A study of *k*-mer patterns within DNA molecules sequenced from a chondrite

Carmel Farage[1], Ido Bachelet[1,2]

1. Augmanity, Rehovot, Israel.
2. Corresponding author. Email for correspondence: dogbach@gmail.com

Abstract

The known tree of life occupies an infinitesimal region of the space of all mathematically possible evolutionary histories, yet our methods implicitly assume that any genetic material we encounter must have originated from somewhere within it, potentially rendering truly foreign genetic systems invisible to detection. Here we studied *k*-mer patterns in DNA molecules sequenced from internal samples within the Zag chondrite in order to associate these sequences to a known taxonomic neighborhood within the tree of life. In an attempt to capture every conceivable pattern, we used the distributions of *k*-mers across the BLAST NT database, in novel genomes, in various control groups including random sequences, and in long sequences of DNA composed entirely of unrepresented *k*-mers (primes) emulating foreign material. Our analysis showed that all >1,000,000 *k*-mers (k = 21) comprising the Zag sequences are primes, with 0 represented in the NT database. Taken together, these results highlight a potential methodological limitation in current frameworks for interpreting newly found genetic sequences, which are implicitly calibrated to the known tree of life.

Introduction

It is generally assumed that the currently known tree of life (ToL) accounts for only a small fraction of all living organisms on earth[1]. Estimates of the actual terrestrial biodiversity vary widely, but even under conservative assumptions, e.g. that the known fraction represents 0.1-1.0% of extant life, the complete ToL would still occupy a negligible fraction of the space of possible DNA sequences with a comparable length distribution. The space of all DNA sequences of length ~ 100 bases contains ~ 10^{60} distinct sequences, whereas even a very conservative estimate of the total number of nucleotides represented in a complete ToL is on the order of 10^{17}-10^{18}. While biological function and evolutionary history strongly constrain which regions of sequence space are accessible, this disparity in scale implies that, in principle, the sequence space could contain an enormous number of independent evolutionary histories - trees of life occupying largely disjoint regions of sequence space and sharing no common ancestry.

New sequences are usually compared to the known database in search for novel genetic material using tools such as BLAST[2] and Kraken2[3], with mapping below a certain threshold indicating that the material is indeed novel. However, it is crucial to recognize that the validity of this practice is predicated upon the assumption that any newly discovered sequence must originate from the ToL i.e. from some organism within it, known or yet unknown. This assumption is arguably scientifically rigorous, but it might also bias our conclusions in the face of hypothetical DNA or RNA originating from a system that is foreign to the one we know.

Much of the newly found genetic material is regarded as novel according to current methods for measuring novelty. This state of matters is suitable for the taxonomic project of filling the blanks in the ToL. However, in terms of detecting foreignness, it has little value as even very novel genetic material would be considered a yet-unknown terrestrial organism or a variant of existing content. Even the most distant branches of the ToL appear to share more similarity than difference[4–9].

This study stemmed in part from the challenge of defining and measuring sequence foreignness in a way that would enable us to see potential domains of biology that may, perhaps, be overlooked. We focused on *k*-mers as components of a layer that is more fundamental to that of genetic meaning, and indeed forms a rudimentary signature that is specific to clades across the ToL. We hypothesized that this insight could provide valuable clues as to the relationships between found material and known material and have been attempting to apply it to the study of new samples, which current methods may be inadequate to analyze.

We recently completed a collaborative effort that resulted in a database of DNA molecules sequenced from internal samples from two masses of the Zag chondrite[10]. This is a fragment of a H-type ordinary chondrite that fell on earth in 1998, which has been reported to contain liquid brine within microscopic halite crystals[11,12], to have formed approximately 4.7 Billion years ago, and to contain complex organic molecules[11,13]. A stringent, multi-method filtering of sequences that mapped to known material highlighted a fraction (approximately 9,000 reads) of unmapped material, which was the focus of this work. Here we studied patterns of *k*-mer representation in this material to associate the filtered sequences to a known taxonomic neighborhood or a specific clade. As frames of reference, we used the distributions of *k*-mers across the NT database, in novel genomes, and in various control groups. We also assembled long sequences of DNA composed entirely of primes[14,15] as an artificial foreign reference group. Our analysis showed that all >1,000,000 *k*-mers (k = 21) comprising the Zag sequences are primes. It was therefore difficult to definitively associate this material to a known clade.

Results

The meteorite known as Zag is a fragment of a H3-6 chondrite, from a yet-unclear parent body[16,17], that fell in Western Sahara near Zag, Morocco, in August 1998[10]. Analyses of the meteorite revealed that it contains inclusions of liquid water within microscopic halite crystals[18], and that it formed approximately 4.7 Billion years ago based on ^{129}I-^{129}Xe radiometric dating[19,20]. More recently, complex organic molecules have been reported from inside the chondrite[11,13,21], which contains micrometer-sized organic matter aggregates that are rare in other meteorites[13]. Meteorites have long been known to contain basic building blocks of life, such as amino acids[22,23] and nucleobases[24–26]. We were interested in the possibility of identifying more complex nucleic acids in two Zag masses, weighing 18.8 and 1.48 kg **(Fig. 1A, 1B)**.

The sample preparation and sequencing workflow is depicted in **Fig. 1C-1F**. Three samples were collected: (1) from the meteorite surface before decontamination (Surface), (2) from the surface after decontamination (Surface-decontaminated) which removed DNA present on the surface, and (3) from a fresh internal surface of the meteorite (Internal) after fracturing the specimen in a clean room. DNA was extracted from each sample using two different methods, Yang-urea ancient DNA extraction method (Yang), and QIAGEN DNeasy PowerSoil Pro kit (PSP). Two

blunt-end DNA libraries were prepared from each extract. The two libraries were prepared using different levels of DNA amplification – standard and high.

