RESEARCH ARTICLE

DNA Data Visualization (DDV): Software for Generating Web-Based Interfaces Supporting Navigation and Analysis of DNA Sequence Data of Entire Genomes

Tomasz Neugebauer[1]*, Eric Bordeleau[2], Vincent Burrus[2], Ryszard Brzezinski[2]

1 Concordia University Libraries, Montreal, Quebec, Canada, **2** Département de Biologie, Faculté des Sciences, Université de Sherbrooke, Sherbrooke, Quebec, Canada

* tomasz.neugebauer@concordia.ca

OPEN ACCESS

Citation: Neugebauer T, Bordeleau E, Burrus V, Brzezinski R (2015) DNA Data Visualization (DDV): Software for Generating Web-Based Interfaces Supporting Navigation and Analysis of DNA Sequence Data of Entire Genomes. PLoS ONE 10 (12): e0143615. doi:10.1371/journal.pone.0143615

Editor: Ruslan Kalendar, University of Helsinki, FINLAND

Received: June 17, 2015

Accepted: November 6, 2015

Published: December 4, 2015

Data Availability Statement: The source code for the software is available at GitHub (https://github.com/photomedia/DDV/). In addition, a dataset was archived on Zenodo (http://dx.doi.org/10.5281/zenodo.33608) which contains all of the files for the visualization interfaces that the authors generated and display at http://www.photomedia.ca/DDV/dnadata/.

Funding: Work at Concordia University was supported by Concordia University Libraries. Work at Sherbrooke University was supported by Discovery Grant and Discovery Acceleration Supplement from

Abstract

Data visualization methods are necessary during the exploration and analysis activities of an increasingly data-intensive scientific process. There are few existing visualization methods for raw nucleotide sequences of a whole genome or chromosome. Software for data visualization should allow the researchers to create accessible data visualization interfaces that can be exported and shared with others on the web. Herein, novel software developed for generating DNA data visualization interfaces is described. The software converts DNA data sets into images that are further processed as multi-scale images to be accessed through a web-based interface that supports zooming, panning and sequence fragment selection. Nucleotide composition frequencies and GC skew of a selected sequence segment can be obtained through the interface. The software was used to generate DNA data visualization of human and bacterial chromosomes. Examples of visually detectable features such as short and long direct repeats, long terminal repeats, mobile genetic elements, heterochromatic segments in microbial and human chromosomes, are presented. The software and its source code are available for download and further development. The visualization interfaces generated with the software allow for the immediate identification and observation of several types of sequence patterns in genomes of various sizes and origins. The visualization interfaces generated with the software are readily accessible through a web browser. This software is a useful research and teaching tool for genetics and structural genomics.

Introduction

Information visualization can amplify cognition by storing massive amounts of information in quickly accessible forms and using visual representations to enhance the detection of patterns [1].Visualization of genomic data augments reasoning and analysis by facilitating the

the Natural Sciences and Engineering Council of Canada [326810 and 412288] to VB and Discovery Grant 2014-05177 to RB. VB holds a Canada Research Chair in Bacterial Molecular Genetics. EB is the recipient of a Cystic Fibrosis Canada-Kin Canada Fellowship and a postdoctoral fellowship from Le Fonds de Recherche du Québec – Santé (FRQS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

complementing of computational methods with human interpretation [2]. Graphical representations of genomic data allow for rapid viewing and identification of characteristics of specific regions of the genome and new encoding patterns with biological significance.

In this paper, we present software that generates interactive, graphical representations of large nucleotide sequence data sets, accessible through an Internet browser. The method is scalable to large chromosomes such as those of *Homo sapiens* and generally speaking, the eukaryotic organisms.

Early attempts to visualize whole genome nucleotide sequences include the one presented by Makino *et al.* [3], who used printed representations on A4 paper to visualize complete nucleotide sequences. Makino *et al.* suggested that the regions in genomes that encode genes are more purine-rich than non-coding sequences. Thus, given their choice of red for A and black for G, Makino *et al.* stipulated that the "reddish/blackish stripes" that were visible on their printed visualization of the genomic sequence of *Mycoplasma pneumoniae* represent putative gene coding regions.

Yoshida *et al.* [4] chose the same four colors for visualizing nucleotides: blue for G, green for C, yellow for T and red for A. Yoshida *et al.* used visualization columns of variable width to locate tandem repeats in the *Escherichia coli* genome.

Seaman & Sanford visualization tool, Skittle [5], is a significant achievement and its literature review offers a critique of other prior efforts such as the DNA Rainbow project [6]. The DNA Rainbow project does not visualize the DNA data using the conventional FASTA data format consisting of less than 80 nucleotides per column; a large image without columns is used instead. The DNA Rainbow images are so large that they are practically difficult if not impossible to view and navigate using conventional technology such as a web browser. Skittle software is designed for a single user interacting with raw nucleotide data visualized with colors. Initially, the use of Skittle required downloading and installing the software on a desktop computer, with no option of exporting the visualization results to the web for collaboration and shared spaces. Skittle now also includes a beta of a web-based version of the application [7]. The implementation and approach to the visualization of the data is different in DDV from Skittle. Among notable differences, Skittle generates a single-column image based on portions of the data requested by the user, whereas DDV pre-generates all of the images required for visualizing the entire DNA data sequence as a multi-column, smoothly zoomable multi-scale image. The multi-column geometry of the visualizations makes it easier for users to visually parse long sequences of data while the use of multi-scale images allows for the integration of community supported web-based viewer technology.

The DNA Data Visualization (DDV) software described in this paper allows a user of a local desktop computer to generate visualization interfaces that can be exported to the web, increasing access and the potential for collaboration around the visualized data. Extending visualization tool functionality to support collaboration increases the scope and applicability of the visualizations. The method we propose in this work pre-generates all of the necessary images and offers an efficient way of navigating these images that display the genome at various levels of detail. It allows for the visualization of long raw nucleotide sequences through a display format that is optimized for access and interaction through the web. The zoom pyramid structure of pre-generated images in our method allows for smooth interaction and browsing of the data. The only requirements for accessing prepared visualizations are Internet access and a modern web browser. DDV includes functions for GC Skew plotting and for computing nucleotide composition density. The web output format of DDV leverages modular tools for biological data visualization and high-resolution image navigation. The functionality of easy selection of sequence fragments allows researchers to continue their analysis using tools external to DDV.

