nature ecology & evolution

ARTICLES

PUBLISHED: 24 APRIL 2017 | VOLUME: 1 | ARTICLE NUMBER: 0127

Random sequences are an abundant source of bioactive RNAs or peptides

Rafik Neme†, Cristina Amador†, Burcin Yildirim, Ellen McConnell and Diethard Tautz*

It is generally assumed that new genes arise through duplication and/or recombination of existing genes. The probability that a new functional gene could arise out of random non-coding DNA is so far considered to be negligible, as it seems unlikely that such an RNA or protein sequence could have an initial function that influences the fitness of an organism. Here, we have tested this question systematically, by expressing clones with random sequences in *Escherichia coli* and subjecting them to competitive growth. Contrary to expectations, we find that random sequences with bioactivity are not rare. In our experiments we find that up to 25% of the evaluated clones enhance the growth rate of their cells and up to 52% inhibit growth. Testing of individual clones in competition assays confirms their activity and provides an indication that their activity could be exerted by either the transcribed RNA or the translated peptide. This suggests that transcribed and translated random parts of the genome could indeed have a high potential to become functional. The results also suggest that random sequences may become an effective new source of molecules for studying cellular functions, as well as for pharmacological activity screening.

The *de novo* evolution of genes and proteins from random sequences has long been considered to be highly unlikely[1,2]. Given that the combinatorial possibilities of even short protein sequences are almost infinite, while the number of actually found folds in solved structures is not more than a few thousand, it seemed for a long time that only a very minute fraction of the sequence space could possibly become bioactive[3–5]. Yet comparative genome and transcriptome analyses have shown that new transcripts and proteins can easily arise *de novo* from random parts of the genome[6–9]. Intriguingly, the highest rates of *de novo* emergence are always found in the evolutionarily youngest lineages[10], suggesting that the constraints for *de novo* evolution cannot be very high. Further, a re-analysis of presumed non-coding transcripts has shown that many of them associate with ribosomes and develop potentially functional open reading frames (ORFs)[11,12].

Here, we have set out to ask whether a random translatable sequence expressed in a cell could cause an effect within the cell that gives it an advantage or disadvantage in growth compared with other cells. We have chosen to use *Escherichia coli* as a test system, for its ease of manipulation in laboratory conditions while maintaining large population sizes. *E. coli* is known to respond to even small selection differences under competitive growth conditions[13].

Assuming that *de novo* gene birth occurs via acquisition of specific elements to produce transcribed and translated protogenes, we decided to mimic the process and provide a population of bacterial cells with artificial protogenes in which all basic elements for transcription and translation are already present. We used an expression vector to express random sequences with the potential to code for peptides under the control of an inducible promoter. Each RNA with its ORF acts as both a novelty in the cell and a marker in a screening-by-sequencing approach. As all other parts of the bacterial machinery are virtually identical in the clonal population, we can expect that the differences in growth are due to competition dynamics[13] and that these should be causally related to the expression of the plasmids in the cell.

In our approach, we do not use restrictive conditions as was done previously in a similar system[14], but allow optimal growth conditions and monitor only frequency changes of clones of a transformed library, rather than expecting fixation of selected variants. We find that a very high fraction of the random sequences that could be statistically assessed in the experiments show changes of frequency over time, thus affecting the growth rate of the cells either positively or negatively. Repetitions of the experiment yield overlapping subsets of clones with the same response modes. We conclude that the majority of randomly generated sequences have reproducible biochemical activity, possibly through interactions with components of the cellular machinery, and are thus relevant for fitness differences in the bacteria.

Results

To generate a library of random sequences with coding potential, we synthesized oligonucleotides including 150 nucleotides with equal representations of all 4 nucleotides at each position during synthesis, and cloned them into the pFLAG-CTC expression vector (Sigma-Aldrich) containing an isopropylthiogalactoside (IPTG)-inducible promoter and a C-terminal FLAG sequence (see Methods). This results in transcribed RNAs with a length of about 700 nucleotides, of which 150 are randomly generated sequence. When translated, the corresponding peptides have a length of 65 amino acids with a central part of 50 amino acids of random sequence, unless the sequence includes a premature stop codon. All of our further analysis focuses only on clones coding for such full-length peptides to make results best comparable.

The initially transformed library was amplified without induction through IPTG. This pool of amplified clones served as a source for all experiments. Two main sets of experiments were performed, one with passage to a new flask after 3 h, the other after 24 h of growth. Inoculation at each cycle was done with a large pool of cells to avoid drift effects through bottlenecks. This meant that only about four to five cell divisions occurred before the stationary phase was reached. Each experiment was done in ten replicates and was

Max-Planck Institute for Evolutionary Biology, August-Thienemannstrasse 2, 24306 Plön, Germany. †Present addresses: Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, 1212 Amsterdam Avenue, New York, NY 10027, USA (R.N.); Technical University of Denmark, Department of Biotechnology and Biomedicine, 2800 Kgs Lyngby, Denmark (C.A.). *e-mail: tautz@evolbio.mpg.de