All sequence data were subjected to a multi-stage quality control and subtraction pipeline prior to downstream analysis. Adapter trimming, quality filtering, and removal of human reads resulted in the exclusion of the majority of raw reads, with approximately 89% and 86% of the R1 and R2 pair reads removed at this stage, consistent with substantial background contamination expected in meteoritic samples handled in terrestrial environments. Following de novo deduplication, approximately 5.1% (R1) and 4.2% (R2) of read pairs remained. To operationally define sequences associated with the internal fracture surface of the Zag chondrite, we further subtracted all reads detected in surface samples (pre- and post-decontamination), no-sample controls, and pre-PCR controls. This subtraction was performed independently for standard- and high-amplification libraries, yielding a conservative "internal" read set defined by absence from all control contexts rather than by any positive taxonomic assignment.

Reads in the internal set were first mapped to curated genome collections using ChocoPhlAn, followed by alignment against the BLAST NT database. A substantial fraction of reads mapped broadly to known bacterial clades, indicating that the pipeline successfully detects conventional terrestrial signals when present. However, a small subset of reads remained unmapped after both steps, comprising 9,252 reads in the high-amplification library and 8,083 reads in the standard-amplification library. These unmapped reads were retained for further analysis without assuming novelty of origin, but rather as sequences not readily assignable within current reference frameworks under stringent alignment criteria.

To analyze *k*-mer distribution in this material as a means to associate it with known taxonomy, we first characterized the basal layer of *k*-mers in the ToL using the BLAST NT database as a source. This database is a non-redundant collection of diverse nucleotide sequences from various sources, including GenBank, EMBL, DDBJ, PDB, and RefSeq[27]. The NT database version used in this study was downloaded in December 2023 and at that point consisted of 102,960,590 records.

We calculated *k*-mer representation for k = 12, 14, 16, 18, 20, and 21 in decreasing NT subsample sizes and observed a pattern which is in agreement with previous findings[28] **(Fig. 2A)**. At k = 12, 14, and 16, 100% of the *k*-mers were represented in NT; at k = 20, only 29.24% of all possible *k*-mers ($4^{20} = 1.099\times10^{12}$) were found to be represented in NT; and at k = 21, only 10.04% of all possible *k*-mers ($4^{21} = 4.398\times10^{12}$) were represented in NT. We defined a parameter, RD_{50}, which equals the database size at which 50% of the *k*-mers (for a certain *k*) are represented, and observed a significant correlation between RD_{50} and k **(Fig. 2B)**, suggesting that all 20-mers and 21-mers might be represented once NT reaches a size of ~ 3.62×10^{16} and ~ 1.37×10^{17} bases, respectively. The representation of 20-mers across the ToL exhibited variation between clades, in a manner suggesting that taxonomic association to a specific clade could be possible for newly found material based on its *k*-mer representation **(Fig. 2C)**.

In addition, we generated various reference groups and measured their representations in NT: (1) a sample of 1,000 sequences from NT, which by definition exhibits 100% representation; (2) a sample of 1,000 randomly-generated sequences, showing average representation of 29.37%, which is approximately equal to the total fraction of 20-mers represented in NT (29.24%); (3) a previously-reported bacterial genome from a potentially new phylum[29], representing novel genetic material, showing average representation of 52.84%. (4) an artificial "foreign" sequence

composed solely from unrepresented 20-mers such that every newly created 20-mer is also unrepresented; this sequence thus exhibited 0% representation and did not map to NT. (5) the foreign sequence was reshuffled to create a fifth reference group, which showed 20% representation.

As mentioned above, we were finally left with 8,667 ± 584 sequences ranging from 75 to 150 bases in length, which assembled into 26.5 ± 0.5 contigs ranging from 78 to 334.5 ± 64.5 bases in length. These sequences did not map to known ones by any tool used and appeared only in the internal sample and not in the surface or surface-decontaminated samples. These sequences contained 1,039,632.5 ± 68,191.5 overlapping *k*-mers ($k = 21$), of which 343,077.5 ± 23,812.5, or approximately 33%, were unique. Analysis of these *k*-mers revealed 100% of them to be primes, i.e. 0% of these *k*-mers were represented in the NT database, which was different than the reshuffled 20-mer prime-constructed group (20%) and the random representations of 21-mers (10.04%). **Figure 3** places this result within a range of estimated representations[30–32] and the empirical ones described above.

Discussion

The continuity of biological research generates a constant flux of new genetic material into the collective database. There seems to be an overarching assumption that all found material, such as DNA and RNA, originates from organisms on known and yet-unknown branches of the terrestrial ToL. While this assumption is not being challenged here, the hypothetical existence of other sources should not be ruled out entirely, and tools should be proposed to identify such material, should it be collected at some point in the future. *k*-mers have been used as the basis for various bioinformatic analysis methods and tools as they enable improvement in computational efficiency and speed[33]. We have been previously applying *k*-mer analysis to DNA and RNA obtained from chemically divergent and biologically extreme environments.

In this work we report *k*-mer analysis of DNA molecules sequenced from an internal sample of the Zag chondrite, which was obtained as a result of a collaborative, open-ended exploratory effort. We report a small fraction of unmapped reads which appears to be composed only of primes of $k = 21$. This pattern did not converge with several reference groups we constructed for the purposes of this study. It was therefore difficult to definitively associate this material to a known clade, which was our original purpose in this analysis.

