Modeling of Global Processes. Nonlinear Dynamics and Humanities

Izvestiya Vysshikh Uchebnykh Zavedeniy. Applied Nonlinear Dynamics. 2023;31(6)

Review DOI: 10.18500/0869-6632-003073

Exact sequence matches in genomic studies

M. Sheinman

Institute for Advanced Studies, Sevastopol State University, Russia
E-mail: ✉msheinman@mail.sevsu.ru

Received 28.09.2023, accepted 5.10.2023, available online 17.11.2023, published 30.11.2023

Abstract. The *purpose* of this article is to review usage of exact sequence matches in different field of genomic studies. *Methods.* The presentation is built in the form of a brief review of clearly non-exhaustive list of works in which the authors inferred biological knowledge using statistical properties of exact matches between different genomic texts or self-matches along the same genomic sequence. *Results.* Often, in genomic studies, different genomic loci exhibit different statistical properties, while their boundaries are not known a priory. In such cases we conclude that studying statistical properties of exact sequence matches is a useful alternative to other methods, for instance, based on arbitrary-size (non-)sliding windowing of the genome. *Conclusion.* This review demonstrates that exact sequences matches are not only an important auxiliary alignment step, but also helpful in other contexts. Their statistical properties are relatively easy to calculate analytically or numerically under various assumptions and compare to empirical data, validating models and fitting the models' parameters.

Keywords: genomics, exact sequence matches, maximal exact matches, k-mers, genome evolution, horizontal gene transfer.

Acknowledgements. I thank P. F. Arndt and F. Massip for useful comments and discussions. Numeric analysis was carried out using the supercomputer cluster "Afalina" in Sevastopol State University. The work was made within the program "Prioritet-2030" of Sevastopol State University (strategic project No. 3, No. 121121700318-1) and project FEFM-2023-0005.

For citation: Sheinman M. Exact sequence matches in genomic studies. Izvestiya VUZ. Applied Nonlinear Dynamics. 2023;31(6):739–756. DOI: 10.18500/0869-6632-003073

Introduction

Genomic sequences evolve via error-prone DNA replication. Therefore, comparing DNA sequences from different origins we often find similar texts. Significant similarity suggests common ancestor of such DNA regions and we define such regions that have a common evolutionary origin as *homologous* [1, 2]. Homologous sequences typically share more and longer exact matches, compared to non-homologous ones.

k-mers. There are different ways to analyze the exact matches. One can look at the shared sequences of specific length k, so-called k-mers. For instance, two sequences: ACG**C**GATT**G**CTAA and ACG**A**GATT**T**CTAA share two 4-mers: GATT and CTAA. There are 6 shared 3-mers: CGA, ACG, GAT, ATT, CTA and TAA. Homologous sequences are expected to share more long k-mers than non-homologous ones.

Similarly, one can compare a sequence to itself searching for self-matches. Using an efficient software, like `Jellyfish` [3], one can obtain all the exact matches of length k along a genomic sequence — all it k-mers. Abundance of genomic k-mers (as well as k-mer peptide abundances in protein sequences [4]) is informative. For $k = 1$ this boils down to GC content, which is an important genomic trait and is often used to characterize within- and between-genome variations [5], especially for bacterial organisms [6]. k=2-mers-dinucleotides — are also used [5, 7], while local abundance of the CG 2-mer is especially interesting (see review about CpG islands in Ref. [8]). $k = 3$-mers abundances reflect codon usage bias: the preferential or non-random use of synonymous codons — an intriguing phenomenon observed in all domains of life [9]. Longer k-mers are discussed in the following sections.

MEMs. Another way to analyze sequence (self-)similarities is to find maximal exact matches (MEMs). These are exact matches between two sequences that cannot be extended either way without introducing mismatches [10]. Since the early days of bioinformatics, MEMs are used to visualize (self-)similarities of genomic sequences using so-called dotplots [11]. In the example above MEMs are given by ACG, GATT and CTAA. Lengths of MEMs reflect degree of similarity: for two random sequences of length $L \gg 1$ an average number of MEMs of length r decays exponentially with r following

$$m(r) = \frac{L^2}{2}(1-p)^2 p^r, \tag{1}$$

where p — the probability of matching nucleotides, which is equal to 1/4 for an i.i.d. sequence with equal proportions of nucleotides. This exponential match length distribution leads to the Gumbel distribution for longest matches in an alignment of i.i.d. sequences, which is commonly used to assess the significance of local alignments [12, 13].

In the following sections I review usage of exact sequence matches in different contexts.

Exact matches as alignment seeds

Aligning two or more genomic (or protein) sequences is arguably the most fundamental tasks in bioinformatics. If two sequences align well, they are likely homologous. A common approach to alignment is the seed and extend approach [14]. In a first step, very short local similarities are found. These similarities are often required to be exact sequence matches. Then they are used as alignment seeds. In a second step, starting from these seeds, alignments are constructed. This construction is often the slowest step, so that the seeds are ought to be sensitive and specific and, yet, their search should be fast [15, 16].

Classical aligners use simple k-mers (exact sequence matches of length k) as seeds [14, 17, 18]. Later, seeds in the form of gapped k-mers were introduced, where bases at certain positions of the k-mer are not required to match [19, 20]. Such spaced seeds (with some variations, like differentiation between different types of mismatches: transitions and transversions) are currently used in alignments like `BLASTZ` [21], `YASS` [22], `DIAMOND` [23], `LASTZ` [24], and `MegaBLAST` version of `BLASTn` [25]. Recently, other modified k-mer methods were suggested [26], like mimimizers [27], syncmers [28] and strobemers [29].

An alternative approach is to use MEMs beyond certain length as alignment seeds. This was implemented in aligners like `MAVID`, [30], `GAME` [31], `CoCoNUT` [32] and `MUMmer` [33, 34]. The

last was the first software system that used suffix trees to find MEMs as potential seeds for an alignment. A suffix tree is a data structure for representing all the subsequences of a sequence [35]. For a sequence of length L it can be represented in space $\mathcal{O}(L)$ and fast algorithms have been found to construct a suffix tree in time $\mathcal{O}(L)$, e.g. [36]. Given the suffix tree of one sequence and another sequence of length L', one can compute all MEMs between the sequences in time $\mathcal{O}(L')$ [33].

Insights into evolutionary history using exact matches

Beyond technical aspects, exact matches can be used directly to infer information about evolutionary history of genomes. The reasons for this are that (i) one can relatively easily obtain them empirically, (ii) in contrast to alignments, where some properties depend on the alignment algorithm and its parameters, exact matches are unambiguously defined, (iii) statistical properties of exact matches within the framework of theoretical models can be often obtained analytically or their numerical calculation is relatively fast. In sum, one can often compare empirical data to theoretical predictions using exact sequence matches, validating a model and fitting its parameters. In the following we review a few such cases.

Sequences similarity estimates using exact matches–alignment-free methods. An important direction in bionformatics is direct phylogeny reconstruction without the alignment step, using exact matches [37, 38]. For phylogenetic reconstruction, a first step is to estimate pairwise evolutionary distances between protein or nucleic-acid sequences [39]. Also, statistical properties of pairwise evolutionary distances between members of a taxon can shed light on its speciation history [40]. The distances are usually inferred using pairwise or multiple-sequence alignments. However, sequence alignment happens to be too slow for huge amount of data. To cope with this problem alignment-free approaches have been being developed. In contrast, some alignment-free approaches are based on k-mer abundances [41–45]. The tools `Cnidaria` [46] and `AAF` [47] use the Jaccard index between two sets of k-mers to estimate the distance between them, while `SlopeTree` [48] defined a distance measure using the decay of the number of k-mer matches between two sequences, as a function of k. Other approaches take the length of maximal (non)exact matches as an input [49–53].

Similar methods are currently used not only for phylogeny, but also in metagenomic analysis [54–57], to identify genome rearrangements [58], in haplotype classification [59], in medical applications [60–63] and other fields. In all these applications, one requires fast estimate of pairwise similarity in large sets of sequence data and analysis of exact matches often allows to circumvent tedious alignment process.