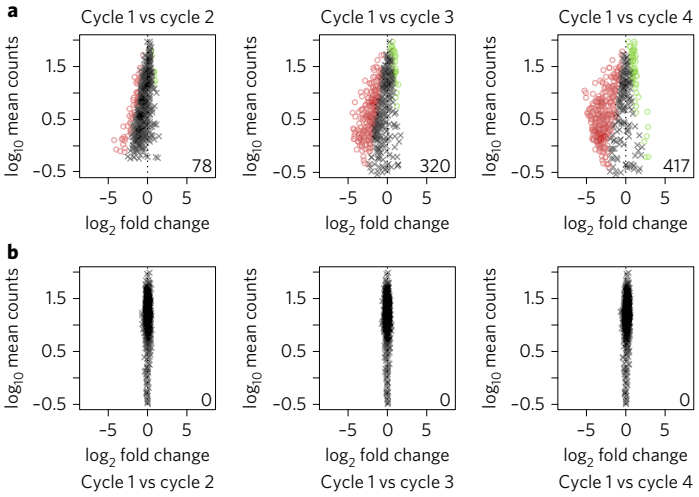

Figure 1 | Induction of expression through IPTG drives changes in clone frequency over time. a, With IPTG. **b**, Without IPTG. Plots of fold change (compared with the first cycle) versus mean counts across pairwise comparisons. Negative fold changes are indicative of depletion compared with the first cycle, and positive fold changes are indicative of enrichment compared with the first cycle. Left, centre and right panels indicate comparisons with the 2nd, 3rd and 4th cycle (24 hours per cycle), respectively. Green and red dots indicate clones with significant fold changes (5% FDR), positive and negative, respectively. Black crosses indicate clones with non-significant fold changes. The number in the lower-right corner of each plot indicates the total number of clones with significant changes. Both experiments were derived from the same stock culture and performed simultaneously. We observe that the induction has a clear effect on the dynamics of competition and differentiation along the experiment.

run for four growth cycles in the presence of IPTG. The frequency of each clone was determined through parallel sequencing at the end of each growth cycle. Counts of sequence reads for each clone were used for statistical analysis using DESeq2[15].

Across all experiments we found a large number of different clones in the library (see Methods). However, most of these occurred only at low frequencies, the majority were detected only once across all experiments. As these rare clones preclude statistical analysis, we focused only on those for which at least five counts were observed in at least one of the parallel replicates. This reduced the number to 1,082 clones that could be evaluated in at least one experiment, and all further statistical analyses were based on these.

In a first test, we asked whether any major frequency changes could be detected when the expression was not induced by IPTG. Figure 1 shows the effect of induction with IPTG compared with replicate cultures of the same experiment without induction. The induced experiment showed major shifts in clone frequency over time, while the non-induced experiment showed only minor, non-significant variation. This proves that expression of the random sequences from the clones is strictly required to cause frequency changes.

A large number of clones showing an increase or decrease in frequency were observed in the experiments under induction conditions. Most of them showed a consistent increase or decrease across the four experimental cycles. Figure 2 shows such examples.

The overall results are summarized in Table 1 and all clones with up or down responses are listed in Supplementary Table 1. In the three experiments with 3 h cycles (E1–E3), we find a significant change in frequency for around 70% of the analysed clones, with about three- to fourfold more clones going down in frequency than going up. The 24 h cycle experiments are more heterogeneous with respect to these overall percentages, but maintain the same overall trend.

The observation of a higher proportion of depleted clones depends on sequencing depth, combined with the statistical requirement of a certain depth of coverage for each clone to detect a significant change. Identifying statistically significant depletion is easier when a clone already has a high starting frequency—and such clones will also be found at lower sequencing depth. Significant increases in frequency, on the other hand, can start at very low initial frequencies and these will become more visible at higher sequencing depth. To test this, we conducted an additional experiment (E7) with four 24 h

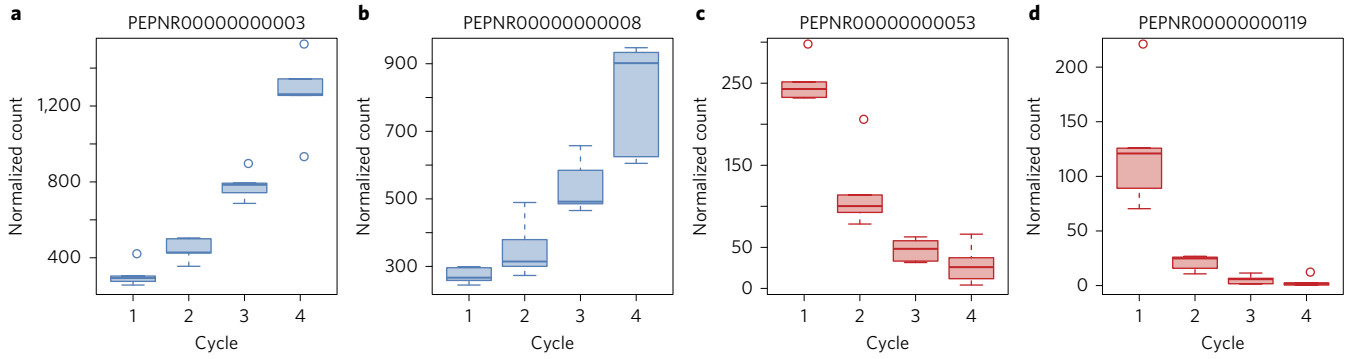

Figure 2 | Examples of four clones with significant changes in frequency over time. Each time point represents a 24 h cycle. **a**,**b**, Clones increasing in frequency (as normalized counts, see Methods). **c**,**d**, Clones decreasing in frequency. Boxplots show the median and the interquartile ranges (outliers as dots) across the ten replicates of each cycle.

