Comparing the Statistical Fate of Paralogous and Orthologous Sequences

Florian Massip,*,†,‡,1 Michael Sheinman,§ Sophie Schbath,* and Peter F. Arndt†

*MaIAGE, Institut National de la Recherche Agronomique, Université Paris-Saclay, 78350 Jouy-en-Josas, France, †Max Planck Institute for Molecular Genetics, Berlin, Germany, ‡Laboratoire Biométrie et Biologie Évolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université de Lyon, Université Lyon 1, Villeurbanne, France, and §Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands

ORCID ID: 0000-0003-1762-9836 (P.F.A.)

ABSTRACT For several decades, sequence alignment has been a widely used tool in bioinformatics. For instance, finding homologous sequences with a known function in large databases is used to get insight into the function of nonannotated genomic regions. Very efficient tools like BLAST have been developed to identify and rank possible homologous sequences. To estimate the significance of the homology, the ranking of alignment scores takes a background model for random sequences into account. Using this model we can estimate the probability to find two exactly matching subsequences by chance in two unrelated sequences. For two homologous sequences, the corresponding probability is much higher, which allows us to identify them. Here we focus on the distribution of lengths of exact sequence matches between protein-coding regions of pairs of evolutionarily distant genomes. We show that this distribution exhibits a power-law tail with an exponent $\alpha = -5$. Developing a simple model of sequence evolution by substitutions and segmental duplications, we show analytically and computationally that paralogous and orthologous gene pairs contribute differently to this distribution. Our model explains the differences observed in the comparison of coding and noncoding parts of genomes, thus providing a better understanding of statistical properties of genomic sequences and their evolution.

KEYWORDS comparative genomics; statistical genomics; DNA duplications; genome evolution

ONE of the first and most celebrated bioinformatic tools is sequence alignment (Needleman and Wunsch 1970; Smith and Waterman 1981; Altschul *et al.* 1990), and algorithms to search for similarity between sequences in a huge database are still actively studied.

For this matter, we need to be able to distinguish sequence alignments that are due to a biological relatedness of two sequences from those that occur randomly. Let us, for simplicity, disregard mismatching nucleotides and insertions and deletions (indels or gaps) in an alignment and consider only so-called maximal exact matches, *i.e.*, sequences that are 100% identical and cannot be extended on both ends. In this case, the length distribution of matches is equivalent to the score distribution and can easily be calculated for an alignment of two random sequences where each nucleotide represents an i.i.d. random variable. We expect the number of matches to be distributed according to a geometric distribution, such that the number, $M(r)$, of exact maximal matches of length r is given by

$$M(r) = p^r(1-p)^2 L_A L_B, \tag{1}$$

where L_A and L_B are the lengths of the two genomes, p^r is the probability that r nucleotides match, and $(1-p)^2$ is the probability that a match is flanked by two mismatches. Here $p = \sum_\alpha f_A(\alpha) f_B(\alpha)$, where $f_X(\alpha)$ is the frequency of nucleotide α in the genome of species X and the sum is taken over all nucleotides. Thus, the number of matches for a given length r is expected to decrease very fast as the length r increases, and for generic random genomes of hundreds of megabase pairs, we do not expect any match >25 bp.

For long matches, real genomes strongly violate the prediction of Equation 1 due to the evolutionary relationships between and within genomes (Salerno *et al.* 2006). Comparing

doi: 10.1534/genetics.116.193912
Manuscript received June 1, 2016; accepted for publication July 26, 2016; published Early Online July 28, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.193912/-/DC1.
[1]Corresponding author: Laboratoire Biométrie et Biologie Évolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université de Lyon, Université Lyon 1, 69622 Villeurbanne, France. E-mail: florian.massip@gmail.com

the genomes of recently diverged species, we find regions in the two genomes that have not acquired even a single substitution. In the following, substitution refers to any genomic change that would disrupt a 100% identical match (for instance, a nucleotide exchange, an insertion, or a deletion). As the divergence time between the two species increases, such matches will get smaller very fast and most remaining long matches will be found either in exons or in ultraconserved elements (Bejerano *et al.* 2004) that both evolve under purifying selection.

Computing the match length distribution (MLD) from the comparison of human and mouse genomes, we thus expect to observe much longer exact matches than in a comparison of two random sequences of the same lengths. The observed MLD for exons and noncoding sequences in the human and mouse RepeatMasked genomes is shown in Figure 1. At the left end of the distribution, *i.e.*, for $r < 25$ bp, the distribution is dominated by random matches, as described by Equation 1. As expected, this MLD deviates from the random model for matches >25 bp.

Interestingly, in this asymptotic regime, the MLD exhibits a power-law tail $M(r) \sim r^{\alpha}$. In the comparison of exonic sequences, the exponent α is close to –5, in contrast to the MLD of noncoding sequences, where the exponent α is close to –4 (Gao and Miller 2011, 2014; Massip *et al.* 2015). This property appears to be impressively reproducible in the comparison of various pairs of species (see Supplemental Material, File S1, Figure S1). In all cases, the value of α was calculated using the maximum-likelihood estimator. To assess the robustness of this estimator, we also performed a bootstrap analysis that showed good agreement with the estimated value of α; see *Materials and Methods*. Note also that discrepancies of the power-law behavior can be observed for very long matches, since such matches are scarce and random noise distorts the distributions.

It is tempting to speculate that this peculiar behavior of exonic sequences is a direct consequence of their coding function. However, we demonstrate in the following that this distribution can be accounted for by a simple evolutionary model that takes into account the generation of paralogous sequences (due to segmental duplications) and orthologous sequences (due to speciation) (Fitch 2000). Further, we assume that paralogous and orthologous exact matches are subsequently broken down by random substitutions. These dynamics can be modeled by a well-known stick-breaking process (Kuhn 1930; Ziff and McGrady 1985) introduced below. Since our model describes the existence of long matching sequence segments in two genomes, it also has to include selection. However, we model selection in a minimal way, since we assume only that regional substitution rates are distributed, such that there are regions that evolve very slowly. Our model can therefore be viewed as a minimal model for evolution of functional sequences, which reproduces certain statistical features of their score distributions. In the next section, we describe the main methods and the data analyzed in this article.