Evolution of DNA repeats and their pseudo-linguistic features. Sequencing of genomes has revealed that genetic texts comprise repeats of different kinds [64]. Repetitive DNA contains many homologous sequences, sharing significant similarities to each other. Hence, statistical properties of genomic sequence differ from those of random ones in this respect [65]. One of these properties, which we discuss here, is that certain k-mers are much more abundant than others (this property is used to identify repeat families [66]). In particular, s — the abundances of long k-mers — exhibits a wide, scale-free distribution: number of k-mers with abundance s scales as $n_k(s) \sim s^{-\alpha}$ with $\alpha \simeq 2$, as shown in Fig. 1, *a*. This phenomenon resembles statistical properties of human texts, where abundances of words also exhibit a scale-free distribution [67]. For human texts such a linguistic feature is often presented as Zipf law [68,69] (see Fig. 1, *a* (Inset)). Despite an incomplete analogy, (k-mers are not genomic "words"), this intriguing similarity between genomes and human texts has led some researchers to analyze genetic sequences from a linguistic perspective [70–72], while many others questioned this approach (see e.g. [73]).

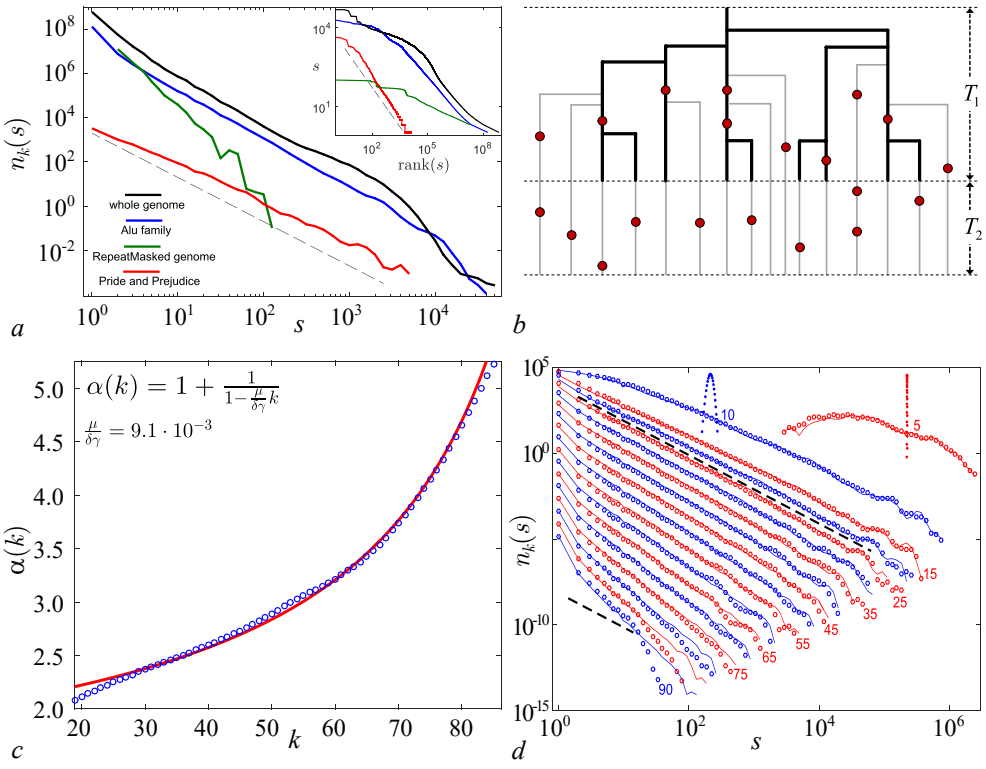

Fig 1. *a* — Distributions of abundances of k-mers for $k = 40$ in human genome. s is the number of copies of a certain k-mer and $n_k(s)$ is the number of different k-mers with abundance s. Distributions for different genomic compartments are presented: the whole genome (solid, black), the whole genome after masking the repeat elements (solid, green) and the Alu family of repeats (solid, blue). For comparison the distribution of word abundances in Pride and Prejudice [77] is also shown (solid, red). The dashed line represents the power-law $n_k(s) \sim s^{-\alpha}$ with $\alpha = 2$. For a randomly shuffled human genome or a random sequence of the same length there is not a single k-mer with $s > 1$. Inset: the corresponding Zipf 's plots for the main figure. For each k-mer (or a word for Pride and Prejudice) its abundance is plotted vs. the rank of its abundance. The dashed line represents the power-law $s \sim 1/\text{rank}(s)$. *b* — Illustration of the analysed model for the dynamics of repeat elements. Each branch represents a sequence of the repeat. Active elements are depicted in thick, black lines, while silent ones are shown in thin, gray lines. During the activity burst, selfish elements duplicate exponentially with time and accumulate mutations (red marks). After the burst sequences do not duplicate anymore but still mutate. *c* — Estimation of parameters of the model using the analytic fit of the empirical data. Circles represent the empirical power-law exponent α as a function of k. The line is the numerical fit of the data points using the analytical prediction (upper-left corner). The resulting estimator is presented below the equation (μ is the mutation rate, γ is the duplication rate and δ is the fraction of active repeats after a duplication). *d* — Distributions of abundances of k-mers, $n_k(s)$, for different values of k, from 5 to 90 in steps of 5, from top to bottom (see numbers in the figure). Circles represent $n_k(s)$ in the empirical data for the Alu family of repeats. Dots represent $n_k(s)$ in a random sequence, of the same length as the empirical one for $k = 5$ (red) and $k = 10$ (blue). Lines represent $n_k(s)$ in simulated Alu elements using the set of parameters in (*c*). The dashed lines represent the power-law decay $n_s \sim s^{-\alpha}$ with $\alpha = 2$. For visibility the values of $n_k(s)$ are normalized differently for each value of k (but in the same way for the empirical and the simulated data), so that the units of the vertical axis are arbitrary. The figure adapted with permission from [75] (color online)

It was shown [74] that a simple "copy-paste" model of selfish DNA *in a fixed-size genome* generates a fat tail in the distribution of k-mer abundances distribution, $n_k(s) \sim s^{-\alpha}$, but the value of the power-law exponent in Ref. [74] was $\alpha(k) = -1$ and k-independent—different from the empirical results, where the exponent was $\alpha \simeq 2$ (see Fig. 1, *a*) and increased with k (see Fig. 1, *d*). After this, it was demonstrated [75] that, taking into account *increase of the genome size* due to spreading of selfish DNA (see Fig. 1, *b*), one can analytically calculate the power-law $\alpha(k)$ (upper-left corner of Fig. 1, *c*). Taking separately one repeat family (Alu family [76] in Ref. [75]), i.e. considering only k-mers from the repeats of the family, the analytical predictions

account very well for the empirical results. Fitting the model parameters to the empirical data per repeat family, the spreading rates of different families was obtained using the analytic formula (see Fig. 1, *c*). Furthermore, numerical simulation of the model with the fitted parameters accurately reproduces abundance distribution of k-mers for all k values (see Fig. 1, *d*). This study, using statistical properties of k-mers, showed that Zipf law in genomic texts is rather a consequence of the evolutionary dynamics of DNA repeats and does not reflect linguistic features of genomic texts. Moreover, evolutionary dynamics of DNA repeats can be well-modelled as exponentially growing process on a large scale.

How to relate time divergence distribution to genomic data using MEMs. Before I proceed to concrete examples, let's discuss here how to validate a model for evolutionary history using genomic sequence data in case that the model predicts distribution of pairwise evolutionary time distances between different loci along the genome(s) and the borders of the loci are not known in advance. In principle, using the molecular clock assumption, the evolutionary time divergence τ between two DNA loci with effective mutation rate μ is related to the density of mismatches at these loci, $\mu\tau \ll 1$ (ignoring back mutations). However, identifying loci with constant mutational density is often challenging, requires setting of arbitrary parameters, like size of the window in the case of windowing the genome to regions of fixed size (see e.g. [78,79]). MEMs analysis circumvents this problem and was used in Refs. [80–86]. The main idea is to use statistical properties of the MEMs lengths instead of the statistical properties of loci time divergences. One can show [87], using the result derived in [88] that the molecular clock instead of fraction of mismatches can be formulated in terms of the MEMs lengths distribution (MLD) r.

