Genome Complexity Browser: visualization and quantification of genome variability

Alexander Manolov1Y\*, Dmitry Konanov1Y, Dmitry Fedorov1, Ivan Osmolovsky1, Elena

Ilina1

# Abstract

Comparative genomics studies may be used to acquire new knowledge about genome architecture - the rules to combine a set of genes in a genome of a living organism. Hundreds of thousands of prokaryote genomes were sequenced and assembled. Still, there is a lack of computational tools able to compare a large number of genomes simultaneously. We developed Genome Complexity Browser, a tool that allows visualization of gene contexts in a graph form and quantification of the variability of different parts of a genome. Graph visualization allows inspecting changes in gene content and neighborhood in hundreds of genomes simultaneously. This may be useful to find conservative and variable parts of operons or to estimate the overall variability of particular genome locus. We introduce a measure called complexity to quantify genome variability. Intraspecies and interspecies comparisons reveal that regions with high complexity values tend to be located in a conservative context in different strains and species.

# Author summary

A comparison of genomes of different bacteria and archaea species revealed that many of them frequently exchange genes. Occasionally such horizontal gene transfer events result in the acquisition of pathogenic properties or antibiotic resistance by the recipient organism. Previously it was observed that the probabilities of gene insertions are not equally distributed along a chromosome. There are loci called hotspots, in which changes occur with much higher frequencies than in other parts of the chromosome. We developed a computational method and a software tool called Genome Complexity Browser, that allows identification of genome variability hotspots and visualization of occurred changes. We performed a comparison of localization of hotspots and revealed that some of them have conservative localization even when comparing different species, while others are transient. Our tool allows users to visually inspect patterns of gene changes in a graph form, which makes such visualizations compact and informative.

# Introduction

Genome is not only a storage of gene sequences. Many sequence motifs and sequence patterns should be located on proper loci to ensure interaction of molecular machines involved in basic cell processes (transcription, genome replication, cell division, etc.) with a chromosome or other replicon [1,2,3]. The interaction of these processes with DNA molecule and interplay between them governs optimal gene localization and orientation [1]. New knowledge about genome architecture - set of imposed constraints and favorable configurations of genomic objects [1] - may help to deepen our knowledge of the basic processes in the cell and facilitate development of artificial genomes in the field of synthetic biology.

Non-random localization of different genes may be important due to several factors.

Genes located near the replication origin have higher copy numbers in fast-dividing cells

– a so-called replication-associated gene dosage effect [4,5]. Folding of the chromosome makes genes located in different parts of the chromosome close to each other in 3D space, which can be beneficial for the gene coding for a regulator and its targets [6,7]. Effects of global regulators (such as H-NS) on gene expression were observed to depend on the location of the target gene [8] and transcriptional propensity also varies depending on the position of the gene on the chromosome [9]. Cooperative effects between RNA polymerases [10] or supercoiling propagation effects may play a role in the transcriptional regulation of neighboring genes.

It was observed that horizontal gene transfer (HGT) events are preferentially localized in hotspots – chromosomal loci in which changes are observed much more frequently than in other regions [11–13]. This might indicate that, although disruptions in genome architecture may result in decreased fitness of an organism, there are some places in the chromosome where changes can be introduced without inducing negative effects. To our knowledge, there is no currently available computational tool to perform quantitative estimation of variability along the chromosome. Such a tool is needed to deepen our knowledge of factors determining genome variability and stability.

Here, we present Genome Complexity Browser (GCB), the tool that allows estimation of local genome variability and visualization of gene rearrangements. Both tasks are performed with the use of a graph representation of gene neighborhood is a set of genomes. Local genome variability is evaluated using the metric introduced in our article, which we called complexity. Complexity profiles may be used to identify hotspots of horizontal gene transfer or other local gene rearrangement events. Graph-based visualization available in GCB allows analysis of patterns of genome changes events and detection of persistent or variable gene combinations (e.g. variable and conservative parts of operons).