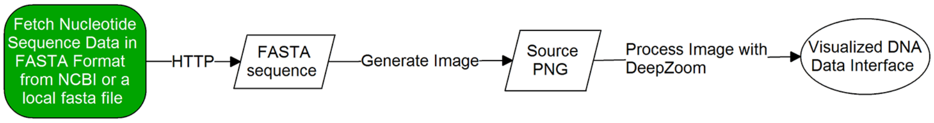

Fetch Nucleotide Sequence Data in FASTA Format from NCBI or a local fasta file
HTTP
FASTA sequence
Generate Image
Source PNG
Process Image with DeepZoom
Visualized DNA Data Interface

Fig 1. Summary of the visualization method. A source PNG image is generated to represent FASTA sequences. These source PNG images are then processed, including the creation of a Deep Zoom image pyramid. The resultant visualized interface includes scripts that compute nucleotide composition density and generate a GC skew graph.

doi:10.1371/journal.pone.0143615.g001

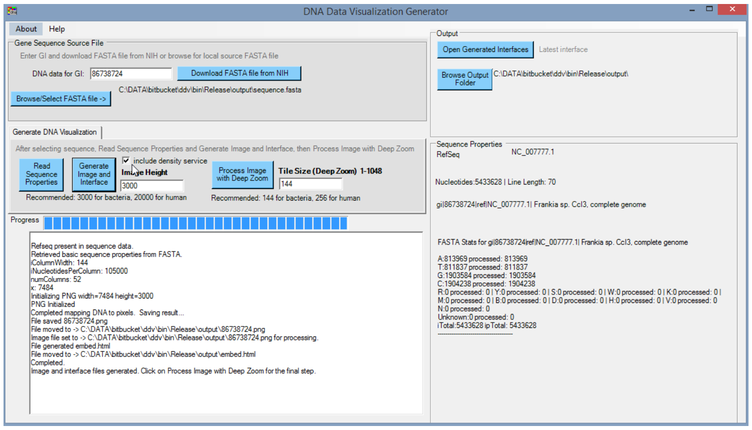
The method creates visualizations that allow for the combination of seamless graphical human inspection and automated computation, an approach that is particularly effective.

Materials and Methods

Visualization method summary

The method is summarized in Fig 1. The first step in the visualization is to download the FASTA formatted nucleotide sequence. A screenshot of the DNA Data Visualization generator (DDV) during operation is shown in Fig 2. There are buttons on the DDV interface for downloading sequence data from National Center for Biotechnology Information (NCBI)'s nucleotide database, given a GI number as input, or the specification of a local FASTA file downloaded manually from another source. These nucleotide sequence data are then processed by clicking on buttons on the interface generator that correspond to the main steps in the algorithm for generating the visualization:

1. Read Sequence Properties: Read name, length and sequence identifiers such as RefSeq and GI numbers,
2. Generate Image and Interface:
    a. Initialize an RGB bitmap of appropriate size for the visualization–this depends on the sequence size.
    b. Process all of the nucleotides in the sequence by painting colors for each nucleotide on the image, generating a master image of large size. The design of this image is illustrated in Fig 3.


DNA Data Visualization Generator
About
Help
Gene Sequence Source File
Download FASTA file from NIH
Browse/Select FASTA file ->
Generate DNA Visualization
Read Sequence Properties
Generate Image and Interface
Process Image with Deep Zoom
Progress
Output
Open Generated Interfaces
Latest interface
Browse Output Folder
Sequence Properties

Enter GI and download FASTA file from NIH or browse for local source FASTA file

DNA data for GI: 86738724

C:\DATA\bitbucket\ddv\bin\Release\output\sequence.fasta

After selecting sequence, Read Sequence Properties and Generate Image and Interface, then Process Image with Deep Zoom

☑ include density service

Image Height: 3000

Recommended: 3000 for bacteria, 20000 for human

Tile Size (Deep Zoom) 1-1048: 144

Recommended: 144 for bacteria, 256 for human

```
Refseq present in sequence data.
Retrieved basic sequence properties from FASTA.
iColumnWidth: 144
iNucleotidesPerColumn: 105000
numColumns: 52
x: 7484
Initializing PNG width=7484 height=3000
PNG Initialized
Completed mapping DNA to pixels.  Saving result...
File saved 86738724.png
File moved to -> C:\DATA\bitbucket\ddv\bin\Release\output\86738724.png
Image file set to -> C:\DATA\bitbucket\ddv\bin\Release\output\86738724.png for processing.
File generated embed.html
File moved to -> C:\DATA\bitbucket\ddv\bin\Release\output\embed.html
Completed.
Image and interface files generated. Click on Process Image with Deep Zoom for the final step.
```

C:\DATA\bitbucket\ddv\bin\Release\output\

RefSeq NC_007777.1

Nucleotides:5433628 | Line Length: 70

gi|86738724|ref|NC_007777.1| Frankia sp. CcI3, complete genome

```
FASTA Stats for gi|86738724|ref|NC_007777.1| Frankia sp. CcI3, complete genome

A:813969 processed: 813969
T:811837 processed: 811837
G:1903584 processed: 1903584
C:1904238 processed: 1904238
R:0 processed: 0 | Y:0 processed: 0 | S:0 processed: 0 | W:0 processed: 0 | K:0 processed: 0 |
M:0 processed: 0 | B:0 processed: 0 | D:0 processed: 0 | H:0 processed: 0 | V:0 processed: 0
N:0 processed: 0
Unknown:0 processed: 0
iTotal:5433628 ipTotal: 5433628
----------------------------------------
```

Fig 2. Screenshot of DNA Data Visualization generator (DDV) after it generates the source PNG for Frankia sp. Ccl3 chromosome [GenBank: NC_007777]. User downloads/selects FASTA data file, selects the image height, and generates the source PNG image. The final step is to click on "Process Image with Deep Zoom" to complete the generation of a visualization interface for this dataset.

doi:10.1371/journal.pone.0143615.g002

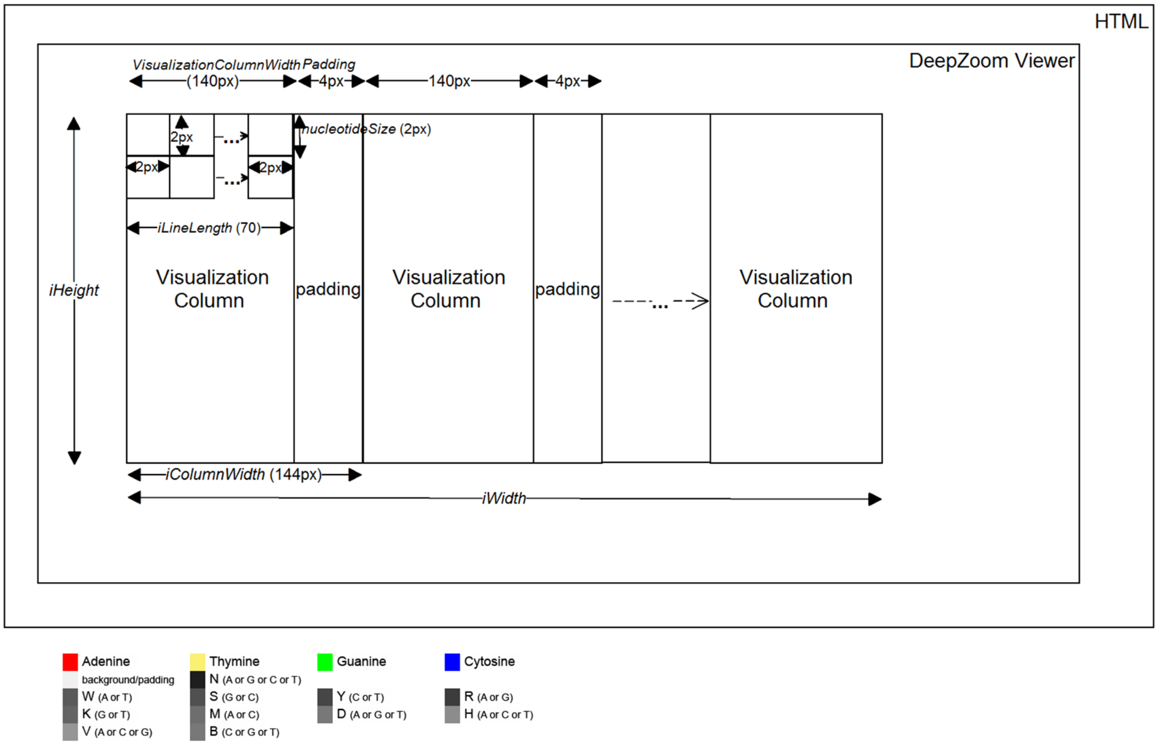


Fig 3. Source PNG image design. Each nucleotide is 2px X 2px, with 70 nucleotides per line. The height (iHeight) is set to 3000 px for bacterial genomes, and the value can be increased for larger data sets. The total width (iWidth) depends on length of the data. Each visualization column is separated by 4px of grey padding.