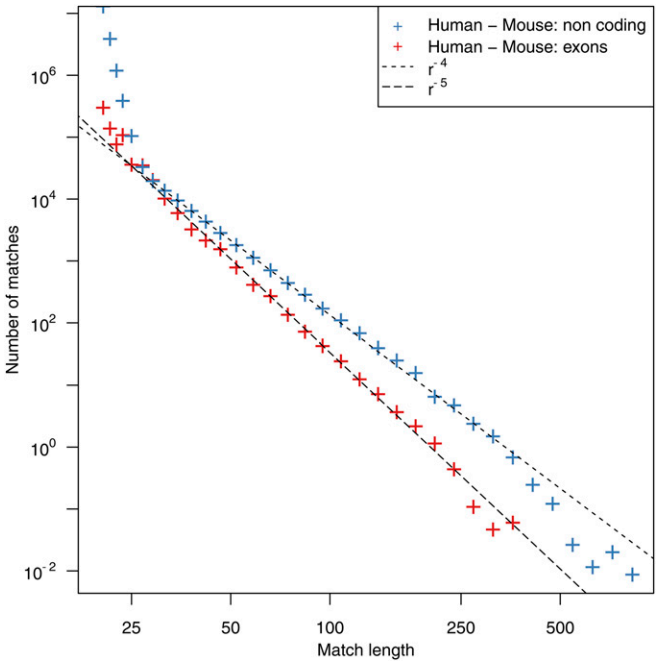

Figure 1 Two MLDs computed from the comparison of different subsets of the human and the mouse genomes. The first MLD was computed from the comparison of the RepeatMasked noncoding part of both genomes (blue crosses) and the second is the result of the comparison of the coding part of these genomes (red crosses). Dashed lines represent power-law distributions with exponents $\alpha = -4$ and $\alpha = -5$. All data are represented using a logarithmic binning to reduce the sampling noise (Newman 2005); see *Materials and Methods* for more details.

Materials and Methods

Computational and statistical analysis

Computing MLDs: To find all identical matches (both in the case of self and comparative alignments), we used the mummer pipeline (Kurtz *et al.* 2004) (version 3.0), which allows us to find all maximal exact matches between a "query" and a reference sequence using a computationally efficient suffix tree approach. For our analyses, we used the following options: -maxmatch such that mummer searches for all matches regardless of their uniqueness; -n that states that only 'A', 'T', 'C', and 'G' can match; -b such that mummer searches for matches on both strands; and -l 20 to filter out matches <20 bp.

The number of matches expected for a random i.i.d. sequence grows quadratically with L. For instance, we expect 3.5×10^{16} matches of length 2 in a comparison of two sequences of length $L = 10^9$ bp (see Equation 1). For this reason, we have to define a threshold for the length of matches that mummer should output especially when comparing entire eukaryotic genomes.

Logarithmic binning: Power laws appear in the tail of distributions, meaning that they are associated to rare events, which are thus subject to strong fluctuations. The high impact of noise in the tail of the distribution can make the assessment of the exponent of the distribution difficult. A way to resolve

this issue is to increase the size of the bins with the value of the horizontal axes and normalize the data accordingly. Namely, the observed values for each bin are divided by the size of the bin. The most common choice to do this is known as the logarithmic binning procedure, which consists of increasing the size of the bin by a constant factor. Note that by doing so, we dramatically reduce the number of data points and some information is lost as we aggregate different data points together in the same bin. Therefore it is often useful to consider both representations, with and without the logarithmic binning. See Newman (2005) for more details on the logarithmic binning procedure and on power-law distributions.

Estimating the value of the power-law exponent: To estimate the value of the exponent of the power-law α, we compute the maximum-likelihood estimator. The estimator $\hat{\alpha}$ is simply the value of α that maximizes the log-likelihood $\mathcal{L}$,

$$\mathcal{L} = \sum_{i=1}^{n} \left[\ln(\alpha - 1) - \ln(x_{\min}) \alpha \ln\left(\frac{x_i}{x_{\min}}\right) \right], \qquad (2)$$

such that

$$\hat{\alpha} = -1 - n \left[\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min}}\right) \right]^{-1}, \qquad (3)$$

while the value of $x_{\min}$ has to be determined visually. This estimator is also sometimes referred to as the Hill estimator (Hill *et al.* 1975).

To estimate the robustness of the value of the exponents found using this method, we proceeded to bootstrap experiments on the human to mouse exome comparison. For each bootstrap, we sampled 5% of the mouse exons and compared them to all human exons. In each experiment, we calculated the exponent of the MLD, using the maximum-likelihood estimator as described in Newman (2005). We repeated this procedure 100 times. Values of a were all in the range [–4.7, –5.2] and the mean value for the exponent was $\alpha = -4.9$.

Data availability

All genomes and their specific annotations (such as repeat elements and exons) were downloaded from the ensembl website (Cunningham *et al.* 2015), using the Perl API (version 80); the corresponding release of the human genome is GRCh38. In all cases, we downloaded the RepeatMasked versions of genomes publicly available in the ensembl databases.

Perl, R, and C++ scripts used to simulate the data and compute the match lentgh distributions are available at https://github.com/Flomass/MaLenDi. The MUMmer pipeline is freely available online (http://mummer.sourceforge.net/).

Results

Theory

The stick-breaking model: Before we turn to the detailed description of our model, let us shortly introduce some relevant

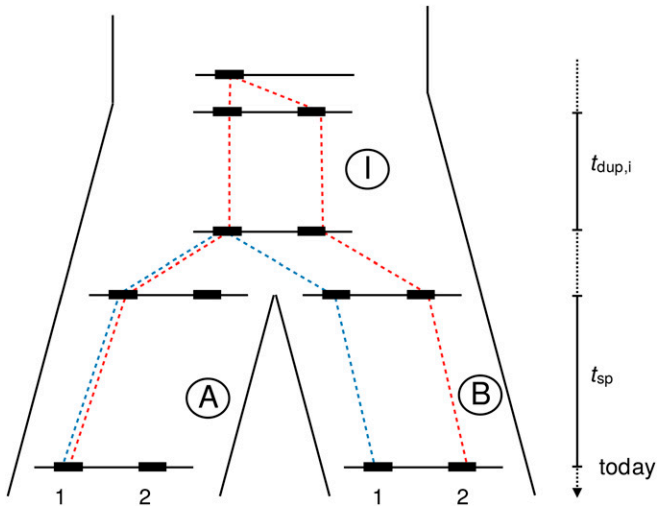

Figure 2 The different contributions to the match length distribution. Sequence 1 was duplicated in the ancestral species *I*. This duplication gives rise to two paralogous sequence pairs: sequence 1 in *A* with sequence 2 in *B* (red dashed line) and sequence *B* 1 and sequence *A* 2. Sequence 1 in *A* is orthologous to sequence 1 in *B* (blue dashed line), and sequence 2 in *A* is orthologous to sequence 2 in *B*. For clarity, we highlight only one pair for each case.

results on random stick breaking. Consider a stick of length K at time $t = 0$, which will be sequentially broken at random positions into a collection of smaller sticks. Breaks occur with rate μ per unit length. The distribution of stick lengths at time t, denoted by $m(r,t)$, follows the integro-differential equation

$$\frac{\partial}{\partial t} m(r,t) = -\mu\, r\, m(r,t) + 2\mu \int_r^{\infty} m(s,t) ds \qquad (4)$$

(Ziff and McGrady 1985; Massip and Arndt 2013), where the first term on the right-hand side represents the loss of sticks of length r due to any break in the given stick and the second term represents the gain of sticks of length r from the disruption of longer sticks. Note that for any stick of length $s > r$, there are two possible positions at which a break would generate a stick of length r.