# Materials and methods

## Graph construction

Input to this step is the set of genomes with inferred orthogroups. The algorithm for graph construction is the following: each orthogroup is represented as a node, and two nodes are connected by a directed edge if the corresponding genes are located sequentially in at least one genome in a set. The weight of the edge is calculated as the number of genomes in which corresponding genes are adjacent (see Fig 1 A, B). Graph objects and their methods are implemented in gene-graph-lib library for Python 3, more information can be found in the library documentation at [https://github.com/DNKonanov/gene](https://github.com/DNKonanov/gene_graph_lib)\_[graph](https://github.com/DNKonanov/gene_graph_lib)\_[lib.](https://github.com/DNKonanov/gene_graph_lib)

**Fig 1. Principal scheme of a graph representation of gene order in a set of genomes and genome variability estimation approach.** To construct a graph, each orthogroup is represented as a node. Nodes are connected by a directed edge if the corresponding genes are arranged sequentially in at least one genome in the set. A. Genomes 1,2,3 represent three different hypothetical genomes. Arrows represent genes, and genes within one orthogroup have the same color and letter designation. B. Graph representation of three genomes shown in A). The weight of the edge (arrow width) is calculated as the number of genomes in which corresponding genes are located sequentially. C. Deviating paths for node X are defined as paths in the graph which bypass the node X and are connected with the section of the reference node chain limited by the window parameter. D. Two examples of counting deviant paths are shown. X is the considered node, deviating paths are shown with blue lines. Complexity value is defined by the number of deviating paths.

Because GCB uses directed graph representation of gene order, all genomes in a set are first coaligned to achieve the same orientation throughout the set. This step is performed automatically, the algorithm for it is given in S1 Listing.

Orthogroups may include paralogues genes [14] and some modifications in the basic process of graph construction are needed in this case. Two methods are suggested in GCB. The first method is that the orthogroup is not included in the graph for genomes containing paralogs (but the same orthogroup will be included in the graph for genomes containing no paralogs), see S1 Fig A. Thus, if some orthogroup contains two paralogs from only one genome, then for only this genome it will not be included in the graph, and for all the others it will. An advantage of this approach is simplicity and clear output, while some genes are missed in the graph. The second approach is to “orthologize” paralogues genes: for each set of paralogous genes with a unique context, a graph node with a unique suffix is created (S1 Fig B). GCB uses the first approach by default, while the second approach is available as an option both in the command line and browser-based versions.

## Genome complexity definition

We introduce a measure called complexity for quantification of a local genome variability. Complexity values are calculated against one reference genome from the set. This genome is extracted from the graph as a simple chain of nodes that is called the reference chain (Fig 1C). To calculate complexity in the node X, nodes from the reference chain in the range ±*window* around the node X are selected, and complexity is defined as the number of distinct paths in the graph that do not contain the node X but start and finish in the nodes from the selected range (deviating paths), see Fig 1C.

Complexity computing is an iterative algorithm that generates a set of possible deviating paths from each node in the reference genome (Algorithm1) and when a new unique deviating path is found, the algorithm adds 1/*window* to complexity value of all nodes in the reference between nodes coinciding with the start and the end of the deviating path (Algorithm2).

The algorithm has the following user-defined parameters: *window* - the size of the area around node X to which deviating paths should be connected (default 20 nodes),

*iterations* - number of random walk processes from each node (default 500).

**Hotspots definition**Genome complexity profiles often contains peaks surrounded by regions with relatively low values. This peaks of complexity corresponds to the regions with high local variability, called variability hotspots. Genes are considered to belong to a hot spot if their complexity exceeds a threshold. We define threshold as the third quartile value plus interquartile range multiplied by the coefficient *k* (*k* equals 1.5, by default), which is arbitrary but commonly used criteria for outlier detection, initially proposed by Tukey [Tukey, 1977]. Coefficient *k* can be altered by the user to obtain the highest or including modestly complex regions. Hotspot region coordinates can be downloaded in a web version and can be obtained with command line version. As there is no rigid mathematical definition of hotspots, users can infer them with their preferred method and threshold after downloading complexity values from a GCB web site or after calculating them with a stand-alone command line version.