Table 1 | Summary across seven different experiments.

	3 h cycles			24 h cycles			
	E1	E2	E3	E4	E5	E6	E7*
Total number of reads	457,212	324,090	678,823	293,638	300,807	812,315	4,142,836
Analysed clones†	623	529	607	616	499	618	1061
Reads for analysed clones	75,927	80,745	265,413	68,285	97,877	326,375	715,675
Significant clones at FDR 0.05	389	322	389	287	148	369	713
Enriched	68	67	68	64	58	67	277
Depleted	321	255	321	223	90	302	436
Bioactive clones (%)	62	61	64	47	30	60	67

*Experiment with only five replicates but deeper sequencing, resulting in tenfold higher read coverage per replicate. †Number of clones with at least five reads in at least one replicate in any experiment. FDR, false discovery rate.

cycles, including five replicates (instead of ten) and a deeper sequencing depth. At this higher sequencing depth, we find an overall fraction of significant clones similar to most of the other experiments (79%), but indeed also a larger number of enriched clones (Table 1).

Using the data from experiment E7 we could also estimate the power to detect depleted versus enriched clones. We performed analyses for subsets at 10% intervals of the total reads in the experiment. Figure 3 shows the fold-change plots for four depths of sampling. We find that from 50% onwards, more and more initially low-frequency clones become significantly enriched. Supplementary Fig. 1 shows the corresponding rarefaction analysis where the detection of depleted clones is more or less complete at 60% coverage, while the detection of enriched clones keeps rising.

Among the clones with significant changes in frequencies, we found a fraction that differs by only a single nucleotide/amino acid change from another clone in the set. These variants were probably created through polymerase chain reaction (PCR)-induced mutations during the oligonucleotide amplification step. A total of 95% of them showed the same direction of change, confirming the high repeatability of the individual effects and suggesting that single nucleotide differences do not have much influence on the effect.

Compositional analysis. To assess whether there are systematic compositional patterns in sequences that have an effect on the cells, we focused on the translated peptides, as they provide potentially more information than nucleotide composition of the underlying RNAs. We have thus compared the amino acid composition and biochemical properties of the amino acids of all analysed peptides with a set of random peptides and with biological controls derived from *E. coli* annotated proteins (Supplementary Fig. 2). We find that for almost every case where the biological control deviates from random expectation, the peptides in our study are closer to the random control. Similarly, there are no major differences between up- and down-regulated clones with respect to amino acid composition in most comparisons. Still, minor differences are found for some comparisons in the relative frequency of certain amino acids, such as less R, D, C, G, S and V, and more N, E, Q, I and T in the real *E. coli* protein sequences. However, these are biochemically rather diverse amino acids in each set, which do not allow much speculation about possible structural differences at present. Moreover, given that a subset of the clones may act via their RNA rather than the coding peptide (see below), such inferences would still be highly speculative.

Validation of individual clones with growth advantage. All of the above experiments were conducted in the context of a large mixture of clones. We were interested in testing some clones individually or in less complex mixtures to see whether the activity patterns could be confirmed. Further, while there could be many ways in which random sequences, especially peptides, could inhibit growth processes, it is of particular interest to study clones that provide cells with a growth advantage. We chose three such clones and isolated the respective plasmids from the library. Western blotting shows that all three express a peptide in dependence of induction by IPTG, albeit at different steady state levels, which could reflect different overall stability (Fig. 4).

To test whether the clones have an advantage with respect to clones harbouring only an empty plasmid, we ran a standard four-cycle experiment as above, amplified the inserts and quantified the DNA on an Agilent gel chip. We find that all three clones