The choice of *k* is of significant importance as it defines the thresholds for proximity and foreignness. As we show, all *k*-mers up to 16 are already known in the database. 17 is the first value of *k* to date at which some *k*-mers are primes. We used *k* of 20 and 21 for specific reasons. 20 is the value that created a sufficiently large pool of primes from which to build "foreign" reference sequences. On the other hand, evidence suggests that *k* of 21 is optimal for comparison across genomes as it balances *k*-mer sharing and specificity[28,34,35], however 20 and 21 are considered similar in terms of sensitivity and precision[34].

A value for *k* could be hinted at from a different direction: what is the minimal length of biological meaning? Short, non-coding RNAs. e.g. miRNAs, which are around 20-22 nt in length, are usually considered to mark the short end of the spectrum[36,37]. Shorter RNA fragments have also been reported[38–41], but the shortest ones (below 17 nt) may still correspond to degradation products. Very short RNAs are known to be functional, for example the 5-nt

minuscule ribozyme[42,43], but these are arguably chemical entities that are not encoded as biological messages. Altogether it is interesting that, according to present understanding, 20- and 21-mers may represent (or be proximal to) a boundary between biological and chemical meaning, which is intricately linked to the question of the potential origin of found material.

As suggested above, hypothetical origins of nucleic acids outside the ToL should not be ruled out completely, with one such source being abiotic synthesis. Although the chemistry of complete abiotic synthesis of nucleic acids is not well understood and has not been recreated to date, it is hypothesized to have occurred on the primitive earth, although the nature of the first nucleic acid species is also unclear. Simple precursor molecules for chemical nucleic acid synthesis are abundant in the universe, and the chemistry for their generation has been described and extensively studied in the past 70 years. The synthesis of ribose, one of the major gaps in our understanding of abiotic nucleic acid synthesis, has been suggested to occur from formaldehyde in the presence of borate and silicate minerals[44–47]; and, while boron in relatively rare in the universe (abundance is approximately 6 orders of magnitude lower than that of carbon and oxygen, and 10 orders of magnitude lower than that of hydrogen), it has been identified in local concentrates in space[48]. Chondrites typically contain boron in the range of 0.1-1 ppm[49,50].

It is not suggested here that nucleic acids are synthesized de-novo inside chondrites: this is unlikely. Rather, we reiterate the hypothesis that there are yet-unknown chemical conditions under which the abiotic synthesis of nucleic acids - regardless of their base and backbone chemistry, chirality, etc. - can proceed forward from simple, abundant precursors, on earth and elsewhere in the present and past. We only suggest that our current sequence analysis tools might be inadequate to identify the products of such a process as abiotic, should such hypothetical products be captured on earth as part of routine biological research.

In conclusion, this work reports *k*-mer patterns observed in DNA molecules sequenced from a chondrite. Since Zag and other chondrite material is available, independent research could and should be carried out in order to confirm or disprove the reported findings. We do not draw any conclusion as to the origin of these molecules. We only wish to highlight a possible methodological limitation which might readily mistake foreign genetic material, should such authentic material be found one day, for terrestrial one.

Methods

Zag material

Meteorite samples used in this study were obtained from Rob Wesel, a recognized dealer documented in the *Meteoritical Bulletin* as having handled material that has been formally classified and described in peer-reviewed meteoritical literature[24].

Processing and sequencing

The meteorite was UV-irradiated (6 J/cm^2 at 254 nm) 40 min on each side before taking it into the cleanroom laboratory. In the cleanroom laboratory, the external surface was carefully and mechanically cleaned using a Dremel tool. The sample was wiped with 0.5% bleach solution and followed by Milli-Q® water, using sterile cotton swabs, and then UV-irradiated again, 30 min on each side. A hammer and chisel were used to chip off small pieces from the UV & bleach-treated meteorite sample, which were then pounded to obtain smaller fragments and powder. A Dremel tool with a diamond cutting disc was used to saw off a piece to expose the internal surface of the meteorite sample. A diamond drill bit was then used to shave off powder from the internal surface, which was collected. Two methods of DNA extraction were used on the samples: the Yang-urea extraction method is a modified silica-based method[51], where 1 M urea was used in place of sodium dodecyl sulfate (SDS) in the extraction buffer. Extraction blanks were included as the negative controls and carried along throughout the whole process. Briefly, the sub-sample was digested with 1 mL of extraction buffer and incubated with rotation at 37 °C for 24 h. Supernatant was collected from the digested sub-sample and concentrated using an Amicon Ultra-4 filter unit (Millipore). The DNA extract was purified with MinElute PCR purification kit (QIAGEN) according to the manufacturer's instructions and eluted in 110 µL EB buffer. The PowerSoil Pro kit (QIAGEN) was also used, according to the manufacturer's instructions and eluted in 100 µL Solution C6.

Double-stranded blunt-end libraries for next-generation sequencing were prepared from 20 µL of DNA extracts according to Meyer and Kircher's protocol[52] with some modifications. MinElute PCR purification kit (QIAGEN) was used to clean up reactions instead of SPRI beads. The extraction blanks and library blanks using water were included as negative controls and carried along throughout the whole process till the quality control step. The library was quantified by real-time PCR (qPCR) in a 25-µL reaction using Maxima SYBR green master mix (Thermo Fisher Scientific), 200 nM of primer IS7 and 200 nM of primer IS8, to determine the amount of indexing cycles, i.e. Cq-value plus 2 cycles. Indexing PCR amplification was performed in a 50-µL reaction using 6 µL of DNA library, 5 units of AmpliTaq Gold DNA polymerase (Thermo Fisher Scientific), 1´ GeneAmp Gold buffer (Thermo Fisher Scientific), 2.5 mM $MgCl_2$, 250 µM of each dNTP, 200 nM of P7-indexing primer and 200 nM of P5-indexing primer, in duplicates. The PCR reaction was performed at 94 °C for 10 min, 24-28 cycles of (94 °C for 30 s, 60 °C for 30 s, 72 °C for 45 s), and 72 °C for 10 min. The duplicates of the DNA library were pooled and purified with AMPure XP beads (Beckman Coulter). The quality of the library was analyzed by Agilent 2200 Tapestation and the Qubit dsDNA HS assay kit (Invitrogen) was used for DNA quantitation.