**Simulations of genomes with predefined variability profiles.**

We supposed that calculated complexity value in some region of a chromosome correlates with the frequency of fixed rearrangements in that region. To verify this assumption and validate the algorithm, sets of model genomes were generated. These simulations included random gene insertions, deletions, HGT events, and inversions. HGT and random insertion probabilities were chosen to be equal to deletion events probability to maintain genome length. The probability of inversion was chosen as 1/100 than others, in agreement with literature data that inversion events are less common than other types of rearrangements, such as deletions and duplication [15]. The localization of these changes through chromosome was chosen according to the predefined variability profile. Next, these model genomes were processed by the complexity computing algorithm and results were compared with input distributions (more details are available in S1 Text). The algorithm for simulations is available in S2 Listing.

## Subgraph generation

To visualize a gene context in the region of interest, a subgraph representing this region is constructed. First, a subset of reference chain nodes representing the region of interest is added to the graph. Next, the algorithm iterates through other genomes from the set and adds deviating paths limited to the selected region. If the length of the path is greater than the *depth* parameter, then the path is cropped and only start and end fragments (tails) of a fixed length (*tails* parameter, *tails <depth*) are added to the subgraph. If the weight of some edge is less than the user-defined *minimal edge weight* parameter, this edge is not added to the subgraph. The subgraph generation algorithm is listed in S3 Listing.

## Web server data acquisition and preparation

To construct a dataset for the web server, we downloaded genomes for 143 prokaryotic species that had more than 50 genomes available in the RefSeq database. For each species, if the number of complete genomes available was higher than 50, then only complete genomes were used. If the number of available genomes was higher than 100, then exactly 100 genomes were randomly selected for further analysis. The only exception was *Escherichia coli* extended genome set, which contained 327 complete genomes available as of November 2017.

All downloaded genomes were reannotated with Prokka ver 1.11 [16] to achieve uniformity. Genes were assigned to orthogroups with OrthoFinder ver. 2.2.6 [17]. Python scripts that are a part of the GCB application were used to parse OrthoFinder output, calculate genome complexity values and generate subgraphs around genome regions of interest.

## Additional methods

Phylogenetic trees were inferred with Parsnp v1.2 [18]. Retention indexes were calculated using RI function from R phangorn library [19]. To estimate similarity to the reference genome in the analysis presented in Figure 6, all genomes were aligned with nucmer [20] and similarity score was calculated as follows: all aligned reference genome ranges were reduced with IRanges R package [21] and their total length was divided by reference genome length, all query genomes were sorted by this value and strains with the top 100 highest values were chosen. Nucmer was used to detect synteny blocks between genomes from the same species, and Mauve [22] was used to detect synteny blocks between genomes from different species. Prophages were detected with Phaster [23]. To obtain Figure 3A we used GCB with following parameters *tails* = 1, *minimal edge* = 5. To produce Figure 3B we used GCB with following parameters *tails* = 1, *minimal edge* = 5. To produce Figure 3C we used GCB with the following parameters: *window* = 20, *tails* = 0, *minimal edge* = 5. Code to make Figures 4-6 is available at [https://github.com/paraslonic/GCBPaperCode.](https://github.com/paraslonic/GCBPaperCode)

# Results

## Software description and availability

The GCB tool is available as a standalone application and as a web server. GCB web server is located at <http://gcb.rcpcm.org and> contains data for 143 prokaryote species. Subset of available genomes was included in the analysis when number of available genomes were greater than 100 (with exception for *E. coli*). Complexity profiles in web version were calculated with window size of 20, 50 and 100 genes. If user wants to perform analysis on a custom set of genomes or with a different window size, then standalone version should be used. All features of web version are available in a standalone version. To use a standalone version user should possess a basic command-line usage skills. No precalculated data is available in standalone version.

In this section we will first describe graphic user interface (GUI) elements, then the use of standalone version.

**Elements of graphic user interface**

GCB browser-based GUI consists of three main parts: 1) the top panel that allows selection of a genome and a region to work with, 2) search panel, 3) the complexity plot, 4) the subgraph visualization form (Figure 2).

**Fig 2. Screenshot of GCB browser-based interface**. GCB GUI is available at gcb.rcpcm.org and when GCB is run as a local server. It consists of panel to select organism, complexity profile panel, subgraph visualization panel with nodes and edges information box.

In the *top panel* user can select genomes set (one per species in a web server, arbitrary in a standalone version), particular genome and a contig (in the case of draft genome or when genome consists of several replicons). When this is selected, complexity profile of selected genome is plotted in a complexity plot panel. User can also specify coordinates of the region of interest to be visualized in a graph form. The size of the region should be limited to dozens or few hundreds of kilobases (depends on region complexity) in order to graph visualization step to be performed.