Namely, for a given time divergence τ (twice the time to their last common ancestor) between two loci of length K the expected number of differences is $\mu\tau K$ and follows Poisson distribution. In contrast, the expected number of MEMs (the distances between subsequent differences) of length r in the regime $K \gg r \gg 1$ is given by

$$m(r|\tau) = \left[2\tau\mu + (\tau\mu)^2(K-r)\right] e^{-\tau\mu r} \simeq K(\tau\mu)^2 e^{-\tau\mu r}. \tag{2}$$

Using Eq. (2) one can relate the empirically observed distribution of MEMs length $m(r)$ to the distribution of loci pairwise time evolutionary distances $P(\tau)$, predicted by a model, using

$$m(r) = \int_0^\infty m(r|\tau)P(\tau)d\tau = L\int_0^\infty (\tau\mu)^2 e^{-\tau\mu r}P(\tau)d\tau = L\frac{d^2\widetilde{P}(\mu r)}{dr^2}, \tag{3}$$

where L is the total length and $\widetilde{P}(\sigma) = \mathscr{L}\{P\}(\sigma) = \int_0^\infty e^{-\sigma\tau}P(\tau)d\tau$ is the Laplace transform of $P(\tau)$. One can see that length distribution of exact sequence matches between two sequences is related to the Laplace transform of their time divergence distribution. Hence, there is a direct relationship between the MLD $m(r)$ and the Laplace transform of the loci pairwise time divergence distribution. In particular, as discussed in detail in Ref. [83], scaling behaviour of pairwise distances distribution for close pairs, $P(\tau) \sim \tau^{\alpha-3}$ as $\tau \to 0$ dictates power-law tail of MEMs length distribution, $m(r) \sim r^{-\alpha}$ as $r \to \infty$. In sum, studying $m(r)$ — a quantity that can be easily computed from empirical data — allows to reconstruct the evolutionary history of the genomes.

Using the same arguments, if different loci mutate with different mutation rate and the rate is distributed as $P_{\text{mut}}(\mu)$, while the evolutionary time distance between all loci τ is the same, MEMs distances are distributed as

$$m(r) = L\frac{d^2\widetilde{P}_{\text{mut}}(\tau r)}{dr^2}. \tag{4}$$

In the following we demonstrate how using this approach one can shed light on different aspects of genomic evolutionary history.

Self MEMs reflect segmental duplication history of genomes. MEMs along genomes exhibit interesting statistical features. In particular, distribution of their lengths, $m(r)$, strongly deviates for the random-sequence prediction (1). The dot-plot, shown in Fig. 2, *a*, demonstrates that the genome contains many paralogous sequences. As discussed in the previous section, repeat families generate such repetitive sequences. However, cleaning the repeats using the repeat masking software one can still observe long similar sequences that look like broken sticks [87] on a dot-plot (see example in Fig. 2, *b*). Length distribution of MEMs along such "sticks" was found [89] to obey $\alpha = -3$ power-law:

$$m(r) \sim r^{-3}, \tag{5}$$

where L is the genome length, as shown in Fig. 2, *c*.

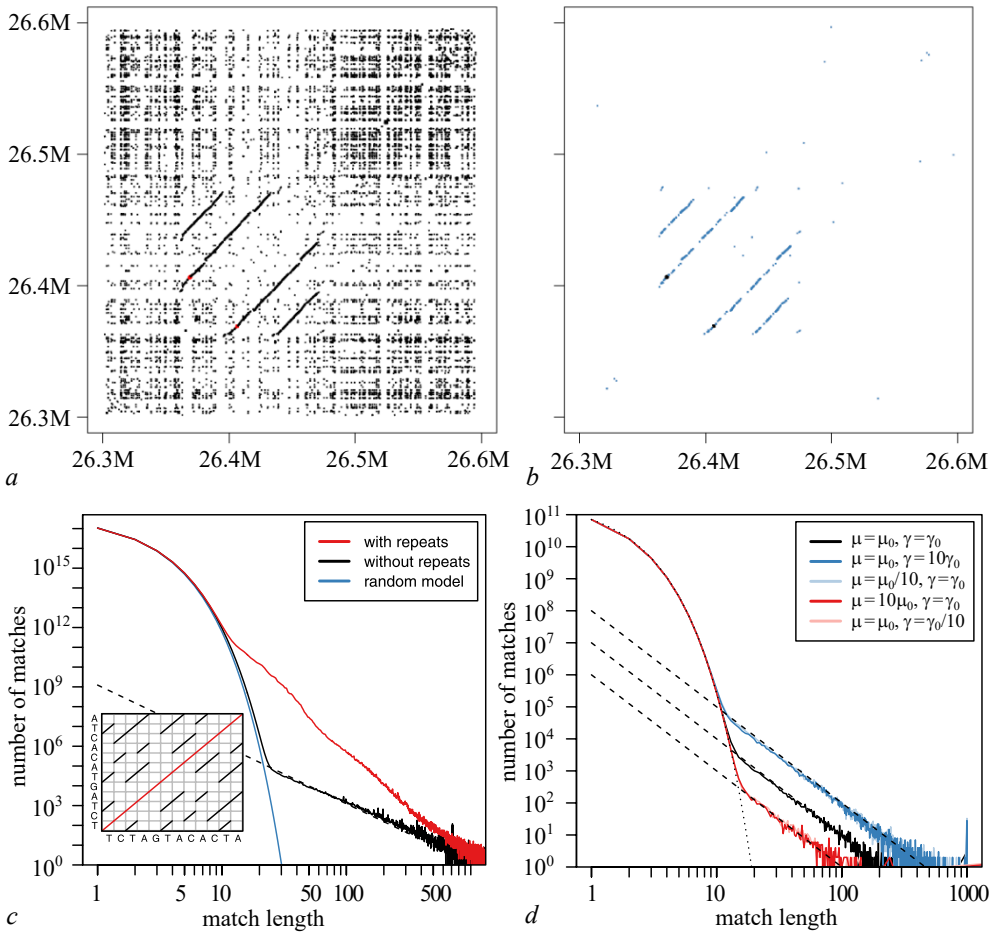

Fig 2. *a*, *b* — Dot-plot of self-MEMs along a part of human chromosome 1 before (*a*) and after (*b*) repeat masking. *c* — The MEMs length distribution (MLD) for a self-alignment of the human genome. The MLD for the complete genome excluding repetitive sequences (in total $L = 1.23$ Gbp) is shown in black and shows the described power law tail. The MLD for a human sequence of the same length but including repetitive elements is shown in red. For small lengths both distributions coincide and are dominated by random sequence matches, which occur in randomly shuffled sequences, that follows Eq. (1) (blue curve). The dashed line represents the theoretical prediction (6), see below. The inset gives an example for an alignment grid of a self-alignment of a sequence of length 12. Matching nucleotides are marked by diagonal lines forming MEMs. The global MEM is shown in red along the main diagonal. Off-diagonal MEMs are depicted in black. The grid is symmetric and only MEMs above the main diagonal are counted. In this example there are six MEMs of length one, three matches of length two, and one match of length three. *d* — The simulated MLD for various values of mutation rate μ and duplication rate γ. Theoretical predictions (6) using the stick-breaking model [81] are shown with dashed lines. The dotted line is based on Eq. (1). The panels (*c*-*d*) adapted from Ref. [81] with permission (color online)

These observations were explained in Ref. [81] using a simple model of neutral duplication and mutation of genomic loci. The model assumes that genomic loci duplicate (copy-paste themselves) to another part of the genome with a constant rate and also undergo point mutations. Using the formalism presented above, if loci of length K duplicate with rate (per bp) γ (such that duplication rate per locus is γK) and mutate with rate μ, the distribution of pairwise time divergences between paralogous loci in a genome in a steady state is uniform and given by $P(\tau) = (2K\gamma)^{-1}$, such that $\widetilde{P}(\sigma) = (\gamma K\sigma)^{-1}$. Then, using Eq. (3), one gets

$$m(r) = \frac{\gamma K}{\mu}\frac{L}{r^3}. \tag{6}$$

Eq. (6) fits the tail of the empirical (see Fig. 2, *c*) and simulated (see Fig. 2, *d*) MLDs and allows to estimate the duplication rate γ in the genome: assuming that mutations occur with a rate of about 1.5% per 10 million years, $\gamma = 4.5$ Mbp per million years have been duplicated in the human lineage, in good agreement with Ref. [90].

Retroduplications generate a different MEMs lengths distribution. Segmental duplication is not the only biological process that produces duplications in eukaryotic genomes. Retroduplication is a well-known biological mechanism which consists of the retrotranscription of an mRNA molecule into the genome. For this reason, retroduplication will solely duplicate transcribed segments of the genome. Besides, this mechanism generates partial duplicates which do not include introns. As retroduplicants also do not contain regulatory elements and promoters, they mostly produce nonfunctional copies, highly similar to the concatenated exons of the functional gene, commonly known as processed pseudogenes [91]. Various functions have been found for such pseudogenes [92,93], even though they often result in evolutionary dead ends.