The initial state is one unbroken stick of length K; *i.e.*, $m(r,0) = \delta(K,r)$. The corresponding time-dependent solution is

$$m(r,\tau) = \begin{cases} \left[2\tau + \tau^2(K-r)\right] \exp(-\tau r) & \text{for } 0 < r < K, \\ \exp(-\tau r) & \text{for } r = K, \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

(Ziff and McGrady 1985), where we define the rescaled time $\tau = \mu t$. Apart from the singularity at $r = K$, which accounts for the possibility that the stick is not even broken once, the distribution is dominated by an exponential function; *i.e.*, there are far more small sticks than long ones. The average stick length is given by $\bar{m}(\tau) = K/\tau$.

The match length distribution of evolving sequences: The above stick-breaking process can be used to describe the breakdown of a long DNA match into several smaller ones

by substitutions in either one of the two copies of the match. In a comparison of two species, A and B, long identical segments are the signature of homology relationships between the two sequences. These homologous sequences either result from the copy of the genetic material during the time of speciation and are then orthologous sequences (see the blue dashed line in Figure 2) or are due to segmental duplications in the ancestral genome, *i.e.*, paralogous sequences (see the red dashed line in Figure 2) (Fitch 2000).

The MLD is then given by the integral

$$M(r) = \int_0^\infty N(\tau) m(r,\tau) d\tau, \tag{6}$$

where $N(\tau)$ is the number of homologous sequences with divergence τ and $m(r,\tau)$ is given in Equation 5; see also Massip *et al.* (2015). The divergence between a pair of orthologous sequences is the sum of two contributions $\tau = \mu_{A,i}\, t_{\mathrm{sp}} + \mu_{B,i}\, t_{\mathrm{sp}}$, where t_{sp} is the time since the two species diverged and i is an index for regions in the genomes. The regional mutation rates $\mu_{A,i}$ and $\mu_{B,i}$ in the two species are themselves distributed and assumed to be independent from each other. We therefore define N_{AB} as

$$N_{AB}(\tau) = \int_0^\tau N_A(\tau_A) N_B(\tau - \tau_A) d\tau_A, \tag{7}$$

where $N_A(\tau)$ [resp. $N_B(\tau)$] is the number of sequences with divergence τ from the last common ancestor I in species A (resp. B). However, if the two regions are paralogous, the divergence τ is a sum of three independent contributions $\tau = \mu_{A,i} t_{\mathrm{sp}} + (\mu_{I,i} + \mu_{I,j}) t_{\mathrm{dup}} + \mu_{B,j} t_{\mathrm{sp}}$, where t_{dup} represents the time elapsed between the segmental duplication and the split of the two species. There are $N_{AIB}(\tau)$ paralogous sequences with divergence τ, with

$$N_{AIB}(\tau) = \int_0^\tau \int_0^{\tau-\tau_A} N_A(\tau_A) N_I(\tau - \tau_A - \tau_B) N_B(\tau_B) d\tau_B d\tau_A. \tag{8}$$

For our purposes we are not interested in the full functional form of the distributions in Equations 7 and 8 but have to consider only their behavior for small $\tau \to 0$, because long matches [and thus the tail of the distribution of the match length distribution $M(r)$] stem from homologous exons that exhibit a small divergence τ. A more general discussion about the functional form of the distribution of pairwise distances can be found in Sheinman *et al.* (2015). We therefore take the Taylor expansion of the distributions $N(\tau)$ around $\tau = 0$. Using Leibniz's formula to take the derivative under the integral sign (Flanders 1973), we find for orthologous exons

$$N_{AB}(\tau) = N_A(0) N_B(0) \tau + \mathcal{O}(\tau^2) \tag{9}$$

(see details in File S1) and subsequently the match length distribution

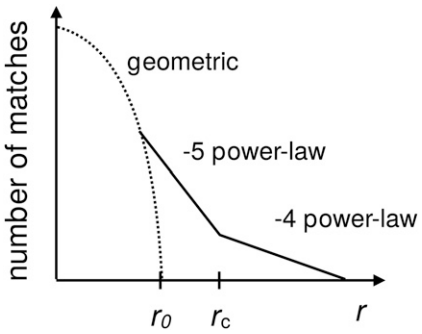

Figure 3 Schematic drawing of the match length distribution in a double logarithmic plot. The two regimes exhibiting a -4 and -5 power law (solid lines) are separated by a crossover point. For very small match lengths the geometric distribution due to random matches, see Equation 1, dominates (dotted line).

$$\begin{aligned} M_{AB}(r) &= \int_0^\infty N_{AB}(\tau) m(r,\tau) d\tau \\ &= N_A(0) N_B(0) \frac{6K - 2r}{r^4} \\ &\sim N_A(0) N_B(0) \frac{6K}{r^4} \end{aligned} \tag{10}$$

(Massip *et al.* 2015), as $K \gg r$. In contrast, expanding Equation 8 around $\tau = 0$ finds

$$N_{AIB}(\tau) = \frac{1}{2} N_A(0) N_I(0) N_B(0) \tau^2 + \mathcal{O}(\tau^3) \tag{11}$$

(see details in File S1). Thus, for paralogous pairs, the number of regions with divergence τ increases as τ^2 in the small τ limit. Therefore the match length distribution exhibits a power-law tail with exponent $\alpha = -5$,

$$\begin{aligned} M_{AIB}(r) &= \int_0^\infty N_{AIB}(\boldsymbol{\tau}) m(r,\tau) d\tau \\ &= N_A(0) N_I(0) N_B(0) \frac{12K - 6r}{r^5} \\ &\sim N_A(0) N_I(0) N_B(0) \frac{12K}{r^5}, \end{aligned} \tag{12}$$

as $K \gg r$.