User can search gene annotations to identify location of genes of interest in a *search panel*. Searching is performed over product feature of annotated genomes (only protein coding sequences are considered).

*Complexity plot panel* shows a visualization of complexity profile of a selected genome. Numeric values of complexity profile can be downloaded as a text file for further analysis (e.g. comparison with profiles for other organisms). It is possible to add visualization of custom data (like GC content, pathogenicity islands, prophage regions, sequence motifs, etc.) by supplying file, formatted as:  <genome position> <numeric value>.

*Graph visualization panel* shows graph representation for a selected region of a genome. Several settings are available to customize subgraph visualization to make it convenient for the analysis, (e.g. by setting minimal edge weight to be shown). Subgraph can be saved as a JPEG image or exported in a JSON format to be visualized with specialized software (e.g. Cytoscape) to prepare publication-ready images.

**Standalone version usage**

For analysis of a custom set of genomes we provide a set of scripts which allows user to: 1) infer orthogroups, 2) generate graph, 3) estimate complexity values in text for and in a SQLite database file, 4) generate and visualize subgraph. Database file can be imported to the local GCB server, which can be run on a standard PC. Fig 2B shows a roadmap of the standalone analysis.

Fasta formatted genomes are expected as the input (both complete and draft genomes may be used, we recommend to include at least one complete genome to be used as a reference). Snakemake script (<https://github.com/paraslonic/orthosnake>) is provided to infer orthogroups. It performs genome annotation with prokka [], generation of protein sequences in fasta format with position and product information in the header, orthogroup inference with orthofinder [].

Python scripts available at <https://github.com/DNKonanov/geneGraph> should be used for further steps of the analysis. Orthofinder output serves as the input to *parse\_og.py* script which produces database and sif-formated file containing graph representation of all genomes. Python script *estimate\_complexity.py* computes complexity profile for specified reference genome. Optionally only subset of genomes may be included in this step of analysis by providing text file containing genome file names (this is useful to exclude some genomes after orthogroups were inferred or to analyze variability in a specific subset from overall set of genomes). Tutorials for performing analysis are available at ….

To estimate the genomic variability profile, the number of genomes should not be too small, a few dozens or hundreds are typical values. The upper limit depends on the computational resources available to infer orthogroups, which is the most computationally difficult step.

## Subgraph visualization

Graph representation of gene order provides a convenient way to inspect visually the context of genes of interest and to identify conservative and variable gene combinations. GCB can construct and visualize subgraph - part of the graph containing the region of interest. Next, we describe examples of subgraphs generated with GCB and visualized with Cytoscape [24].

**Subgraph visualization reveals conservative and variable parts of operons.**

Fig 3A shows the subgraph representing the gene context of the capsule gene cluster (chromosomal coordinates 3111444-3128026 in NCBI sequence NC 011993.1) in 327 complete genomes of *Escherichia coli*. From this visualization, it can be seen that the operon has two conserved parts and one variable part. Variable part consists of genes from the serotype-specific synthesis region while neighboring conserved parts correspond to regions involved in polysaccharide export [25]. Capsule is considered as important virulence factor [26] for *E. coli* and many other species. Its variation is essential in avoidance of immune response and phage infection [27,28].

While the existence of variable and conserved parts of this operon was previously known, new information about the architecture of other operons may be obtained with such type of analysis that can be done with GCB.

**Subgraph visualization reveals a genome variability in a particular locus**

Fig 3 B, C shows the visualization of subgraphs of regions containing two operons: hemin uptake (hmu) and propanediol utilization (pdu). The occurrence of *E. coli* harboring these operons in the intestinal microbiome was previously shown to be associated with Crohn’s disease [29–31]. These operons have different phylogenetic **Fig 3. Subgraph visualization allows identification of conservative and variable parts of operons and reveals genome variability in a particular locus.** Subgraphs of the genomic regions in which A) capsule gene cluster, B) hemin uptake and C) propanediol utilization operon are located. A) graph visualization reveals the conservative and variable parts of the capsule gene operon. Visualizations in B) and C) indicate that both operons are located in a conservative context. Genomes that do not contain hemin uptake operon do not contain other genes in the same context. In the case of propanediol utilization operon, in several genomes alternative and highly variable gene sets are present.

