scientific reports

Check for updates

OPEN Mutations as Levy flights

Dario A. Leon[1,2]✉ & Augusto Gonzalez[3,2]

Data from a long time evolution experiment with *Escherichia Coli* and from a large study on copy number variations in subjects with European ancestry are analyzed in order to argue that mutations can be described as Levy flights in the mutation space. These Levy flights have at least two components: random single-base substitutions and large DNA rearrangements. From the data, we get estimations for the time rates of both events and the size distribution function of large rearrangements.

Life is coded in the DNA molecule, and the combined effect of random mutations and natural selection leads to biological evolution. Point mutations provide a kind of fine tuning of the genome, allowing the optimization of protein functions, for example. On the other hand, radical remodeling by genetic recombination events is thought to be the source of global changes, leading even to new biological species[1].

In order to describe mutations, one shall determine the rate at which they occur and the "spatial" distribution function, that is their distribution along the DNA. To the best of our knowledge, there are precise measurements of the mutation rates in many situations[2–4], as well as precise indications of sites or regions in the DNA prone to mutations[5]. However, there are no results concerning the length distribution function of mutations.

In the present paper, we use data from a long-term evolution experiment (LTEE) with *E. coli* populations[6] in order to get the rate of both, point mutations and large chromosomal rearrangement events in the evolution of this bacterium. Data on single-nucleotide polymorphisms (SNPs) in mixed-population samples, taken from generation 2000–40,000, come from sequencing these samples and aligning to the genome sequence of the ancestral strain[7]. On the other hand, large chromosomal rearrangements in clones harvested from these samples are identified by means of a combination of optical techniques, genome sequencing and PCR analysis[8]. The emergence of a mutator phenotype, which increases the rate of SPMs by 100 times but does not affect large rearrangements, indicates that these are essentially different processes. However, although there are many types of large rearrangements, responding to different scales and mechanisms, all of them can be accommodated into a global distribution function for the lengths of the modified DNA segments, exhibiting a scale-free power-like behavior.

On the other hand, the rate of single point mutations (SPMs) and other DNA rearrangements in germline cells in humans has been precisely measured[4,9]. With regard to the length distribution function of mutations, we shall use data from a recent large study on copy number variants in subjects with European ancestry[10]. As in the bacterial case, a scale-free distribution function arises in the scale range spanned by the experiment. In both, bacteria and human germline cells, we show that the fit can be extended to the small-length range of the data by means of a stable Levy distribution[11].

From an abstract point of view, mutations can be described as a succession of transformations in the DNA molecule—a Markov chain[12]. The chain configuration at the step $i + 1$ is written in terms of the configuration at step i as: $X_{i+1} = X_i + \delta_i$, where δ_i is the introduced modification. The main result of the paper is that in δ_i we shall distinguish at least two kinds of transformations: SPMs and large rearrangements, the lengths of the latter are distributed according to a stable Levy law. That is, mutations are a kind of Levy flights[13].

Data on bacterial SPMs

In an evolution experiment, random fluctuations are filtered by natural selection. The evolution dynamics in the LTEE is schematically represented in Fig. 1. Cell lineages with neutral or deleterious mutations are usually truncated, whereas beneficial mutations confer evolutionary advantage to clones and, thus, higher probability to continue. Once they appear, beneficial mutations are fixed in more than 50% of the population after a fixing time. The number of cell lineages, that is number of cells passing to the next day in the evolution, is kept fixed to around five millions in the experiment.

We draw in Fig. 2 the data on SPMs, taken from Ref.[7]. A population, called Ara-1 in the experiment, is sampled at generations 2000, 5000, 10,000, 15,000, 20,000, 30,000, and 40,000. The two latter points are not included in the figure because of a mutator phenotype, which appears at generation 27,000 and leads to a 100-fold increase of the mutation rate.