DNA libraries from the samples were sequenced along with libraries prepared from no-sample controls from the ancient DNA lab (NSC-aDNA) and pre-PCR lab (NSC-prePCR) to determine the level of background DNA. In total, twenty libraries were prepared and sequenced on one NovaSeq SP lane, 2×150 bp. Sequencing of libraries was performed by SciLifeLab SNP & SEQ Technology Platform. Data analysis was performed by the SciLifeLab Ancient DNA facility.

The lab in which this segment of the work took place follows strict cleaning routines, maintenance, and lab work based on ancient DNA guidelines[53,54]. Personnel wear a disposable body suit with hood, a hair net, a mask with visor, shoe covers, and at least 3 layers of gloves at all times. Entry into the clean laboratories is prohibited if personnel have been in a post-PCR laboratory, unless they have showered and changed into a clean set of clothes. The clean laboratories and equipment are frequently decontaminated with bleach, and/or RNase and DNA AWAY™ Surface Decontaminant (Thermo Scientific), and/or UV irradiation.

Sequencing data retrieval

The NT database from BLAST was downloaded from NCBI's FTP site[55] in FASTA format in December 2023. Subsamples of decreasing fractions (0.1, 0.01, 0.001, 0.0001, 0.00001 and 0.000001) were extracted using SeqKit tool (v2.5.1)[56]. The assembly of a newly discovered microbe was kindly provided by Wild Biotech[29]. From this assembly, 1,000 sequences of length 176 bases were extracted for subsequent analysis. Additionally, random and shuffled sequences were generated with a Python script to produce 1,000 sequences of length 176 bases. The random sequences were generated with a probability of 0.25 for each nucleotide, whereas the shuffled sequences maintained the same nucleotide probability distribution as the "distant-nt" sequences.

k-mer counting

All FASTA files were processed with Jellyfish, a tool designed for counting *k*-mers in a sliding window manner. Jellyfish generates output in a binary format, which can be translated into a human-readable format, queried for specific *k*-mers, and used to produce statistics of the counted *k*-mers. The NT FASTA file was used as input to count *k*-mers of lengths 12, 14, 16, 18, and 20. The representation of *k*-mers was calculated as the number of distinct *k*-mers divided by the number of possible *k*-mers of length *k* (4^k).

Assembly of foreign sequences

The sequences were constructed using a Python script in an iterative manner, starting with a sample of unrepresented 20-mers from the NT. The initial sample was obtained by extending unrepresented 17-mers from a previous search (results not reported) to 20-mers. This step produced a pool of approximately 38 million unrepresented 20-mers as a starting point. The building algorithm attempts to elongate each *k*-mer in the pool by appending one nucleotide at a time. It then checks whether the newly generated *k*-mer at the 3' end exists in the pool of unrepresented *k*-mers. If the *k*-mer exists, the algorithm continues with this new *k*-mer and repeats the one-base elongation process until the longest possible sequence is reached. If the *k*-mer does not exist, the elongation stops and reinitiates with the next *k*-mer.

k-mers that are missing from the unrepresented pool are either present in the NT database or absent from the current sample pool. Therefore, when the elongation process stops at a length close to the current longest sequence, the missing *k*-mers are queried in the Jellyfish 20-mer count file. The resulting *k*-mers that are truly missing from the NT database are added to the sample pool, and another search iteration begins.

Phylogenetic analysis

The phylogenetic tree of life was obtained from Nature Microbiology (reference: https://www.nature.com/articles/nmicrobiol201648). Records assigned to each genus on the tree were retrieved from a 10% subsample of the NT FASTA file using the SeqKit tool (v2.5.1), and their *k*-mers were counted using the Jellyfish tool. The visualization of the tree and the log representation of *k*-mers was performed using iTOL (v6.9.1)[57].

Acknowledgements

Processing of DNA and data analysis were performed by the SciLifeLab Ancient DNA facility. Sequencing was performed by the SNP & SEQ Technology Platform in Uppsala, part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP & SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. The authors wish to thank Joe Davis and Prof. George Church for valuable ideas and discussions; Dr. Haider Hassan and the team at Bridge Informatics for their technical expertise, valuable discussions, and work carried out in this project; Dr. David Tolpin for valuable help with Bayesian analysis of the data; and Mr. Marius Nacht for his generosity and vision.

Author contributions

C. F. and I. B. designed the study, analyzed data, and wrote the manuscript.

Declaration of competing interests

The authors declare no competing interests.