distribution: hmu operon is preferentially present in B2 phylogroup (S2 Fig A, retention index = 1), pdu operon can be found in phylogenetically distinct strains of *E. coli*, and its presence is in low agreement with the phylogenetic tree (S2 Fig B, retention index = 0.26). Hmu operon is located on the 3691615-3700567 positions and pdu operon on the 2083448-2101340 positions of NCBI Reference Sequence NC 011993.1).

The graph visualization reveals that hmu operon is located in a conservative context, which means that the neighboring genes are the same in all strains in which it is present (Fig 3 B). The edge that bypasses the operon indicates that in some genomes the genes to the left and right of the operon are adjacent. Graph visualization also indicates that one of the genes (hemin transport system permease, HmuU) or its close homologs are present in two alternative contexts.

Pdu operon is harbored by only a fraction of all considered strains (27 out of 327), but also located in a conservative context (Fig 3 C). Some variations in pdu operon are visible and reflect different operon variants [31]. Unlike the case of hemin uptake locus, here alternative gene sets are present. These alternative sets include genes of iron transport (FepC, FcuA, HmuU), DNA mobilization (retroviral integrase core domain, transposase DDE Tnp ISL3). What is especially interesting is the high variability of that alternative sets, with many overlapping changes being visible in the subgraph visualization.

In the next section, we will describe a quantitative measure of this observable difference in complexity of subgraphs.

## Complexity is a measure of genome variability

In a set of genomes with identical gene content and localization, each node in the resulting graph will have two edges. Any gene rearrangements (deletion, translocation, insertion) result in the addition of new edges. We hypothesized that the number of distinct paths in a subgraph representing a genomic region will monotonically depend on the frequency of fixed gene rearrangements in that region.

We implemented the algorithm (Algorithm 1,2 in Methods) to count the number of distinct random walks in a subgraph representing a given region of the reference genome, the value which we further call complexity of the region. By selecting subregions with the sliding window and calculating complexity value for them we get the complexity profile of the reference genome. The size of the sliding window can be set by the user, the default value of 20 was used for the results described below.

**Method verification**

We verified our approach by performing simulations of genome evolution through gene transfer events, by comparison with the previously developed method and by analysis of known variability hotspot regions (integrons).

Complexity profile for *E. coli* calculated with GCB is in good agreement with hotspots evaluated in [32], see Fig 4A. Complexity values within hotspots identified in [32] are significantly greater than outside of the hotspots (p-value *<*10−16, Mann-Whitney test, see Fig 4B).

**Fig 4. Complexity value can be used as a genome variability measure.** A) Comparison of complexity profile for *E. coli K12* with hotspots identified in [32] (blue rectangles underneath complexity profiles). B) Comparison of complexity values for genes inside and outside hotspots from [32]. C) Comparison of initial variability profiles with complexity profiles calculated from genomes after 3000 evolution simulation steps. D) Complexity profile for *Vibrio cholerae N16961* chromosome II with noticeable high levels at the integron region.

We performed a number of simulations in which we suggested that the probability of genomic rearrangement events (HGT, deletion, translocation) is non uniformly distributed along the chromosome. This may reflect the unequal probability of changes or their fixation. The algorithm to make simulations is listed in S2 Listing. We used three patterns to generate profiles of such probabilities: sinusoidal, rectangular, and sawtooth and performed 10 independent simulations for each pattern. The number of rearrangement iterations was 3000 for each model. The results of our method were in good correspondence with the predefined distribution (R-square *>*0.74, Spearman correlation *>*0.69, FDR corrected p-value *<*10−300), see Fig 3C. Results for each simulation are available in S1 Text.

Integron and they have high complexity values (see Fig )

Integrons are gene acquisition systems capable of integration, excision and rearrangement of gene cassettes and are examples of genome variability hotspots. We observe that integron regions have high complexity values. Fig 3D shows integron region of *V. cholera* as an example. This integron was dubbed superintegron because of its high length of about 120 kbp and more than a hundred of integrated cassettes [33].