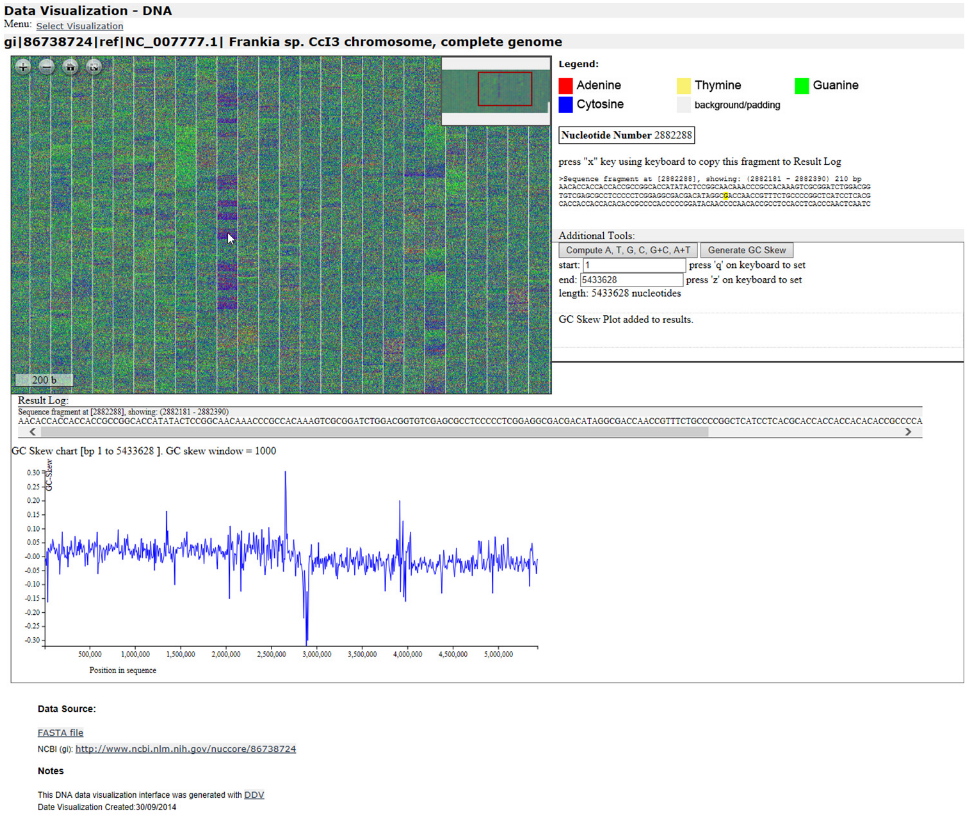
doi:10.1371/journal.pone.0143615.g003

3. Process Image with Deep Zoom:
    a. Generate the Deep Zoom Images (DZI) pyramid and XML from the master image using DeepZoomTools.dll [5].
    b. Generate the HTML, CSS and JavaScript files for the completed DNA Data Visualization interface (Fig 4).

DNA Data Interface Navigation

The end user interacts with the generated visualization through a web browser, using the navigation buttons (zoom in, zoom out, full screen, home) in the top left corner or the visual navigator in the top right corner. A screenshot example of a generated DNA data set interface for a bacterial genome [GenBank: NC_007777] is presented in Fig 4. The user can zoom in and out with the mouse wheel or navigation buttons. A scale bar on the bottom left shows length in bp units while a viewport navigator shows zoom position and can also be used for panning.

While pointing at a particular nucleotide on the visualization with the mouse pointer, the surrounding 210 bp sequence fragment is displayed as text on the interface. This sequence fragment can also be copied to the results by pressing the “x” key. Pressing the ‘q’ and ‘z’ keys on the keyboard marks the beginning and end of the portion of the sequence that can be sent for computation of % G+C and for the determination of the respective frequencies of the four nucleotides. The end user can also request a GC Skew plot for the sequence.



Data Visualization - DNA

Menu: Select Visualization

gi|86738724|ref|NC_007777.1| Frankia sp. CcI3 chromosome, complete genome

Legend:

🟥 Adenine 🟨 Thymine 🟩 Guanine
🟦 Cytosine ⬜ background/padding

Nucleotide Number 2882288

press "x" key using keyboard to copy this fragment to Result Log

```
>Sequence fragment at [2882288], showing: (2882181 - 2882390) 210 bp
AACACCACCACCACCGCCGGCACCATATACTCCGGCAACAAACCCGCCACAAAGTCGCGGATCTGGACGG
TGTCGAGCGCCTCCCCCTCGGAGGCGACGACATAGGCGACCAACCGTTTCTGCCCCGGCTCATCCTCACG
CACCACCACCACACACCGCCCCACCCCCGGATACAACCCCAACACCGCCTCCACCTCACCCAACTCAATC
```

Additional Tools:

Compute A, T, G, C, G+C, A+T | Generate GC Skew

start: 1 press 'q' on keyboard to set
end: 5433628 press 'z' on keyboard to set
length: 5433628 nucleotides

GC Skew Plot added to results.

Result Log:

Sequence fragment at [2882288], showing: (2882181 - 2882390)
AACACCACCACCACCGCCGGCACCATATACTCCGGCAACAAACCCGCCACAAAGTCGCGGATCTGGACGGTGTCGAGCGCCTCCCCCTCGGAGGCGACGACATAGGCGACCAACCGTTTCTGCCCCGGCTCATCCTCACGCACCACCACCACACACCGCCCCA

GC Skew chart [bp 1 to 5433628]. GC skew window = 1000

Data Source:

FASTA file

NCBI (gi): http://www.ncbi.nlm.nih.gov/nuccore/86738724

Notes

This DNA data visualization interface was generated with DDV
Date Visualization Created:30/09/2014

Fig 4. Screenshot of generated interface for visualized data set: *Frankia* sp. CcI3 genome [GenBank: NC_007777], 5433628 bp. Screenshot taken after computing nucleotide density for the whole sequence, zooming in, generating GC-Skew plot for the sequence, and selecting a sequence fragment at bp 2882288. The interface includes a scale bar (bottom left of Deep Zoom viewer) shown in bp units, and a viewport navigator (top right of Deep Zoom viewer) that shows zoom position and can also be used for panning.

doi:10.1371/journal.pone.0143615.g004

Implementation Details

The full implementation is summarized in Fig 5. The DNA Data Visualization generator (DDV) is implemented in Visual Studio C# and compiled with. Net Framework 4. This application and all of its dependencies are available for download. Running DDV requires an operating system that supports. NET Framework 4, such as Windows Vista, Windows 7 or 8. The generated visualizations can be viewed on any operating system that supports a web browser with JavaScript. During initialization, DDV attempts to check if the correct version of. NET is installed, displaying a message directing the user to the free web installer [8] when necessary. DDV uses the Microsoft DeepZoomTools.dll in step 3.1 to generate the image tiles. DDV-generated visualizations are placed under the output folder which can then be placed on a web server for sharing and collaboration. DDV places all of the shared files into the root "output" folder, and the sequence-specific files in subfolders under "dnadata".

