into distinct CTDGs. Examination of human genes showing tissue-specific enhancement of their expression by CTDGFinder identified members of several well-known gene clusters (e.g., cytochrome P450s and olfactory receptors) and revealed that they were unequally distributed across tissues. By formalizing and automating CTDG identification, CTDGFinder will facilitate understanding of CTDG evolutionary dynamics, their functional implications, and how they are associated with phenotypic diversity.**

***Key words:*** **synteny, tandem gene duplication, genetic linkage, functional divergence.**

## Introduction

Gene duplications are among the most frequent types of mutational changes in genomes (Reams and Roth 2015) and arguably the largest source of novel gene functions (Lynch and Conery 2003; Zhang 2003; Andersson and Hughes 2009). Gene duplication can occur by many different mechanisms (Zhang 2003), including transposition (Freeling et al. 2008), polyploidization (Grant et al. 2000; Carretero-Paulet and Fares 2012), and recombination (Krause and Pestka 2015). Recombination-based gene duplication results in tandem gene duplication, in which the gene duplicates lie adjacent to each other and are closely spaced on the chromosome (Wu and Maniatis 1999; Glusman et al. 2000; Kawasaki and Weiss 2003). Chromosomal regions containing multiple homologs that have arisen through tandem gene duplication are common features of genomes, and are often described as clusters of tandemly duplicated genes (CTDGs) (Krumlauf 1992; Martin et al. 2000; Noonan et al. 2004; Alam et al. 2006; MacLean et al. 2006; Yagi 2008).

Notable examples of CTDGs include the vertebrate protocadherin gene clusters (Wu et al. 2001; Noonan et al. 2004), the vertebrate and invertebrate olfactory receptor gene clusters (Hallem et al. 2006; Niimura 2009; Niimura et al. 2014), the vertebrate natural killer cell receptor gene clusters (Kelley et al. 2005), and the Hox gene clusters found in one or more copies across metazoans (Hoffman et al. 1995; Ferrier and Holland 2001; Glusman et al. 2000; Martin et al. 2000; Noonan et al. 2004). Many CTDGs contribute to traits that are highly variable, such as the composition of snake venom (Vonk et al. 2013), the architecture of animal body plan formation (Pendleton et al. 1993), the olfactory repertoire (Glusman et al. 2000) or the immune response (Martin et al. 2000).

Given the many CTDGs from diverse gene families found in a wide diversity of organisms, the absence of a formal definition of what constitutes a "cluster of tandemly duplicated genes" is surprising. The standard, informal definition that unites the known examples of CTDGs is that they represent groups of duplicated genes that are closely spaced (Graham 1995), although CTDGs sometimes also contain nonhomologous genes (Hallast et al. 2008). As practical as this definition may be, it is subjective. For example, should two genes located adjacent to each other on a chromosome be considered a cluster? Answering this question is challenging without considering the probability of observing two duplicates next to each other in the chromosome, which in turn







Article

requires knowledge of the number and distribution of duplicates in the genome as well as comparison of the intergenic distance between genes in the cluster with those in the rest of the chromosome.

An examination of the organization of the Hox gene cluster across diverse metazoans, which is often portrayed as a conserved, organized, and temporally and spatially clustered set of duplicated genes (Pendleton et al. 1993; Garcia-Fernàndez 2005; Lemons and McGinnis 2006), is a good case in point. Whereas genes in vertebrate Hox gene clusters are typically closely spaced and encoded on the same strand, Hox gene clusters in other animal phyla show striking differences in the number of genes that are members of the gene cluster, in their intergenic spacing, as well as in their general organization (Duboule 2007). For example, the Hox gene cluster in sea urchin *Strongylocentrotus purpuratus* contains two non-Hox genes, and its constituent genes exhibit long intergenic distances and are encoded in both strands (Cameron et al. 2006; Duboule 2007) (fig. 1). In the fruit fly *Drosophila melanogaster*, the Hox "gene cluster" is actually composed by two distinct clusters; the ANT-C cluster, which contains 11 non-Hox genes and five Hox genes encoded in both strands (five non-Hox genes and four Hox genes are in the 3′ strand and the others in the 5′ strand), and the BX-C cluster, which contains three Hox genes in the 3′ strand and two non-Hox genes, one in each strand (fig. 1). In contrast, the HoxD gene cluster in the mouse *Mus musculus*—one of the four Hox clusters in this organism—is composed by nine contiguous homologous genes in the same strand (fig. 1).

In this study, we propose a formal definition for a CTDG that takes into account the sequence similarity of its members and their intergenic distances in the context of their chromosomal and genomic background to statistically assess whether neighboring gene duplicates form a CTD. We further implement our definition of CTDG in CTDGFinder, a computational tool for the identification of CTDGs, and use it to examine the statistical validity of well-known gene clusters as well as explore the CTDG landscape across different human tissues.

## Results

### Defining a Cluster of Tandemly Duplicated Genes

We define a cluster of tandemly duplicated genes (CTDG) in a given genome as a genomic region that contains a statistically significant higher number of tandemly duplicated genes from a specific gene family than the average background genomic region of the same length.

### The CTDGFinder Algorithm

To formally define and identify CTDGs we developed the CTDGFinder algorithm (fig. 2), which uses sequence similarity and the density of the distribution of duplicated genes across the genome to statistically assess and demarcate the presence of CTDGs in a genome. CTDGFinder is written in Python and is freely available from https://github.com/biofilos/ctdg_finder. Briefly, given a query reference protein sequence, or a set of homologous reference protein sequences, and a

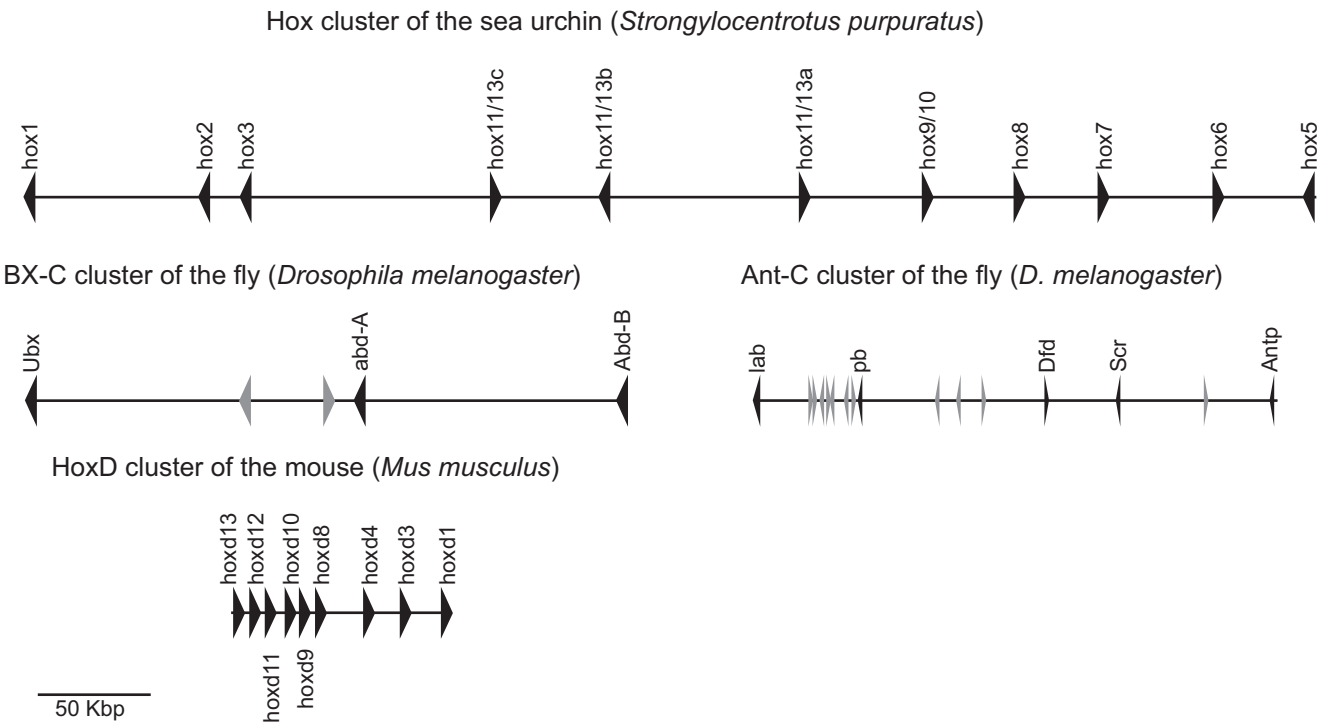


**FIG. 1.** An illustration of the variation in the genomic organization of Hox "gene clusters" in three different animal species. The Hox gene cluster of the sea urchin (*Strongylocentrotus purpuratus*), the Bithorax complex (BX-C), and Antennapedia complex (Ant-C) clusters of the fruit fly (*Drosophila melanogaster*) and the HoxD cluster of the mouse (*Mus musculus*). Genes belonging to the Hox gene family are shown in black and intervening non-Hox genes in gray.