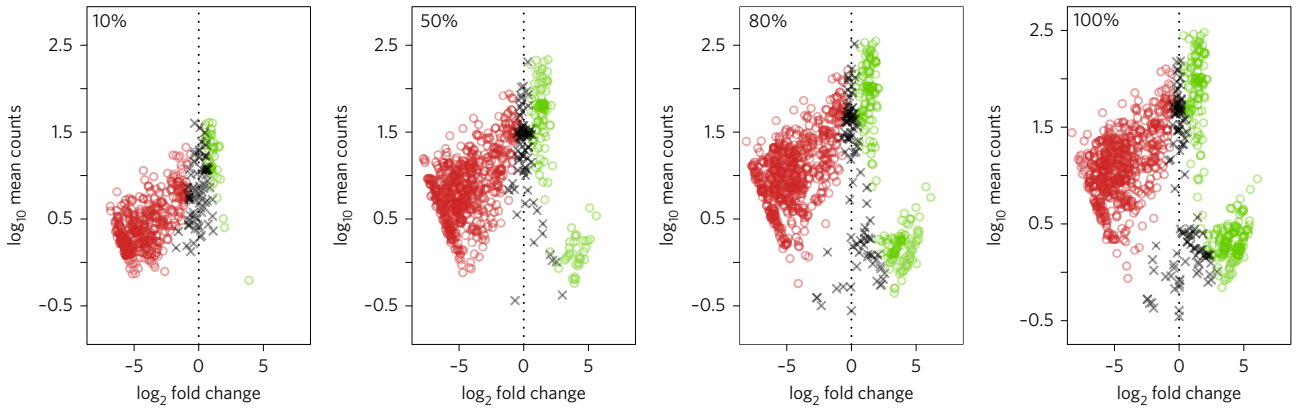

Figure 3 | Assessment of read depth on detection power. Progression of significant fold changes with sampling depth in experiment E7 (from 10 to 100% of the reads). Red circles (left of the dotted line) indicate clones significantly decreasing over time; green circles (right of the dotted line) indicate clones significantly increasing over time; black crosses indicate clones with non-significant fold changes. Significance was set at 5% FDR.

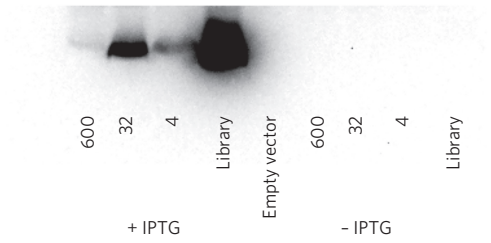

Figure 4 | Expression of peptides. Western blot with antiFLAG antibody for the three individual clones and the whole library. Left side: after induction of the promoter with IPTG; right side: control without induction.

show an increase in frequency over time (Fig. 5a). We tested further whether they would also show this in competition with each other in different combinations. This is indeed the case; all are better than the empty vector (Fig. 5b).

Given that the bioactivity could be conveyed by either the transcribed RNA or the translated peptide, we produced versions of these clones harbouring a stop codon directly at the start of the random part of the sequence, that is, only the first four amino acids that are common among all clones would be translated. These mutated clones were also tested in pairwise competition assays with the empty vector. Only one of the clones (clone 600) showed a clear difference between the mutated and the non-mutated version (Fig. 5c), which would suggest that only this clone exerts its effect via the encoded peptide, whereas the two other clones might act through their RNA alone. To study this in more detail, we did an experiment with a direct competition of each clone with its stop codon counterpart, but with the same qualitative results (Supplementary Fig. 3).

Discussion

Our experiments show that an unexpectedly large fraction of random RNA or peptide sequences are bioactive, at least in the sense of influencing relative growth rates in *E. coli* cells. The results imply that it could be either the RNA itself, or the corresponding translated protein that conveys the bioactivity. Although two of our three individually tested clones suggest that the RNA function could be more important than the protein function, this constitutes at present only a small sample and may not be indicative of the true ratio between RNA and peptide functions. However, this observation fits well with the notion that an active RNA may precede an active peptide during *de novo* gene evolution of genes[6,10–12].

Previous studies have shown that specific biochemical activities or specific resistance against stress conditions can be recovered from random peptides using very large sample sizes and directed selection experiments[14,16]. However, this occurred only at low frequencies and required many rounds of selection. Our results show