Navigation of the deep zoom image is implemented with the open source OpenSeadragon JavaScript library [9]. OpenSeadragon has an active development community and features compatibility with desktop and mobile devices. In addition, BioJS [10] is used for sequence fragment display and D3.js [11] for GC Skew visualization. DDV also includes the minimal Civetweb web server [12] with PHP [13], ensuring that the generated visualizations are able to

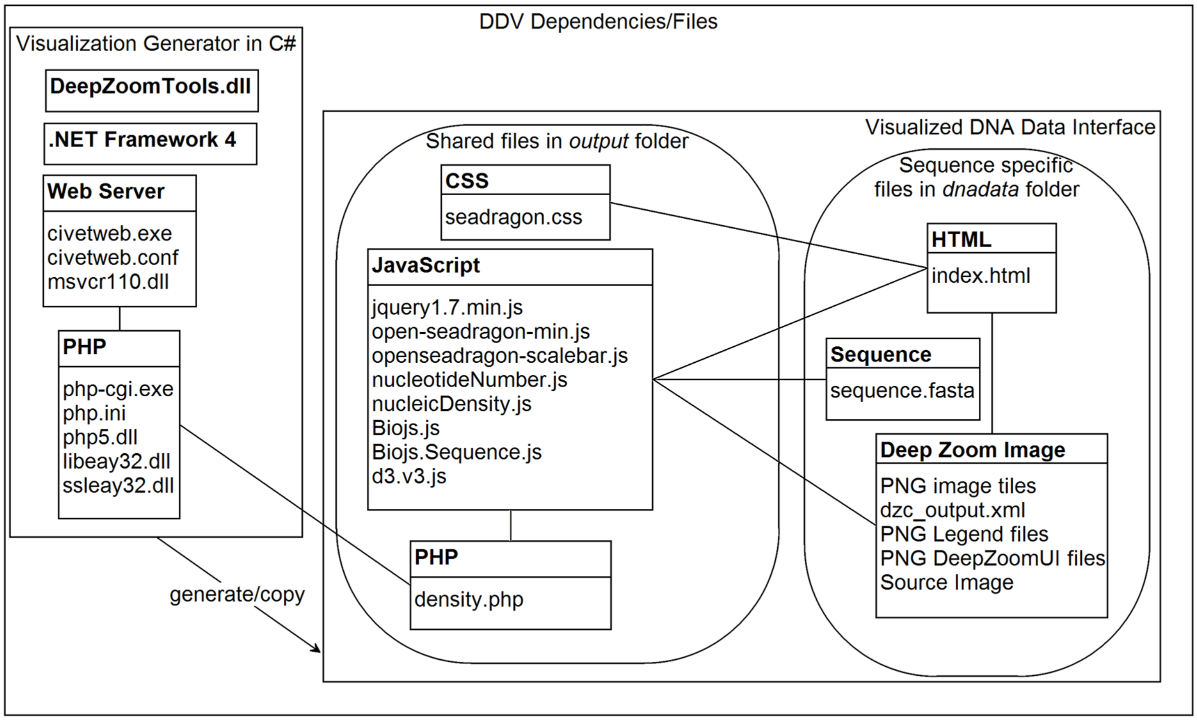


Fig 5. DDV generator dependencies and visualized DNA Data interface files. DDV is a C# application that generates sequence specific files for each DNA dataset, as well as the shared CSS, JavaScript, and PHP files in the output folder. The included civetweb serves the PHP files when the user is viewing the visualized interface on a local PC.

doi:10.1371/journal.pone.0143615.g005

use PHP to compute nucleotide composition density from a Windows desktop computer as well as a hosted web server when the visualizations are placed online.

The images generated in step 2 are very large, as they contain the information from the entire chromosome in one image. For example, the image for *Clostridium difficile* 630 [GenBank: NC_009089] bacterial chromosome is 5 900px X 3 000px while the representation of the longest chromosome 1 [GenBank: NC_000003] of *H*. *sapiens* is 50 972px X 20 000px. Such large images would be difficult to display on the web without further processing. The geometric dimensions of these images and the colors used to represent the nucleotides are shown in Fig 3. The three additive primary colors: red, green, blue and the fourth color, yellow, are used to represent the four nucleotides. Light grey is used as the background color, black and dark shades of grey are used to represent various possible coded 'unknowns' in the FASTA format: N, W, S, Y, R, K, M, D, H, V, B. The legend on generated visualizations shows only those unknowns that actually appear in the sequence.

In Step 3, for bacterial chromosomes, tile size of 144px is selected so that at 1:1 level of magnification, each tile contains approximately 5 040 contiguous nucleotides (70 nucleotides/line × 72 lines = 5 040 nucleotides). The image pyramid is used by the web interface to offer the multiple views at various levels of magnification. This step creates thousands of small tiles at various levels of magnification which are loaded on demand while the user navigates the image. The *C*. *difficile* 630 [GenBank: NC_009089] bacterial chromosome (4 290 252 nt) is mapped to a source PNG image with dimensions 5 900px by 3 000px, which results in a Deep Zoom Image with 15 levels of magnification, ranging from 1 tile in levels 0 to 8, and gradually increasing to 861 tiles arranged in 41 columns by 21 rows at the highest level 14. In total, when users are navigating the *C*. *difficile* 630 visualization, they are actually navigating 1 195 image tiles. The second example, *H. sapiens* chromosome 1 [GenBank: NC_000003], the largest DNA molecule to which the method has been applied so far (247 249 719 nt) has the source image dimensions of 50 972px by 20 000px. Due to its much larger size, a tile size of 256px was adopted. These parameters result in a Deep Zoom folder/image structure with 17 levels of magnification, ranging from 1 image tile in levels 0 to 8, increasing to 15 800 image tiles arranged in 200 columns by 79 rows at the highest level 16. Therefore, when users are navigating the human chromosome 1 visualization, they are actually navigating a total of 21 155 image tiles.