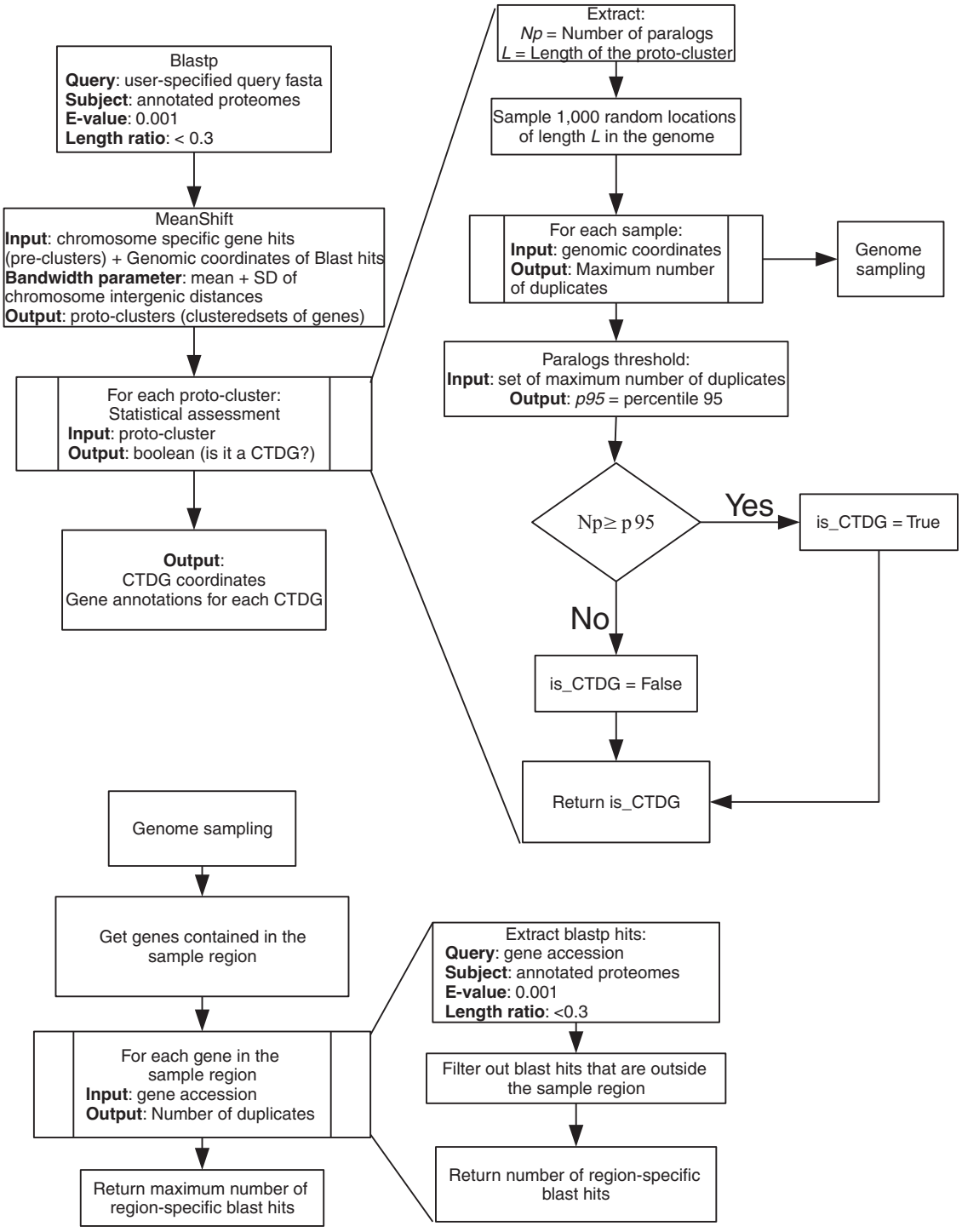


**FIG. 2.** Overview of the CTDGFinder algorithm.

subject genome or set of genomes, CTDGFinder uses the BLAST algorithm (Altschul et al. 1990) to identify sequences that are statistically significantly similar (homologs) in the subject genome(s). The sets of closely spaced duplicated genes identified on each of the chromosomes or genomic scaffolds of a given subject genome are considered candidate clusters.

Whether the genes in these candidate clusters comprise one or more genuine clusters of tandemly duplicated genes (CTDGs) is evaluated and determined by the meanshift algorithm (Comaniciu et al. 2002). Specifically, the meanshift algorithm treats a parameter space as an empirical density function, and its objective is to find the region(s) of the parameter space with the highest density (or densities). In the context of the distribution of duplicated genes across a given chromosome, the meanshift algorithm identifies the genomic region(s) with the highest density (or densities) of duplicated genes of the protein reference sequence(s). Statistical assessment is performed by comparing the number of duplicated genes present in the genomic region(s) identified by the meanshift algorithm against an empirical

distribution of duplicated genes. Such an empirical distribution is obtained by counting the highest number of duplicated genes from any gene family contained in each of 1,000 randomly sampled genomic regions from the same genome of length equal to that of the cluster candidate from the meanshift step. Genomic regions with more duplicated genes than the genome-wide 95th percentile of this empirical distribution are considered CTDs (an overview of the algorithm is shown in fig. 2).

## CTDGFinder Recovers Several Well-Known CTDGs

To evaluate the performance of CTDGFinder, we first examined whether it was able to identify a diverse set of previously characterized, well-known gene clusters in the human and mouse genomes. To capture as much sequence diversity as possible, we retrieved all the genes from each of eight published gene clusters in either the mouse (*Mus musculus*) or in the human (*Homo sapiens*) genome and used them as a combined query in CTDGFinder to identify CTDGs in the same or in the other genome. The lists of mouse and human gene clusters used, and of the mouse and human genes used as queries for CTDGFinder are described in supplementary table S1, Supplementary Material online.

### *CTDGFinder Performance When Searching the Same Genome*

Using previously reported genes from eight gene clusters found in the human and/or mouse genomes as queries, CTDGFinder correctly identified the CTDGs in their corresponding genomes (table 1). For example, all five genes reported to be part of the growth hormone gene cluster on the human chromosome 17 (Su et al. 2000), all 26 protocadherin genes on the mouse chromosome 18 (Kohmura et al., 1998), all 11 HoxA genes in the human (chromosome 7; Krumlauf 1994) and mouse (chromosome 6; Krumlauf 1994) genomes, all five beta-globin genes in the human (chromosome 11; Levings and Bungert 2002) and mouse (chromosome 5; Weaver et al. 1981; Bulger et al. 1999) genomes, and all seven duplicated genes that are part of the luteinizing hormone beta (LHB) gene cluster on the human chromosome 19 (Hallast et al. 2008) were identified as statistically significant CTDGs by CTDGFinder.

CTDGFinder also correctly identified known additional CTDGs in searches of the same genome (table 1). For example, using all 11 HoxA genes from human chromosome 7 as a query, CTDGFinder correctly identified CTDGs corresponding the human HoxA, HoxB, HoxC, and HoxD clusters in the human genome; the same results were obtained for mouse (table 1).

The correspondence between the identified CTDGs and the previously described gene clusters was very good (table 1) and the few observed differences fell into three categories. The first concerned differences associated with the exclusion of nonhomologous genes that have historically been annotated as part of the gene cluster. For example, the RUVBL2 and NTF5 genes have historically been considered parts of mammalian LHB gene clusters, but their inclusion simply reflects the knowledge that those genes flank the LHB genes in human and chimpanzee.

The second category included cases in which the CTDGs contained one or a few additional genes not originally described as part of the gene cluster. For example, the HoxB CTDGs in both the human (chromosome 17) and mouse (chromosome 11) genomes included Hoxb13, which was not originally described as part of the mammalian HoxB gene cluster (Krumlauf 1994), but was subsequently added to it (Zeltser et al. 1996). Similarly, using the 52 protocadherin genes in human chromosome 5 as a query (Wu and Maniatis 1999), CTDGFinder identified a CTDG that additionally contained gene PCDHB16, whereas using the four Siglec genes on mouse chromosome 7 as a query (Kohmura et al. 1998), CTDGFinder identified a 7-gene CTDG that additionally contained genes 4931406B18Rik, Iglon5, and Vsig10L (table 1).

The third and arguably most interesting category included cases in which CTDGFinder split a previously described gene cluster into two distinct sub-clusters, which may or may not be both CTDGs. For example, CTDGFinder identified two distinct galectin CTDGs, a 4-gene and a 3-gene one, instead of the single 7-gene cluster previously reported to reside on human chromosome 19 (Than et al. 2009). CTDGFinder identified two separate CTDGs because the average intergenic distances of the genes in the 3-gene and 4-gene CTDs are 13 and 38 kb, respectively, which are significantly smaller than the 790 kb that separates the two CTDs. Interestingly, examination of the tissue expression patterns of the genes in the two galectin CTDGs using data from the ProteinAtlas project (Uhlen et al. 2015) showed that expression of the genes in the 4-gene CTDG was enhanced in the placenta and bone marrow, whereas expression of the genes in the 3-gene CTDG was enhanced in the skin and digestive tract (supplementary table S6, Supplementary Material online). Similarly, genes previously reported as part of a single 9-gene Siglec cluster on human chromosome 19 (Cao et al. 2009) were identified by CTDGFinder as parts of two separate CTDGs, containing four and seven genes (with intergenic distances of 38 and 29 kb), respectively, separated by 142 kb In addition to the genes reported in the literature, the 4-gene CTDG additionally contained the SiglecL1 gene, and the 7-gene CTDG additionally contained the loc105372490 gene. Interestingly, the two CTDGs directly correspond with the A and B subclusters that resulted from an inverse duplication of the CD33rSiglec cluster in eutherian mammals (Cao et al. 2009). No significant difference in tissue expression patterns was found between the CTDGs (supplementary table S6, Supplementary Material online). Finally, using the 26 genes in the prolactin cluster on mouse chromosome 13 (Simmons et al. 2008) as a query, CTDGFinder identified a single 24-gene CTDG (table 1). An additional candidate gene cluster located 13.7 Mbp away and comprised of the genes Prl2c3, Prl2c2, and Prl2c5 was also identified, but it was below the 95th percentile of the empirical distribution of paralogs for a genomic region of that size and was not recognized as a CTDG.