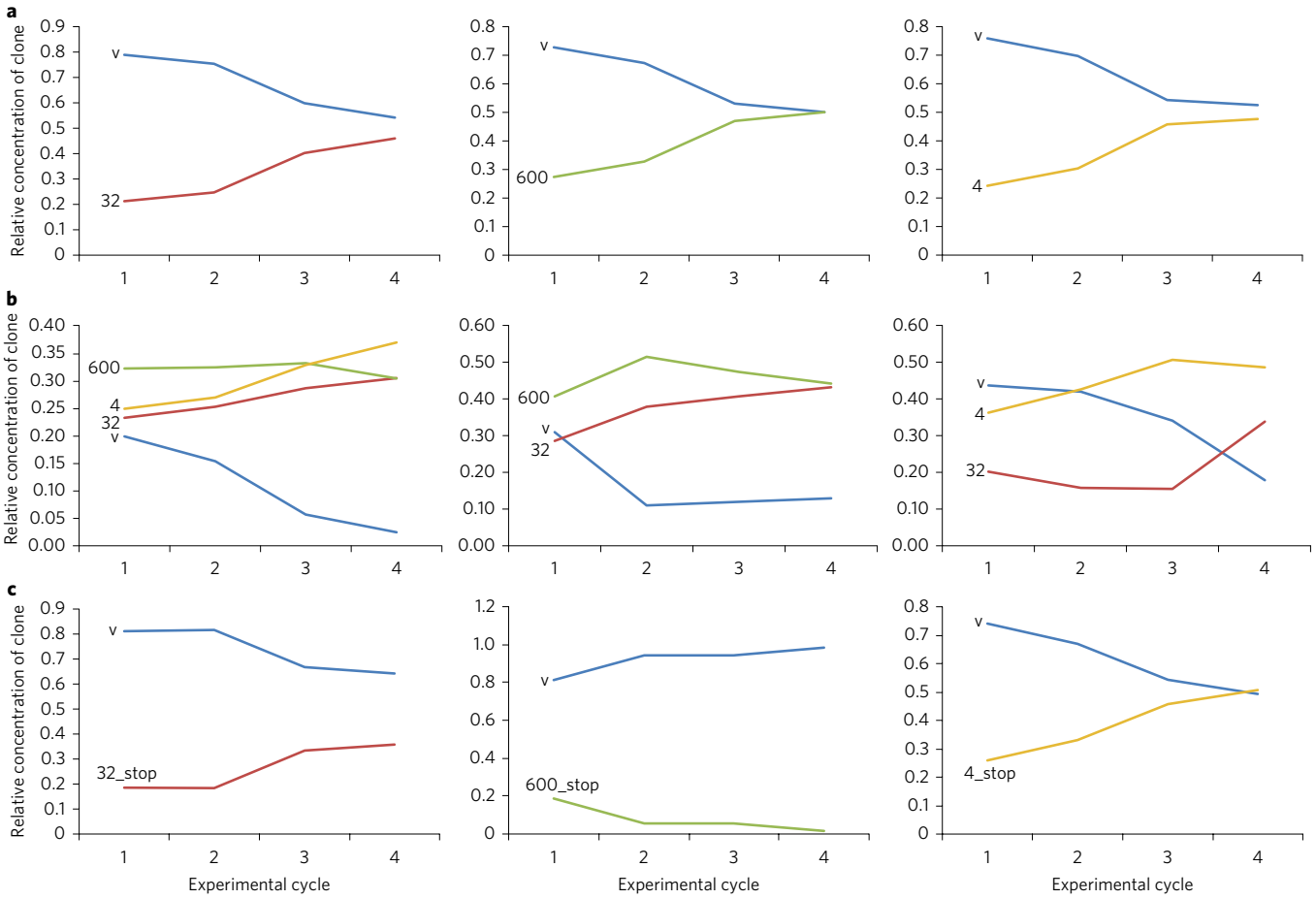

Figure 5 | Growth competition experiment with three selected clones. Relative concentrations of each clone (*y* axis) are plotted against the four experimental cycles (*x* axis). **a**, Competition against empty vector separately with each of the clones. **b**, Competition in different combinations of vector and clones. **c**, Competition between vector and stop codon version of the respective clones. v, vector; 4, clone 4; 32, clone 32; 600, clone 600. Note that in each experiment we mixed the cells of the corresponding clones in equal starting amounts, but the subsequent PCR favoured different fragments to different extents. Hence, only the trajectories are comparable, not the absolute values.

that a non-directional approach in which only proxies for biological fitness are considered, rather than specific activities, recovers a much higher fraction of bioactive RNAs and peptides, and thus allows a better understanding of the functional potential of the unexplored random sequence space for molecular innovation.

Almost any random RNA could fold into a higher order structure, or interact with other RNAs via base pairing, although the free energies and interaction would be expected to be weak. For peptides, one could expect that they interact via charged or hydrophobic interactions with other molecules. They would not need to fold into a stable structure to do this. Many proteins exist that are partly or fully made up of intrinsically disordered protein regions[17,18]. One can assume that such disordered peptides or proteins can associate with molecular complexes and can influence their activity. This can be more or less specific, that is, only a single complex, or many complexes are affected[19,20]. But as long as the specific effect or activity is reproducible, it becomes functionally—and thus also evolutionarily—relevant.

Negative effects of expressed peptides may not necessarily be very specific, given that a strong promoter is used in the expression vector and that some peptides might simply aggregate and thus harm the cell. However, as our first frequency measurement is taken at the end of cycle 1, where cells have already grown under induction conditions, we expect that the strongly deleterious peptides are already mostly lost. Hence, we consider even the negative effects as possibly specific, in the sense that they do not simply block the whole cell physiology.

We find a very high reproducibility in our experiments, both with respect to trends across cycles, and between and within experimental set-ups. Further, we show for three clones that their effects are also measurable in isolation. Evidently, it will now be of interest to study many more isolated clones in a range of conditions, to assess whether they are broadly active, or generate a growth effect only within a limited range of conditions.

Our present results suggest that a large fraction of all possible random sequences may have some biochemical activity of biological relevance, at least in *E. coli*, but possibly in any other cellular organisms too. Still, it will require testing more libraries with different clone compositions to confirm this notion. Similarly, although we find more deleterious than advantageous clones, it is too early to speculate on the exact relative frequencies between them. Our results show that this ratio is heavily dependent on sequencing depth, and we expect that it may also depend on growth conditions and other environmental variables. Furthermore, different cells might show different responses. It will therefore be necessary to carry out similar experiments with other bacterial and/or eukaryotic cells with a range of different conditions.