[1]University of Modena & Reggio Emilia, 41125 Modena, Italy. [2]Institute of Cybernetics, Mathematics and Physics, 10400 Havana, Cuba. [3]University of Electronic Science and Technology, Chengdu 610051, People's Republic of China. ✉email: dario@icimaf.cu
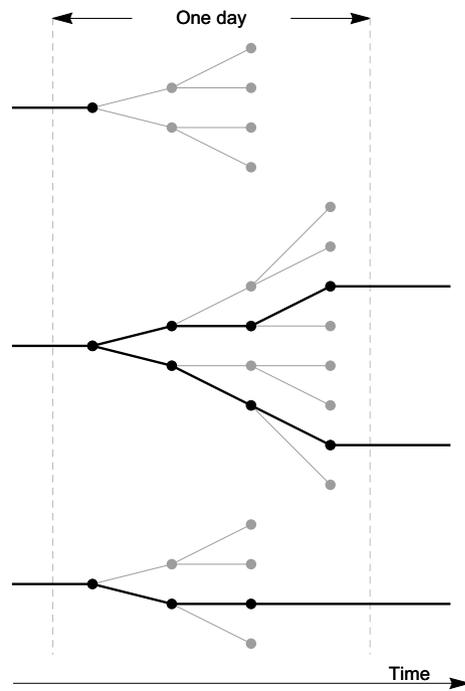
Figure 1. Phylogenetic representation of one day evolution in the LTEE. After a few clonal divisions (2–3 in the figure, 6–7 in the experiment) individuals are randomly selected to pass to the next day. Most lineages are truncated, whereas those with higher fitness have better possibilities to continue to the next day.
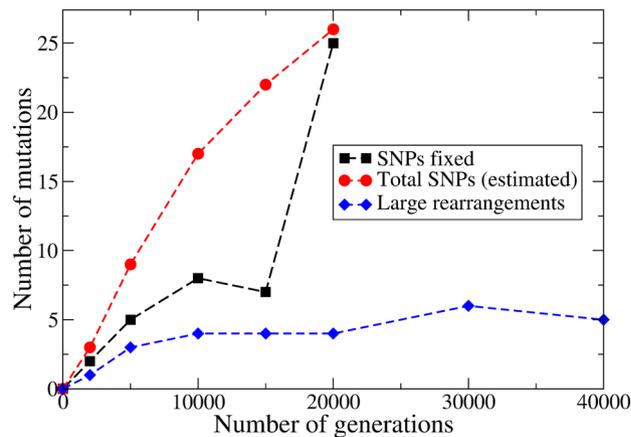

Figure 2. The number of observed mutations as a function of time (number of generations) in a population named Ara-1 of the LTEE[7]. Data from generation 0 (ancestral strain, taken as reference) to 20,000 are included in the figure. Fixed SNPs (black squares), the estimated mean number of SNPs in clones (red dots, coming from calculations in the Supplementary Materials), and the number of large rearrangements (blue diamonds) are shown.

Alignment of 36-base reads in mixed population samples yielded 40- to 60-fold coverage, allowing to determine frequencies of SPMs above 4% in the population. Authors report "fixed" SPMs, meaning that their frequency, f, is above 96%, as well as so called single-nucleotide polymorphisms, SNPs, where $4\% < f < 96\%$.

The data labeled "fixed" in the figure, corresponding to mutations with $f \geq 96\%$, show a linear increase at short times with a slope 1.0×10^{-3} mutations/generation. The data labeled "mean", on the other hand, correspond to our estimation for the mean number of mutations one may detect in a clone (see Supplementary Material for details). The slope of the mean curve at short times is a little higher, around 1.8×10^{-3} mutations/generation, which may be taken as an estimation of the SPM rate, p_{SPM}.

The value obtained for p_{SPM} should be compared with the total point mutation rate, that in the LTEE was estimated to be $p_{SPM} = 10^{-4}$–10^{-3} mutations per generation for the whole genome[14].

These are effective rates. A detailed simulation of the evolutionary dynamics requires including competition between clones, drift processes, etc[15,16].

Summarizing the present section, we may say that, for SPMs along a cell lineage, we estimate $p_{SPM} \sim 1.8 \times 10^{-3}$ mutations per generation. The latter is obtained from the slope of the model curve in Fig. 2 near the origin.

Finally, a very significant point is that, in the studied population, a mutator genotype becoming dominant after generation 27,000 increases by a factor of 100 the point mutation rate.

The next section is devoted to large rearrangements.