In Step 3.2, the nucleotideNumber.js JavaScript converts the position currently pointed by the user on the Deep Zoom viewer into the corresponding nucleotide number on the sequence. This is computed based on properties of the coordinate system of Deep Zoom viewport as well as the geometry of the source PNG image (Fig 3), using the following formula:

$$\begin{aligned} Nucleotide = iLineLength \times \left(\frac{\mathrm{iWidth}}{\mathrm{nucleotideSize}}\right) \times \left\lfloor\frac{\mathrm{x} \ \times \ \mathrm{iWidth}}{\mathrm{nucleotideSize}}\right\rfloor + iLineLength \\ \times \left\lfloor\frac{\mathrm{y} \ \times \ \mathrm{iWidth}}{\mathrm{nucleotideSize}}\right\rfloor + \left\lfloor\frac{(\mathrm{x} \ \times \ \mathrm{iWidth}) \ \mathrm{mod} \ \mathrm{ColumnWidth}}{\mathrm{nucleotideSize}}\right\rfloor + 1 \end{aligned}$$

where iWidth is the width of the source PNG (pixels), iHeight is the height of the source PNG (pixels), nucleotideSize is the width of 1 nucleotide square on the source image (2 pixels), iLineLength is the number of nucleotides per line in column (70 nucleotides), VisualizationColumnWidth is the width of one visualization column (140 pixels), Padding is the separation between visualization columns (4 pixels), ColumnWidth is the sum of VisualizationColumnWidth and Padding (144 pixels), while (x, y) are the coordinates of the cursor position on the page minus the position of the viewer. In addition, the user is pointing at a nucleotide on the image as opposed to background or padding if and only if the following four conditions are true:

$$\left\{ \begin{array}{l} 0 < x < 1 \\ 0 < y < \left(\dfrac{iHeight}{iWidth}\right) \\ Nucleotide \leq totalNucleotides \\ VisualizationColumnWidth \leq ((x \times iWidth)\,\mathrm{mod} ColumnWidth) \leq ColumnWidth \end{array} \right\}$$

where totalNucleotides is the total number of nucleotides in the sequence.

%G+C nucleic acid composition computation. The generated DNA Data Visualization Interface includes a PHP implemented function for computing %G+C and exact nucleic acid composition density of a selected portion of the visualized sequence, and to displays the computed results on the interface.

GC Skew Graph. The generated DNA Data Visualization Interface includes a JavaScript function for computing GC Skew data and plotting it on the interface using D3.js [11] library. The skew window is set depending on the number of nucleotides in the sequence, ranging from 50 for sequences of less than 100 000 nucleotides to 10 000 for sequences longer than 10 000 000 nucleotides.

Partial sequence data display and select. The generated DNA Data Visualization Interface includes the ability to display, select and copy 210 bp portions of the sequence currently pointed at on the interface. This is implemented with the help of BioJS [10]. A simple copy-paste operation allows for sending the selected sequence to external web services for further analysis.

Relative scale bar display. The generated visualizations also leverage OpenSeadragon's scale bar. The bar is customized so that it shows the relationship between width on the image and the number of nucleotides, as the user zooms in and out.

Results and Discussion

The DNA Data Visualization generator (DDV), as well as its source code, are available for download. All of the visualizations generated with this software and discussed below are also accessible with a web browser. As discussed below, the visualization method presented in this work allows immediate identification and observation of several types of sequence patterns in genomes of various sizes and origins, often not easily identified by other methods. The GenBank sequence files used to generate visualizations presented on Figs 4 and 6–11, along with URLs of the corresponding generated visualizations are listed in Table 1.

Compositional asymmetry of DNA strands

Many bacterial genomes show a deviation from the classical base composition in DNA strands, [A] = [T] and [G] = [C], in a DNA replication-dependent way. Thus, G > C bias is observed in the leading strand which is replicated co-directionally with the replication fork, while C > G bias is observed in the lagging strand [14]. The bias is changing at the points of origin and termination of the DNA replication. A visually expressive example of this tendency is the main

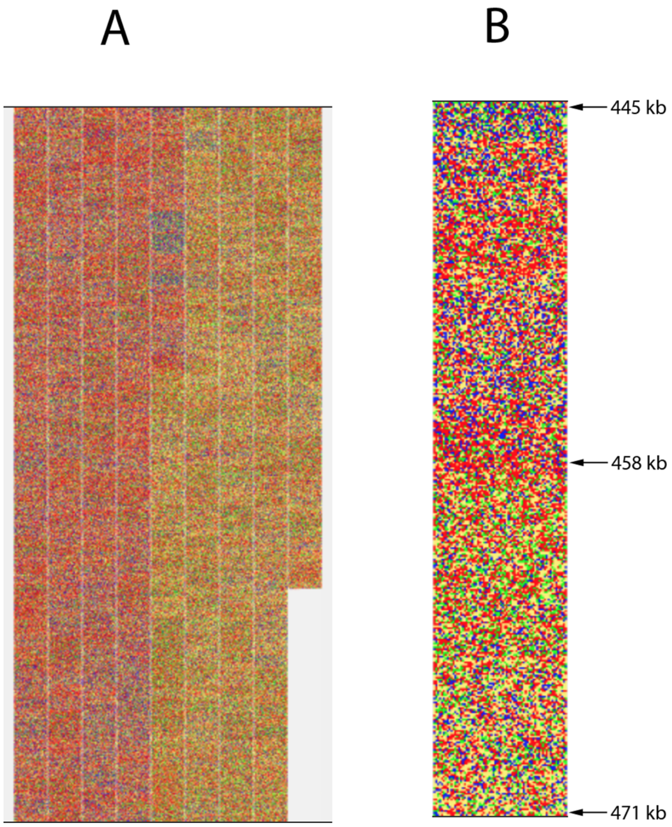


Fig 6. Visualization of the major linear chromosome of *Borrelia burgdorferi* B31 [GenBank: NC_001318]. A) Whole linear chromosome of 910724 bp; B) enlargement of the segment surrounding the origin of replication (localized between coordinates 458036 and 458227). Arrows indicate the distance from the beginning of the chromosome sequence.

doi:10.1371/journal.pone.0143615.g006

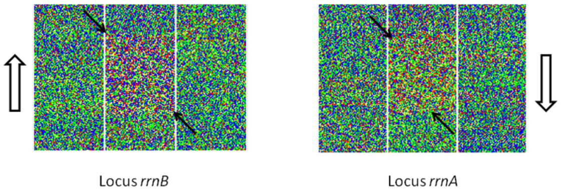


Fig 7. Ribosomal RNA gene clusters. Zoomed-in fragments of the visualization of the *Streptomyces coelicolor* A3(2) [GenBank: NC_003888] linear chromosome showing the *rrnB* and *rrnA* ribosomal gene clusters (coordinates 1916451 to 1921599 and 4530650 to 4535576). Arrows indicate the limits of each cluster. Empty arrows indicate the direction of chromosomal DNA replication fork movement and the direction of rRNA genes transcription.

doi:10.1371/journal.pone.0143615.g007

component of the genome of *Borrelia burgdorferi* [GenBank: NC_001318], the Lyme disease causing agent. It consists of a linear chromosome of almost 1 M base pairs [15]. This chromosome is replicated bi-directionally from an origin localized close to the center. The visualization of the chromosome by our method allows immediate identification of the position of the putative origin of replication on the chromosome in the segment between 458 200 and 458 400 (A and B in Fig 6). Using the tools provided with DDV software, calculation of base frequency showed that both the left and the right part of the genome have almost identical values for [A +T] (71%) and [G+C] (29%). However the GC skew, estimated from the ratio (C-G)/(C+G), as suggested by Lobry [16], switches from positive (0.211) to negative (-0.217) when calculated