Depending on the number of orthologous sequences Q_{ortholog} and paralogous sequences Q_{paralog}, we will be able to distinguish two regimes: one where the MLD follows an $\alpha = -4$ power law and one where it follows an $\alpha = -5$ power law. From Equations 10 and 12, it is straightforward to find that the crossover point r_c between those regimes (see Figure 3) is at

$$r_c = 2N_I(0). \tag{13}$$

Recall that $N_I(0)$ is defined as the number of paralogous segments that have not mutated even a single time since the duplication event at the time of the split. Thus, this term is proportional to the ratio of the duplication rate over the mutation rate. If $N_I(0) \gg 10$, there are significantly more

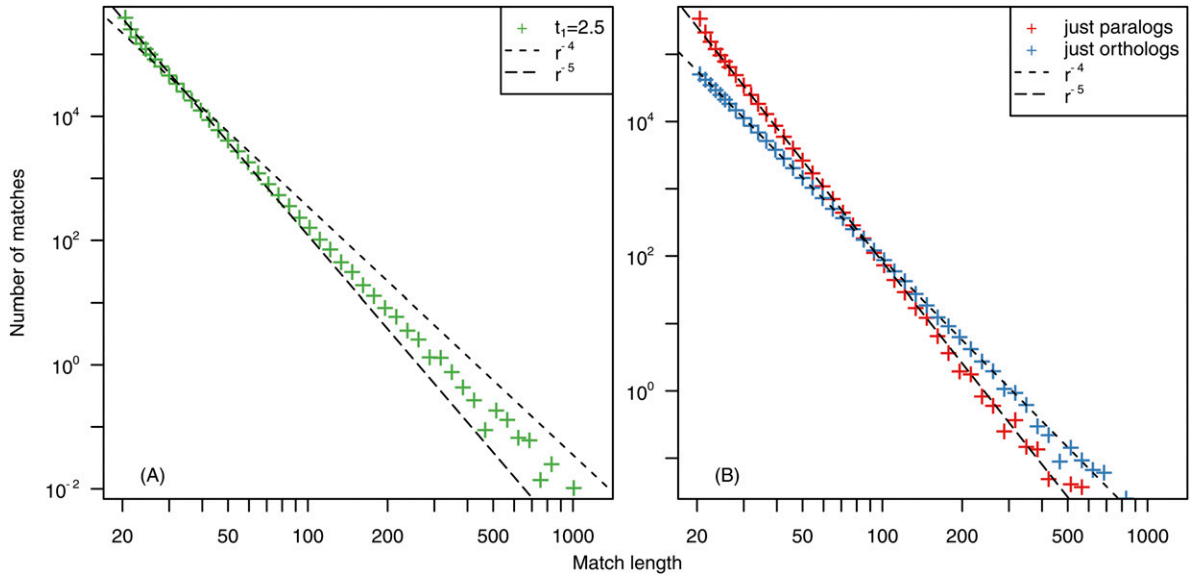

Figure 4 The MLD computed from 10,000 simulated sequences according to the procedure described in the main text. Data are represented using the logarithmic binning. On the left, we show the MLD computed from all possible matches, while on the right, we represent two different MLDs: one computed from paralogous matches only (red) and one computed from the orthologous matches only (blue). We can see that the two different MLDs cross close to the expected crossing value $r_c = 100$.

paralogous sequences compared to orthologous ones and the crossover value, r_c, is large. Then, only the $\alpha = -5$ power-law tail will be observed. On the other hand, if $N_I(0) \ll 10$, then the crossing point r_c is expected to be <20 such that the $\alpha = -5$ power law holds only for lengths where the distribution is already dominated by random matches. In contrast to previous models (Massip *et al.* 2015), this model does take into account the contribution of paralogous sequences and can explain both power-law behaviors and therefore predicts the crossing point between the two regimes.

Numerical validation

Our theoretical considerations predict a complex behavior of the match length distribution under the described evolutionary dynamics. The key ingredients are segmental duplications, generating paralogous sequences in an ancestral genome, and point mutations that break identical pairs of homologous sequences of the two genomes into smaller pieces. To illustrate our theoretical predictions concerning the two power laws, as well as the existence of the crossover point r_c, we simulated the evolution of sequences according to the discussed scenario.

We describe the evolution of a genome of length L according to two simple processes, point mutations and segmental duplications. Point mutations exchange one base pair by another one and occur with rate μ per base pair and unit of time. To mimic the existence of regions under different degrees of selective pressure, we allow for regional differences of the point mutation rates. Segmental duplications copy a contiguous segment of K nucleotides to a new position where it replaces the same amount of nucleotides, such that the total length of the genome stays constant. Segmental duplications occur with rate λ per base pair.

Our simulation has two stages (see Figure 2). At time $t = 0$, we generate a random i.i.d. sequence $\mathcal{S}$. During a time t_0, this sequence evolves according to the two described processes. In this first stage, the mutation rate is the same for all positions. At the end of this stage, the sequence represents the common ancestral genome of two species. At the beginning of the second stage, we copy the entire sequence of the common ancestor to generate the genomes of the two species A and B. These sequences are then subdivided into M continuous regions of equal length. In each such region j, the point mutation rates $\mu_{A,j}(\text{resp.}\mu_{B,j})$ are the same for all sites i and are drawn from an exponential distribution with mean μ(*i.e.*, the point mutation rate during the first stage). We chose the exponential distribution because it stipulates the least information under the given constraints. For more details about the simulation procedure, see File S1, Appendix A.

We show the result of the comparison of simulated sequences in Figure 4, left. We obtain a power-law tail in the match length distribution, which for match length $20 < r < 100$ has an exponent $\alpha = -5$ and an exponent $\alpha = -4$ for longer matches $r > 100$. For simulated sequences, we can easily classify homologous sequences into orthologous and paralogous sequences (while for natural sequences, paralogs and orthologs are not easily distinguishable due to genomic rearrangements). We show the MLD obtained from the comparison of paralogs and the MLD obtained from the comparison of orthologs for simulated sequences in Figure 4, right. We can clearly observe that orthologous sequences generate an $\alpha = -4$ power-law distribution while paralogous matches generate an $\alpha = -5$

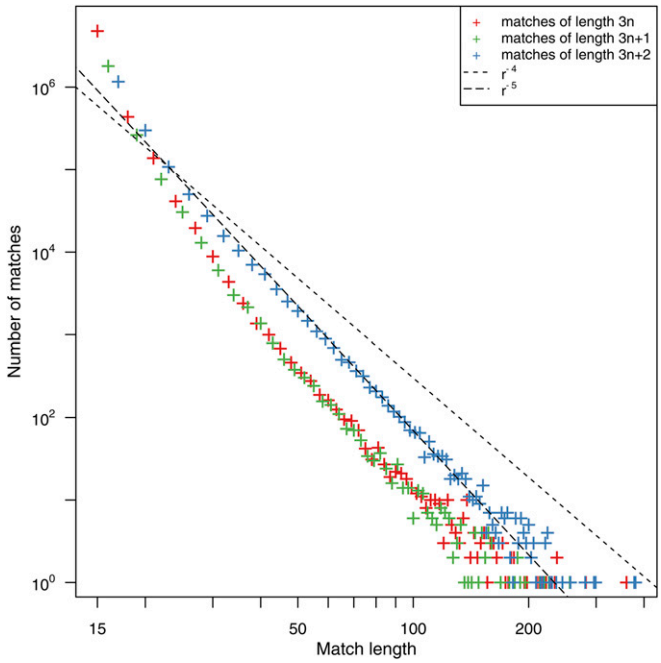

Figure 5 The MLD computed from the comparison of the human and the mouse exome, represented without logarithmic binning. Three different colors are used to represent matches of length $3n$, $3n+1$, and $3n+2$. Dashed lines represent power-law distribution with exponents $\alpha = -4$ and $\alpha = -5$.

power-law distribution. We can further easily identify the crossing point r_c as the value of r for which we obtain more matches from the comparison of orthologs than from the comparison of paralogs.