As an example, consider large family of 113 processed pseudogenes of the ribosomal protein RPL21 in the human genome [82]. Its distance matrix and a compatible phylogenetic tree in Fig. 3, *a*. The matrix and the tree suggest that all these pseudogenes were actually generated by retrotranscription of a single functional gene. Following this mechanism, a gene of length K duplicates with rate γK, while its duplicates (processed, nontranscribed pseudogenes) do not

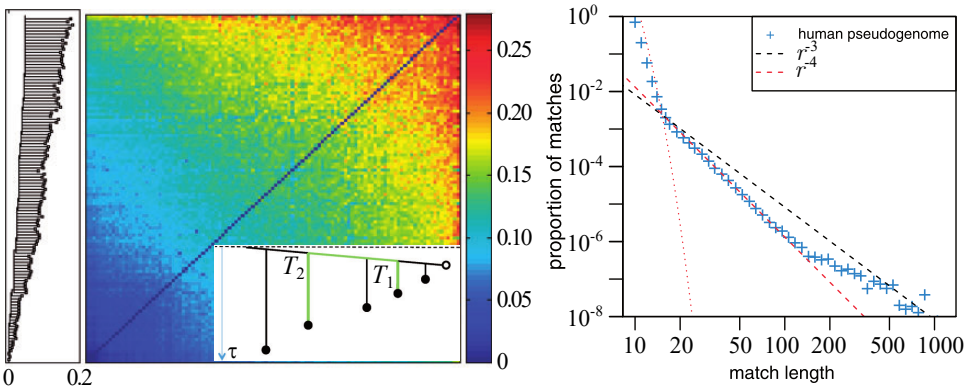

Fig 3. *a* — Distance matrix of 113 processed pseudogenes of the RPL21 gene and their phylogenetic tree. Inset: An example of the rooted tree of a pseudogene family (filled circles) stemming from one gene (open circle). The gene evolves much slower than its pseudogenes and the pseudogenes do not retroduplicate. The evolutionary distance between two leaves (green path) is the sum of the evolutionary distance covered by each pseudogene since its retroduplication event and the evolutionary distance covered by the gene between the two retroduplication events. All circles represent contemporary sequence segments. *b* — The MLD computed from the self-alignment of the human processed pseudogenome. The total length of this genome is $L = 6,433,368$ bp. The red dotted line represents the expected distribution for random sequences, and the red and black dashed lines represent power laws with exponent $\alpha = -4$ and $\alpha = -3$, respectively. Adapted from [82] with permission (color online)

duplicate. Since the evolutionary pressure on the pseudogenes is expected to be much weaker (if any), we assume that the gene and its pseudogenes exhibit different effective mutation rates. This results in a tree similar to the one shown in Fig. 3, *a* (Inset). The evolutionary time that separates two leaves on such a tree is a sum of three times: The evolutionary time elapsed after the first retroduplication event, the evolutionary time elapsed after the second retroduplication event, and the evolutionary time elapsed in the source gene between the two retroduplications (see the green path of the tree in Fig. 3, *a* (Inset)).

One can show [82] that the distribution of the pairwise distances on such a tree scales as $P(\tau) \sim \tau$ as $\tau \to 0$. Thus, its Laplace transform scales as $\widetilde{P}(\sigma) \sim \sigma^{-2}$ as $\sigma \to \infty$, so that, using Eq. (3), one obtains $m(r) \sim r^{-4}$. In sum, MEMs of processed pseudogenes (retroduplicants) are expected to generate MLD with a power law tail with exponent $\alpha = -4$. Indeed, concatenating all the annotated processed pseudogenes of the human genome to construct the so-called human "processed pseudogenome", the MLD computed from this processed pseudogenome shows a good agreement with this prediction (see Fig. 3, *b*). The deviation of the power law in the very tip of the MLD can be explained either by subsequent segmental duplication of retroduplicated loci or by selective constrains on the retroduplicants making them more conserved than expected by the neutral model.

MEMs of different genomes from the same population reflect demographic history

Genetic diversity within a population is shaped by many factors: mutagenesis initially introduces genomic variation into the genome of a single individual, which is then subject to natural selection and genetic drift. In a neutral evolution the average diversity depends on the mutation rate and the effective population size. For diploid organisms with constant effective population size N_e it is given by

$$\theta = 4\mu N_e. \tag{7}$$

The evolutionary pairwise time distances in such a population (time is measured in generations) is distributed exponentially [94, 95]

$$P(\tau) = \frac{1}{4N} e^{-\frac{\tau}{4N}}. \tag{8}$$

In addition to described evolutionary processes, genetic recombination shuffles genetic material of different individuals into a single genome, such that, assuming random mating [95], comparing two genomes from the populations, one gets loci with different evolutionary distance, following Eq. (8). Therefore, using Eqs. (8,2,3), lengths of MEMs of two haploid genomes from the diploid populations are distributed as [86]

$$m(r) = \frac{2\theta L^2}{(1+\theta r)^3}. \tag{9}$$

In practice, instead of taking genomes of two individuals one can find MEMs of sister chromosomes from the same individual, i.e. the distances between sequential heterozygous sites [80, 86]. Thus, Eq. (9) predicts distribution of distances between sequential heterozygous sites of an individual from a neutrally evolving fixed-size population of diploid organisms [86]. Often, MEMs of sister chromosomes in the literature are referred to as runs of homozygosity (ROH) and beyond inferring population history are also used to calculate genomic inbreeding, decipher genetic architecture of complex traits and diseases [96].

Interestingly, Eq. (9) agrees very well only with empirical distances between sequential heterozygous sites of individuals from African descent [86]. Other individuals exhibit enrichment

of long MEMs relative to the theoretical prediction and African individuals [80,86]. The most plausible explanation is that non-Africans underwent a population bottleneck while moving out of Africa [97]. To account for such non-fixed population size demographic history, suitable theoretical framework was developed in Ref. [86]. Using this approach $m(r)$ for population with a bottleneck was obtained and fitting the parameters to the empirical data the timing and the bottleneck strength were estimated. In sum, MEMs length distribution can be used to analyse demographic history using unphased genomic data of a single individual. It is still unclear how this approach compares to other methods based on the Sequential Markovian Coalescent (e.g. Ref. [98]) and how to extend it to polyploid organisms.

MEMs identify evolutionary conserved sequences. Distant organisms rarely share long MEMs. In higher organimsm such MEMs (for instance, the ones shared between human, mouse, and rat genomes and longer than 200bp) are usually interpreted as ultra-conserved elements, [99]. Although functions of such elements is still mostly unclear, their existence clearly demonstrates that effective mutation rate varies along the genome possibly due to different selective pressure acting on different loci of the genome [100].

MEMs reveal statistical properties of evolutionary conserved sequences. Variation of selective pressure along a genome generates a certain distribution of loci effective mutation rate. The selective pressure on a locus is affected by the fitness effect of a mutation at this locus and by the effective population size of the taxon [101, 102]. The distribution of fitness effects can in principle be assessed [103], but these methods require in general to conduct complex experiments in controlled environments. However, studying MEMs one can directly model the mosaic distribution of effective mutation rates under simple assumptions, following Refs. [82,83,85].

The evolutionary distance between orthologous regions of two taxa A and B along locus i is given by $\frac{\tau}{2}(\mu_A^i + \mu_B^i)$, where τ is the time divergence (twice the time to the last common ancestor of A and B) and μ_A^i is the effective mutation rate along locus i in taxon B. If we assume that (i) different loci have different effective mutation rates, such that μ_A^i and μ_B^i are distributed with certain non-trivial probability distributions along the genome that does not vanish at zero and (ii) μ_A^i and μ_B^i are not correlated, the density of the average mutation rate $\mu^i = (\mu_A^i + \mu_B^i)/2$ scales as $P(\mu^i) \sim \mu^i$ as $\mu^i \to 0$. Using this consideration and Eq. (4), the MEMs length distribution of two taxa that evolve under assumptions (i) and (ii) has a power-law tail with exponent $\alpha = -4$: $m(r) \sim r^{-4}$. This prediction is validated for different pairs of taxa (bacterial pairs and eukaryotic pairs) (see [82,83,85,89,104]), demonstrating how generic and robust are the taken model assumptions. In fact, one should distinguish between purely orthologous genomic regions and paralogous ones. For the last ones segmental duplication might have occurred before splitting of the considered pair of taxa. The MEMs along such loci is predicted and found empirically to follow a power-law tail with exponent $\alpha = -5$: $m(r) \sim r^{-5}$. See detailed discussion about this in Ref. [83].