We note that our findings may also have practical implications. Assuming that one can show that a given RNA or peptide interacts with a specific cellular process, such molecules could be seen as new probes for studying cell physiology and growth. Random expression libraries could also be used for screening approaches similar to the ones that have been developed for short hairpin RNA libraries (shRNA[21]), to identify specific RNAs or peptides that influence particular pathways or physiological states. This could also lead to new procedures to identify pharmaceutically relevant molecules.

Conclusion

The results presented here suggest that an unexpectedly high fraction of randomly combined nucleotide or amino acid chains are biologically active and can influence fitness. This lends credence to the possibility that *de novo* evolution of genes has played a significant role in evolutionary history. But the finding also raises many new questions, including whether any of these molecules would be able to form a stable fold on its own, or whether they can only act in interactions with stable molecules or complexes. We can now experimentally address these and many other questions related to *de novo* gene evolution and their implications for the cell and the organism.

Methods

Construction of a library of random inserts in an expression vector. We opted for a length of 150 nucleotides of random sequence to be expressed, corresponding to 50 amino acids when fully translated. This length was chosen as it should give peptides a higher chance to interact with other components of the cell machinery. At the same time it is short enough to allow full-length sequencing by an Illumina-based approach. We used the pFLAG-CT expression vector (Sigma-Aldrich, catalogue no. E8408) for cloning. This includes the strong P*tac* promoter (a hybrid of the *trp* and *lac* promoters from *E.coli*) regulated by the presence of the *lac*O sequences and inclusion of the *lac* repressor gene (*lac*I) on the plasmid. It drives a transcript that includes a ribosome binding site and a start codon, as well as a C-terminal FLAG sequence with a stop codon. The oligonucleotide including the randomized sequence was cloned between the *Hind*III and *Sal*I sites of the multiple cloning site. The insert sequence was synthesized as an oligonucleotide of the following sequence:

5′-ACGTCCAAGCTTAGC(N150)GCATTGGTCGACGTA-3′

whereby (N150) represents 150 nucleotides that were synthesized as equimolar mixes of A, C, G and T at every position. This oligonucleotide pool was purified on a 8% acrylamide gel and amplified using the following primers:

Oligo forward: 5′-ACGTCCAAGCTTAGC-3′ / Oligo reverse: 5′-TACGTCG ACCAATGC-3′

The double stranded product was digested with *Hind*III and *Sal*I, and ligated into the digested vector. This results in the predicted peptide sequence:

*ATGAAGCTT*AGC NNN GCATTG*GTCGACTACAAGGACGATGACGA CAAGTGA*

*MetLysLeu*Ser (aa50) AlaLeu*ValAspTyrLysAspAspAspAspLysSTOP*

whereby (aa50) denotes 50 amino acids in random combinations. Positions in italics represent the parts provided by the vector, including the FLAG sequence. The ligation products were transformed into *E.coli* DH10B cells by electro-transformation. The initially transformed cells were grown without IPTG induction to stationary phase to generate the library stock, which was frozen after adding 20% glycerol.

Growth competition experiments. The design of the experiments was aimed to identify clones that would consistently show a frequency change across multiple cycles of growth, whereby all cycles are run under the same culture conditions. The following general setup was used: (1) generate a pre-culture by inoculating 25 ml LB medium (Invitrogen; catalogue no. 12780-052) supplemented with 50 μg ml^{-1} ampicillin (AMP; Roth; catalogue no. K029.1) with 500 μl from the library stock and grow overnight at 37 °C at 250 r.p.m. shaking conditions in an Erlenmeyer flask. (2) inoculate up to ten replicates with 500 μl each of the pre-culture in 5 ml fresh LB medium + AMP + IPTG (Sigma; catalogue no. I1284; 1 mM final concentration) in 14 ml tubes with snap lid (Falcon, 17 × 100 mm, catalogue no. 352057); grow overnight at 37 °C at 250 r.p.m. shaking conditions; this is cycle no. 1. (3) Repeat the last step, but use 500 μl each of the culture from the previous step until cycle no. 4.

This set-up means that the cells from cycle no. 1 have already grown under induction conditions, that is, the frequency of their clones may already have been changed in comparison with the starting frequency in the library or the pre-culture. We chose this set-up to ensure that all frequency estimates from the sequencing of the clones (see below) come from samples grown under the same conditions. Further, given the inoculation with a high number of cells (approximately 10^8) at each cycle, one assures that there is no dilution effect with respect to the number of clones in the library (approximately 10^6). The inoculation with the high number of cells implies also that there are only about three to four generations until the new stationary phase is reached.

A total of seven experiments were performed according to this scheme, all with four cycles; three (E1–E3) where each cycle lasted 3 h and three (E4–E6) where each cycle lasted 24 h (a prolonged stationary phase), with ten replicates each. A further experiment (E7) was done with 24 h cycles and five replicates.