**High complexity values associated with prophages and genomic islands but not limited to them**

Fig 5 shows the complexity profile along the chromosome of adherent-invasive *E. coli LF82* [34]. As expected, the regions with a high density of essential genes have low complexity values. On the contrary, pathogenicity islands and prophage regions have relatively high complexity values. At the same time, there are chromosomal loci with high complexity values that have no recognizable signs of mobile elements. A subgraph of the region with the highest complexity values (located at 2,115,791-2,164,382) do not contain recognizable genes with mobility associated functions.

**Fig 5. Regions with high complexity values are mostly associated with mobile elements.** Complexity profile for the *E. coli LF82* chromosome is shown, 327 other *E. coli* genomes were used to calculate complexity values. Orange color bars denote prophages, red color bars denote pathogenicity islands. For essential genes, low complexity values are observed. Some most variable regions lack features of mobile elements.

**Complexity profiles have conservative features on the inter- and intraspecies levels**

Genome complexity analysis can be used to compare variability profiles of different species or intraspecies structures (e.g. phylogroups). We performed comparison of the complexity profiles for 146 species and observed that when genomes are similar enough (synteny blocks covers most part of the genomes), then complexity profiles are also similar (S3 Fig).

Fig 6A shows comparison of complexity profiles of the four *Bacillus* species, Fig 6B for their phylogenetic relation. It can be seen that regions with high complexity values are associated with prophages (denoted with an orange bar below the complexity profile) and have conservative location. Some regions that lack integrated viruses also have high complexity values and conservatively located in genomes of different species (i.e. the one located at 2.5 Mbp in *B. subtilis*), while others are only highly variable at one species only (e.g. the one located at 2.8 Mbp in *B. velezensis*).

**Fig 6. Regions with high complexity values are mainly located in a conservative context, when both intra and interspecies comparisons are performed.** A) Complexity profiles and synteny blocks of the four *Bacillus* species; B) phylogenetic tree of the four *Bacillus* species; C) complexity profiles and synteny blocks of the five *E. coli* phylogroups, green triangle denotes conserved region with high variability not associated with prophages and genome islands; D) phylogenetic tree of the *E. coli* genomes selected for the analysis, for each phylogroup one reference strain and 100 closest genomes was selected.

Fig 6C shows a comparison of complexity profiles for different *E. coli* phylogroups [35]. For each of the five large phylogroups (A, B1, B2, D, E) we selected one reference strain and 100 most similar strains from 5466 RefSeq genomes (both finished and draft assemblies), see Fig 6D for the phylogenetic tree of selected genomes. Complexity profiles for each reference genome were inferred using genomes from corresponding clade only. This comparison reveals that many of the regions with high variability rate are located in the same context in the genomes of the strains belonging to the different phylogroups. The majority of them contain prophages, but some do not include phage-associated genes. Transient hotspots (with high complexity in some clades and low complexity values in others) can also be observed.  
 We described above a region in *E. coli* genome with high variability rate and without identifiable mobile genetic elements (designated with green triangle in Fig 6C). As can be seen from Fig6C this variability hotspot is present in A, B1, B2, D and to lesser extent in E phylogroup. Phylogroup E consisted of genomes closely related to *O157:H7 Sakai* strain and contains the largest genomes, mainly due to expansion of bacteriophages. We observed that only in this phylogroup this region has integrated prophage. Prophage integration can explain variability in E phylogoup, but what is a driving force of high variability of the region in other phylogoups remain to be elucidated.

**Method applicability**

Complexity profiles and subgraphs could be obtained for any set of genomes for which orthogroups could be inferred. Graph representation, window-based variability estimation and subgraph visualization performs best when the set contains closely related genomes in which local variability does not overwhelmed by chromosomal rearrangements longer then the chosen window size. From our experience and estimates from other studies [Brilli, BMC Genomics 2013] this means that genomes within 0.05 phylogenetic distance (approximately, a species boundary [Mash 2016]) should be used. Complexity profiles of different sets of genomes can be compared with same limitations, large amounts of genome rearrangements make comparison not informative. We observed that some species have no obvious peaks in the complexity profile (for example naturally competent *P. fluorescens*), which makes comparative analysis not as informative as for species with clear regions of low and high complexity, see SFig XA. Meanwhile naturally competent *N. gonorrhoeae* gives quite clear and comparable profiles (SFig XB).