In the previous section the value of this crossing point between the two regimes was predicted to be $r_c = 2N_I(0)$ (see Equation 13), where $N_I(0)$ is the number of paralogous segments with a divergence $\tau = 0$ just before the species split. In our simulation procedure, $N_I(0)$ is simply the number of sequences that have been duplicated but that have not been mutated yet at the splitting time $t = t_0$. This number is known to be $N_I(0) = \lambda L/\mu K$ (Massip and Arndt 2013). In our simulations, $\lambda = 0.05$, $\mu = 1$, $K = 1000$, and $L = 10^7$ and therefore we predict $r_c = 100$, which is in good agreement with our observations in Figure 4. The results of our simulations are thus in good agreement with our analytical predictions.

Discussion

We developed a simple model that accounts for power-law tails in the length distribution of exact matches between two genomes. Our model assumes regional differences of the selective pressure such that the substitution rates in a region are drawn from a certain distribution. However, for naturally evolving exons the selective pressure varies also on shorter length scales. For instance, some nucleotides for many codons can be synonymously substituted by another one, mostly at third codon positions. Therefore, these substitutions at third codon positions occur with a higher rate than nonsynonymous ones. Hence, exons are expected to break preferentially at positions $3n$, with $n \in \mathbb{N}$, such that the matches with 100% identity would have lengths $3n+2$ with integer n. Classifying genomic matches according to the remainder that is left when dividing their length by 3, we observe an almost 10-fold overrepresentation of matches with length $3n+2$ over matches of lengths $3n$ and $3n+1$; see Figure 5. This suggests that the match-breaking process is dominated by the synonymous mutation rate

Using the presented model, the puzzling observation of an $\alpha = -4$ power-law tail in the MLD in the comparison of the human and mouse genomes and a corresponding $\alpha = -5$ power-law tail in the comparison of their exomes can be explained. Although the sequences stem from the same species, the relative amount of paralogous to orthologous sequence segments is different in the two data sets, which subsequently leads to different crossover points r_c. Because of the selective pressure on coding exons, the number of nonmutated paralogous sequences at the time of species divergence $N_I(0)$ is higher (relative to the number of orthologous sequences) in the exonic data set than in the noncoding data set. Thus, the crossover point in exomes r_c is larger than the longest observed match and only the $\alpha = -5$ power law can be observed.

The opposite is true for matches in the alignment of noncoding sequences. Quantitatively, in this set, paralogous sequences play a lesser role and therefore only the $\alpha = -4$ power law is observed (see Figure 1). This is surprising, as the duplication rate is thought to be roughly the same in the coding and noncoding parts of genomes. To confirm this paradoxical observation, we classified matches according to the uniqueness of their sequences in both genomes. Assuming that unique matches are more likely to be orthologous, this gives us a rough classification of homologs into orthologs and paralogs, although matches unique in both sequences can be paralogs. After the classification of all matches, our analysis made apparent that matches unique in both genomes dominate the MLD in the comparison of the noncoding parts of the genomes, while matches with several occurrences in either (or both) of the genomes dominate the distribution in the case of the comparison of exomes (see File S1, Figure S2). Moreover, we computed the MLD from the set of nonunique matches of the noncoding part of the genomes. In this comparison, the contribution of paralogs is expected to be much higher than in the full set. As expected, this MLD also exhibits an $\alpha = -5$ power law (see File S1, Figure S2), confirming that the relative contribution of orthologs and paralogs is responsible for the shape of the MLD. These differences in the proportion of paralogous sequences in the coding and noncoding DNA are likely due to the fact that paralogs are more often retained in the coding part than in the noncoding part of genomes. Since there are many more noncoding sequences in both genomes, we also observe at least

10 times more matches in the comparison of noncoding sequences than in the comparison of exomes.

The presented model does not account for changes in the divergence rates after a duplication, a phenomenon that is well documented following a gene duplication (Scannell and Wolfe 2008; Han *et al.* 2009; Panchin *et al.* 2010; Pegueroles *et al.* 2013). To assess the impact of this phenomenon on the MLD, we performed simulations where the two paralogous segments are assigned different and independent mutation rates. Interestingly, these simulations yield results qualitatively similar to those of the simpler model introduced above (see File S1, Figure S3). This new condition does not affect the value of the number of paralogous sequences that have not diverged at the time of the split [*i.e.*, the value of Nj (0)] and thus the shape of the distribution.

The model we present is very simple, and more realistic models of genome evolution include many more evolutionary processes (Dalquen *et al.* 2012). For instance, we could include a transition/transversion bias in the mutational process, variations of mutation rates in time, a codon usage bias, or different rates of duplication within and between chromosomes. Since in the end we consider just identical matching sequences and want to explain the power-law tail in the MLD, all these additional model details are not expected to affect the results.

In this article, we demonstrate that on the genome-wide scale, the length distribution of identical homologous sequence segments in a comparative alignment is nontrivial and exhibits a power-law tail, and we propose a simple model able to explain such distributions. While paralogous sequences, which had been duplicated before the species diverged, generate a power-law tail with exponent $\alpha = -5$, orthologous sequences generate a power-law tail with exponent $\alpha = -4$. Depending on the relative amount of paralogous to orthologous sequences there is a crossover between these two power-law regimes. The exponent of the power-law tail in the comparative MLD can therefore be a litmus test for the abundance of paralogous relative to orthologous sequences, while it is usually difficult to distinguish between orthologous and paralogous sequences using classical bioinformatic methods (Studer and Robinson-Rechavi 2009; Dalquen *et al.* 2013; Gabaldón and Koonin 2013). If paralogous sequences dominate, the crossover occurs for a large value of r and the apparent exponent is equal to -5; otherwise it is equal to -4.

Our method is very easy to apply. In particular, it does not require that genomes are fully assembled as long as the continuous sequences are >1 kbp, comparable to the longest matches one expects. A natural extension of our method would be to apply it to sequences from metagenomic samples to assess relative amounts of paralogous and orthologous sequences. However, we would also have to consider horizontal gene transfer, which is common among prokaryotes and generates homologous sequence segments even between unrelated genomes. Our computational model can easily be extended to take into account these and other more complex biological processes, using, for instance, already developed tools (Dalquen *et al.* 2012). This would allow us to assess their impact on our results and will be the subject of future work.