Sampling and sequencing. For all experiments and after each cycle, 2 ml was used for plasmid extraction (QIAGEN Plasmid Mini Kit; catalogue no. 12125). Library preparation was done using PCR amplicon sequencing, targeted to the periphery of the insert on the plasmid, and providing the primers for subsequent sequencing, using the following primers:

Forward primer: 5′-ATGATACGGCGACCACCGAGATCTACACNNNNNN-NNTATGGTAATTGTCATCATAACGGTTCTGGCAAATATTC-3′

Reverse primer: 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNNNAGT CAGTCAGCCCTGTATCAGGCTGAAAATCTTCT-3′

The hexanucleotide with Ns indicates a region where sequencing barcodes were placed to distinguish the different replicates. After PCR, the samples were pooled together. Sequencing was carried out with an Illumina MiSeq sequencer, following the standard amplicon sequencing protocol, and using the MiSeq Reagent Kit v3 for 600 cycles (catalogue no. MS-102-3003) to produce 300 bp paired-end reads from each sequenced fragment.

Single clone recovery. To obtain single clones from the library, we used PCR primers based on the determined sequences of the clones facing outward of each other. Stop codons at the desired positions were engineered by modifying one of the primers at its 5′-end to code for a stop codon. Amplification then yields the full vector that needs only to be religated. However, to ensure that the vector had not suffered a mutation, we recloned the inserts of the recovered clones into a new common vector. All inserts obtained in this way were resequenced to confirm the original sequence.

We used Western blots to check for the expression of the respective peptides. A quantity of 1.4 ml of overnight culture was spun down and resuspended in 100 µl Laemmli buffer with 5% β-mercaptoethanol, then samples were incubated at 99 °C for 5 min and debris was centrifuged down. A quantity of 30 µl was loaded onto a 4–20% tris-glycine gel (Bio-Rad) and run for 1 h 40 min at 70 V. The proteins were then transferred to polyvinylidene difluoride membrane for 15 min at 13 V using a Bio-Rad semi-dry electroblot unit. The membrane was washed 2 × 10 min with gentle shaking in phosphate buffered saline (PBS; 140 mM sodium chloride, 10 mM sodium hydrogen phosphate, 2.7 mM potassium phosphate, 1.8 mM potassium dihydrogen phosphate pH 7.3) with 0.1% tween 20 (PBST) and then blocked in 5% powdered milk (1% fat) dissolved in PBST with shaking at room temperature for 1 h. The monoclonal mouse anti-FLAG M2 antibody (F1804 Sigma) was added, diluted 1 in 2,000 in 2.5% milk PBST. The membrane was incubated overnight with shaking in a cold room (approximately 6 °C). The membrane was washed 3 × 10 min in PBST with shaking. Goat-anti mouse horse radish peroxidase (HRP)(A16072 Thermo-Fisher) diluted 1 in 2,500 in 2.5% milk PBST was added and incubated with shaking at room temperature for 1 h. The membrane was washed 3 × 10 min in PBST with shaking. Enhanced chemiluminescence (ECL; Clarity Western ECL from Bio-Rad) was pipetted onto the blot (approximately 3 ml per blot) and incubated for 5 min, then blotted with thick filter paper and protected from light. The membrane was then imaged using a digital imager (Alpha Innotech) with increasing exposures until bands were well visible.

Single clone competition experiments. The competition experiments with individual clones or combinations thereof were done under the same conditions as for the 24 h cycles. But instead of proceeding with a sequencing step (see above), we amplified the inserts by PCR and ran the products on an Agilent Biochip (DNA 7500). To distinguish the sizes of the clones with inserts, we digested them first with diagnostic restriction enzymes. The Agilent software for quantification of the bands was then used to obtain concentration differences of the fragments between time points.

Data processing and analysis. FASTQ paired-end reads were collapsed into a single FASTA sequence each using USEARCH[22]. Whenever conflicting bases were detected between pairs, those with the best quality score were retained. The translated peptide sequence was obtained from each merged sequence using getorf from the EMBOSS suite[23]. Only those ORFs starting and ending with the expected sequences (see above), and having exactly 65 amino acid residues (50 from the randomized sequences and 15 from the vector, including the tag) were retained for further analyses.

A non-redundant database was constructed with USEARCH[22] for all experiments using protein sequences at 100% identity, that is, similar non-identical sequences are treated as independent units. This implies that this database includes translated sequences with possible sequencing errors or PCR-induced mutations.

It was possible to estimate the error rates per sequencing run using the first 85 nucleotides of plasmid sequences in the reads. We cropped the reads to this length using Unix shell scripts, mapped them to the reference plasmid sequence with NextGenMap[24] and determined the percentage of mismatches using samtools fillmd[25] to assess substitutions as a proxy for errors. We found error rates in the range of 0.12–0.56%. Given these low rates, we did not try to curate the database further. The sequences of each replicate in each experiment were matched to the database using DIAMOND[26]. This provided a quantitative representation of each sequence in each cycle and each replicate, as well as across experiments. These counts were used to compare the changes in frequency of each clone over time.