In Ref. [85], it was shown that using this approach on can unravel conserved sequences and horizontally transferred ones (see below more detailed discussion about the latter ones) and construct calibrated phylogenetic trees of bacterial taxa (*Enterobacteriaceae* family was taken as an example). In sum, statistical properties of MEMs between different taxa shed light on the evolution of their genomes before and after their split, including non-neutral effects due to selective pressure.

MEMs identify and quantify horizontal gene transfers. In contrast to higher organisms, in bacterial domain of life, an ubiquitous source of long MEMs between distant taxa is horizontal gene transfer (HGT): transfer of genetic material from one organism to another [105, 106]. This happens via a variety of mechanisms: conjugation, transduction, and

transformation [107]. In fact, to some extent, HGT present in all domains of life [108], but in the bacterial one, exchange of genetic material is a key driver of evolution (see e.g. [109]).

Since the discovery of HGT [110], several methods have been developed to infer HGT (see review in [111]). Arguably, the simplest one is to search long MEMs between two genomes, much longer that one would expect based on their average genome-wide genomic divergence [84,112,113]. An example of how this works one can see on Fig. 4 [84]. In the dot plot comparing the genome sequences of *Escherichia coli* and *Salmonella enterica* from the *Enterobacteriaceae* family (Fig. 4, *a*), there are many exact matches shorter than 300 bp along the diagonal, revealing a conservation of the genomic architecture at the family level. Filtering out matches shorter than 300 bp (Fig. 4, *b*) completely eliminates the diagonal line, suggesting that exact matches in the orthologous sequences of these genomes are invariably short. Because very long exact sequence matches are extremely unlikely in orthologs, those that do occur are most likely xenologs [114]: sequences that are shared due to HGT event. As an example, Fig. 4, *c* shows a dot plot

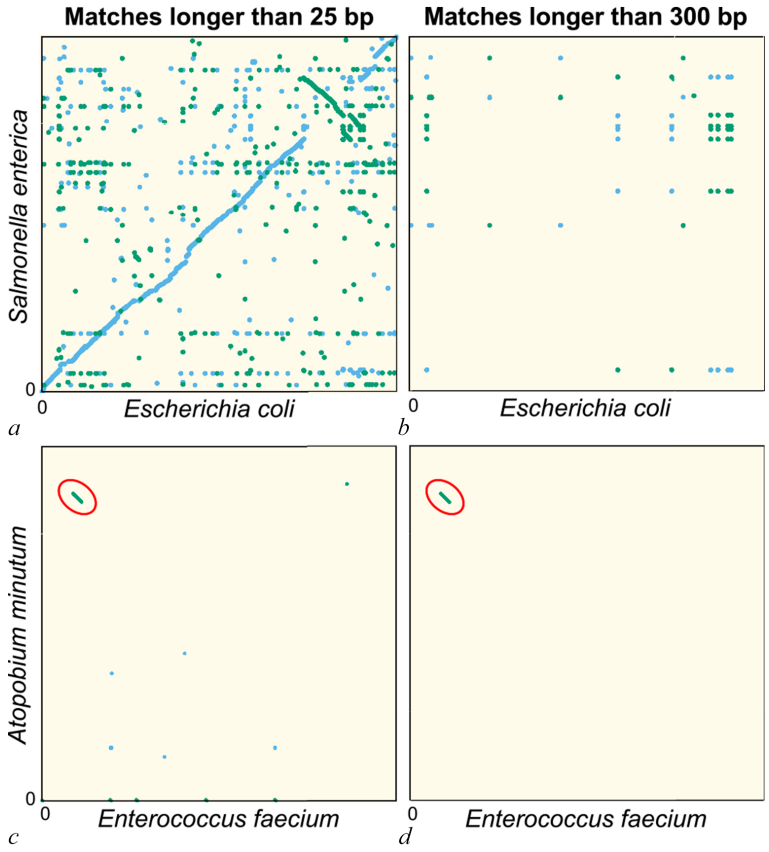

Fig 4. Dot plots of MEMs found in pairs of distant bacteria. On panels *a* and *b* resp. *c* and *d*, each dot/line on the grid represents a MEM at locus x of the genome of *Escherichia coli* (resp. *Enterococcus faecium*) and locus y of the genome of *Salmonella enterica* (resp. *Atopobium minutum*). Blue dots/lines indicate MEMs between the forward strands of the two species, and green dots/lines those between the forward strand of *E. coli* (resp. *E. faecium*) and the reverse complement strand of *S. enterica* (resp. *A. minutum*). *a*, *b* — Full genomes of *E. coli* K-12 substr. MG1655 (U00096.3) and *S. enterica* (NC_003198.1), which both belong to the family of *Enterobacteriaceae*. Panel *a* shows all MEMs longer than 25 bp. The sequence similarity and synteny of both genomes, by descent, is evident from the diagonal blue line. Panel *b* only shows MEMs longer than 300 bp. *c*, *d* — Same as panels *a*, *b*, but for the first 1.4 Mbp of *E. faecium* (NZ_CP013009.1) and *A. minutum* (NZ_KB822533.1), which belong to different phyla, showing few MEMs longer than 25 bp (panel *c*). Yet, a single MEM of 19,117 bp is found, as indicated with red ellipses in panels *c*, *d*. The most parsimonious explanation for this long MEM is an event of horizontal gene transfer. Adapted from Ref. [84] with permission (color online)

comparable to Fig. 4, *a*, but now comparing the genomes of distant *Enterococcus faecium* and *Atopobium minulum*. No diagonal line is present because these genomes belong to different phyla and therefore have low sequence identity. Nevertheless, an exact match spanning 19,117 bp is found (diagonal green line highlighted by the red ellipse). The most parsimonious explanation for such a long match is a recent HGT event. In addition, the GC content of the match (55%) deviates strongly from that of both genomes (38.3% and 48.9%, respectively), another indication that this sequence originates from HGT [111]. Alignment of this exact match with all non-redundant GenBank CDS translations using `blastx` [14], one finds very strong hits to VanB-type vancomycin resistance histidine, antirestriction protein (ArdA endonuclease), and an LtrC-family phage protein that is found in a large group of phages that infect Gram-positive bacteria [115]. Together, this suggests that the sequence was transferred by transduction and established in both bacteria aided by natural selection acting on the conferred vancomycin resistance. More general enrichment/depletion analysis (in Ref. [84]) of genes located along such long MEMs in different bacteria pairs indicated which groups of genes are more/less prone to HGT (see also [116] for comparison with other studies).

In addition to HGT events identification, long MEMs allow to estimate pairwise rate of HGT between two taxa using a simple model [84]. The model assumes that HGT acts as a continuous process on the evolutionary time-scale with rate (per bp) ρ between two taxa. Then the evolutionary time divergence between transferred genomic loci in the two taxa is uniformly distributed: $P(\tau) \sim \rho$. Using this and Eq. (3) one obtains that the distribution of MEMs lengths r between two taxa along horizontally transferred loci is

$$m(r) = \frac{A}{r^3}, \tag{10}$$

where the prefactor $A \propto \rho/\mu$ (μ is the mutation rate). Maybe not so surprisingly, if you have read through the review to this point, the empirical distributions of MEMs length nicely validate this prediction, as on can see in Fig. 5, *a*. Fitting prefactor A for different pairs of taxa, one