This study shows that even very simple models can often successfully be applied to seemingly very complex phenomena in biology. We were able to present a minimal model for the evolution of homologous sequences that includes effects due to segmental duplications and evolution under selective constraints—the two processes that are responsible for a power-law tail in the length distribution of identical matching sequences.

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent *et al.*, 2004 Ultraconserved elements in the human genome. Science 304: 1321–1325.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis *et al.*, 2015 Ensembl 2015. Nucleic Acids Res. 43: D662–D669.

Dalquen, D. A., M. Anisimova, G. H. Gonnet, and C. Dessimoz, 2012 Alf—a simulation framework for genome evolution. Mol. Biol. Evol. 29: 1115–1123.

Dalquen, D. A., A. M. Altenhoff, G. H. Gonnet, and C. Dessimoz, 2013 The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PLoS One 8: e56925.

Fitch, W. M., 2000 Homology: a personal view on some of the problems. Trends Genet. 16: 227–231.

Flanders, H., 1973 Differentiation under the integral sign. Am. Math. Mon. 80: 615–627.

Gabaldón, T., and E. V. Koonin, 2013 Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. 14: 360–366.

Gao, K., and J. Miller, 2011 Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments. PLoS One 6: e18464.

Gao, K., and J. Miller, 2014 Human–chimpanzee alignment: ortholog exponentials and paralog power laws. Comput. Biol. Chem. 53: 59–70.

Han, M. V., J. P. Demuth, C. L. McGrath, C. Casola, and M. W. Hahn, 2009 Adaptive evolution of young gene duplicates in mammals. Genome Res. 19: 859–867.

Hill, B. M., 1975 A simple general approach to inference about the tail of a distribution. Ann. Stat. 3: 1163–1174.

Kuhn, W., 1930 Über die Kinetik des Abbaues hochmolekularer Ketten. Ber. Dtsch. Chem. Ges. 63: 1502–1509.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12.

Massip, F., and P. F. Arndt, 2013 Neutral evolution of duplicated DNA: an evolutionary stick-breaking process causes scale-invariant behavior. Phys. Rev. Lett. 110: 148101.

Massip, F., M. Sheinman, S. Schbath, and P. F. Arndt, 2015 How evolution of genomes is reflected in exact DNA sequence match statistics. Mol. Biol. Evol. 32: 524–535.

Needleman, S. B., and C. D. Wunsch, 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: 443–453.

Newman, M. E., 2005 Power laws, pareto distributions and zipf's law. Contemp. Phys. 46: 323–351.

Panchin, A. Y., M. S. Gelfand, V. E. Ramensky, and I. I. Artamonova, 2010 Asymmetric and non-uniform evolution of recently duplicated human genes. Biol. Direct 5: 54.
Pegueroles, C., S. Laurie, and M. M. Albà, 2013 Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Mol. Biol. Evol. 30: 1830–1842.
Salerno, W., P. Havlak, and J. Miller, 2006 Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments. Proc. Natl. Acad. Sci. USA 103: 13121–13125.
Scannell, D. R., and K. H. Wolfe, 2008 A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res. 18: 137–147.
Sheinman, M., F. Massip, and P. F. Arndt, 2015 Statistical properties of pairwise distances between leaves on a random yule tree. PLoS One 10: e0120206.
Smith, T. F., and M. S. Waterman, 1981 Identification of common molecular subsequences. J. Mol. Biol. 147: 195–197.
Studer, R. A., and M. Robinson-Rechavi, 2009 How confident can we be that orthologs are similar, but paralogs differ? Trends Genet. 25: 210–216.
Ziff, R. M., and E. D. McGrady, 1985 The kinetics of cluster fragmentation and depolymerisation. J. Phys. Math. Gen. 18: 3027.

Communicating editor: E. Eskin

Appendix A: Simulating the Evolution of DNA Sequences

To simulate our evolutionary models, we proceeded as follows. A sequence of nucleotides $\mathcal{S} = (s_1, \ldots, s_L)$ of length L with $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is evolved through time in small time intervals Δt. The time intervals Δt are small enough such that for all considered evolutionary processes E of our model, which are assumed to occur with rate ρ_E, we have $\rho_E L \Delta t \ll 1$. At each step, random numbers u_i^E for all positions i and possible evolutionary processes E are drawn from a uniform distribution. The event E then occurs at position i if $u_i^E < \rho_E \Delta t$. These steps are repeated until the desired time t has elapsed.

Sequences evolve according to two simple processes, point mutations and segmental duplications. Point mutations exchange one nucleotide by another and occur with rate μ per bp and unit of time. Note that to mimic the existence of regions under different degrees of selective pressure we allow for regional differences of the point mutation rates. Segmental duplications copy a contiguous segment of K nucleotides starting at position c and paste them to a different position v, such that the K nucleotides at positions v to $v + K - 1$ are replaced by the ones from position c to $c + K - 1$. As a consequence, the total length of the sequence L stays constant in time. The segmental duplication process occurs with rate λ per bp and per unit of time.

The evolutionary scenario of our simulation has two stages, as shown in Fig. 2. At time $t = 0$, we start with a random iid sequence $\mathcal{S}$ with equal proportions of all 4 nucleotides. During a time interval of length t_0, this sequence evolves according to the two described processes. In this first stage, the mutation rate is the same for all positions. At the end of this stage, the sequence represents the common ancestor of the two species.

At the beginning of the second stage, we duplicate the entire sequence of the common ancestor to generate the genomes of the two species A and B. These sequences are then subdivided into M continuous regions of equal lengths. The point mutation rates $\mu_{A,j}$ (resp. $\mu_{B,j}$) are the same for all sites in a given region j and are independently drawn from the same exponential distribution of mean μ, i.e. the point mutation rate during the first stage.

For simplicity, the length of the M continuous regions is set to $M = K$ and the segmental duplication rates in both species λ during the second stage are set to zero. Both species then evolve independently for a divergence time t_{sp}, and we compute the MLD from a comparison of the sequences of the two species A and B. Note that even when we chose finite duplication

rates after the split (i.e. $\lambda > 0$ in the second stage), we obtained qualitatively similar MLDs.

To control for the potential impact of our choice to keep the genome size constant on our results, we also simulated the evolution of sequences where duplicated segments where added to the sequences (thus generating growing genomes). In that case, duplicates were added at the very end of the sequence, such that duplicates do not disrupt pre-existing matches. This control experiment yields qualitative similar results, in agreement with our theoretical considerations (data not shown).