Data on large chromosomal rearrangements in the LTEE

Data on large chromosomal rearrangements are provided in Ref.[8]. Due to experimental limitations, authors can not reliably detect rearrangements smaller than 5 Kilo base pairs (Kbp). On the other hand, they perform measurements on clones, that is representatives of a population, which may exhibit strong deviations from mean values.

Let us stress that mutations are rare events. The data reported in Refs.[7] and[8] are the results of 20 years of evolution and 40,000 bacterial generations. However, only around 100 large chromosomal rearrangements are registered in the 12 populations under study. With such scarce data we can not pretend a precise description of the mutation distribution function. Only qualitative and semi-quantitative results can be extracted.

The first set of results involve a time sequence of clones of the population Ara-1, as in the previous section. That is, samples at generations 2000, 5000, 10000, 15,000, 20,000, 30,000, 40,000, and 50,000.

We shall estimate the time rate, p_{LR}, and size probability distribution, $\pi_{LR}(l)$, of such events. The experiments report on different kinds of rearrangements: deletions, insertions, translocations, and inversions.

We included in Fig. 2 the detected number of large rearrangement events as a function of time (number of generations). Half of these rearrangements seem to be fixed, in the sense that they are detected also at later times in different clones. From the slope at short times, we get a rough estimation for the rate of large changes, $p_{LR} \sim 5 \times 10^{-4}$ large chromosomal rearrangements/generation, a value three times smaller than p_{SPM}.

Because of the experimental resolution, not all the rearrangements are registered, specially short-length ones. Thus, our estimation for p_{LR} is a lower bound, and the actual rate could be similar to p_{SPM}.

A second very important point is related to the fact that this figure does not show any abrupt increase of p_{LR} after generation 27,000, where the mutator phenotype becomes dominant. This fact stresses the differences between the mechanisms leading to SPMs and large rearrangements in the bacterial chromosome.

Figure 3, upper panel, on the other hand, reflects the size statistics. We use a log-log plot. The x-axis is the size, l, and the y-axis is the number of rearrangements with size greater or equal than l. In the interval $5 \times 10^3 < l < 1.5 \times 10^6$, the data are very well fitted by the function C/l^ν, with $\nu = 0.42$ and C a normalization constant (Pearson correlation coefficient $r = 0.96$).

This dependence can be understood as coming from a probability $\sim \nu/l^{1+\nu}$ for a large rearrangement of size l to occur. Indeed, the number of events with size greater or equal than l is thus computed as:

$$C\nu \int_l^\infty \frac{\mathrm{d}x}{x^{1+\nu}} = \frac{C}{l^\nu}. \tag{1}$$

These results are based on the 9 detected large DNA rearrangement events in the Ara-1 population in 50,000 generations. Below, we shall consider a larger data with better statistics. The data comes from clones harvested from the 12 independently evolving populations in the LTEE, sampled at generation 40,000. There are 110 detected large rearrangements in these clones. The results are shown in Fig. 3 center panel.

First, we shall stress that the points for $5 \times 10^3 < l < 1.7 \times 10^6$ show a remarkable scaling with $\nu = 0.49$ (Pearson correlation coefficient $r = 0.99$). The slope of the experimental curve changes for $l < 5$ Kbp. This fact may be partially due to the limitations of the experimental techniques that can not detect all of the rearrangements for these l values, as mentioned by the authors. But it could also be related to a saturation of the distribution function for small values of l. On the other hand, in the right hand side of the figure l is near the bacterial DNA size, $L \approx 5 \times 10^6$ bp.

The idea behind this figure is to show that the length distribution function of all the evolving populations is similar. A common exponent near 1/2 seems to describe all the populations.

Up to this point we have concluded that SPMs and large rearrangements have essentially different mechanisms for their time rates, but the many types of large rearrangements can be accommodated into a common distribution function. For this latter property to hold, the distribution function should obey the central limit theorem, that is the sum of independent sub-processes should preserve the function.

Motivated by these facts, that is the central limit theorem and saturation in the low-length region, we tried a stable Levy distribution $L_{1/2}(\alpha l)$ in order to fit the observed distribution of points. In general[11], the Levy probability density distribution

$$L_\nu(y) = \frac{2}{\pi} \int_0^\infty \exp(-q^\nu) \cos(qy) \mathrm{d}q \tag{2}$$

behaves as $1/y^{\nu+1}$ for large values of its argument.