References

1. Wiens, J. J. How many species are there on Earth? Progress and problems. *PLoS Biol.* **21**, e3002388 (2023).
2. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
3. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257 (2019).
4. Ledvina, H. E. & Whiteley, A. T. Conservation and similarity of bacterial and eukaryotic innate immunity. *Nat. Rev. Microbiol.* **22**, 420–434 (2024).
5. Takemata, N., Samson, R. Y. & Bell, S. D. Physical and Functional Compartmentalization of Archaeal Chromosomes. *Cell* **179**, 165–179.e18 (2019).
6. Alvarez-Ponce, D. & McInerney, J. O. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. *Genome Biol. Evol.* **3**, 782–790 (2011).
7. Isenbarger, T. A. *et al.* The most conserved genome segments for life detection on Earth and other planets. *Orig. Life Evol. Biosph.* **38**, 517–533 (2008).
8. Moody, E. R. R. *et al.* The nature of the last universal common ancestor and its impact on the early Earth system. *Nat Ecol Evol* (2024) doi:10.1038/s41559-024-02461-1.
9. Ban, N. *et al.* A new system for naming ribosomal proteins. *Curr. Opin. Struct. Biol.* **24**, 165–169 (2014).
10. Grossman, J. n. The Meteoritical Bulletin, No. 83, 1999 July. *Meteorit. Planet. Sci.* **34**, A169–A186 (1999).
11. Chan, Q. H. S. *et al.* Organic matter in extraterrestrial water-bearing salt crystals. *Sci Adv* **4**, eaao3521 (2018).
12. [No title]. https://www.lpi.usra.edu/meetings/metsoc99/pdf/5050.pdf.
13. Kebukawa, Y. *et al.* A novel organic-rich meteoritic clast from the outer solar system. *Sci. Rep.* **9**, 3169 (2019).
14. Hampikian, G. & Andersen, T. Absent sequences: nullomers and primes. *Pac. Symp. Biocomput.* 355–366 (2007).
15. Vergni, D. & Santoni, D. Nullomers and High Order Nullomers in Genomic Sequences. *PLoS One* **11**, e0164540 (2016).
16. Gaffey, M. J. & Gilbert, S. L. Asteroid 6 Hebe: The probable parent body of the H-type ordinary chondrites and the IIE iron meteorites. *Meteorit. Planet. Sci.* **33**, 1281–1295 (1998).
17. [No title]. https://arxiv.org/pdf/1705.10515.
18. [No title]. https://www.lpi.usra.edu/meetings/metsoc99/pdf/5050.pdf.
19. Whitby, J., Burgess, R., Turner, G., Gilmour, J. & Bridges, J. Extinct (129)I in halite from a primitive meteorite: evidence for evaporite formation in the early solar system. *Science* **288**, 1819–1821 (2000).
20. Gilmour, J. D., Pravdivtseva, O. V., Busfield, A. & Hohenberg, C. M. The I‑Xe chronometer and the early solar system. *Meteorit. Planet. Sci.* **41**, 19–31 (2006).
21. Complex mixture of organic matter in a xenolithic clast from the Zag meteorite revealed by coordinated analyses using AFM-IR, NanoSIMS and STXM/XANES. *Icarus* **400**, 115582 (2023).
22. Kvenvolden, K. *et al.* Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. *Nature* **228**, 923–926 (1970).

23. Meierhenrich, U. J., Muñoz Caro, G. M., Bredehöft, J. H., Jessberger, E. K. & Thiemann, W. H.-P. Identification of diamino acids in the Murchison meteorite. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9182–9186 (2004).
24. Callahan, M. P. *et al.* Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13995–13998 (2011).
25. Extraterrestrial nucleobases in the Murchison meteorite. *Earth Planet. Sci. Lett.* **270**, 130–136 (2008).
26. Oba, Y. *et al.* Identifying the wide diversity of extraterrestrial purine and pyrimidine nucleobases in carbonaceous meteorites. *Nat. Commun.* **13**, 2008 (2022).
27. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
28. Bussi, Y., Kapon, R. & Reich, Z. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLoS One* **16**, e0258693 (2021).
29. Levin, D. *et al.* Diversity and functional landscapes in the microbiota of animals in the wild. *Science* **372**, (2021).
30. Monzoorul Haque, M., Ghosh, T. S., Komanduri, D. & Mande, S. S. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* **25**, 1722–1730 (2009).
31. Ghosh, T. S., Monzoorul Haque, M. & Mande, S. S. DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics* **11 Suppl 7**, S14 (2010).
32. Gerlach, W. & Stoye, J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res* **39**, e91 (2011).
33. A survey of k-mer methods and applications in bioinformatics. *Comput. Struct. Biotechnol. J.* **23**, 2289–2303 (2024).
34. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**, 236 (2015).
35. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
36. Lee, R. C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862–864 (2001).
37. Sim, G. *et al.* Determining the defining lengths between mature microRNAs/small interfering RNAs and tinyRNAs. *bioRxiv* (2023) doi:10.1101/2023.10.27.564437.
38. Li, Z. *et al.* Characterization of viral and human RNAs smaller than canonical MicroRNAs. *J. Virol.* **83**, 12751–12758 (2009).
39. Li, Z. *et al.* Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Res.* **40**, 6787–6799 (2012).
40. Baldrich, P. *et al.* Plant Extracellular Vesicles Contain Diverse Small RNA Species and Are Enriched in 10- to 17-Nucleotide 'Tiny' RNAs. *Plant Cell* **31**, 315–324 (2019).
41. Sim, G. *et al.* Manganese-dependent microRNA trimming by 3'→5' exonucleases generates 14-nucleotide or shorter tiny RNAs. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2214335119 (2022).
42. Turk, R. M., Chumachenko, N. V. & Yarus, M. Multiple translational products from a five-nucleotide ribozyme. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 4585–4589 (2010).