**Table 1.** Comparison of Eight Human and Mouse CTDGs Inferred by CTDGFinder with the Corresponding Gene Clusters Previously Described in the Literature.

| Query Genes | Subject Genome | | Number of Genes in CTDG Predicted by CTDGFinder | Notes/Differences between Predicted CTDG and Previously Described Gene Cluster | References |
|---|---|---|---|---|---|
| | Species | Gene Cluster (chromosome) | | | |
| The 11 HoxA genes in chromosome 7 from *Homo sapiens* | *H. sapiens* | HoxA (7) | 11 | None | (Krumlauf 1992, 1994; Zeltser et al. 1996) |
| | *H. sapiens* | HoxB (17) | 10 | None | |
| | *H. sapiens* | HoxC (12) | 9 | None | |
| | *H. sapiens* | HoxD (2) | 9 | None | |
| The 11 HoxA genes in chromosome 6 from *Mus musculus* | *M. musculus* | HoxA (6) | 11 | None | |
| | *M. musculus* | HoxB (11) | 10 | None | |
| | *M. musculus* | HoxC (15) | 9 | None | |
| | *M. musculus* | HoxD (2) | 9 | None | |
| The seven LHB genes in chromosome 19 from *Homo sapiens* | *H. sapiens* | LHB (19) | 7 | Genes RUVBL2 and NTF5 were not used in the search because they are not homologous to LHB genes | (Hallast et al. 2008) |
| The five GH genes in chromosome 17 from *Homo sapiens* | *H. sapiens* | GH (17) | 5 | None | (Su et al. 2000) |
| The 26 prolactin genes in chromosome 13 from *Mus musculus* | *M. musculus* | Prolactin (13) | 24 | In the 24-gene CTDG, Prl2c1 was identified between Prl2a1 and Prl4a1, instead of the reported Plf. In addition to the reported gene duplicates, Gm3821 was also identified as part of the CTDG. The remaining three genes (Prl2c4, Prl2c2 and Prl2c5) were identified as an independent cluster candidate but did not pass the statistical threshold to be considered a CTDG | (Simmons et al. 2008) |
| The seven galectin genes in chromosome 19 from *Homo sapiens* | *H. sapiens* | Galectins (19) | 4/3 | Genes previously reported as part of a single gene cluster were identified by CTDGFinder as parts of two separate CTDGs containing four and three galectin genes, respectively. LGALS17 was not included in the list of query genes and was not identified as part of the CTDG because it is annotated as a pseudogene | (Than et al. 2009) |
| The two galectin genes in chromosome 7 from *Mus musculus* | *M. musculus* | Galectins (7) | 2 | None | (Houzelstein et al. 2004) |
| The 52 protocadherin genes in chromosome 5 from *Homo sapiens* | *H. sapiens* | Protocadherins (5) | 53 | In addition to the genes reported in the literature, PCDHB16 was identified as part of the CTDG | (Wu and Maniatis 1999) |
| The 26 protocadherin genes in chromosome 18 from *Mus musculus* | *M. musculus* | Protocadherins (18) | 26 | None | (Kohmura et al. 1998) |
| The five globin-beta genes in chromosome 11 from *Homo sapiens* | *H. sapiens* | Globins (11) | 5 | None | (Levings and Bungert 2002) |
| The five globin-beta genes in chromosome 7 from *Mus musculus* | *M. musculus* | Globins (7) | 5 | None | (Weaver et al. 1981; Bulger et al. 1999) |
| The nine SIGLEC genes in chromosome 19 from *Homo sapiens* | *H. sapiens* | SIGLECs (19) | 4/7 | Genes previously reported as part of a single gene cluster were identified as parts of two separate CTDGs, containing 4 and 7 genes, respectively. In addition to the genes reported in the literature, the 4-gene CTDG additionally contained SiglecL1, and the 7-gene CTDG additionally contained loc105372490 | (Cao et al. 2009) |
| The four SIGLEC genes in chromosome 7 from *Mus musculus* | *M. musculus* | SIGLECs (7) | 7 | In addition to the genes reported in the literature, the CTDG additionally contained 4931406B18Rik, Iglon5 and Vsig10L | (Angata et al. 2004) |

**Table 2.** Comparison of Mouse CTDGs Using Human Queries Inferred by CTDGFinder with Gene Clusters Previously Described in the Literature.

| Query Genes | Subject Genome | | Number of Genes in CTDG Predicted by CTDGFinder | Notes/Differences between Predicted CTDG and Previously Described Gene Cluster | References |
|---|---|---|---|---|---|
| | Species | Gene Cluster (chromosome) | | | |
| The 11 HoxA genes in chromosome 7 from *H. sapiens* | *M. musculus* | HoxA (6) | 11 | None | (Krumlauf 1992, 1994; Zeltser et al. 1996) |
| | *M. musculus* | HoxB (11) | 10 | None | |
| | *M. musculus* | HoxC (15) | 9 | None | |
| | *M. musculus* | HoxD (2) | 9 | None | |
| The seven galectin genes in chromosome 19 from *H. sapiens* | *M. musculus* | Galectins (7) | 2 | None | (Houzelstein et al. 2004) |
| The 52 protocadherin genes in chromosome 5 from *H. sapiens* | *M. musculus* | Protocadherins (18) | 24 | Pcdh1 and Pcdh12 were not found (these two genes were found as part of the CTDG when mouse genes were used as the query) | (Kohmura et al. 1998) |
| The five globin-beta genes in chromosome 11 from *H. sapiens* | *M. musculus* | Globins (7) | 5 | None | (Weaver et al. 1981; Bulger et al. 1999) |
| The nine SIGLEC genes in chromosome 19 from *H. sapiens* | *M. musculus* | SIGLECs (7) | 6 | In addition to the genes reported in the literature, 4931406B18Rik and Vsig10L were identified as part of the CTDG | (Angata et al. 2004) |

### *CTDGFinder Performance When Searching Different Genomes*

Given that CTDGFinder correctly identified several well-known gene clusters in searches of the same genome, including known additional homologous CTDGs (table 1), we next sought to examine the performance of CTDGFinder on the human genome when using gene queries from the mouse genome (table 2) as well as on the mouse genome when using gene queries from the human genome (table 3).

In general, CTDGFinder correctly identified CTDGs for several well-known gene clusters in the human and mouse genomes. For example, using the 11 HoxA genes from the human genome as a query, CTDGFinder correctly identified a CTDG corresponding to the mouse HoxA gene cluster (table 2); similarly, using the 11 mouse HoxA genes as a query, CTDGFinder identified the human HoxA gene cluster (table 3). The same was true for the mouse galectin and beta-globin gene clusters, which were identified using their human homologs (table 2), as well as for the human beta-globin gene cluster, which was identified by CTDGFinder using mouse homologs (table 3). CTDGFinder also correctly identified known additional homologous CTDGs in a given genome. For example, using all 11 HoxA genes from human chromosome 7, CTDGFinder correctly identified CTDs corresponding to the HoxA, HoxB, HoxC, and HoxD clusters in the mouse genome (table 2). Likewise, using the mouse HoxA genes as a query, CTDGFinder correctly identified CTDGs corresponding to the human HoxA, HoxB, HoxC, and HoxD gene clusters (table 3).

Similarly to the results of the performance of CTDGFinder when searching the same genome, the differences between the inferred CTDGs and the previously described gene clusters when searching other genomes fell into three categories. The first category concerned differences associated with the exclusion of nonhomologous genes that have historically been considered to be parts of specific gene clusters, the second included single or a few gene differences between the CTDGs and their corresponding gene clusters, and the third cases in which a previously described gene cluster was split by CTDGFinder into two distinct CTDGs (tables 2 and 3).