In Fig. 3 bottom panel, the integrated probability density is plotted. As the length runs from the maximal value, l_{max}, to its minimum, l_{min}, the number of rearrangements rises from 1 to the total value, N_{LR}. The integrated distribution, red crosses in Fig. 3 bottom panel, may thus be written as:
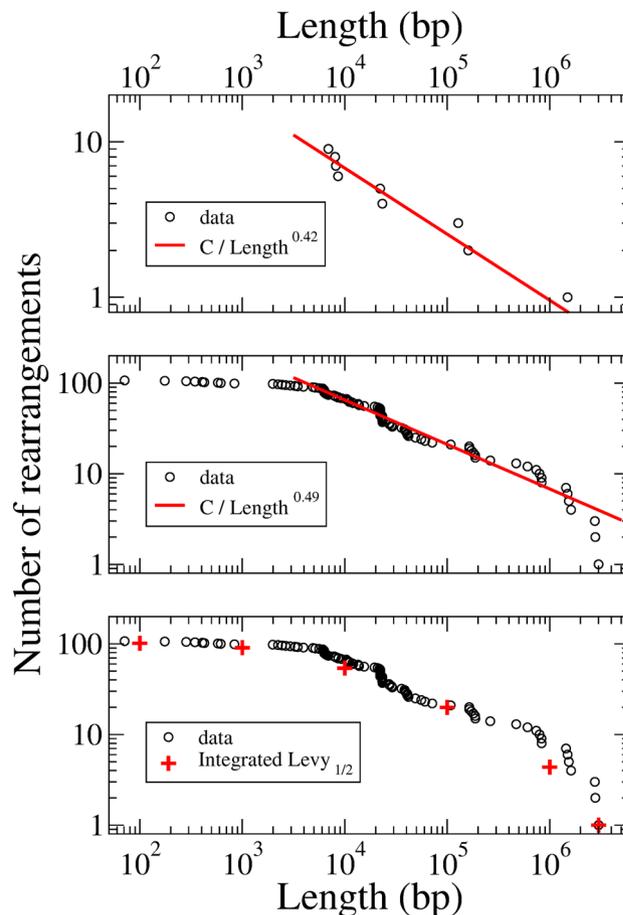
Figure 3. Top: Log-log plot of the integrated size distribution function of large (greater than 5 Kbp) chromosomal rearrangements in clones of the Ara-1 population. The red line is a fit with a $1/l^\nu$ dependence. Center: log–log plot of the integrated size distribution of large rearrangements in clones obtained from the 12 independently evolving populations in the LTEE, sampled at generation 40,000. The red line is a fit with the function $1/l^\nu$ for l in the interval $5 \times 10^3 < l < 1.7 \times 10^6$. Bottom: A fit to the observed distribution in a wider interval by means of the integral Levy 1/2 distribution given in Eq. (3).

$$f(l) = 1 + (N_{LR} - 1) \frac{\int_l^{l_{max}} L_{1/2}(\alpha y)\, \mathrm{d}y}{\int_{l_{min}}^{l_{max}} L_{1/2}(\alpha y)\, \mathrm{d}y}. \tag{3}$$

The parameter $\alpha = 10^{-4}$ provides a very good fit.

Summarizing the section, we may say that large rearrangements are observed in bacterial cell lineages at rates $p_{LR} \sim 5 \times 10^{-4}$ per generation. This is a lower bound for p_{LR}, the actual value could be closer to p_{SPM}. No changes in p_{LR} are reported after generation 27,000 in the Ara-1 population, when a mutator phenotype leads to a 100-fold increase of p_{SPM}, which means that the mechanisms responsible for SPMs and LRs are very different. The observed rearrangements are well described by a stable Levy distribution $L_{1/2}(\alpha l)$, which in the reliable size interval, $5 \times 10^3 < l < 1.7 \times 10^6$, shows a dependence $\sim 1/l^{3/2}$.

The rate of mutations in human germline cells

We shall consider mutations in the human germline cells. Somatic mutations, although relevant in aging processes, cancer, etc are less constrained by evolution and may be dictated by different rules.