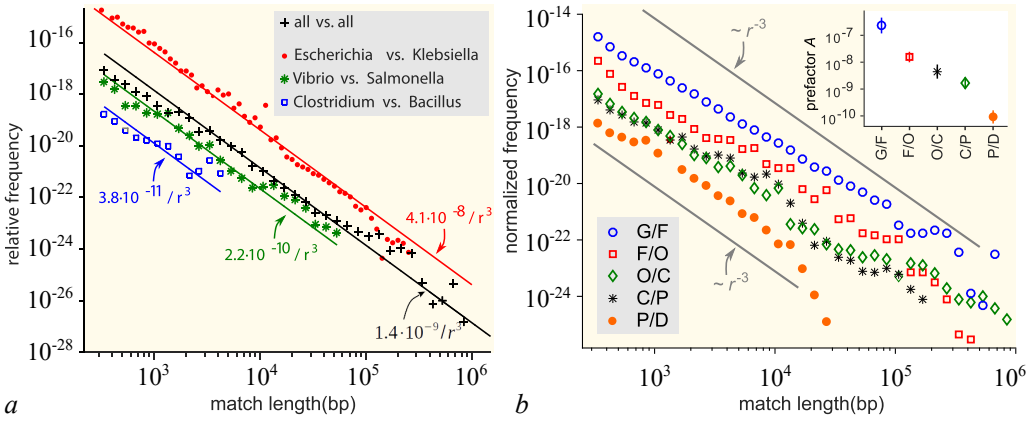

Fig. 5. *a* — MEMs length distributions in pairs of genomes from different genera (see legend). Each distribution is normalised to account for differences in the number of available genomes in each genus. Only the tails of the distributions (length $r \geqslant 300$) are shown. Solid lines are fits of power-laws with exponent $\alpha = -3$ Eq. (10) with just a single free parameter A. *b* — MEMs lengths resulting from comparison of all pairs of genera at a given taxonomic distance. G/F (blue circles): pairs of genera that belong to the same family. F/O (red squares): pairs of genera that belong to the same order, but to different families. O/C (green diamonds): Pairs of genera from the same class, but different orders. C/P (black stars): Same phylum, different classes. P/D (red circles): Same domain, different phyla. Grey lines indicate power-laws $m \sim r^{-3}$ for comparison. Inset: Fitted prefactor A for each of the distributions in the main figure. The prefactor decreases by orders of magnitude as the taxonomic distance increases. Adapted from Ref. [84] with permission (color online)

can estimate HGT rate between these pairs. Using this approach one can also average HGT rate for different groups of taxa. For instance, as shown in Fig. 5, *b*, the average HGT rate strongly depends on the taxonomic distance and differs by 3 orders of magnitude for same-family genera, compared to bacteria pairs from different phyla.

Summary

In this article different ways to exploit exact sequence matches between genomic sequences were reviewed. It was shown that one can use exact matches (i) as starting points of alignments, (ii) to infer evolutionary relations of homologous sequences, (iii) to classify genomes and (iv) to validate and fit models of evolution of genomic sequences. Exact matches are clearly defined and can be relatively easily obtained from empirical data without setting more or less ambiguous parameters. This review illustrated application of exact matches in different fields of genomic studies.

References

1. Reeck G, de Haën C, Teller D, Doolittle R, Fitch W, Dickerson R, Chambon P, McLachlan A, Margoliash E, Jukes T, Zuckerkandl E. "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. Cell. 1987;50(5):667. DOI: 10.1016/0092-8674(87)90322-9.
2. Koonin E. Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics. 2005;39:309–338. DOI: 10.1146/annurev.genet.39.073003.114725.
3. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–770. DOI: 10.1093/bioinformatics/btr011.
4. Claverie J, Sauvaget I, Bougueleret L. Computer generation and statistical analysis of a data bank of protein sequences translated from GenBank. Biochimie. 1985;67(5):437–443. DOI: 10.1016/s0300-9084(85)80261-3.
5. Karlin S, Mrázek J. Compositional differences within and between eukaryotic genomes. Proceedings of the National Academy of Sciences. 1997;94(19):10227–10232. DOI: 10.1073/pnas.94.19.10227.
6. Mahajan S, Agashe D. Evolutionary jumps in bacterial GC content. G3 Genes|Genomes|Genetics. 2022;12(8):jkac108. DOI: 10.1093/g3journal/jkac108.
7. Karlin S. Global dinucleotide signatures and analysis of genomic heterogeneity. Current Opinion in Microbiology. 1998;1(5):598–610. DOI: 10.1016/S1369-5274(98)80095-7.
8. Angeloni A, Bogdanovic O. Sequence determinants, function, and evolution of CpG islands. Biochemical Society Transactions. 2021;49(3):1109–1119. DOI: 10.1042/BST20200695.
9. Parvathy S, Udayasuriyan V, Bhadana V. Codon usage bias. Molecular Biology Reports. 2022;49(1):539–565.
10. Gusfield D. Algorithms on strings, trees, and sequences: Computer science and computational biology. Acm Sigact News. 1997;28(4):41–60.
11. Gibbs A, McIntyre G. The diagram, a method for comparing sequences: Its use with amino acid and nucleotide sequences. European Journal of Biochemistry. 1970;16(1):1–11. DOI: 10.1111/j.1432-1033.1970.tb01046.x.
12. Karlin S, Altschul S. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences of the United States of America. 1990;87(6):2264–2268. DOI: 10.1073/pnas.87.6.2264.
13. Karlin S, Altschul S. Applications and statistics for multiple high-scoring segments in

molecular sequences. Proceedings of the National Academy of Sciences of the United States of America. 1993;90(12):5873–5877. DOI: 10.1073/pnas.90.12.5873.
14. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403–410. DOI: 10.1016/S0022-2836(05)80360-2.
15. Brown D. A survey of seeding for sequence alignment. In: Bioinformatics Algorithms: Techniques and Applications. New York: Wiley; 2007. P. 117–142. DOI: 10.1002/9780470 253441.ch6.
16. Ebel M, Migliorelli G, Stanke M. Global, highly specific and fast filtering of alignment seeds. BMC Bioinformatics. 2022;23(1):225. DOI: 10.1186/s12859-022-04745-4.
17. Wilbur W, Lipman D. Rapid similarity searches of nucleic acid and protein data banks. Proceedings of the National Academy of Sciences of the United States of America. 1983; 80(3):726–730. DOI: 10.1073/pnas.80.3.726.
18. Lipman D, Pearson W. Rapid and sensitive protein similarity searches. Science. 1985; 227(4693):1435–1441. DOI: 10.1126/science.2983426.
19. Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics. 2002;18(3):440–445.
20. Burkhardt S, Kärkkäinen J. Better filtering with gapped q-grams. Fundamenta Informaticae. 2003;56(1–2):51–70.
21. Schwartz S, Kent W, Smit A, Zhang Z, Baertsch R, Hardison R, Haussler D, Miller W. Human-mouse alignments with BLASTZ. Genome Research. 2003;13(1):103–107. DOI: 10.1101/gr.809403.
22. Noé L, Kucherov G. Improved hit criteria for DNA local alignment. BMC Bioinformatics. 2004;5:149. DOI: 10.1186/1471-2105-5-149.
23. Buchfink B, Xie C, Huson D. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2015;12(1):59–60. DOI: 10.1038/nmeth.3176.
24. Harris R. Improved Pairwise Alignment of Genomic DNA. University Park, PA United States: The Pennsylvania State University; 2007. 84 p.
25. Morgulis A, Coulouris G, Raytselis Y, Madden T, Agarwala R, Schäffer A. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24(16):1757–1764. DOI: 10.1093/bioinformatics/btn322.
26. Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal P, Yang H, Xue V, Knyazev S, Singer B, Balliu B, Koslicki D, Skums P, Zelikovsky A, Alkan C, Mutlu O, Mangul S. Technology dictates algorithms: recent developments in read alignment. Genome Biology. 2021;22(1):249. DOI: 10.1186/s13059-021-02443-7.
27. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18): 3094–3100. DOI: 10.1093/bioinformatics/bty191.
28. Edgar R. Syncmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences. PeerJ. 2021;9:e10805. DOI: 10.7717/peerj.10805.
29. Sahlin K. Effective sequence similarity detection with strobemers. Genome Research. 2021;31(11):2080–2094. DOI: 10.1101/gr.275648.121.
30. Bray N, Pachter L. MAVID: Constrained ancestral alignment of multiple sequences. Genome Research. 2004;14(4):693–699. DOI: 10.1101/gr.1960404.
31. Choi J, Cho H, Kim S. GAME: A simple and efficient whole genome alignment method using maximal exact match filtering. Computational Biology and Chemistry. 2005;29(3): 244–253. DOI: 10.1016/j.compbiolchem.2005.04.004.
32. Abouelhoda M, Kurtz S, Ohlebusch E. CoCoNUT: an efficient system for the comparison and analysis of genomes. BMC Bioinformatics. 2008;9:476. DOI: 10.1186/1471-2105-9-476.
33. Delcher A, Kasif S, Fleischmann R, Peterson J, White O, Salzberg S. Alignment of whole genomes. Nucleic Acids Research. 1999;27(11):2369–2376. DOI: 10.1093/nar/27.11.2369.