The most conspicuous differences between the identified CTDGs and the previously described gene clusters were observed in the mouse and human Siglec gene clusters. Specifically, using the nine Siglec genes on human chromosome 19 as a query, CTDGFinder identified a 6-gene CTDG on mouse chromosome 7 that contained two additional genes (4931406B18Rik and Vsig10L) in addition to those previously described for the mouse Siglec gene cluster (Angata et al. 2004). Furthermore, using the four Siglec genes on mouse chromosome 7 as a query, CTDGFinder identified two separate CTDGs 69.0 kb away from each other, containing five and seven genes (with average intergenic distances of 44.5 and 29.4 kb), respectively, that correspond to the A and B subclusters previously identified by Cao et al. (2009; table 2). In addition to the genes previously reported (Cao et al. 2009), the 5-gene CTDG additionally contained SiglecL1 and VSIG10L, and the 7-gene CTDG additionally contained

**Table 3.** Comparison of Human CTDGs Using Mouse Queries Inferred by CTDGFinder with Gene Clusters Previously Described in the Literature.

| Query Genes | Subject Genome | | Number of Genes in CTDG Predicted by CTDGFinder | Notes/Differences between Predicted CTDG and Previously Described Gene Cluster | References |
|---|---|---|---|---|---|
| | Species | Cluster (chromosome) | | | |
| The 11 HoxA genes in chromosome 6 from *M. musculus* | *H. sapiens* | HoxA (7) | 11 | none | (Krumlauf 1992, 1994; Zeltser et al. 1996) |
| | *H. sapiens* | HoxB (17) | 10 | none | |
| | *H. sapiens* | HoxC (12) | 9 | none | |
| | *H. sapiens* | HoxD (2) | 9 | none | |
| The two galectin genes in chromosome 7 from *M. musculus* | *H. sapiens* | Galectins (19) | 4/3 | Genes previously reported as part of a single gene cluster were identified by CTDGFinder as parts of two separate CTDGs containing four and three galectin genes, respectively | (Than et al., 2009) |
| The 26 protocadherin genes in chromosome 18 from *M. musculus* | *H. sapiens* | Protocadherins (5) | 55 | In addition to the genes reported in the literature, PCDHB16, PCDH1, PCDH12 were identified as part of the CTDG | (Wu and Maniatis 1999) |
| The five globin-beta genes in chromosome 7 from *M. musculus* | *H. sapiens* | Globins (11) | 5 | none | (Levings and Bungert 2002) |
| The four SIGLEC genes in chromosome 7 from *M. musculus* | *H. sapiens* | SIGLECs (19) | 5/7 | Genes previously reported as part of a single gene cluster were identified as parts of two separate CTDGs, containing five and seven genes, respectively. In addition to the genes reported in the literature, the 5-gene CTDG additionally contained SiglecL1 and VSIG10L, and the 7-gene CTDG additionally contained loc105372490 | (Cao et al. 2009) |

loc105372490 (table 3). Interestingly, the human gene VSIG10L was identified as part of this CTDG only when using mouse Siglec genes as queries (tables 1 and 3). This is because VSIG10L shows statistically significant sequence similarity only to the mouse gene Siglecg but not to any human Siglec genes.

## Identifying CTDGs across Placental Mammals

Given that CTDGFinder performed very well in recovering several well-known CTDGs in the human and mouse genomes, we next used all the paralogs from each of six published gene clusters in the human genome and used them as a combined query in CTDGFinder (supplementary table S1, Supplementary Material online) to identify CTDGs from the same seven gene families (galectin, Hox, beta-globin, Siglec, protocadherin, LHB, and growth hormone/prolactin) in 20 other mammalian genomes (fig. 3; supplementary table S2, Supplementary Material online).

Overall, CTDGFinder identified all the CTDGs that are expected to be present and conserved in these 20 mammalian genomes (supplementary table S2, Supplementary Material online). For example, running CTDGFinder with the HoxA genes from human as a query identified four Hox clusters in all the studied genomes, with the exceptions of HoxA and HoxB in vole (*Microtus ochrogaster*), HoxB in orangutan (*Pongo abelii*), and HoxC in opossum (*Monodelphis domestica*) (supplementary table S2, Supplementary Material online). The reason for these exceptions is that these clusters are present in nonassembled scaffolds that are not part of the standard genome assemblies provided by GenBank.

Using the human beta-globins as the query, CTDGFinder also identified both alpha- and beta-globin CTDGs in all species, except in vole where only the alpha-globin CTDG was found, and in horse (*Equus caballus*) and rabbit (*Oryctolagus cuniculus*) where only the beta-globin CTDG was found (alpha- and beta-globin CTDGs were differentiated by constructing their phylogeny; see methods). The number of genes contained in the beta-globin CTDG varied across species (fig. 3; supplementary tables S2 and S3, Supplementary Material online). The largest beta-globin CTDG was found in goat (*Capra hircus*) and stemmed from a previously reported beta-globin cluster duplication (Hardies et al. 1984), followed by the rat (*Rattus norvegicus*) 8-gene CTDG. Most species contained either 5-, 4-, or 3-gene CTDGs, but opossum (*Monodelphis domestica*) contained a 2-gene CTDG. This variation is likely due to both gene duplicates gain and loss as well as errors in annotation.

Running CTDGFinder with the 52 reported genes from the human protocadherin cluster (Wu and Maniatis 1999) as a query identified one protocadherin CTDG per species, with the number of gene duplicates per CTDG ranging from 15 in dog (*Canis lupus familiaris*), to 53 in human (supplementary table S2, Supplementary Material online). The only exception was pig, where a 13-gene CTDG (average intergenic distance of 13.1 kb) and a 20-gene CTDG (average intergenic distance of 16.0 kb) were found on the same chromosome but separated by 213 kb.

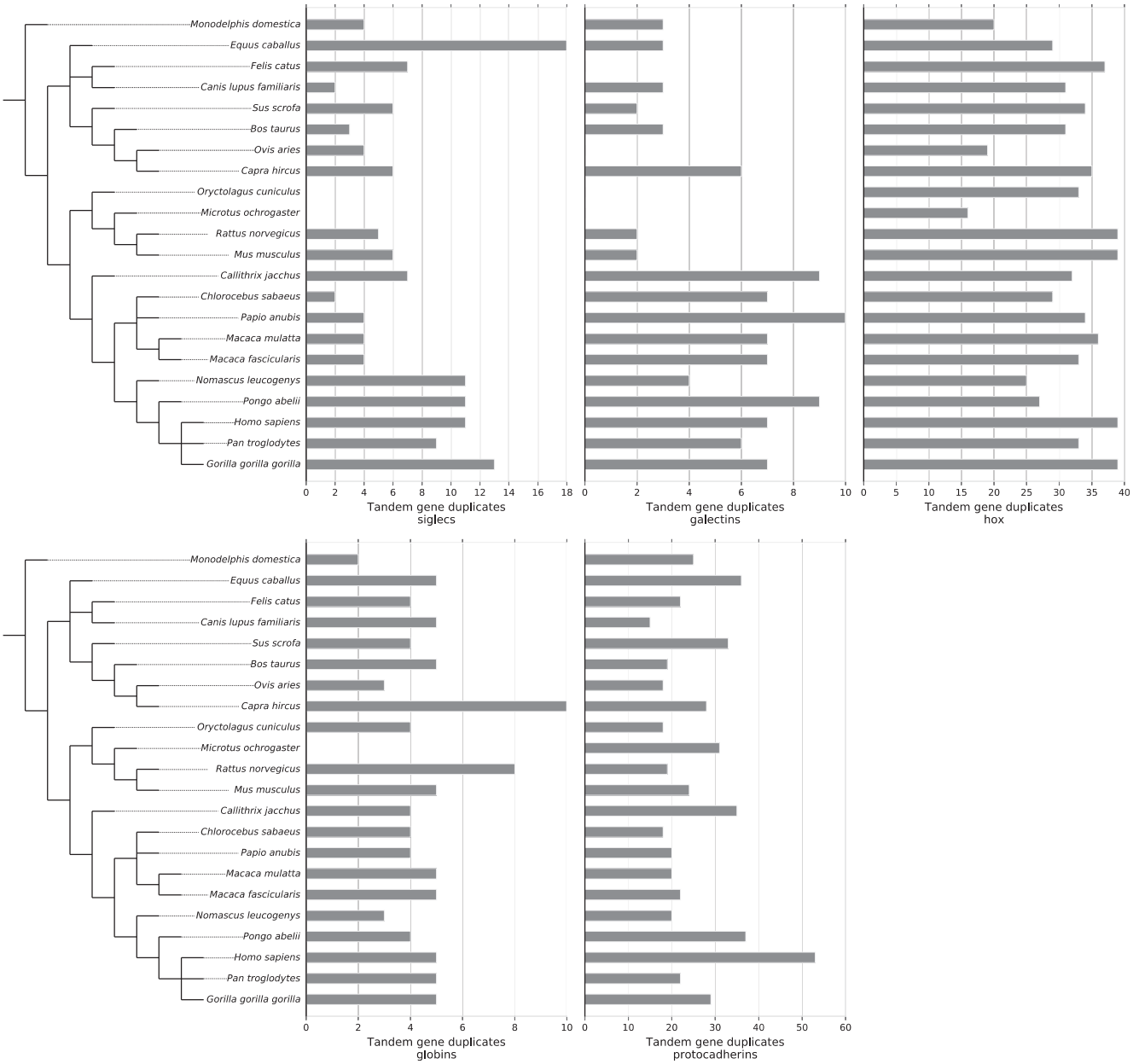


**FIG. 3.** Distribution of CTDGs in five selected gene families across 22 mammalian genomes. For gene families with more than one CTDG in a given organism, numbers of gene duplicates reflect the total number found in all CTDGs associated with that gene family.

The galectin CTDG in all species was identified by CTDGFinder using the 7-gene human galectin cluster reported by Than et al. (2009) as a query. Orthologous galectin CTDGs were extracted from all the galectin CTDGs using a phylogenetic tree rooted on the clade containing the galectins LGALS1 and LGALS2, following a previously reported galectin phylogeny (Houzelstein et al. 2004) (fig. 3; supplementary tables S2 and S3, Supplementary Material online). No orthologous CTDGs were found in cat (*Felis catus*), vole, rabbit and sheep, although vole and cat contained other homologous galectin CTDGs. One orthologous 3-gene CTDG was found in cow, dog and horse, whereas mouse, rat and pig contained one orthologous 2-gene CTDG. All other mammals contained two orthologous galectin CTDGs containing 2-3, and 3-7 genes respectively (supplementary table S2, Supplementary Material online).