The natural unit of time in the present case, instead of cell generations, are organism generations, that is births. The data is summarized in Fig. 1 of paper[9], where single nucleotide variants (SNVs) are distinguished from rearrangements such as small indels (mean length 2 bp), mobile elements insertions (MEIs, mean length 200 bp), copy number variants (CNVs, mean length 10^6 bp), and aneuploidies (mean length 10^8 bp). There are around 60 SNVs per birth, five times the number of all other mutations taken together.

Although there is not a complete understanding of mechanisms causing these kinds of mutations, one should expect different acting mechanisms and, thus, independent random processes. In a model of mutations we shall consider, as in bacteria, at least two independent processes: SNVs and chromosome rearrangements. The former, of Brownian character, acting with a rate of 60 mutations per birth; and the latter, with a rate of around
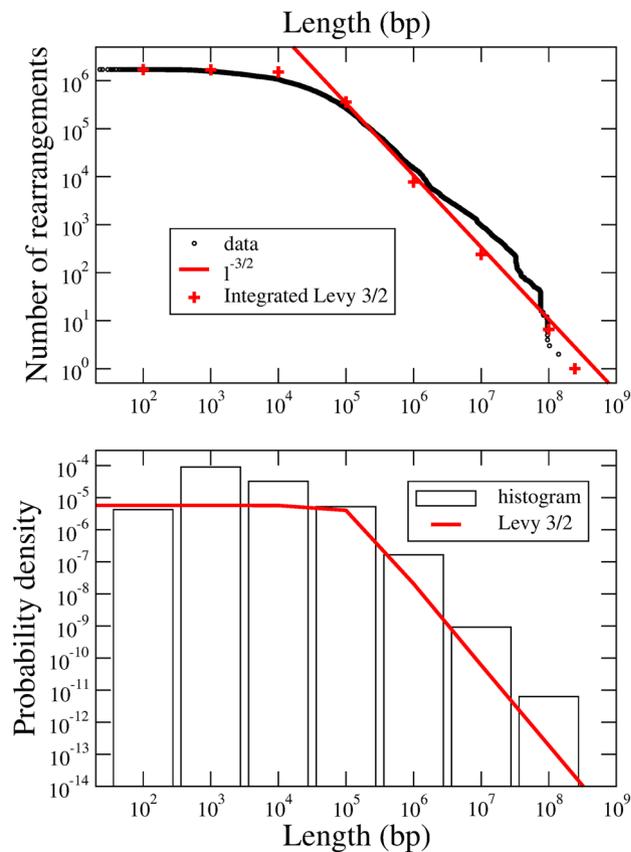
Figure 4. Top: Log–log plot of the integrated size distribution of CNVs in germline cells. The red line is a fit with the function $1/l^{3/2}$ for $l > 10^5$ bp whereas the red crosses come from the integral Levy 3/2 distribution. Bottom: A direct comparison of $\alpha L_{3/2}(\alpha l)$ with a log-contracted histogram of the data.

10 mutations per birth and a length distribution function which shall be determined. In the next section it will become apparent that CNV events can be described by a stable Levy function.

The length distribution function of CNVs

We use data from a recent study of CNVs in more than 100000 subjects of European ancestry[10]. Typical CNVs have lengths below 1 Mbp and frequencies below 0.01 in the studied cohort. However, the authors provide data for more than 1.7 millions of rearrangements which lengths range from 10 to 2.4×10^8 bp. Not all mutations are detected with the same fidelity in this wide range. One should expect short-length mutations to be under counted.

The low frequencies of mutations indicate that they are mostly of neutral or deleterious character.

The results are presented in Fig. 4 top panel. This figure is similar to Fig. 3 for bacteria. The x axis is the length, l, of the mutated segment, and the y axis is the number of rearrangements with lengths greater than or equal to l. Thus, this is an integrated probability distribution and we expect it to be described by a formula like Eq. (3), in which $L_{1/2}$ is replaced by L_ν, and the parameters l_{max}, l_{min} and N_{LR} are actualized accordingly.