34. Marçais G, Delcher A, Phillippy A, Coston R, Salzberg S, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Computational Biology. 2018;14(1):e1005944. DOI: 10.1371/journal.pcbi.1005944.
35. Weiner P. Linear pattern matching algorithms. In: 14th Annual Symposium on Switching and Automata Theory (SWAT 1973). New York: IEEE; 1973. P. 1–11. DOI: 10.1109/SWAT.1973.13.
36. Ukkonen E. On-line construction of suffix trees. Algorithmica. 1995;14(3):249–260. DOI: 10.1007/BF01206331.
37. Zielezinski A, Vinga S, Almeida J, Karlowski W. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. 2017;18(1):186. DOI: 10.1186/s13059-017-1319-7.
38. Bernard G, Chan C, Chan Y, Chua X, Cong Y, Hogan J, Maetschke S, Ragan M. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Briefings in Bioinformatics. 2019;20(2):426–435. DOI: 10.1093/bib/bbx067.
39. Felsenstein J. Inferring Phylogenies. Oxford University Press; 2004. 580 p.
40. Sheinman M, Massip F, Arndt P. Statistical properties of pairwise distances between leaves on a random yule tree. PLoS ONE. 2015;10(3):e0120206. DOI: 10.1371/journal.pone.0120206.
41. Qi J, Luo H, Hao B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Research. 2004;32(Suppl_2):W45–W47. DOI: 10.1093/nar/gkh362.
42. Sims G, Jun S, Wu G, Kim S. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(8):2677–2682. DOI: 10.1073/pnas.0813249106.
43. Reinert G, Chew D, Sun F, Waterman M. Alignment-free sequence comparison (I): Statistics and power. Journal of Computational Biology. 2009;16(12):1615–1634. DOI: 10.1089/cmb.2009.0198.
44. Wan L, Reinert G, Sun F, Waterman M. Alignment-free sequence comparison (II): Theore tical power of comparison statistics. Journal of Computational Biology. 2010;17(11):1467–1490. DOI: 10.1089/cmb.2010.0056.
45. Song K, Ren J, Zhai Z, Liu X, Deng M, Sun F. Alignment-free sequence comparison based on next-generation sequencing reads. Journal of Computational Biology. 2013;20(2):64–79. DOI: 10.1089/cmb.2012.0228.
46. Aflitos S, Severing E, Sanchez-Perez G, Peters S, de Jong H, de Ridder D. Cnidaria: fast, reference-free clustering of raw and assembled genome and transcriptome NGS data. BMC Bioinformatics. 2015;16(1):352. DOI: 10.1186/s12859-015-0806-7.
47. Fan H, Ives A, Surget-Groba Y, Cannon C. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics. 2015;16(1):522. DOI: 10.1186/s12864-015-1647-5.
48. Bromberg R, Grishin N, Otwinowski Z. Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. PLoS Computational Biology. 2016;12(6):e1004985. DOI: 10.1371/journal.pcbi.1004985.
49. Ulitsky I, Burstein D, Tuller T, Chor B. The average common substring approach to phylogenomic reconstruction. Journal of Computational Biology. 2006;13(2):336–350. DOI: 10.1089/cmb.2006.13.336.
50. Leimeister C, Morgenstern B. kmacs: the k -mismatch average common substring approach to alignment-free sequence comparison. Bioinformatics. 2014;30(14):2000–2008. DOI: 10.1093/bioinformatics/btu331.
51. Horwege S, Lindner S, Boden M, Hatje K, Kollmar M, Leimeister C, Morgenstern B. Spaced words and kmacs: fast alignment-free sequence comparison based on inexact word matches. Nucleic Acids Research. 2014;42(W1):W7–W11. DOI: 10.1093/nar/gku398.
52. Haubold B, Pfaffelhuber P, Domazet-Loso M, Wiehe T. Estimating mutation distances from

unaligned genomes. Journal of Computational Biology. 2009;16(10):1487–1500. DOI: 10.1089/cmb.2009.0106.
53. Morgenstern B, Schöbel S, Leimeister C. Phylogeny reconstruction based on the length distribution of k-mismatch common substrings. Algorithms for Molecular Biology. 2017; 12(1):27. DOI: 10.1186/s13015-017-0118-8.
54. Břinda K, Sykulski M, Kucherov G. Spaced seeds improve k-mer-based metagenomic classification. Bioinformatics. 2015;31(22):3584–3592. DOI: 10.1093/bioinformatics/btv419.
55. Ondov B, Treangen T, Melsted P, Mallonee A, Bergman N, Koren S, Phillippy A. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016;17(1): 132. DOI: 10.1186/s13059-016-0997-x.
56. Lu J, Breitwieser F, Thielen P, Salzberg S. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science. 2017;3(1):e104. DOI: 10.7717/peerj-cs.104.
57. Linard B, Swenson K, Pardi F. Rapid alignment-free phylogenetic identification of metagenomic sequences. Bioinformatics. 2019;35(18):3303–3312. DOI: 10.1093/bioinformatics/btz068.
58. Hosseini M, Pratas D, Morgenstern B, Pinho A. Smash++: an alignment-free and memory-efficient tool to find genomic rearrangements. GigaScience. 2020;9(5):giaa048.
59. Navarro-Gomez D, Leipzig J, Shen L, Lott M, Stassen A, Wallace D, Wiggs J, Falk M, van Oven M, Gai X. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. Bioinformatics. 2015;31(8):1310–1312. DOI: 10.1093/bioinformatics/btu825.
60. Lees J, Harris S, Tonkin-Hill G, Gladstone R, Lo S, Weiser J, Corander J, Bentley S, Croucher N. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Research. 2019;29(2):304–316. DOI: 10.1101/gr.241455.118.
61. Břinda K, Callendrello A, Cowley L, Charalampous T, Lee R, MacFadden D, Kucherov G, O'Grady J, Baym M, Hanage W. Lineage calling can identify antibiotic resistant clones within minutes. bioRxiv. 2018;(403204).
62. Zhang Q, Jun S, Leuze M, Ussery D, Nookaew I. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. Scientific Reports. 2017;7(1):40712. DOI: 10.1038/srep40712.
63. Ahlgren N, Ren J, Lu Y, Fuhrman J, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Research. 2017;45(1):39–53. DOI: 10.1093/nar/gkw1002.
64. Consortium IHGS. Initial sequencing and analysis of the human genome. Nature. 2001; 409(6822):860–921. DOI: 10.1038/35057062.
65. Peng C, Buldyrev S, Goldberger A, Havlin S, Sciortino F, Simons M, Stanley H. Long-range correlations in nucleotide sequences. Nature. 1992;356(6365):168–170. DOI: 10.1038/356168a0.
66. Price A, Jones N, Pevzner P. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl_1):i351–i358. DOI: 10.1093/bioinformatics/bti1018.
67. Estoup J. Gammes sténographiques: méthode et exercices pour l'acquisition de la vitesse. Institut Sténographique; 1916.
68. Zipf G. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, Mass.: Addison-Wesley Press; 1949. 573 p.
69. Newman M. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 2005;46(5):323–351. DOI: 10.1080/00107510500052444.
70. Mantegna R, Buldyrev S, Goldberger A, Havlin S, Peng C, Simons M, Stanley H. Linguistic features of noncoding DNA sequences. Physical Review Letters. 1994;73(23):3169–3172. DOI: 10.1103/PhysRevLett.73.3169.
71. Gimona M. Protein linguistics – a grammar for modular protein assembly? Nature Reviews Molecular Cell Biology. 2006;7(1):68–73. DOI: 10.1038/nrm1785.