Draft genomes (consisting of fragmented genome regions called contigs) may be used for complexity estimation. We performed comparisons of complexity profiles inferred with 100 complete or 100 draft genomes with same complete genome as reference and observed significant similarity with pearson correlation coefficient equals 0.87 (Fig 7). The code for the analysis is available at github. The impact of draft genomes may be higher in the case of highly stable genomes with large scale rearrangements as main source of variability (e.g. Mycobacterium tuberculosis). Subgraph visualization suffers from genome fragmentation because false negatives may be introduced by contig boundaries (for example no context of region representing some particular contig could be identified).

**Fig 7 Method applicability and benchmark.** A) complexity profiles computed with 100 complete genomes (at the top) or 100 draft genomes (at the bottom) of *E. coli* are very similar, B) Correlation of complexity values obtained with either 100 *E. coli* genomes or subsets with lower number of genomes. C) Time needed for the graph construction and complexity estimation based on different number of genome.

Small number of genomes included in the analysis may lower the accuracy of complexity profile estimation. We recommend to use not less than several dozens. Fig 7 shows the correlation of complexity values obtained with either 100 *E. coli* genomes or subsets with lower number of genomes. When more than 40 genomes are included in the analysis, pearson correlation coefficient becomes greater than 0.9.

Time needed for the analysis depends on the number of genomes. Fig 7 shows time needed for the graph construction and complexity evaluation steps for a different number of genomes (up to 1000). Main time consuming step for the overall analysis is the orthogroup inference. When no computational cluster is available and the number of genomes is large, other method than orthofinder may be considered. In this case users will be needed to format orthogroup information as: <orthology\_gourp\_id>: gene\_id1 gene\_id2 …, with one line per orthogroup (orthofinder output format).

Local variability hotspots identified with GCB are often coincide with prophages and genomic islands. We compared identified hotspots to curated literature-based dataset [Bertelli et al., 2019] and obtained 0.65 mean F1 score (which is comparable to existing tools such as IslandPath-DIMOB, GIHunter and IslandViewer 4), and 0.54 mean precision. Comparison to the automatically generated dataset gives much lower accuracy scores, see S2 Text for detailed information. We conclude that complexity analysis should be considered as a helper in explorative analysis of genomic islands or prophages, but should not be used as a genomic island predictions without additional analysis. This is because hotspots may be of a different (often unknown) origin.

# Discussion

Synteny visualization tools (e.g. Mauve [22], BRIG [36], genePlotR) are often used for genome comparative studies. Such tools allow visual inspection of large and small genome rearrangements. The number of genomes which can be effectively compared with such approach ranges from a few to several dozens. Meanwhile, hundreds and even thousands of genomes are available for some species now. This large amount of genomes may be used to gain new information about genome variability, genome architecture, and structure of operons. To analyze efficiently large sets of genomes we propose a graph-based approach, in which genes are represented with nodes that are connected depending on their co-localization (neighborhood).

Graphs were previously applied to analyze genome changes in a form of breakpoint graphs [37], which is useful to reconstruct possible ancestral states but, in our opinion, are not convenient for visualization properties. They also were used to represent known genome variants to increase mapping quality [38]. To our knowledge, gene neighborhood graph visualization is available only in FindMyFriends R package beside GCB.

Graph representation of a set of genomes and selecting a subgraph representing a region of interest facilitates answering the following questions. Is a gene (operon) located in the same context in all genomes? If not, then what alternatives are present? Which parts of a gene set (operon) are conservative and which are variable? Which genomes contain some particular combination of genes?

Hotspots of genome variability were described for a number of bacterial species. In [32] the authors analyzed HGT hot spots for 80 bacterial species. They concluded that many hotspots lack mobile genetic elements and proposed that homologous recombination is mainly responsible for the variability of those loci. The factors that determine the location of hot spots, their emergence, and elimination, are still an open question.

We implemented a method for quantification of local genome variability based on the number of unique paths in a subgraph. To our knowledge, it is the first tool that allows quantification of genome variability based on a user-defined set of genomes. GCB provides a way to study dynamics of variability hot spots, changes in their intensity and location on different levels ranging from intraspecies structures like phylogroups or ecotypes to interspecies and intergenus comparisons.