The figure shows that the distribution can be well fitted by a $\nu = 3/2$ stable Levy function, $L_{3/2}(\alpha l)$, with a scale parameter $\alpha = 10^{-5}$. A tail $\sim l^{-3/2}$ is apparent in the integrated distribution function for lengths greater than $1/\alpha \sim 10^5$ bp.

The bottom panel of this figure contains a direct comparison between the $\alpha L_{3/2}(\alpha l)$ probability density and a histogram in which the $x-$axis is log contracted such that each bin spans a decade. Fluctuations are apparent in the figure, specially in the short-length region, where mutations are most likely to be undetected. This is the reason why we decided to fit the smoother integrated probability function, instead of the probability density. The exponent $\nu \approx 3/2$ seems to be a robust determination.

To summarize the section we may state that a model of mutations in human germline cells should contain at least two processes: Brownian SNVs with a rate of around 60 mutations per birth, and chromosome rearrangements with a stable Levy 3/2 distribution function for the lengths, which become scale-free for $l > 10^5$ bp. The rate of the latter events is probably well below 10 events per birth. Additional short- and intermediate-lengths measurements should be conducted in order to precise whether they can be described merely by a modification of the scale α in the Levy 3/2 function or should be included as independent random processes.

Discussion

The data on SPMs and large rearrangements in bacterial DNA in the course of 50,000 generations of evolution seem to support a picture in which both kinds of events occurs with similar rates[17]. The size distribution of large rearrangements can be fitted with a stable Levy function with exponent $\nu \approx 1/2$ and scale $\alpha \approx 10^{-4}$.

This is a kind of Levy flight picture for mutations along a cell lineage in which small deviations and radical changes in the genome are combined.

In a way, our paper is similar to Refs.[18,19], where the Levy flight theory of foraging is tested against experimental data.

The picture is not complete, however, because of the lack of experimental data on chromosomal rearrangement, in the range $1 < l < 5 \times 10^3$ bp. Notice that the inverse of α coincides with both the lower range of experimental observations and the length above which the integrated distribution reaches its asymptotic behavior $\sim 1/l^\nu$.

For $l > 1/\alpha$ there are no additional scales, and the distribution function is roughly scale-free. This is somehow unexpected. Naively, one would expect a scale of the order of a few Kbp and a rapidly decaying distribution function for rearrangement lengths larger than the scale. The biological mechanism by which such a scale-free distribution is generated should be further clarified.

We stress that a power-like scale-free distribution is observed also for the distances (spacers) between highly conserved fragments in several genomes[20]. We guess that an evolutionary model for mutations in which Levy flights are constrained to respect conserved fragments would lead to a power-like distribution function for distances between fragments.

The LTEE is a clean clonal evolution experiment. In wild conditions, horizontal gene transfer through recombination events is expected to play an important role[21]. The lengths of recombined fragments seem to be distributed along a power law also[22].

From an abstract perspective, a scale-free distribution for large rearrangements is a good strategy[18]. In the described experiment, where the population size is controlled and nutrients are limited, biological evolution can be viewed as an optimization problem. The mean fitness in the population is the cost function. Mutations provide the mechanism for surveying the parameter space, and natural selection picks up the best representatives in the population. A local search alone, like the SPMs or short length rearrangements, could trap mutation trajectories around a local maximum in the fitness landscape. An optimal search algorithm shall include large rearrangements of any size, that is a scale-free size distribution.

The near optimal character of the search algorithm is confirmed in the experiment by what authors call "parallel mutations"[8], that is very similar fixed mutations in independently evolving populations.

We notice, by the way, that the idea of a Levy search has been implemented in computational optimization techniques[11].

We also checked our statement about the Levy nature of mutations in eukaryotes, in particular in human germline cells. Recent precise data on CNVs allowed the determination of the length distribution function in scales larger than $10^4 - 10^5$ bp. We could fit the distribution to a stable Levy 3/2 function, which shows a scale-free behavior up to the typical chromosome length. However, in the short-length region, related to small indels and MEIs, one expects that the data is incomplete, and we can not distinguish whether this lower scale region can be described simply by a modification of the parameter α or independent random processes should be included in the theory.

The fact that the biological complex processes leading to mutations, probably originated from many different mechanisms, exhibit scaling in a very wide range of lengths should be based on very general laws. Our idea to use a stable Levy function in order to fit the data is motivated by such arguments. Stable functions, respecting the central limit theorem, are very good candidates.