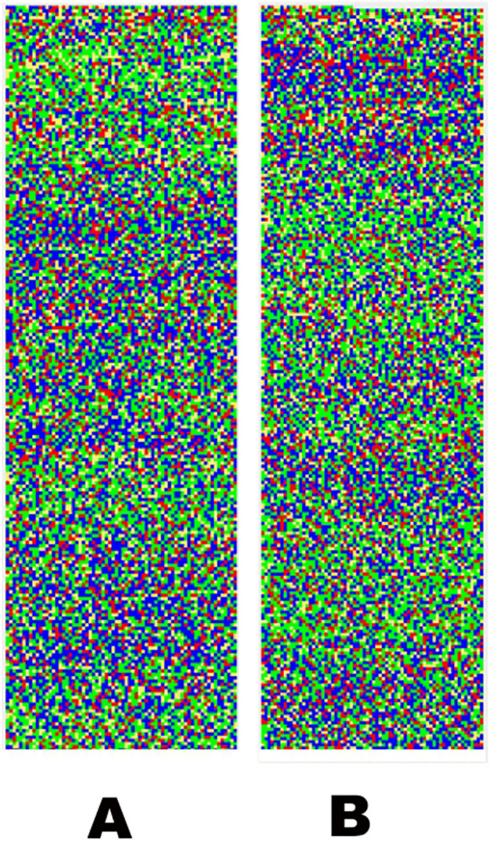


Fig 8. Visualization of the 16.5 kb-terminal segments of the *Streptomyces davawensis* JCM 4913 chromosome [GenBank: HE971709]. The segments are parts of the 33.3 kb LTIRs of this genome. A: image of the left end of the chromosome; B: 180°-rotated image of the right end of the chromosome.

doi:10.1371/journal.pone.0143615.g008

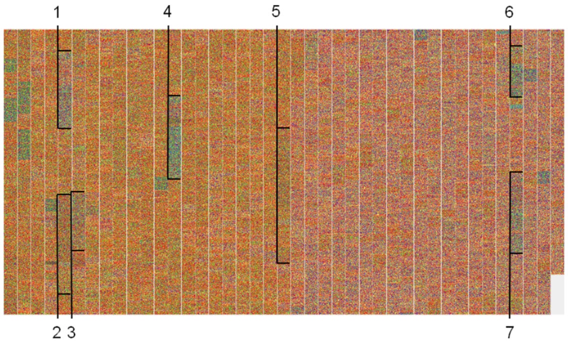


Fig 9. Visualization of the *Clostridium difficile* 630 chromosome [GenBank: NC_009089]. Segments corresponding to the seven known conjugative transposons are indicated by brackets. 1: CTn1; 2: CTn2; 3: CTn3, also known as Tn*5397*; 4: CTn 4; 5: CTn5; 6: CTn6; 7:CTn7.

doi:10.1371/journal.pone.0143615.g009

for 10kb window localized, respectively, upstream and downstream from the putative origin of replication (B in Fig 6), confirming that the position of the origin of replication has been correctly identified.

Ribosomal RNA gene clusters in G+C-rich actinobacterial genomes

Actinobacteria form a branch of Gram+ bacteria with G+C-rich genomes. As a consequence, the coding sequences of protein-encoding genes show a particular codon usage pattern maximizing the use of G or C in the third position of degenerate triplets. In some genes, the third codon position G+C content can reach 98.3% [17, 18]. However, due to functional constraints, the ribosomal RNA gene clusters are relatively A+T-rich [19, 20]. The genome of the model actinomycete, *Streptomyces coelicolor* A3(2) [GenBank: NC_003888], includes six ribosomal RNA gene clusters, named *rrnA–F* [21]. All these clusters are clearly distinguishable on our visualization and two of them are shown on Fig 7. While the overall G+C ratio of the *S. coelicolor* genome is of 72.1%, it decreases to about 57% in rRNA gene clusters. All these clusters respect the rule saying that highly expressed genes are transcribed in the same direction as the movement of the DNA replication fork of the chromosome [22, 23]. Accordingly, a difference in GC skew is observed depending on the localization of the rRNA gene clusters relative to the DNA replication origin (~4 271 kb from the left end of the chromosome). While the global A+T and G+C proportions are similar for both clusters shown on Fig 7 (42.5% and 57.5% respectively), the (C-G)/(C+G) ratio is positive for *rrnB* but negative for *rrnA*, reflecting their respective positions on both sides of the origin of replication. On the visualization, this is translated into more abundant blue pixels for *rrnB* and green pixels for *rrnA*.

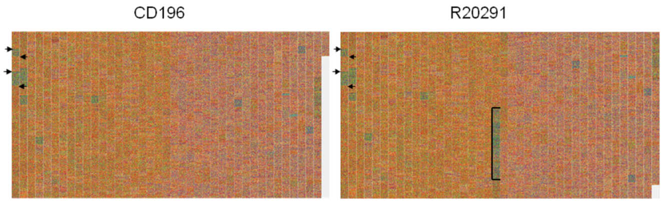


Fig 10. Comparison of the chromosomes of *Clostridium difficile* strains CD196 [GenBank: NC_013315] and R20291 [GenBank: NC_013316]. The region comprising transposons Tn*6104*, Tn*6105* and Tn*6106* is indicated by brackets. The two first corresponding RNA gene clusters are indicated by arrows for both strains.

doi:10.1371/journal.pone.0143615.g010

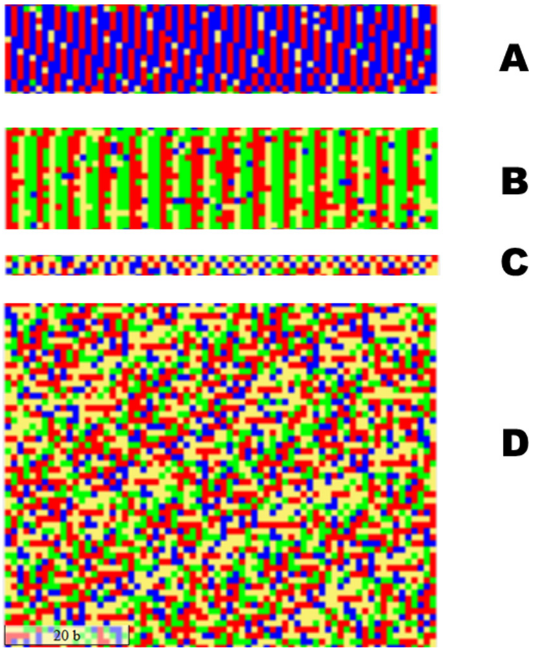


Fig 11. Examples of visualizations of tandem repeats in human chromosome 21 [GenBank: NC_000021.9]. The scale represents a segment of 20 nucleotides. A: microsatellite consisting of CA repeats (nucleotide coordinates 6565371–6566350); B: GGAAT tandem repeats, also known as DNA satellite III (7259491–7260680); C: imperfect [TCTA][TCTG] 4bp repeat in the D21S11 locus included in CODIS database (19181961–19182170). D: 171-base pairs repeat, also known as alphoid DNA satellite or α-satellite (7265231–7269080).

doi:10.1371/journal.pone.0143615.g011

Long terminal inverted repeats in linear chromosomes

Bacterial linear plasmids and chromosomes often include inverted repeat sequences at their ends, sometimes longer than 1 000 kb [24]. They are typically present in chromosomes of the members of the order *Actinomycetales* [14, 18, 21, 24, 25]. As an example, the recently sequenced genome of *Streptomyces davawensis* JCM 4913 [GenBank: HE971709] has long terminal inverted repeats (LTIRs) of 33.3 kb each [26]. While the G+C content of the LTIRs (69.0%) is similar to that of the entire chromosome (70.6%), the respective G and C density varies along the LTIR, what is reflected by locally more dense blue or green color on the visualization (Fig 8). Furthermore, when showed side-by-side with one segment rotated at 180° to the other, the visualized LTIRs reveal well aligned areas corresponding to tracts of higher density of complementary nucleotides (blue as opposed to green; Fig 8), as expected for inverted repeat sequences.