Statistical procedure. The number of times a clone was observed was recorded and the counts for each time point were determined. Clones of very low frequency cannot be statistically analysed. For this reason, we required occurrence of at least five times or more in any one replicate of an experiment to consider a statistical analysis of a given clone. The further statistical analysis is based on procedures designed for differential gene expression[15], but is applicable to any type of count data, in particular those derived from high-throughput sequencing experiments. The analysis was done using the R package DESeq2[15].

Clones with significantly different frequencies between the first time point and the last time point were recorded. By including comparisons at the other two time points, they were further categorized into increasing or decreasing in frequency across time points when the tendency was consistent.

Comparison across experiments. Results from different experiments were compared by collecting all clones with any significance across all experiments and recording the direction of the fold change (enrichment or depletion). Clones with opposing effects in any two experiments were not further considered to avoid inflation from false positives, and these remained only a minor fraction of the overall detected clones.

Rarefaction analyses. Experiment E7, which was sequenced intensively, was used to estimate the effects of sampling on the discovery of enrichment or depletion. All replicates were normalized at 50,000 clones each, thus giving the whole experiment a total number of 1 million sequences. Random subsamples were obtained at 10% intervals. Subsampled experiments were analysed as described above.

Sequence properties. To assess whether the enriched or depleted peptides in the experiments behave like random sequences or biological protein sequences, we simulated random 150 nucleotide RNA sequences, and added the vector information to obtain a translation like the one performed experimentally. We also obtained all protein sequences from *E. coli* deposited in the GenBank and fragmented them into lengths of 65 amino acids to act as biological controls. We extracted simple compositional properties derived from sequence information using the package protr[27] and ran Wilcoxon rank tests to compare properties of experimental peptides with random and biological sequences.

Data availability. The data tables with read counts and corresponding statistics for each experiments are provided at Dryad http://dx.doi.org/10.5061/dryad.6f356. Sequence files are available at the European Nucleotide Archive (ENA) under the project number PRJEB19640.

Received 22 October 2016; accepted 1 March 2017; published 24 April 2017

References

1. Jacob, F. Evolution and tinkering. *Science* **196,** 1161–1166 (1977).
2. Tautz, D. The discovery of *de novo* gene evolution. *Perspect. Biol. Med.* **57,** 149–161 (2014).
3. Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature* **357,** 543–544 (1992).
4. Lupas, A. N., Ponting, C. P. & Russell, R. B. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134,** 191–203 (2001).
5. Orengo, C. A. & Thornton, J. M. Protein families and their evolution—a structural perspective. *Annu. Rev. Biochem.* **74,** 867–900 (2005).
6. Carvunis, A. R. *et al.* Proto-genes and *de novo* gene birth. *Nature* **487,** 370–374 (2012).
7. Reinhardt, J. A. *et al. De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* **9,** e1003860 (2013).
8. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of *de novo* genes in *Drosophila* melanogaster populations. *Science* **343,** 769–772 (2014).
9. Neme, R. & Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. *eLife* **5,** e09977 (2016).
10. Tautz, D. & Domazet-Loso, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12,** 692–702 (2011).
11. Xie, C. *et al.* Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8,** e1002942 (2012).
12. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *Elife* **3,** e03523 (2014).
13. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14,** 827–839 (2013).
14. Stepanov, V. G. & Fox, G. E. Stress-driven *in vivo* selection of a functional mini-gene from a randomized DNA library expressing combinatorial peptides in *Escherichia coli*. *Mol. Biol. Evol.* **24,** 1480–1491 (2007).
15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
16. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410,** 715–718 (2001).
17. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *BBA-Proteins Proteom.* **1804,** 1231–1264 (2010).
18. Tompa, P., Schad, E., Tantos, A. & Kalmar, L. Intrinsically disordered proteins: emerging interaction specialists. *Curr. Opin. Struct. Biol.* **35,** 49–59 (2015).
19. Cumberworth, A., Lamour, G., Babu, M. M. & Gsponer, J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochem. J.* **454,** 361–369 (2013).
20. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A million peptide motifs for the molecular biologist. *Mol. Cell* **55,** 161–169 (2014).

21. Sims, D. *et al.* High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol.* **12,** R104 (2011).
22. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
23. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16,** 276–277 (2000).
24. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29,** 2790–2791 (2013).
25. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
26. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2015).
27. Xiao, N., Cao, D. S., Zhu, M. F. & Xu, Q. S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **31,** 1857–1859 (2015).

Acknowledgements

We thank S. Künzel for sequencing and E. Özkurt for contributions during her rotation project. The project was financed through an ERC advanced grant to D.T. (NewGenes—322564).

Author contributions

R.N. and D.T. designed the experiment, C.A. constructed the library, C.A., B.Y. and E.M. conducted the experiments, R.N. did the bioinformatic analysis, and R.N. and D.T. wrote the paper.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.T.

How to cite this article: Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides *Nat. Ecol. Evol.* **1,** 0127 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The work described in this publication is subject to patent application by the Max-Planck Society.