We compared variability profiles between different species and, in the case of *E. coli*, between different phylogroups. We observed that, as a rule, when genomes are close enough for the large synteny blocks to be detected (with blast or nucmer tool), then complexity profiles look similar: the regions with high complexity values are surrounded with low complexity regions forming the same conservative context in different groups of organisms. The analysis of complexity profiles of *E. coli* revealed that many hotspots are located in the prophage or pathogenicity islands integration sites, and site-specific mechanisms could govern their conservative location. Some hotspots lack such factors and reasons for their conservative location are still to be elucidated.

The here proposed approach is not universal, for example, it is not suited for the detection of large genomic rearrangements (larger than window parameter, usually several dozens genes) or changes in noncoding parts of the genome. Our methodology has also some limitations coming from its dependence on orthology inference accuracy. Here we used orthofinder tool [17], which uses MCL graph clustering algorithm based on gene length normalized blast scores. We find this tool to be optimal in terms of efficiency and accuracy. Still, paralogous genes may be attributed to one orthogroup which makes graph representation of the context problematic. We observed that on the average 0.5% of all orthogroups per genome contains at least one paralogues gene; but among all orthogroups inferred for the species the proportion of orthogroups with paralogs is almost 16% (see S1 Table for information for each species individually, code is available at <https://github.com/paraslonic/GCBPaperCode/tree/master/GeneCount>). We implement two possible ways of dealing with paralogous genes in GCB: the default approach is to ignore them, the other is to perform an artificial orthologization process (each paralogous gene with unique left and right context is denoted with a suffix and added to the graph). From our experience, the optimal strategy is to work in the default mode for explorative analysis and verify all conclusions in the orthologization mode. The graph layout process is also hard to automate. We use two layout algorithms (Dagre and Graphviz), but manual manipulations are often needed to make a clear layout, and Cytoscape (or other graph manipulation software) is desirable to make publication-ready images.

Despite the above-mentioned drawbacks, we find the here proposed method of complexity analysis informative as it successfully identifies known rearrangement hot spots (prophages, integrons et al.), and we hope that GCB with its capacity of visualization and complexity assessment will find its application in the area of comparative genomics studies.

# Conclusion

We have developed a novel tool, called Genome Complexity Browser (GCB), to analyze sets of genomes in order to quantify genome variability and visualize gene context variants. We use graph representation of genes neighbourhood to make visualizations and estimate local genome variability. GCB browser-based interface enables simultaneous analysis of genome variability profile and pattern of changes occurred in a particular locus.

We precalculated data for 143 prokaryotic organisms and web server gcb.rcpcm.org can be used to browse them. Command line tool and stand-alone server application make possible for user to analyze any particular set of genomes.

We observed that there are genome regions with high variability which have conservative localization in intraspecies and interspecies comparisons. Some of them are free of genes with identifiable mobility functions.

# Supporting information

**S1 Fig. Comparison of two methods to deal with the paralogs.** A) The

graph obtained with the default approach to ignore groups with several representatives in the particular genome; B) graph obtained with paralogs “orthologization” approach.

**S2 Fig. Phylogenetic tree of 327 E. coli strains with pdu and hmu operon presence information.** Red bars denote genomes in which complete gene set from the operon is present, green bars denote genomes in which more than the half of operon genes are present. A) Hmu operon is in good correspondence with the phylogenetic tree of *E. coli*; B) pdu operon presence is poorly correlated with the phylogenetic tree of *E. coli*.

**S3 Fig. Comparison of complexity profiles of 146 prokaryotic species.**

Complexity profiles are shown on the same scale for all organisms. Synteny blocks are shown in green color. The phylogenetic tree was built based on 16S rRNA sequence.

**S1 Text Simulations of genomes with predefined variability profiles.**

Details concerning method and results of genome simulations with predefined variability profile with further complexity analysis.

**S1 Listing. Algorithm to coalign genomes in the set.**

**S2 Listing. Algorithm to simulate genomes with predefined variability distribution.**

**S3 Listing. Algorithm to generate subgraph representing genome region of interest.**