Horizontal gene transfer events

Bacterial genome evolution is extensively driven by horizontal gene transfer [27]. Large DNA fragments (up to 600kb) can be acquired by conjugation, transformation and transduction and integrated into the native chromosome. This foreign DNA typically has a G+C composition that is different from the host chromosome. Therefore, such mobile elements can be discovered using our method by examining a single bacterial chromosome. As an example, *C. difficile* 630 [GenBank: NC_009089] (G+C ratio of 29.1%), the leading cause of hospital-acquired diarrhea, harbors seven conjugative transposons (G+C ratio of 32.7 to 42.3%) [28, 29, 30]. All seven mobile elements were identified at first sight with accuracy (rarely more than one CDS apart)

Table 1. Source data files used to generate visualizations presented in Figs 4 and 6–11, along with URLs of the corresponding generated visualizations.

Figure	GenBank source FASTA file	Sequence Reference	Visualization
4	http://www.ncbi.nlm.nih.gov/nuccore/86738724?report=fasta	[GenBank: NC_007777]	http://www.photomedia.ca/DDV/dnadata/nuccore86738724/
6	http://www.ncbi.nlm.nih.gov/nuccore/15594346?report=fasta	[GenBank: NC_001318]	http://www.photomedia.ca/DDV/dnadata/nuccore15594346/
7	http://www.ncbi.nlm.nih.gov/nuccore/32141095?report=fasta	[GenBank: NC_003888]	http://www.photomedia.ca/DDV/dnadata/nuccore32141095/
8	http://www.ncbi.nlm.nih.gov/nuccore/408526205?report=fasta	[GenBank: HE971709]	http://www.photomedia.ca/DDV/dnadata/nuccore408526205/
9	http://www.ncbi.nlm.nih.gov/nuccore/126697566?report=fasta	[GenBank: NC_009089]	http://www.photomedia.ca/DDV/dnadata/nuccore126697566/
10	http://www.ncbi.nlm.nih.gov/nuccore/260681769?report=fasta	[GenBank: NC_013315]	http://www.photomedia.ca/DDV/dnadata/nuccore260681769/
10	http://www.ncbi.nlm.nih.gov/nuccore/260685375?report=fasta	[GenBank: NC_013316]	http://www.photomedia.ca/DDV/dnadata/nuccore260685375/
11	http://www.ncbi.nlm.nih.gov/nuccore/224589813?report=fasta	[GenBank: NC_000021.9]	http://www.photomedia.ca/DDV/dnadata/nuccore568815577/

doi:10.1371/journal.pone.0143615.t001

(Fig 9). However, other mobile elements such as the likely mobilizable transposon Tn*5398* and prophages 1 and 2 were not identified owing to their G+C content similar to the chromosome [28, 31]. Moreover, comparison of two bacterial chromosomes allows for rapid identification of recent mobile element acquisition events. Visualization of the two closely related "hypervirulent" ribotype 027 *C. difficile* strains CD196 [GenBank: NC_013315] and R20291 [GenBank: NC_013316] (Fig 10) reveals the acquisition of the three transposon Tn*6104*, Tn*6105* and Tn*6106* carried on the conjugative transposon Tn*6103* in R20291, as previously reported [29].

Finally, echoing the observations aforementioned, ribosomal RNA gene clusters are also visible, because of their relatively high G+C ratio (~50%) compared to the rest of the entire chromosome (29.1%) (Fig 10).

Human chromosomes

The 22 human autosomes as well as the two sexual chromosomes X and Y have been visualized using our method, demonstrating the scalability of the method to very large data sets. The black segments correspond mainly to heterochromatic areas of the chromosomes that are occupied by extensive tandem repeats where sequence has not been determined in detail [32] and, at least at present, constitute gaps in the genomic sequence. The determination of the sequences covered by the gaps is an active area of research [33].

It was immediately apparent that this visualization method makes it relatively easy to quickly identify regions of variable nucleotide composition density. As the image is zoomed out, the nucleotide composition information which is encoded as color is compressed through graphical algorithms that scale down images by averaging the pixel color by considering neighboring pixels. The result is that areas of high G+C concentration become even more apparent, as a combination of blue, green and cyan (sum of the two) as the user zooms out. In a similar way, the higher concentrations of A+T appear as red, yellow and orange respectively.

Tandem repeats can also be observed. Yoshoida *et al.* [4] found that the width of the visualization column does not have to be the same as the period of the repeat, an observation which is confirmed in our visualization of human chromosomes. Repeats of variable width are visible to the eye even when their period does not match the set width of 70 nucleotides per column in our visualization. Examples of dinucleotidic (A), pentanucleotidic (B) and 171bp-long repeats

(D) are shown on Fig 11. Visualization of the entire chromosome was based on the genomic contig GRCh38 primary assembly [GenBank:NC_000021.9]. Areas with tandem repeats were visually identified by their regular patterns, extracted from the genomic sequence and analyzed with the Tandem Repeats Finder program [34]. The imperfect 4bp repeats (C in Fig 11) represent the D21S11 locus included in the CODIS database for forensic applications [35].

Future Development

There are many web services available online that could be integrated with DDV, such as National Library of Medicine's BLAST web service [36]. The requirement that FASTA files accepted by DDV have 70 nucleotides per line stems from the fact that this is the common and default FASTA format returned by NCBI's eFetch. This requirement will be broadened to accept more formats with future development of the FASTA parser functionality in DDV. Leveraging the integration of an additional C# bioinformatics library such as .NET Bio [37] into DDV is a promising development strategy that could be used for this purpose.

DDV currently requires Windows operating system for generating visualizations, so the development of Unix and Mac OS versions of the generating software are among current development plans. Microsoft recently announced the release of .NET core as open source and cross platform [38], which simplifies the porting of DDV to other platforms and its long term sustainability. DDV is itself free and open source, and it is dependent almost entirely on open source components. DDV uses the DeepZoomTools.dll that is redistributable, but currently not open source. However, while DDV is ported to other operating systems in the future, this dll can be replaced with one of the alternative tools for the creation of DZI images, such as the free open source VIPS [39].

Conclusions

We present a novel method for generating visual representations of nucleotide sequences. The method presented is especially practical for visualizing and navigating the DNA sequence data of whole genomes or chromosomes. We confirmed that the visualizations generated allow for the immediate identification and observation of several types of sequence patterns. This software is capable of generating interactive graphical representations of large nucleotide sequence data sets that are accessible through a web browser. In generating the visualization of *H. sapiens'* DNA data, we have also shown that this method scales to large data sets.