43. Yarus, M. The meaning of a minuscule ribozyme. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **366**, 2902–2909 (2011).
44. Lambert, J. B., Gurusamy-Thangavelu, S. A. & Ma, K. The silicate-mediated formose reaction: bottom-up synthesis of sugar silicates. *Science* **327**, 984–986 (2010).
45. Furukawa, Y., Horiuchi, M. & Kakegawa, T. Selective stabilization of ribose by borate. *Orig. Life Evol. Biosph.* **43**, 353–361 (2013).
46. Ricardo, A., Carrigan, M. A., Olcott, A. N. & Benner, S. A. Borate minerals stabilize ribose. *Science* **303**, 196 (2004).
47. Hirakawa, Y., Kakegawa, T. & Furukawa, Y. Borate-guided ribose phosphorylation for prebiotic nucleotide synthesis. *Sci. Rep.* **12**, 1–7 (2022).
48. Stephenson, J. D., Hallis, L. J., Nagashima, K. & Freeland, S. J. Boron enrichment in martian clay. *PLoS One* **8**, e64624 (2013).
49. Mason, B. H. *Handbook of Elemental Abundances in Meteorites*. (Gordon & Breach Publishing Group, 1971).
50. Boron in chondritic meteorites. *Geochim. Cosmochim. Acta* **52**, 2311–2319 (1988).
51. Yang, D. Y., Eng, B., Waye, J. S., Dudar, J. C. & Saunders, S. R. Technical note: improved DNA extraction from ancient bones using silica-based spin columns. *Am J Phys Anthropol* **105**, 539–543 (1998).
52. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**, db.prot5448 (2010).
53. Fulton, T. L. & Shapiro, B. Setting Up an Ancient DNA Laboratory. *Methods Mol Biol* **1963**, 1–13 (2019).
54. Shapiro, B. A. & Hofreiter, M. *Ancient DNA: Methods and Protocols*. (2012).
55. Index of /blast/db/FASTA. https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA.
56. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
57. Website. https://academic.oup.com/bioinformatics/article/23/1/127/188940.

Legends

Figure 1.

A-B, photographs of Zag meteorite fragments used in this study. B, photograph taken during processing at SciLifeLab. C-D, schematic diagram of the sample preparation and sequencing workflow. NSC, non-specific control. aDNA, ancient DNA. E, distribution of sequences following analysis. F, sequence length distribution.

Figure 2

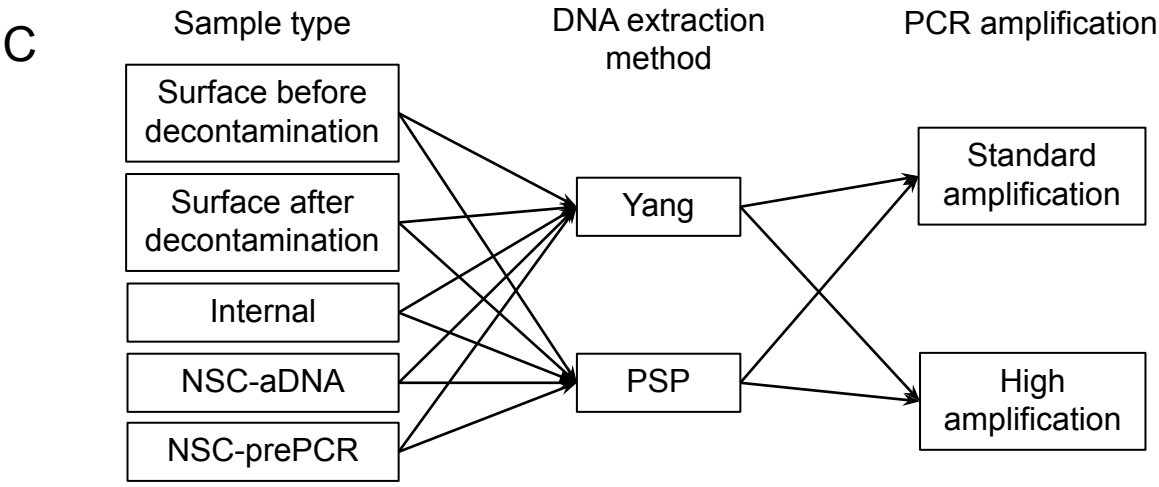
A, *k*-mer representations in the NT database. B, observed correlation between RD_{50}, defined as the database size at which 50% of the *k*-mers (for a certain *k*) are represented, and the value of *k*. C, the tree of life (ToL) with outer rings depicting *k* mer representation distribution for *k* of 20 and 21 (outer rings, color coded by log representation; color keys on the right).