Using the protein sequences from the previously reported Siglec cluster in human (Cao et al. 2009), CTDGFinder identified a Siglec CTDG in cow, dog, goat, green monkey (*Chlorocebus sabaeus*), horse, cat, rhesus macaque (*Macaca mulatta*), mouse, sheep, chimpanzee (*Pan troglodytes*), baboon (*Papio Anubis*), orangutan, rat, and pig (supplementary table S2, Supplementary Material online). The number of paralogs in these CTDGs ranged from two in dog and green monkey, to 11 in orangutan and 18 in horse. No Siglec CTDG was found in rabbit and vole. Two Siglec CTDGs were found in marmoset (*Callithrix jacchus*), gorilla (*Gorilla gorilla*), crab-eating macaque (*Macaca fascicularis*), human, and gibbon (*Nomascus leucogenys*). In all these cases, the number of genes in the first (ordered by genomic coordinates) CTDG (ranging between two and six gene duplicates) was smaller or equal to

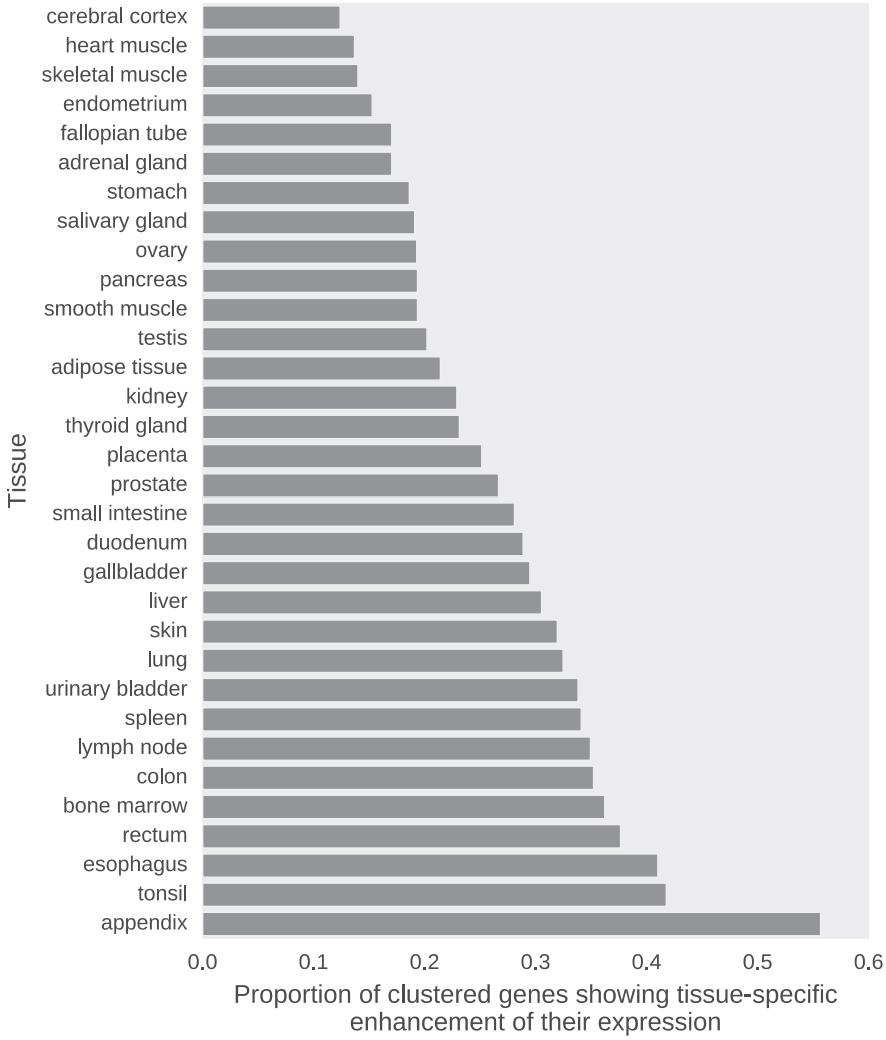


**FIG. 4.** Relative abundance of clustered genes (i.e., genes contained in CTDGs) in each tissue showing tissue-specific enhancement of their expression across 32 human tissues and organs. Genes showing tissue-specific enhancement were those that exhibited "at least five-fold higher mRNA levels in a particular tissue as compared with average levels in all tissues" (Uhlen et al. 2015).

the number of genes in the second CTDG (2–9 gene duplicates) (fig. 3; supplementary table S2 and S3, Supplementary Material online).

Finally, CTDGFinder accurately identified CTDGs for clusters that are known to be taxonomically restricted to certain lineages. For example, using the genes from the human growth hormone and LHB clusters, CTDGFinder identified CTDGs for growth hormone and LHB in primate genomes (supplementary table S2, Supplementary Material online), consistent with the previously described primate-specific presence of these gene clusters (Su et al. 2000; Hallast et al. 2008). However, a small CTDG of two LHB gene duplicates was identified in cow and horse; interestingly, the duplications that led to the formation of these CTDGs appear to have been independent of the duplications that led to the formation of the primate LHB cluster. Similarly, using the genes from the prolactin cluster in mouse as a query, CTDGFinder was able to identify CTDGs for previously described prolactin clusters in certain rodent (rat and vole) and bovid species (cow) (Wallis 1991; Simmons et al. 2008) (supplementary table S2, Supplementary Material online). Among the bovids, a previously unreported 12-gene prolactin CTDG was identified in goat (*Capra hircus*).

## What Fraction of Genes Showing Enhanced Expression in Human Tissues Is Clustered?

To illustrate the potential of CTDGFinder as a tool for furthering our understanding of the function of CTDGs in the human genome, we used all 3,450 genes showing tissue-specific enhancement of their expression in 32 human tissues and organs (genes showing tissue-specific enhancement were those exhibiting "at least five-fold higher mRNA levels in a particular tissue as compared with average levels in all tissues") (Uhlen et al. 2015) as individual queries in CTDGFinder to examine whether they were clustered or not. On average, 25% of these enhanced genes were part of CTDGs. However, clustered genes were unequally distributed across tissues, with the percentage of clustered genes per tissue ranging from 56% (49/88) in the appendix, to 12% (74/597) in the cerebral cortex (fig. 4).

Examination of the enhanced genes that reside within CTDGs identified members of several well-known gene clusters. In general, gene families that are known to be parts of gene clusters like olfactory receptors (Niimura 2009), phospholipase A (Tischfield et al. 1996), golgins (Locke et al. 2003), histones (Albig et al. 1997), cytochrome P450s (Hoffman et al. 1995), aquaporins (Finn et al. 2014), myosin chains (Weiss et al. 1999), Hox genes (Krumlauf 1992), and protocadherins (Wu and Maniatis 1999), were enhanced in different tissues (supplementary table S4, Supplementary Material online). For example, nine of the 49 clustered genes in the appendix are chemokine ligands or receptors, six are leukocyte immunoglobulin-like receptors, and four are SIGLECs. Similarly, 20 of the 75 clustered genes in the bone marrow are annotated as part of the histone cluster 1, 32 of 72 clustered genes in the lymph node are involved in the immune system (e.g., chemokines, MHC, T-cell and B-cell associated), and six of 12 clustered genes in the endometrium belong to the Hox gene family (clusters A and D). Finally, in the placenta, clustered genes encode for proteins important in tissue structure and remodeling like metallopeptidases (2 of 41 clustered genes), collagen (3 genes), and gap junction proteins (3 genes), as well as immunity related genes (one interleukin receptor and one immunoglobulin receptor) (supplemental table S4, Supplementary Material online).

## Usage Recommendations

CTDGs are features of the spatial arrangement of genes in chromosomes. For that reason, genomes with high quality assemblies (e.g., genomes with assembled chromosomes) and gene annotations will give optimal results. However, genomes with fragmented assemblies can also be used, as long as they contain the coordinates of the genes present in their corresponding scaffolds along with their protein sequence information. Although CTDGFinder works for any genome, its application in genomes with highly fragmented assemblies might be problematic because of the lack of continuity of nonassembled chromosomes. This fragmentation might also influence the statistical sampling process by lowering the number of gene duplicates per sample (because the sampled regions will be shorter), leading to spurious genome-wide 95th percentile thresholds. Similarly, application of CTDGFinder in poorly annotated genomic regions (i.e., regions where many actual genes are unannotated), especially ones containing clusters of tandemly duplicated genes, will be problematic, due to inaccurate information about the spatial arrangement of genes.

To further explore our algorithm's behavior with assemblies of different quality, we examined CTDGFinder's performance in chromosome-level and scaffold-level assemblies and their corresponding gene annotations of the human genome (chromosome-level assembly: 32,010 genes in 260 linkage groups; 23 of these linkage groups are chromosome-level in size, the remaining 237 have fewer than 238 genes, and 173 linkage groups have fewer than 10 genes; scaffold-level assembly: 32,010 genes in 335 linkage groups; none of these linkage groups are chromosome-level in size and 202 have fewer than 10 genes each). We used the same queries as previously (see section "Identifying CTDGs across placental mammals") and employed CTDGFinder to identify CTDGs in Hox, protocadherin, globin, galectin, and Siglec gene families. In all but one case, the same number of TDGs was found for each species irrespective of the assembly level for each species. The only exception was in the protocadherin gene family, where PCDHGA2 was found as part of a CTDG using the chromosome-level assembly, but not when using a scaffold-level assembly because the gene was not present in this assembly's annotation. Furthermore, comparison of the mean 95th percentile values showed that they were slightly lower when using scaffold-level assemblies than when using chromosome-level assemblies, but such a difference was not statistically significant (table 4), suggesting that the process used in our statistical sampling is robust to differences in the quality of the assembly.