72. Loose C, Jensen K, Rigoutsos I, Stephanopoulos G. A linguistic model for the rational design of antimicrobial peptides. Nature. 2006;443(7113):867–869. DOI: 10.1038/nature05233.
73. Csűrös M, Noé L, Kucherov G. Reconsidering the significance of genomic word frequencies. Trends in Genetics. 2007;23(11):543–546. DOI: 10.1016/j.tig.2007.07.008.
74. Sindi S, Hunt B, Yorke J. Duplication count distributions in DNA sequences. Physical Review E. 2008;78(6):061912. DOI: 10.1103/PhysRevE.78.061912.
75. Sheinman M, Ramisch A, Massip F, Arndt P. Evolutionary dynamics of selfish DNA explains the abundance distribution of genomic subsequences. Scientific Reports. 2016;6(1): 30851. DOI: 10.1038/srep30851.
76. Schmid C. Alu: structure, origin, evolution, significance, and function of one-tenth of human DNA. Progress in Nucleic Acid Research and Molecular Biology. 1996;53:283–319. DOI: 10.1016/S0079-6603(08)60148-8.
77. Austen J. Pride and Prejudice. Whitehall, London: T. Egerton; 1813. 279 p.
78. Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. eLife. 2021;10:e65366. DOI: 10.7554/eLife.65366.
79. Dixit P, Pang T, Studier F, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(29):9070–9075. DOI: 10.1073/pnas.1510839112.
80. Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genetics. 2013;9(6):e1003521. DOI: 10.1371/journal.pgen.1003521.
81. Massip F, Arndt P. Neutral evolution of duplicated DNA: An evolutionary stick-breaking process causes scale-invariant behavior. Physical Review Letters. 2013;110(14):148101. DOI: 10.1103/PhysRevLett.110.148101.
82. Massip F, Sheinman M, Schbath S, Arndt P. How evolution of genomes is reflected in exact DNA sequence match statistics. Molecular Biology and Evolution. 2015;32(2):524–535. DOI: 10.1093/molbev/msu313.
83. Massip F, Sheinman M, Schbath S, Arndt P. Comparing the statistical fate of paralogous and orthologous sequences. Genetics. 2016;204(2):475–482. DOI: 10.1534/genetics.116.193912.
84. Sheinman M, Arkhipova K, Arndt P, Dutilh B, Hermsen R, Massip F. Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. eLife. 2021;10:e62719. DOI: 10.7554/eLife.62719.
85. Sheinman M, Arndt P, Massip F. Modeling the mosaic structure of bacterial genomes to infer their evolutionary history. bioRxiv. 2023;(2023.09.22.558938). DOI: 10.1101/2023.09.22.558938.
86. Arndt P, Massip F, Sheinman M. An analytical derivation of the distribution of distances between heterozygous sites in diploid species to efficiently infer demographic history. bioRxiv. 2023;(2023.09.20.558510). DOI: 10.1101/2023.09.20.558510.
87. Arndt P. Sequential and continuous time stick-breaking. Journal of Statistical Mechanics: Theory and Experiment. 2019;2019(6):064003. DOI: 10.1088/1742-5468/ab1dd8.
88. Ziff R, McGrady E. The kinetics of cluster fragmentation and depolymerisation. Journal of Physics A: Mathematical and General. 1985;18(15):3027–3037. DOI: 10.1088/0305-4470/18/15/026.
89. Gao K, Miller J. Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments. PLoS ONE. 2011;6(7):e18464. DOI: 10.1371/journal.pone.0018464.
90. Bailey J, Eichler E. Primate segmental duplications: crucibles of evolution, diversity and disease. Nature Reviews Genetics. 2006;7(7):552–564. DOI: 10.1038/nrg1895.
91. Vanin E. Processed pseudogenes: characteristics and evolution. Annual Review of Genetics. 1985;19:253–272. DOI: 10.1146/annurev.ge.19.120185.001345.

92. Okamura K, Nakai K. Retrotransposition as a source of new promoters. Molecular Biology and Evolution. 2008;25(6):1231–1238. DOI: 10.1093/molbev/msn071.
93. Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nature Reviews Genetics. 2009;10(1):19–31. DOI: 10.1038/nrg2487.
94. Kingman J. The coalescent. Stochastic Processes and their Applications. 1982;13(3):235–248. DOI: 10.1016/0304-4149(82)90011-4.
95. Hein J, Schierup M, Wiuf C. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford: Oxford University Press; 2004. 296 p.
96. Ceballos F, Joshi P, Clark D, Ramsay M, Wilson J. Runs of homozygosity: windows into population history and trait architecture. Nature Reviews Genetics. 2018;19(4):220–234. DOI: 10.1038/nrg.2017.109.
97. Henn B, Cavalli-Sforza L, Feldman M. The great human expansion. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(44):17758–17764. DOI: 10.1073/pnas.1212380109.
98. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475(7357):493–496. DOI: 10.1038/nature10231.
99. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent W, Mattick J, Haussler D. Ultraconserved elements in the human genome. Science. 2004;304(5675):1321–1325. DOI: 10.1126/science.1098119.
100. Snetkova V, Pennacchio L, Visel A, Dickel D. Perfect and imperfect views of ultraconserved sequences. Nature Reviews Genetics. 2022;23(3):182–194. DOI: 10.1038/s41576-021-00424-x.
101. Sturtevant A. Essays on evolution. I. On the effects of selection on mutation rate. The Quarterly Review of Biology. 1937;12(4):464–467.
102. Silander O, Tenaillon O, Chao L. Understanding the evolutionary fate of finite populations: The dynamics of mutational effects. PLoS Biology. 2007;5(4):e94. DOI: 10.1371/journal.pbio.0050094.
103. Eyre-Walker A, Keightley P. The distribution of fitness effects of new mutations. Nature Reviews Genetics. 2007;8(8):610–618. DOI: 10.1038/nrg2146.
104. Gao K, Miller J. Human–chimpanzee alignment: Ortholog exponentials and paralog power laws. Computational Biology and Chemistry. 2014;53:59–70. DOI: 10.1016/j.compbiolchem.2014.08.010.
105. Boto L. Horizontal gene transfer in evolution: facts and challenges. Proceedings of the Royal Society B: Biological Sciences. 2010;277(1683):819–827. DOI: 10.1098/rspb.2009.1679.
106. Puigbò P, Lobkovsky A, Kristensen D, Wolf Y, Koonin E. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biology. 2014;12:66. DOI: 10.1186/s12915-014-0066-4.
107. Soucy S, Huang J, Gogarten J. Horizontal gene transfer: building the web of life. Nature Reviews Genetics. 2015;16(8):472–482. DOI: 10.1038/nrg3962.
108. Van Etten J, Bhattacharya D. Horizontal gene transfer in eukaryotes: Not if, but how much? Trends in Genetics. 2020;36(12):915–925. DOI: 10.1016/j.tig.2020.08.006.
109. Boucher Y, Cordero O, Takemura A, Hunt D, Schliep K, Bapteste E, Lopez P, Tarr C, Polz M. Local mobile gene pools rapidly cross species boundaries to create endemicity within global Vibrio cholerae populations. mBio. 2011;2(2):e00335–10. DOI: 10.1128/mBio.00335-10.
110. Freeman V. Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. Journal of Bacteriology. 1951;61(6):675–688. DOI: 10.1128/jb.61.6.675-688.1951.
111. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. PLoS Computational Biology. 2015;11(5):e1004095. DOI: 10.1371/journal.pcbi.1004095.
112. Smillie C, Smith M, Friedman J, Cordero O, David L, Alm E. Ecology drives a global

network of gene exchange connecting the human microbiome. Nature. 2011;480(7376): 241–244. DOI: 10.1038/nature10571.
113. Groussin M, Poyet M, Sistiaga A, Kearney S, Moniz K, Noel M, Hooker J, Gibbons S, Segurel L, Froment A, Mohamed R, Fezeu A, Juimo V, Lafosse S, Tabe F, Girard C, Iqaluk D, Nguyen LTT, Shapiro BJ, Lehtimäki J, Ruokolainen L, Kettunen PP, Vatanen T, Sigwazi S, Mabulla A, Domínguez-Rodrigo M, Nartey YA, Agyei-Nkansah A, Duah A, Awuku YA, Valles KA, Asibey SO, Afihene MY, Roberts LR, Plymoth A, Onyekwere CA, Summons RE, Xavier RJ, Alm EJ. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell. 2021;184(8):2053–2067.e18. DOI: 10.1016/j.cell.2021.02.052.
114. Darby C, Stolzer M, Ropp P, Barker D, Durand D. Xenolog classification. Bioinformatics. 2017;33(5):640–649. DOI: 10.1093/bioinformatics/btw686.
115. Quiles-Puchalt N, Tormo-Más MA, Campoy S, Toledo-Arana A, Monedero V, Lasa I, Novick R, Christie G, Penadés J. A super-family of transcriptional activators regulates bacteriophage packaging and lysis in Gram-positive bacteria. Nucleic Acids Research. 2013;41(15):7260–7275. DOI: 10.1093/nar/gkt508.
116. Dmitrijeva M, Tackmann J, Rodrigues J, Huerta-Cepas J, Coelho L, von Mering C. A global survey of eco-evolutionary pressures acting on horizontal gene transfer. Research Square. 2023:25. DOI: 10.21203/rs.3.rs-3062985/v1.