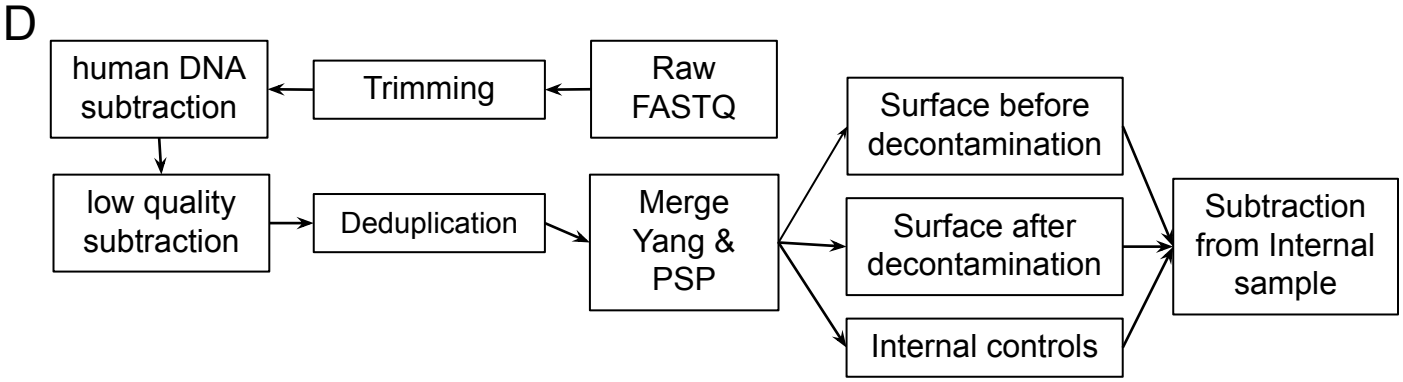
Figure 3

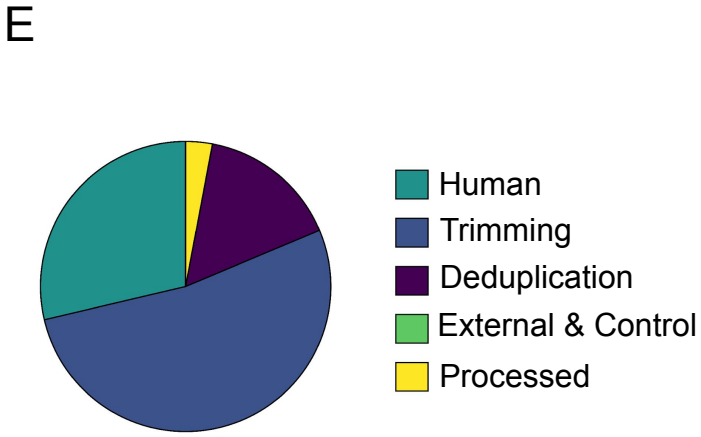
A range of estimated and empirical representations of putative new samples in the NT database, based on BLAST of 1,000 random sequences from the new sample. A new genome from a different phylum is based on Levin et al[29]. The representation of Zag is denoted by a star.