**Table 4.** Impact of Assembly Quality on the 95th Percentile Threshold.

| Species | Assembly Level | 95th Percentile | T-test P-value (of the 95th percentiles for all the clusters identified using each assembly level) |
|---|---|---|---|
| Human | Chromosomes | 2.47 | 0.700 |
| | Scaffolds | 2.34 | |

### *Data input*

CTDGFinder uses a series of annotation files and one precomputed all vs. all BLASTP results output to run. Specification and helper scripts, as well as instructions on how to use them can be accessed on the "extras" directory of the CTDGFinder repository (https://github.com/biofilos/ctdg_finder/tree/master/extras). CTDGFinder requires protein sequences as its input and cannot be used with genomic DNA sequences. If genomic DNA sequences were used (via TBLASTN), multiple BLAST hits could either be due to the presence of duplicated genes or to the presence of duplicated domains in one protein sequence. Furthermore, users should take into account that gene duplicates from different gene families can vary in their evolutionary dynamics. All the sequences (and their BLAST hits) from a query sequence set provided by the user are assumed to belong to a gene family, and all the hits from them will be treated as hits of that gene family. With this in mind, the query sequence set should include as many sequences belonging to the gene family under study as possible. In that way, as much sequence diversity as possible will be accounted for, including highly diverged gene sequences from a gene family, increasing the chances of a comprehensive search for CTDGs. However, CTDGFinder can also be run using only one sequence as query. In this case, CTDGFinder will perform well if the sequence divergence between the genes in the gene family under study is low, but not so well if the sequence divergence between gene family members is very high (highly divergent genes might be missed in the BLAST search).

To help the user in the delimitation of a query sequence set (especially in highly divergent gene families), CTDGFinder can be run in iterative mode (option –iterative). In this mode, CTDGFinder will perform the BLAST search and length overlap filter steps using the query sequence set provided by the user, and then all the complete sequences resulting from the first BLAST step will be extracted, and used as a new query. A good case in point is the Hox cluster. Using a query containing the genes from the HoxA cluster without EVX1 (a very distant homolog to the Hox genes) in iterative mode, CTDGFinder identified all the Hox CTDs including EVX1 and EVX2. Whereas powerful, it is important to note that, since the iterative mode uses all the proteins from a BLAST search as new queries, it might also result in spurious results because the sequence search space will be so big that sequences with poor sequence similarity can get included in CTDGs.

Finally, CTDGFinder is coded to use several CPU cores, if needed (this is useful when testing of various E-value thresholds and query sequence inputs is required). As a reference, the mean running wall time of CTDGFinder when using the queries reported in supplementary table S1, Supplementary Material online for human and mouse was 59.38 s in default mode, and 5.05 min in –iterative mode (this time estimate does not include calculation of pre-computed all vs. all BLAST results output).

### *Sensitivity of the BLAST E-Value Threshold*

CTDGFinder uses the E-value extracted from BLASTP (plus the length overlap filter encoded in CTDGFinder) to identify potential gene duplicates. Selecting an appropriate *E*-value threshold depends on the variability of the gene family in question. To explore the effects of varying the E-value threshold on CTDG inference, CTDGFinder was run using *E*-values ranging from $1e^{-10}$ to 1 for all the gene families under study in human and mouse. The number of CTDGs and total number of gene duplicates (TDGs) identified for the galectin, globins, Hox, and protocadherin gene families did not change in either human or mouse (number of CTDGs and TDGs did not change for the primate-specific LHB CTDG) (supplementary fig. S1, Supplementary Material online). The number of CTDGs and TDGs in the GH gene family did not change in human, but changed in mouse. This is because a query containing GH was able to identify the distantly related (rodent and bovid specific) prolactins at high (0.1, 1) *E*-values. Similarly, using a query containing prolactin sequences yielded variable results on both human and mouse. The number of prolactin CTDGs was observed to change in mouse because there is a small cluster of four genes that was seen at the edge of several percentile 95 thresholds. It might seem counter-intuitive that decreasing an *E*-value will increase the number of clusters. However, if the genes in question have enough sequence similarity, but the number of gene duplicates is close to the 95th percentile of the empirical distribution of gene duplicates, a low *E*-value might cause such empirical distribution to filter out more potential duplicates, thus resulting in a lower 95th percentile threshold. For this reason, it might be advisable to increase the number of sampled regions used to build the empirical distribution when using very low *E*-values. Finally, the Siglec gene family showed variations in both the number of CTDGs and TDGs. This is a consequence of the high sequence variation between the Siglec CTDGs and the genes in the Siglec family.

### *Identifying CTDGs across Metazoans*

In order to test the performance of CTDGFinder across very distantly taxa, we used it to identify known Hox CTDGs in representative metazoan species (the fruit fly *Drosophila melanogaster*, the tunicate *Ciona intestinalis*, the lancelet *Branchiostoma floridae*, the zebrafish *Danio rerio*, the lizard *Anolis carolinensis*, the chicken *Gallus gallus*, and the human *Homo sapiens*). Because of the phylogenetic range encompassed in the analysis, we used the iterative mode (–iterative) of CTDGFinder, and an *E*-value threshold of 0.001.

In *the fruit fly*, two CTDGs including the homeobox-containing genes *lab, pb, zen2, zen, bcd* (mean intergenic distance: 15 kb), and *Dfd, Scr, ftz, Antp* (mean intergenic distance: 26 kb) were identified by CTDGFinder. Interestingly, the genes *Ubx, abd-A*, and *Abd-B* (mean intergenic distance: 89 kb) were identified as a candidate cluster in the BLASTP step, but did not pass the statistical significance threshold in the MeanShift step. In the tunicate, four CTDGs, each comprised of two genes were identified. These included: a CTDG on chromosome 5 containing the homeobox-containing gene LOC778697 and an uncharacterized Hox-D3-like gene; a CTDG on chromosome 1 containing the Hox2 and Hox3 genes (intergenic distance: 7.4 kb; Hox4 was rejected from being a member of that CTDG, presumably because it is 19 kb away from Hox3); two CTDGs on chromosome 7, one containing the distal-less one and two genes and the other containing the Hox12 and Hox13 genes (the genes Hox1, Hox5/6, Hox6/7, and Hox10 were not present in the gene annotation we used because they reside in unplaced chromosomes, or unassembled scaffolds). The assembly of the lancelet used in this analysis was highly fragmented (composed by 398 genomic regions); however, CTDGFinder identified 12 genes in five CTDGs in the scaffold NW_003101559 that corresponds to the location of the Hox gene cluster. Two Hox CTDGs were found in the lizard corresponding to the Hox-C (10 genes) and Hox-B (8 genes) CTDGs. Genes belonging to the Hox-A and Hox-D CTDGs in the lizard were found to be in unplaced scaffolds. In the zebrafish, two Hox-A (5 genes in chromosome 16, and seven genes in chromosome 19) CTDGs, two Hox-B (4 genes in chromosome 12, and 12 genes in chromosome 3) CTDGs, two Hox-C (4 genes in chromosome 11, and four genes in chromosome 23) CTDGs, and a Hox-D CTDG (8 genes in chromosome 9) were found. In the chicken, the CTDGs Hox-A (11 genes in chromosome 2), Hox-B (9 genes in chromosome 27), Hox-C (3 genes in chromosome 33), and Hox-D (10 genes in chromosome 7) were found. Finally, all four Hox CTDGs were found in human as reported in table 1.

## Discussion

From the classic globin and Hox gene clusters (Efstratiadis 1980; Krumlauf 1994) to the more recently described venom-

related gene clusters in snakes (Ikeda et al. 2010; Vonk et al. 2013), the study of protein families whose genes are closely spaced on chromosomes has greatly enhanced our understanding of physiology, development, and the genetic basis of phenotypic diversity (Holland and Garcia-Fernàndez 1996; Hallem et al. 2006; Rawn and Cross 2008; Whittington et al., 2008). In the post-genomic era, however, the fact that such gene clusters are still defined arbitrarily and in different ways in each gene family or genome makes attempts for the kinds of comparative analyses required to understand the dynamics of gene cluster evolution and function problematic.

Our formal definition and identification of clusters of tandemly duplicated genes (CTDGs; implemented as CTDGFinder) through a statistical approach that takes into account both intergenic distance and homology solves this problem. This approach not only enables the comparison of the genomic arrangement of well-known clusters in different model organisms, but also the discovery of novel gene clusters or novel arrangements in genomes. For example, by using previously known sequences of the prolactin gene cluster in mouse, CTDGFinder was able to find novel prolactin CTDGs in goat and sheep (supplementary table S2, Supplementary Material online), providing further evidence supporting the independent evolution of prolactin clusters in ruminants and rodents (Miller and Eberhardt 1983). Thus, coupled with robust phylogenetic analysis at the gene and species levels, CTDGFinder could be very useful for distinguishing between different types of clusters (e.g., primary, secondary, and independently evolved) comprised of paralogous genes and the evolutionary history of their assemblies (Ferrier 2016).