Appendix B: Calculation of the derivative of N_{AIB}

In this section we describe the Taylor expansion that leads to Eq. (11) from the main text. The Taylor expansion for N_{AIB} in the neighborhood of $\tau = 0$ results in

$$N_{AIB} = N_{AIB}(0) + N'_{AIB}(0)\tau + N''_{AIB}(0)\frac{\tau^2}{2} + N'''_{AIB}(0)\frac{\tau^3}{6} + \mathcal{O}(\tau^4)\,. \tag{S1}$$

From Eq. (8) in the main text, it follows that the first term always vanishes. Using Leibniz formula to take the derivative under the integral sign in Eq. (8), we find for the first derivative

$$\begin{aligned} N'_{AIB}(\tau) = &\int_0^\tau \left(\int_0^{\tau_2} N_B(\tau_B) N_A(\tau_2 - \tau_B) N'_I(\tau - \tau_2)\, d\tau_B \right) d\tau_2 \\ &+ \int_0^\tau N_I(0) N_B(\tau_B) N_A(\tau - \tau_B)\, d\tau_B\,. \end{aligned} \tag{S2}$$

It follows that the first derivative of $N_{AIB}(\tau)$ at $\tau = 0$ vanishes. For the next term, we get

$$\begin{aligned} N''_{AIB}(\tau) = &N_A(0) N_I(0) N_B(\tau) \qquad\qquad (S3)\\ &+ \int_0^\tau N_I(0) N_B(\tau_B) N'_A(\tau - \tau_B)\, d\tau_B \\ &+ \int_0^\tau \left(\int_0^{\tau_2} N_B(\tau_B) N_A(\tau_2 - \tau_B) N''_I(\tau - \tau_2)\, d\tau_B \right) d\tau_2 \\ &+ \int_0^\tau N_B(\tau_B) N'_I(0) N_A(\tau - \tau_B)\, d\tau_B + N_A(0) N_I(0) N_B(\tau)\,. \qquad\qquad (S4) \end{aligned}$$

Here, all terms but the first one vanish for $\tau = 0$. Similarly, we can calculate the third derivative of $N(\tau)$, and we find that for $\tau = 0$

$$N'''_{AIB}(\tau) = N_A(0) N_I(0) N'_B(\tau) + N_A(0) N'_I(0) N_B(\tau) + N'_A(0) N_I(0) N_B(\tau), \tag{S5}$$

such that Eq. (S1) finally takes the form

$$\begin{aligned} N_{AIB}(\tau) = &\frac{1}{2} N_A(0) N_B(0) N_I(0) \tau^2 \\ &+ \frac{1}{6} \left(N'_A(0) N_B(0) N_I(0) + N_A(0) N'_B(0) N_I(0) + N_A(0) N_B(0) N'_I(0) \right) \tau^3 \\ &+ \mathcal{O}(\tau^4) \,. \end{aligned} \tag{S6}$$

Therefore, as long as

$$\tau \ll 3 \frac{N_A(0) N_B(0) N_I(0)}{N'_A(0) N_B(0) N_I(0) + N_A(0) N'_B(0) N_I(0) + N_A(0) N_B(0) N'_I(0)}, \tag{S7}$$

$N_{AIB}(\tau)$ is expected to scale as τ^2 and subsequently $M(r) \sim r^{-5}$.

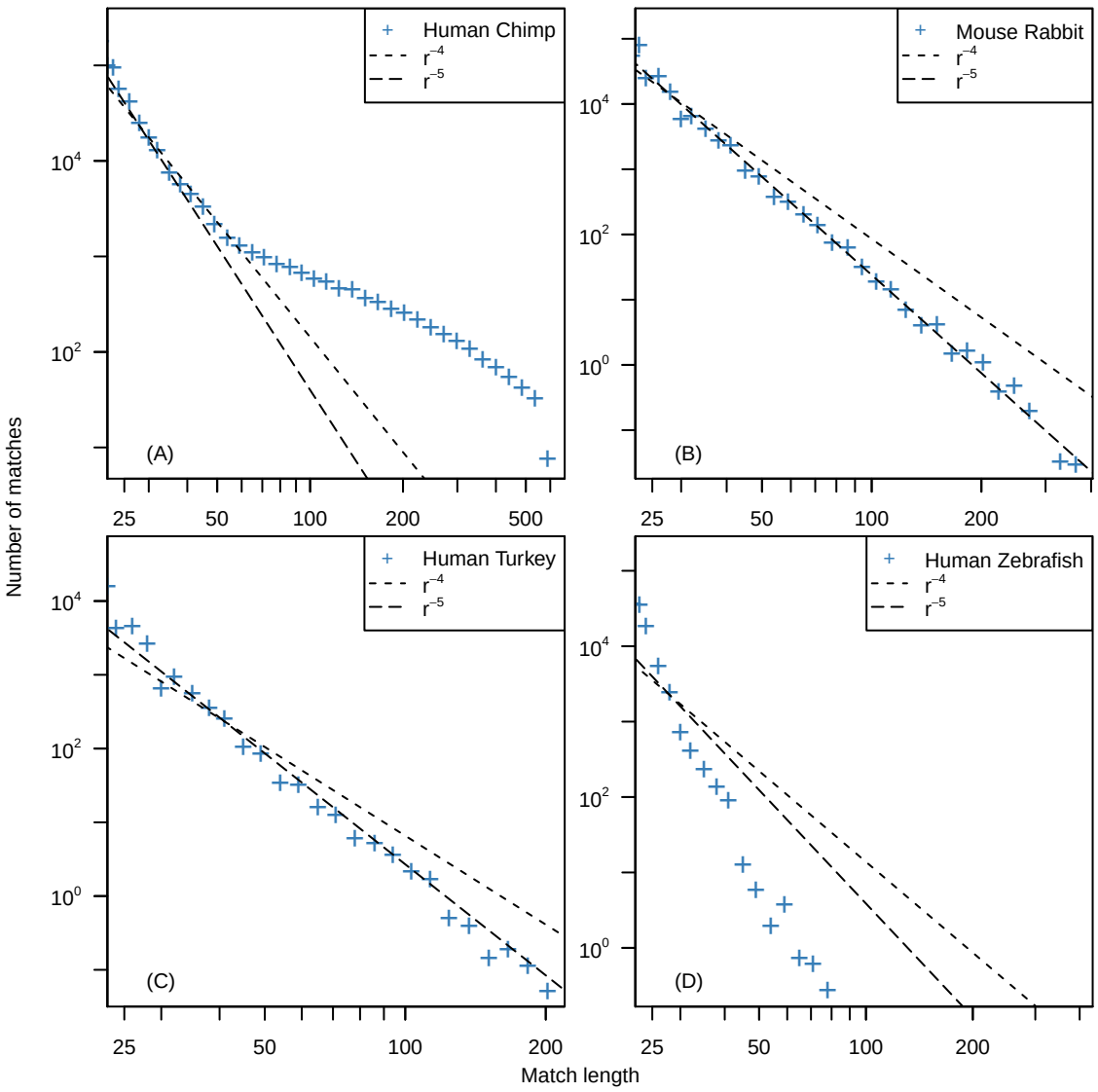

Figure S1. MLDs computed from the comparison of the exome of several species. In all four panels, dashed lines represent power-law distribution with exponent $\alpha = -4$ and $\alpha = -5$, and empirical data are represented using logarithmic binning. MLDs represented are computed from the comparison of the exomes of (A) Human and Chimp, (B) Mouse and Rabbit (C) Human and Turkey (D) Human and Zebrafish.