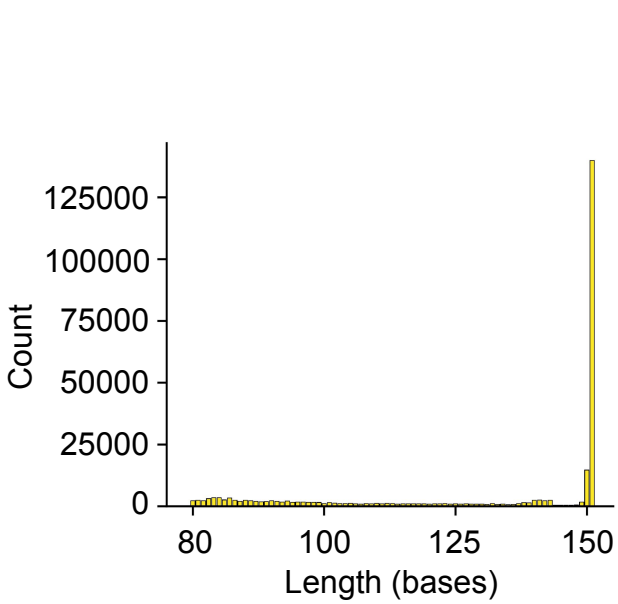
A

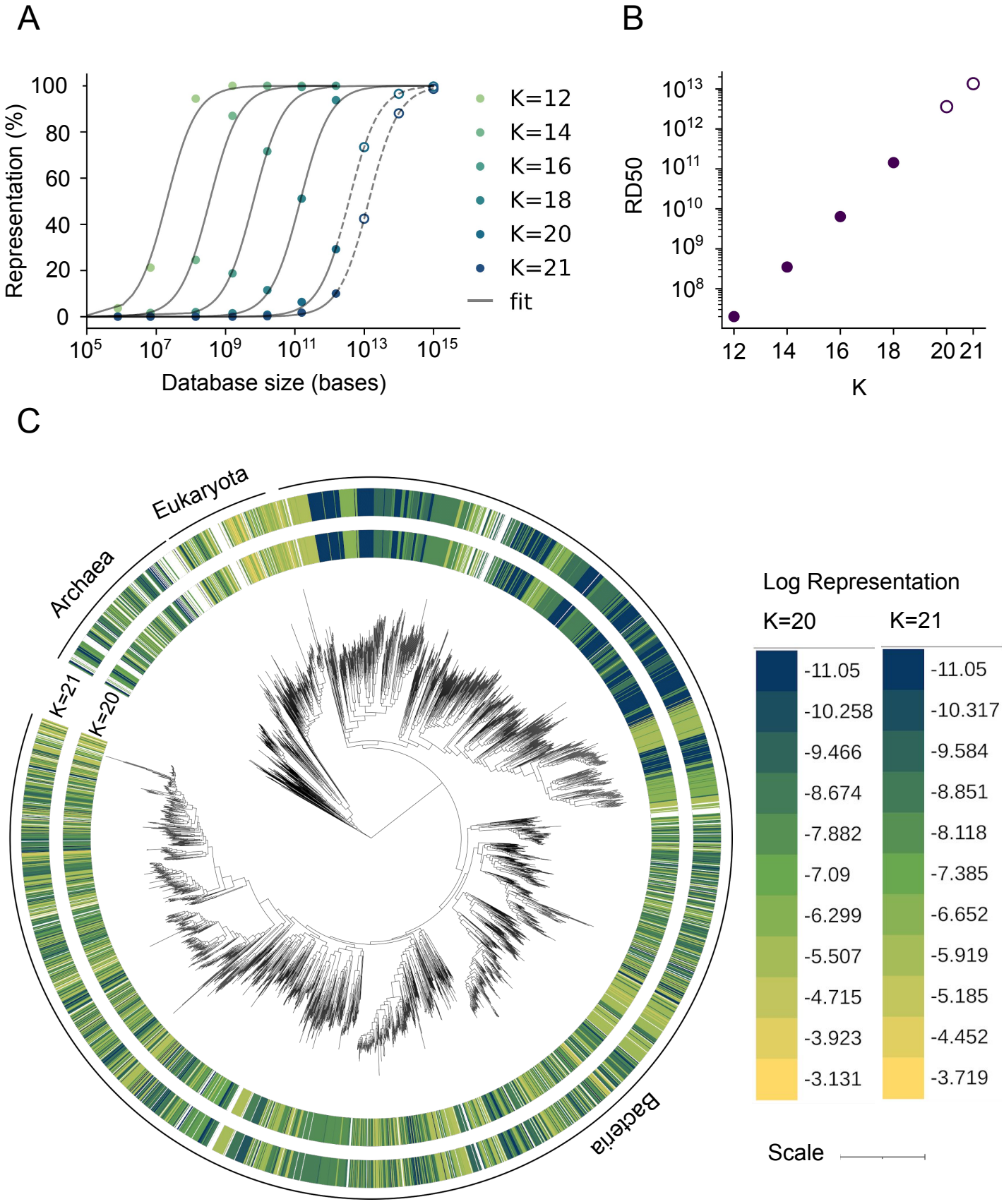
B

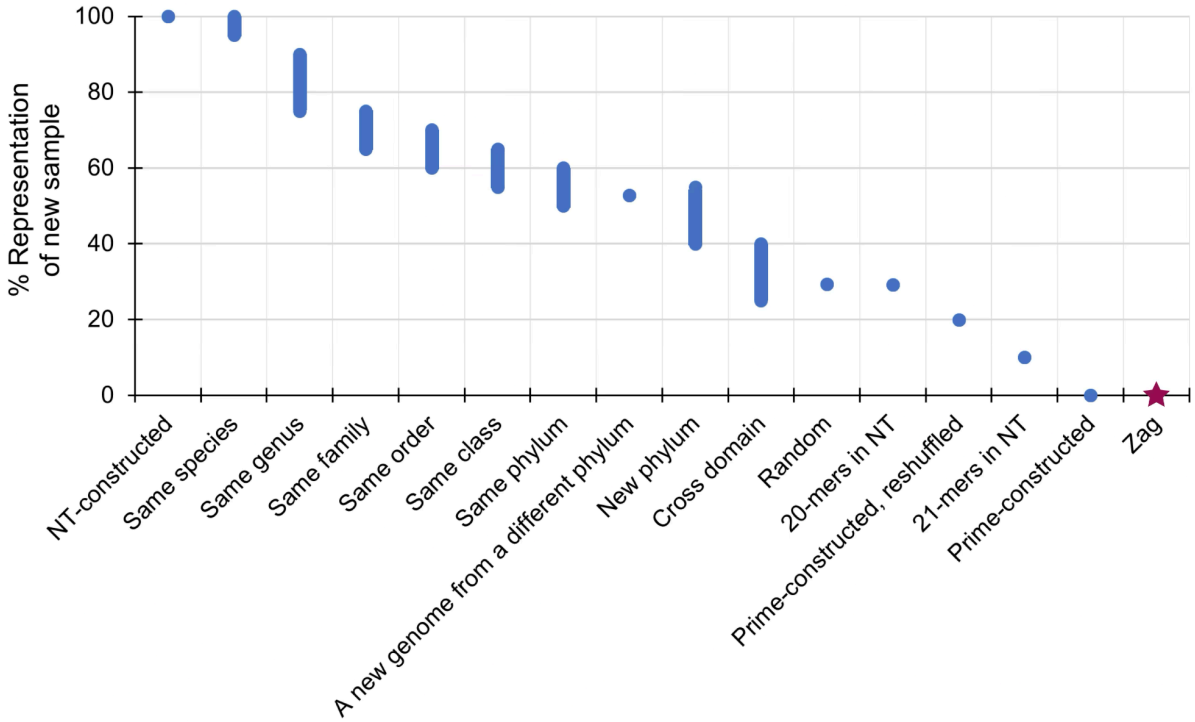

C
Sample type
DNA extraction method
PCR amplification
Surface before decontamination
Surface after decontamination
Internal
NSC-aDNA
NSC-prePCR
Yang
PSP
Standard amplification
High amplification


D
human DNA subtraction
Trimming
Raw FASTQ
low quality subtraction
Deduplication
Merge Yang & PSP
Surface before decontamination
Surface after decontamination
Internal controls
Subtraction from Internal sample


E
Human
Trimming
Deduplication
External & Control
Processed


F
Count
125000
100000
75000
50000
25000
0
80
100
125
150
Length (bases)


A
Representation (%)
100
80
60
40
20
0
10^5
10^7
10^9
10^{11}
10^{13}
10^{15}
Database size (bases)
K=12
K=14
K=16
K=18
K=20
K=21
fit
B
RD50
10^{13}
10^{12}
10^{11}
10^{10}
10^9
10^8
12
14
16
18
20
21
K
C
Eukaryota
Archaea
Bacteria
K=21
K=20
Log Representation
K=20
K=21
-11.05
-10.258
-9.466
-8.674
-7.882
-7.09
-6.299
-5.507
-4.715
-3.923
-3.131
-11.05
-10.317
-9.584
-8.851
-8.118
-7.385
-6.652
-5.919
-5.185
-4.452
-3.719
Scale


% Representation of new sample
100
80
60
40
20
0
NT-constructed
Same species
Same genus
Same family
Same order
Same class
Same phylum
A new genome from a different phylum
New phylum
Cross domain
Random
20-mers in NT
Prime-constructed, reshuffled
21-mers in NT
Prime-constructed
Zag