Using a formal definition for CTDGs also has the potential to help identify interesting gene arrangement and orientation features in clusters and inform our understanding of the evolutionary steps that explain their current assembly. For example, according to CTDGFinder, the single Siglec gene cluster on the human chromosome 19 is actually comprised of two distinct CTDGs (tables 1 and 3). Interestingly, the two inferred CTDGs correspond precisely to the two sub-clusters (A and B) that Cao and co-workers previously identified and inferred to have been generated through a large-scale inverse duplication of the ancestral Siglec locus in a vertebrate ancestor (Cao et al. 2009). Such inversions can help stabilize the size of a gene cluster by reducing the effectiveness of recombination to add or remove additional gene duplicates to the existing cluster (Passananti et al. 1987; Cao et al. 2009).

Availability of a formal definition and means of characterizing CTDGs will also facilitate efforts to understand the functional implications of clustering. For example, CTDGFinder inferred two galectin CTDGs on the human chromosome 19 (table 1), rather than a single cluster (Than et al. 2014). This split into two CTDGs appears to be informative for function; genes on the one CTDG show enhanced gene expression in placenta and bone marrow, whereas genes on the other CTDG are enhanced in skin and digestive track (supplementary table S6, Supplementary Material online). At a broader level, CTDGFinder can be used to understand whether tissue-specific expression of clustered genes is evenly distributed across human tissues and organs (fig. 4), which in turn could be associated with the types and functions of genes expected to function in them (e.g., many secreted proteins are known to be expressed in tissues such as the liver and salivary glands) (Uhlen et al. 2015). Additionally, given that CTDGFinder can simultaneously analyze the genomes of multiple species, a formal definition enables the investigation of the types of functional categories of genes (e.g., immunity, metabolism) that tend to be clustered in the genomes of organisms from diverse lineages and which potentially could be implicated in the generation of interesting lineage-specific phenotypes.

Finally, a formal definition for CTDGs has the potential to greatly aid in understanding the relationship between the specific genomic organization of a given gene cluster with its mechanism of regulation. For example, using a diverse and rich body of data on animal Hox gene clusters, Duboule (2007) has argued that Hox gene cluster organization in different animals (fig. 1) has strong implications for how the activity of Hox genes is regulated in these organisms. The first step toward answering this question, not just for the animal Hox gene clusters, but for the wide diversity of gene families forming gene clusters, is the availability of a clear, precise, unambiguous, and easy to implement definition, such as the one provided by this study.

## Methods

### Gene Annotation

Genome annotation for all the mammalian species with assembled chromosomes available from the NCBI was downloaded fromftp.ncbi.nlm.nih.gov/genomes/ (supplementary table S5, Supplementary Material online). Genes were annotated using CDS (coding sequence) features from their annotation GenBank files. If two or more CDS features had the same start or end coordinate, the longest gene was selected. If two genes in the same strand had overlapping coordinates but different start coordinates, the gene with the start coordinates downstream from the other gene was removed.

### CTDGFinder Validation on Previously Described (Reference) Gene Clusters

Representative protein sequences from the prolactin, growth hormone, Hox, galectin, luteinizing hormone beta (LHB), beta-globin, Siglecs, galectin, and protocadherin gene clusters were downloaded from the NCBI. Gene family cluster analysis was performed for each downloaded set of sequences using the CTDGFinder algorithm described below.

### Phylogenetic Analysis of Orthologous CTDGs

In order to identify orthologous CTDGs, phylogenetic trees were built using the protein sequence from all the CTDGs in a gene family. The best substitution model was selected using ProtTest (Darriba et al. 2011). The phylogenetic tree was estimated using the rapid bootstrap function implemented in RAxML, version 8.2.4 (Stamatakis 2006), and 100 bootstrap pseudoreplicates were performed. The tree was rooted using evolutionary information about each family, and the leaves containing genes from known CTDGs in human and mouse were used to identify the clade with CTDG orthologs

(selected clade). Given that CTDGs in human and mouse could have experienced gene deletions or accelerated rates of evolution, sister branches of the selected clade were also inspected for the presence of members of the known CTDGs. The set of CTDG orthologs was defined as all the CTDGs with all their genes included in the selected clade.

### CTDGFinder Programing Environment

CTDGFinder was coded in Python 3.4.3 using the Pandas library v0.16.2 for table manipulation, Numpy v1.9.2 (Oliphant 2007) for numerical arrays manipulation, BioPython v1.65 (Cock et al. 2009) for sequence manipulation and communication with the NCBI servers, scikit-learn v0.16.1 (Pedregosa et al. 2011) for statistical analysis, and matplotlib v1.4.3 (Hunter 2007) and Bokeh v.0.9.2 for graphics generation. CTDGFinder is freely available from https://github.com/biofilos/ctdg_finder.

### Assessing the Degree of Clustered of Genes with Human Tissue-Enhanced Expression

The "RNA gene data" expression dataset was downloaded from the Protein Atlas (tissue dataset) (Uhlen et al. 2015), and genes that were tissue-enhanced were selected. Protein sequences from these selected genes were downloaded using the BioMart API from Ensembl. In cases where several protein sequences were mapped to the same gene identifier, only the longest protein sequence was used. CTDGFinder was run for each of the selected genes. To evaluate whether a given gene was a member of a CTDG, its coordinates were contrasted with those of all the clustered genes found by CTDGFinder. If the coordinates of a selected gene overlapped with those of a clustered gene, the gene was annotated as clustered.

### Gene Family Cluster Analysis (CTDGFinder Algorithm)

Each query protein or protein set was used as a query for BLASTP (default *E*-value: 0.001 was used; it can be changed by the user) against all the proteomes of the species under study. BLAST hits that were either more than three times the length of the query or less than one third the length of the query were removed (length ratio less than 0.3). The set of BLASTP hits in each chromosome (pre-clusters) was processed using the meanshift algorithm as implemented in scikit-learn (Pedregosa et al., 2011). Since the meanshift algorithm is implemented to identify clusters in more than one dimension, the start coordinates of the pre-clusters were considered to be the *x* coordinates and the *y* coordinates were set to zero. The mean plus the standard deviation of all intergenic distances in the chromosome where a given pre-cluster resides were used as the bandwidth parameter for the meanshift implementation. In the meanshift step, the pre-clusters can be ignored, subdivided, or confirmed as proto-clusters. Its output consists of a set of proto-clusters per chromosome.

In order to extract clusters that have more gene duplicates than expected in the genome, 1,000 random regions of the length of each proto-cluster were taken from the genome for the following analysis: an all vs. all BLASTP search of the complete set of proteomes under study was used to extract the gene duplicates of the genes in the region, and the maximum number of hits was considered the maximum number of gene duplicates. Proto-clusters containing more gene duplicates than the 95th percentile from the genome-wide sample set were considered clusters of tandemly duplicated genes (CTDGs).

## Supplementary Material

Supplementary figure S1 and tables S1–S6 are available at *Molecular Biology and Evolution* online.

## Acknowledgments

We thank members of the Rokas lab for critical feedback on the work described in this study. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. JFO was supported by the Graduate Program in Biological Sciences at Vanderbilt University. This work was supported in part by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative and by the National Science Foundation (DEB-1442113 to A.R.).

## References

Alam SMK, Ain R, Konno T, Ho-Chen JK, Soares MJ. 2006. The rat prolactin gene family locus: species-specific gene family expansion. *Mamm Genome*. 17(8):858–877.

Albig W, Kioschis P, Poustka A, Meergans K, Doenecke D. 1997. Human histone gene organization: nonregular arrangement within a large cluster. *Genomics* 40(2):314–322.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215(3):403–410.

Andersson DI, Hughes D. 2009. Gene amplification and adaptive evolution in bacteria. *Annu Rev Genet*. 43:167–195.

Angata T, Margulies EH, Green ED, Varki A. 2004. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc Natl Acad Sci U S A*. 101(36):13251–13256.

Bulger M, van Doorninck JH, Saitoh N, Telling A, Farrell C, Bender MA, Felsenfeld G, Axel R, Groudine M. 1999. Conservation of sequence and structure flanking the mouse and human beta-globin loci: the beta-globin genes are embedded within an array of odorant receptor genes. *Proc Natl Acad Sci U S A*. 96(9):5129–5134.

Cameron RA, Rowen L, Nesbitt R, Bloom S, Rast JP, Berney K, Arenas-Mena C, Martinez P, Lucas S, Richardson PM., et al. 2006. Unusual gene order and organization of the sea urchin hox cluster. *J Exp Zool B Mol Dev Evol*. 306B(1):45–58.

Cao H, De Bono B, Belov K, Wong ES, Trowsdale J, Barrow AD. 2009. Comparative genomics indicates the mammalian CD33rSiglec locus evolved by an ancient large-scale inverse duplication and suggests all Siglecs share a common ancestral region. *Immunogenetics* 61(5):401–417.

Carretero-Paulet L, Fares MA. 2012. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Mol Biol Evol*. 29(11):3541–3551.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, De Hoon MJL, et al. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25(11):1422–1423.