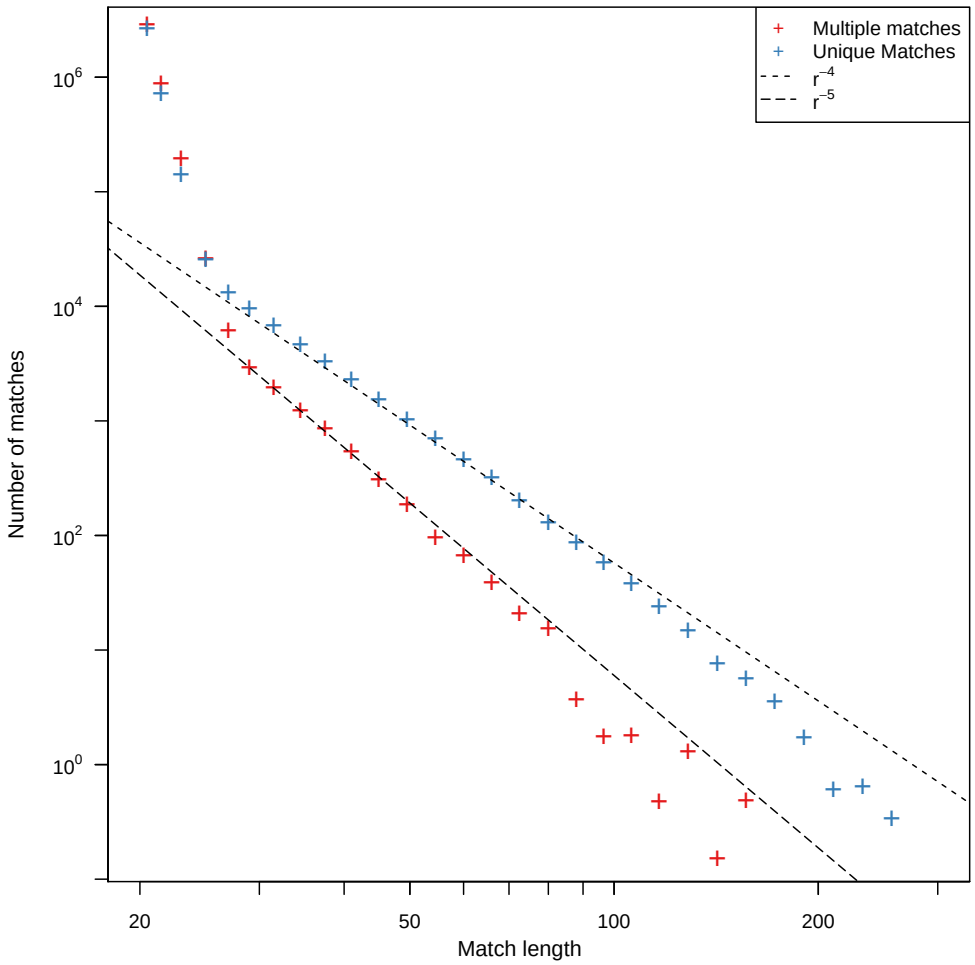

Figure S2. MLD computed from the comparison of subsets of the non-coding part of Human and Mouse genomes. Also the non-coding part of these genomes contains paralogous sequence segments from segmental duplications before the species split. To enrich for such sequences we partitioned the genomes into two subsets. Using self-alignments of the two species, we first created two libraries containing the sequences of all exact matches within their non-coding part. This library is thus be enriched for sequences duplicated before and after the speciation. Subsequently the two libraries are compared and their MLD (multiple matches, red data points) shows the expected -5 power-law. As a control, the complements of the two libraries show an -4 power-law (unique matches, blue data points).

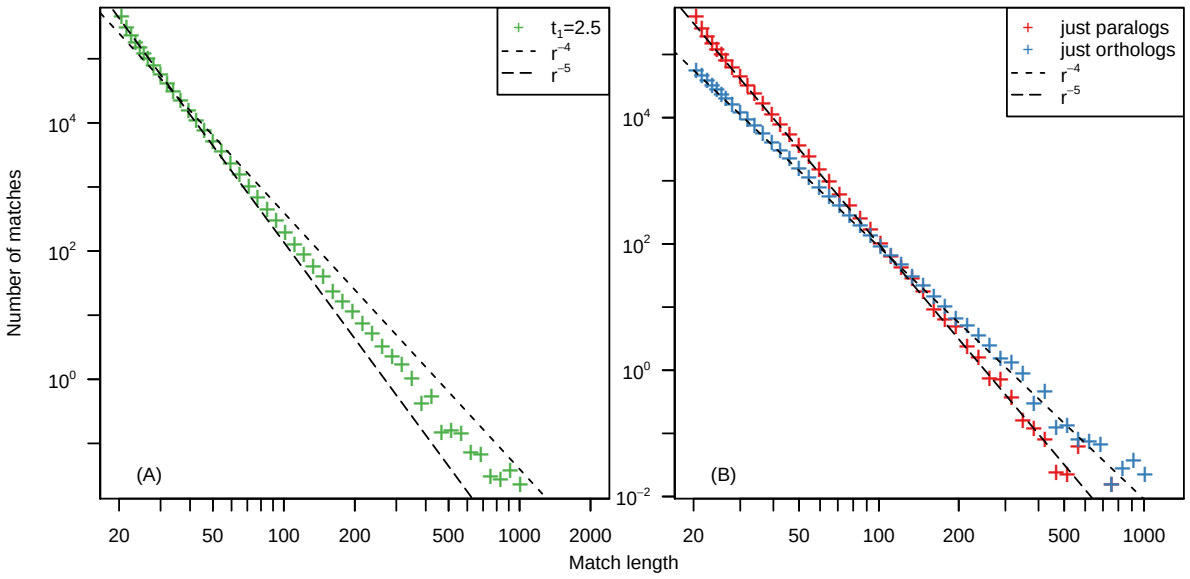

Figure S3. The MLD computed from 1000 simulated sequences according to the procedure described in the main text, but in this case, after any duplication event, the two copies are randomly assigned a new mutation rate, drawn from the mutation rate distribution (i.e. an exponential distribution of mean 1). Data are represented using the logarithmic binning. On the left panel, we show the MLD generated computed from all possible matches, while on the right panel, we represent two different MLDs: one computed from paralogous matches only (red), and one computed from the orthologous matches only (blue). One can see that the results of these simulations hold qualitatively and quantitatively similar results as the one presented in the main text on Fig. 4.