We notice that very general arguments have been suggested to explain the observed power-like (Pareto) distribution function for gene expression[23] in cells. Our paper is similar in spirit to this one.

Differently from the conclusions of the LTEE experiment, the low frequencies observed in CNVs indicate that most of these mutations exhibit neutral or deleterious character, and indeed they are shown to be strongly correlated to diseases or disorders[10].

The data, although limited, seems to suggest exponents 1/2 and 3/2 for bacteria and human germinal cells, respectively. New questions arise as, for example, whether the exponents, and not only the mutation rates, may vary under different selective pressure, whether the change from 1/2 to 3/2 reflects a trend in evolution[24], etc. On the other hand, it is known that the optimal value for the exponent in Levy searches is equal to one[25]. The obtained exponents are close to this value. There are also known limitations of Levy searches, in particular to find close minima[26]. The question arise as to whether the addition of SPMs as an independent process in our model (and, probably, other short-length processes) is a way of correcting such limitations.

A probable next step in our research would be to describe the very important somatic mutations, involved in aging processes and cancer. Somatic stem cells in human tissues have been shown to reach numbers above 10^8, and their replication rates may lead to 10^4 cell generations along a lifespan[27], a number comparable to the number of generations in the controlled LTEE with *E. coli*.

Massive sequencing of tumors are already available, see for example[28], and the importance of somatic mutations in cancer is widely recognized. A catalogue of somatic mutations in cancer exists (https://cancer.sanger.ac.uk/cosmic), which may provide the data for checking the Levy hypothesis. The idea that large rearrangement hits on particular genes may lead to cancer is very plausible. In particular, hits on very important genes, such as p53[29]. Correlations between CNVs and relevant genes have been tested[10].

If the Levy nature of mutations is generally confirmed, it could have practical implications in modeling carcinogenesis. The key obstacle is to relate mutations to cellular fitness[21]. In gene expression space[30], however,

the high and low fitness regions are apparent. Normal tissues and tumors are grouped in disjoint high fitness regions. We have tried[30] a local plus Levy jumps model for the motion in this space that seems to reproduce the data on cancer risk in a set of tissues.

Data availability

The information about the data we used, the procedures and results are integrated in a public repository that is part of the project "Processing and Analyzing Mutations and Gene Expression Data in Different Systems": https://github.com/DarioALeonValido/evolp. The data we use for bacteria[6] and copy number variations[10] are replicated in paths `../evolp/bases_external/LTEE/mutations/` and `../evolp/bases_external/CNV/` respectively. To process each data set we include specific scripts in `../evolp/Levy_mutations/`.