Comaniciu D, Meer P, Member S. 2002. Mean shift: a robust approach toward feature space analysis. Guang Pu Xue Yu Guang Pu Fen Xi 24(5):603–619.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* 27(8):1164–1165.
Duboule D. 2007. The rise and fall of Hox gene clusters. *Development (Cambridge, England)* 134(14):2549–2560.
Efstratiadis A. 1980. The structure and evolution of the human $\beta$-globin gene family. *Cell* 21(3):653–668.
Ferrier DEK, Holland PWH. 2001. Ancient origin of the Hox gene cluster. *Nat Rev Genet*. 2(1):33–38.
Ferrier DEK. 2016. Evolution of homeobox gene clusters in animals: the giga-cluster and primary vs. secondary clustering. *Front Ecol Evol*. 4(April):1–13.
Finn RN, Chauvigné F, Hlidberg JB, Cutler CP, Cerdà J. 2014. The lineage-specific evolution of aquaporin gene clusters facilitated tetrapod terrestrial adaptation. *PLoS ONE* 9(11):e113686.
Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res*. 18(12):1924–1937.
Garcia-Fernàndez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet*. 6(12):881–892.
Glusman G, Sosinsky a, Ben-Asher E, Avidan N, Sonkin D, Bahar A, Rosenthal A, Clifton S, Roe B, Ferraz C., et al. 2000. Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. *Genomics* 63(2):227–245.
Graham GJ. 1995. Tandem genes and clustered genes. *J Theor Biol*. 175(1):71–87.
Grant D, Cregan P, Shoemaker RC. 2000. Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proc Natl Acad Sci U S A*. 97(8):4168–4173.
Hallast P, Saarela J, Palotie A, Laan M. 2008. High divergence in primate-specific duplicated regions: human and chimpanzee chorionic gonadotropin beta genes. *BMC Evol Biol*. 8(1):195.
Hallem EA, Dahanukar A, Carlson JR. 2006. Insect odor and taste receptors. *Annu Rev Entomol*. 51(1):113–135.
Hardies SC, Edgell MH, Hutchison CA. 1984. Evolution of the mammalian beta-globin gene cluster. *J Biol Chem*. 259(6):3748–3756.
Hoffman SM, Fernandez-Salguero P, Gonzalez FJ, Mohrenweiser HW. 1995. Organization and evolution of the cytochrome P450 CYP2A-2B-2F subfamily gene cluster on human chromosome 19. *J Mol Evol*. 41(6):894–900.
Holland PWH, Garcia-Fernàndez J. 1996. Hox genes and chordate evolution. *Dev Biol*. 173(2):382–395.
Houzelstein D, Gonçalves IR, Fadden AJ, Sidhu SS, Cooper DNW, Drickamer K, Leffler H, Poirier F. 2004. Phylogenetic analysis of the vertebrate galectin family. *Mol Biol Evol*. 21(7):1177–1187.
Hunter JD. 2007. Matplotlib: A 2D graphic environment. *Comput Science Engg*. 9(3):90–95.
Ikeda N, Chijiwa T, Matsubara K, Oda-Ueda N, Hattori S, Matsuda Y, Ohno M. 2010. Unique structural characteristics and evolution of a cluster of venom phospholipase A2 isozyme genes of Protobothrops flavoviridis snake. *Gene* 461(1–2):15–25.
Kawasaki K, Weiss KM. 2003. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proc Natl Acad Sci U S A*. 100(7):4060–4065.
Kelley J, Walter L, Trowsdale J. 2005. Comparative genomics of natural killer cell receptor gene clusters. *PLoS Genet*. 1(2):129–139.
Kohmura N, Senzaki K, Hamada S, Kai N, Yasuda R, Watanabe M, Ishii H, Yasuda M, Mishina M, Yagi T. 1998. Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. *Neuron* 20(6):1137–1151.
Krause CD, Pestka S. 2015. Cut, copy, move, delete: the study of human interferon genes reveal multiple mechanisms underlying their evolution in amniotes. *Cytokine* 76(2):480–495.
Krumlauf R. 1992. Evolution of the vertebrte hox homeobox genes. *BioEssays* 14(2):245–252.
Krumlauf R. 1994. Hox genes in vertebrate development. *Cell* 78(2):191–201.
Lemons D, McGinnis W. 2006. Genomic evolution of Hox gene clusters. *Science (New York, N.Y.)* 313(5795):1918–1922.
Levings PP, Bungert J. 2002. The human $\beta$-globin locus control region. *Eur J Biochem*. 269(6):1589–1599.
Locke DP, Archidiacono N, Misceo D, Cardone MF, Deschamps S, Roe B, Rocchi M, Eichler EE. 2003. Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol*. 4(8):R50.
Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 3(1–4):35–44.
MacLean J. a, Lorenzetti D, Hu Z, Salerno WJ, Miller J, Wilkinson MF. 2006. Rhox homeobox gene cluster: recent duplication of three family members. *Genesis (New York, N.Y. : 2000)* 44(3):122–129.
Martin AM, Freitas EM, Witt CS, Christiansen FT. 2000. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics* 51(4–5):268–280.
Miller WL, Eberhardt NL. 1983. Structure and evolution of the growth hormone gene family. *Endocr Rev*. 4(2):97–130.
Niimura Y. 2009. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. *Hum Genomics* 4(2):107–118.
Niimura Y, Matsui A, Touhara K. 2014. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res*. 24(9):1485–1496.
Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol Evol*. 1(2006):34–44.
Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res*. 14(3):354–366.
Oliphant TE. 2007. Python for scientific computing. *Comput Science Engg*. 9(3):10–20.
Passananti C, Davies B, Ford M, Fried M. 1987. Structure of an inverted duplication formed as a first step in a gene amplification event: implications for a model of gene amplification. *EMBO J*. 6(6):1697–1703.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O 2011. Scikit-learn: Machine Learning in Python. *J Machine Learning Res*. 12:2825–2830.
Pendleton JW, Nagai BK, Murtha MT, Ruddle FH. 1993. Expansion of the Hox gene family and the evolution of chordates. *Proc Natl Acad Sci U S A*. 90(13):6300–6304.
Rawn SM, Cross JC. 2008. The evolution, regulation, and function of placenta-specific genes. *Annu Rev Cell Dev Biol*. 24:159–181.
Reams AB, Roth JR. 2015. Mechanisms of gene duplication and amplification. *Cold Spring Harb Perspect Biol*. 7(2):a016592.
Simmons DG, Rawn S, Davies A, Hughes M, Cross JC. 2008. Spatial and temporal expression of the 23 murine Prolactin/Placental Lactogen-related genes is not associated with their position in the locus. *BMC Genomics* 9:352.
Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
Su Y, et al. 2000. The human growth hormone gene cluster locus control region supports position-independent pituitary- and placenta-specific expression in the transgenic mouse. *J Biol Chem*. 275(11):7902–7909.
Than NG, Romero R, Goodman M, Weckle A, Xing J, Dong Z, Xu Y, Tarquini F, Szilagyi A, Gal P., et al. 2009. A primate subfamily of galectins expressed at the maternal-fetal interface that promote immune cell death. *Proc Natl Acad Sci U S A*. 106(24):9731–9736.
Than NG, Romero R, Xu Y, Erez O, Xu Z, Bhatti G, Leavitt R, Chung TH, El-Azzamy H, LaJeunesse C., et al. 2014. Evolutionary origins of the placental expression of chromosome 19 cluster galectins and their complex dysregulation in preeclampsia. *Placenta* 35(11):855–865.
Tischfield JA, Xia YR, Shih DM, Klisak I, Chen J, Engle SJ, Siakotos AN, Winstead MV, Seilhamer JJ, Allamand V., et al. 1996. Low-molecular-weight, calcium-dependent phospholipase A2 genes are linked and map to homologous chromosome regions in mouse and human. *Genomics* 32(3):328–333.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A., et al. 2015. Tissue-based map of the human proteome. *Science* 347(6220):1260419.

Vonk FJ, Casewell NR, Henkel CV, Heimberg AM, Jansen HJ, McCleary RJR, Kerkkamp HM, Vos RA, Guerreiro I, Calvete JJ., et al. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc Natl Acad Sci U S A*. 110(51):20651–20656.

Wallis M. 1991. The expanding growth hormone/prolactin family. *J Mol Endocrinol*. 9:185–188.

Weaver S, Mary B, Hutchison CA, Hill C, Carolina N, Jahn CL. 1981. The adult P-globin genes of the Single type mouse C57BL. *Cell* 24(May):403–411.

Weiss A, McDonough D, Wertman B, Acakpo-Satchivi L, Montgomery K, Kucherlapati R, Leinwand L, Krauter K. 1999. Organization of human and mouse skeletal myosin heavy chain gene clusters is highly conserved. *Proc Natl Acad Sci U S A*. 96(6):2958–2963.

Whittington CM, Papenfuss AT, Bansal P, Torres AM, Wong ESW, Deakin JE, Graves T, Alsop A, Schatzkamer K, Kremitzki C, et al. 2008. Defensins and the convergent evolution of platypus and reptile venom genes. *Genome Res*. 18(6):986–994.

Wu Q, Maniatis T. 1999. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* 97(6):779–790.

Wu Q, Zhang T, Cheng JF, Kim Y, Grimwood J, Schmutz J, Dickson M, Noonan JP, Zhang MQ, Myers RM, Myers RM. 2001. Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. *Genome Res*. 11(3):389–404.

Yagi T. 2008. Clustered protocadherin family. *Development Growth and Differentiation* 50(SUPPL. 1):S131–S140.

Zeltser L, Desplan C, Heintz N. 1996. Hoxb-13: a new Hox gene in a distant region of the HOXB cluster maintains colinearity. *Development* 122(8):2475–2484.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol*. 18(6):292–298.