References

1. Alberts, B. *et al. Molecular Biology of the Cell* (Garland Science, New-York, 2008).
2. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
3. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *PNAS* **99**(2), 803–808. https://doi.org/10.1073/pnas.022629899 (2002).
4. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1), 297–304 (2000).
5. Rogozin, I. B. & Pavlov, Y. I. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat. Res.* **544**(1), 65–85. https://doi.org/10.1016/S1383-5742(03)00032-2 (2003).
6. Lenski, R. E. Summary data from the long term evolution experiment. http://myxo.css.msu.edu/ecoli/summdata.html (2019).
7. Barrick, J. E. & Lenski, R. E. Genome-wide mutational diversity in an evolving population of *Escherichia coli. Cold Spring Harb. Symp. Quant. Biol.* **54**, 1. https://doi.org/10.1101/sqb.2009.74.018 (2009).
8. Raeside, C. *et al.* Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli. mBio* **5**, e01377-14. https://doi.org/10.1128/mBio.01377-14 (2014).
9. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet.* **29**, 575–584. https://doi.org/10.1016/j.tig.2013.04.005 (2016).
10. Li, Y. R. *et al.* Rare copy number variants in over 100,000 european ancestry subjects reveal multiple disease associations. *Nat. Commun.* **11**, 255. https://doi.org/10.1038/s41467-019-13624-1 (2020).
11. Lee, C. Y. & Yao, X. Evolutionary programming using mutations based on the levy probability distribution. *IEEE Trans. Evol. Comput.* **8**(1), 1–13. https://doi.org/10.1109/TEVC.2003.816583 (2004).
12. Koroliuk, V. S., Portenko, N. I., Skorojod, A. V. & Turbin, A. F. *Handbook on Probability Theory and Mathematical Statistics* (Nauka, Moscow, 1978).
13. Shlesinger, M. F., Zaslavsky, G. & Frish, U. E. *Levy Flights and Related Phenomena in Physics, Lecture Notes in Physics* Vol. 450 (Springer, Berlin, 1995).
14. Lenski, R. E., Winkworth, C. L. & Riley, M. A. Evolutionary programming using mutations based on the levy probability distribution. *Mol. Evol.* **56**, 498. https://doi.org/10.1109/TEVC.2003.816583 (2003).
15. Wiser, M. J., Ribeck, N. & Lenski, R. E. Long-term dynamics of adaptation in asexual populations. *Science* **342**, 1364. https://doi.org/10.1073/pnas.0226298990 (2013).
16. Leon, D. A. & Gonzalez, A. Modeling evolution in a long time evolution experiment with *E. coli.* arXiv:1804.02660 (2020).
17. Tenaillon, O. *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165–170. https://doi.org/10.1073/pnas.0226298992 (2016).
18. Viswanathan, G. M. *et al.* Optimizing the success of random searches. *Nature* **401**, 911–914. https://doi.org/10.1073/pnas.0226298993 (1999).
19. Humphries, N. E. & Sims, D. W. Optimal foraging strategies: Levy walks balance searching and patch exploitation under a very broad range of conditions. *J. Theor. Biol.* **358**, 179–193. https://doi.org/10.1073/pnas.0226298994 (2014).
20. Polychronopoulos, D., Sellis, D. & Almirantis, Y. Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics. *PLoS ONE* **9**(5), e95437. https://doi.org/10.1073/pnas.0226298995 (2014).
21. Fisher, D., Lässig, M. & Shraiman, B. Evolutionary dynamics and statistical physics. *J. Stat. Mech: Theory Exp.* **2013**, N01001. https://doi.org/10.1073/pnas.0226298996 (2013).
22. Sakoparnig, T., Field, C. & van Nimwegen, E. Whole genome phylogenies reflect long-tailed distributions of recombination rates in many bacterial species. https://doi.org/10.1101/601914 (2020).
23. Kuznetsov, V. A., Knott, G. D. & Bonner, R. F. General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics* **161**(3), 1321–1332 (2002).
24. Pigliucci, M. Is evolvability evolvable?. *Nat. Rev. Genet.* **9**(1), 75–82. https://doi.org/10.1073/pnas.0226298997 (2008).
25. Lomholt, M. A., Tal, K., Metzler, R. & Joseph, K. Lévy strategies in intermittent search processes are advantageous. *Proc. Natl. Acad. Sci.* **105**(32), 11055–11059. https://doi.org/10.1073/pnas.0226298998 (2008).
26. Palyulin, V. V., Chechkin, A. V. & Metzler, R. Lévy flights do not always optimize random blind search for sparse targets. *Proc. Natl. Acad. Sci.* **111**(8), 2931–2936. https://doi.org/10.1073/pnas.0226298999 (2014).
27. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Supplementary materials at www.sciencemag.org/content/347/6217/78/suppl/. *Science* **347**, 78. https://doi.org/10.1126/science.1260825 (2015).
28. The ICGC/TCGA Pan Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93, https://doi.org/10.1038/s41586-020-1969-6 (2020).
29. Vogelstein, B., Sur, S. & Prives, C. p53: The most frequently altered gene in human cancers. *Nat. Educ.* **3**(9), 6 (2010).
30. Herrero, R., Leon, D. A. & Gonzalez, A. Levy model of cancer. arXiv:1507.08232v4 (2020).

Acknowledgements

Author contributions

A.G. conceived and coordinated the work. D.A.L. and A.G. processed the experimental data on bacteria. A.G. processed the experimental data on humans. D.A.L. is in charge of the GitHub repository. Both authors analyzed and interpreted the results, contributed to the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-88012-1.

Correspondence and requests for materials should be addressed to D.A.L.

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

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