Availability and Requirements

- **Project name:** DNA Data Visualization (DDV)
- **Project home page:** http://www.photomedia.ca/DDV/
- **Source code:** https://github.com/photomedia/DDV
- **Generated interfaces dataset:** http://dx.doi.org/10.5281/zenodo.33608
- **Operating system(s):** Windows
- **Programming languages:** C#, JavaScript, PHP
- **Other requirements:** .NET Framework 4.0 or higher
- **License:** BSD 3-Clause https://github.com/photomedia/DDV/blob/master/DDV-license.txt
- **Any restrictions to use by non-academics:** none

Producing visualizations with DDV requires a Windows operating system with .NET Framework version 4 or higher. However, any modern browser that is capable of supporting JavaScript is sufficient for end users to access and use the generated visualizations. This was tested on various browsers, including Safari (tested on version 6), Chrome (tested on version 42), Firefox (tested on version 18 and higher), Internet Explorer (tested on version 8 and 9); running on different operating systems such as Mac OS X, Windows Vista, Windows 7, Windows 8, Ubuntu and Android.

Acknowledgments

Work at Concordia University was supported by Concordia University Libraries. Work at Sherbrooke University was supported by Discovery Grant and Discovery Acceleration Supplement from the Natural Sciences and Engineering Council of Canada [326810 and 412288] to VB and Discovery Grant 2014–05177 to RB. VB holds a Canada Research Chair in Bacterial Molecular Genetics. EB is the recipient of a Cystic Fibrosis Canada-Kin Canada Fellowship and a postdoctoral fellowship from Le Fonds de Recherche du Québec–Santé (FRQS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the experiments: TN EB VB RB. Performed the experiments: TN EB VB RB. Analyzed the data: TN EB VB RB. Contributed reagents/materials/analysis tools: TN. Wrote the paper: TN EB VB RB. Conceived, designed and implemented the visualization method, developed the software, generated visualizations: TN. Applied the visualization technique to the analysis of bacterial and human chromosomes, suggested improvements to the visualization interface: EB VB RB.

References

1. Isenberg P, Elmqvist N, Scholtz J, Cernea D, Ma KL, Hagen H. Collaborative visualization: definition, challenges, and research agenda. Inf Vis. 2011; 10:310–326. doi: 10.1177/1473871611412817
2. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T. Visualizing genomes: techniques and challenges. Nat Methods. 2010; 7:S5–S15. doi: 10.1038/nmeth.1422 PMID: 20195257
3. Makino S, Naoki A, Suzuki M. Visual presentation of complete genomic DNA sequences, and its application to identification of gene-coding regions. Proc Japan Acad. 1999; 75(10): Ser. B. doi: 10.2183/pjab.75.311
4. Yoshida T, Obata N, Oosawa K. Color-coding reveals tandem repeats in the *Escherichia coli* genome. J Mol Biol. 2000; 298: 343–349. doi: 10.1006/jmbi.2000.3667 PMID: 10772854
5. Seaman JD, Sanford JC. Skittle: a 2-dimensional genome visualization tool. BMC Bioinformatics. 2009; 10:452. doi: 10.1186/1471-2105-10-452 PMID: 20042093
6. DNA rainbow [Internet]. Available: http://www.dna-rainbow.org
7. Dnaskittle [Internet]. Available: http://dnaskittle.com/
8. Microsoft .NET Framework 4 (Web Installer) [Internet]. Available: http://www.microsoft.com/en-ca/download/details.aspx?id=17851
9. OpenSeadragon [Internet]. Available: http://openseadragon.github.io/
10. BioJS library of JavaScript components to represent biological data [Internet]. Available: https://github.com/biojs/biojs
11. D3.js–Data Driven Documents [Internet]. Available: http://d3js.org/
12. Civetweb embedded c/c++ web server [Internet]. Available: https://github.com/sunsetbrew/civetweb
13. PHP [Internet]. The PHP Group. Available: http://www.php.net/

14. Képès F, Jester BC, Lepage T, Rafiei N, Rosu B, Junier I. The layout of a bacterial genome. FEBS Lett. 2012; 586:2043–2048. doi: 10.1016/j.febslet.2012.03.051 PMID: 22483986

15. Picardeau M, Lobry JR, Hinnebusch B. Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. Mol Microbiol. 1999; 32: 437–445. doi: 10.1046/j.1365-2958.1999.01368.x PMID: 10231498

16. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 1996; 13: 660–665. Available: http://mbe.oxfordjournals.org/content/13/5/660.full.pdf+html PMID: 8676740

17. Wright F, Bibb MJ. Codon usage in the G+C-rich *Streptomyces* genome. Gene. 1992; 113: 55–65. doi: 10.1016/0378-1119(92)90669-G PMID: 1563633

18. Hopwood DA. Soil to genomics: the *Streptomyces* chromosome. Ann Rev Genet. 2006; 40: 1–23. doi: 10.1146/annurev.genet.40.110405.090639 PMID: 16761950

19. Pernodet JL, Bocard F, Alegre MT, Gagnat J, Guérineau M. Organization and nucleotide sequence analysis of a ribosomal RNA gene cluster from *Streptomyces ambofaciens*. Gene. 1989; 79: 33–46. doi: 10.1016/0378-1119(89)90090-5 PMID: 2777089

20. Żarko-Postawka M, Hunderuk M, Mordarski M, Zakrzewska-Czerwińska J. Organization and nucleotide sequence analysis of the ribosomal gene set (rrnB) from *Streptomyces lividans*. Gene. 1997; 185:231–237. doi: 10.1016/S0378-1119(96)00649-X PMID: 9055820

21. Bentley SD, Chater KF, Cerdeño-Tàrraga AM, Challis GL, Thomson NR, James KD, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature. 2002; 417:141–147. doi: 10.1038/417141a PMID: 12000953

22. French S. Consequences of replication fork movement through transcription units in vivo. Science. 1992; 258:1362–1365. doi: 0.1126/science.1455232 PMID: 1455232

23. Rocha EPC, Danchin A, Viari A. Universal replication biases in bacteria. Mol Microbiol. 1999; 32:11–16. doi: 10.1046/j.1365-2958.1999.01334.x PMID: 10216855

24. Weaver D, Karoonuthaisiri N, Tsai HH, Huang CH, Ho ML, Gai S, et al. Genome plasticity in *Streptomyces*: identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome. Mol Microbiol. 2004; 51:1535–1550. doi: 10.1111/j.1365-2958.2003.03920.x PMID: 15009883

25. Włodarczyk M, Giersz D. [Linear plasmids of bacteria]. Post Mikrob. 2006; 45:5–18. Polish.

26. Jankowitsch F, Schwarz J, Rückert C, Gust B, Szczepanowski R, Blom J, et al. Genome sequence of the bacterium *Streptomyces davawensis* JCM 4913 and heterologous production of the unique antibiotic roseoflavin. J Bacteriol. 2012; 194:6818–6827. doi: 10.1128/JB.01592-12 PMID: 23043000

27. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005; 3:722–732. doi: 10.1038/nrmicro1235 PMID: 16138100

28. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet. 2006, 38:779–786. doi: 10.1038/ng1830 PMID: 16804543

29. Brouwer MS, Warburton PJ, Roberts AP, Mullany P, Allan E. Genetic organisation, mobility and predicted functions of genes on integrated, mobile genetic elements in sequenced strains of *Clostridium difficile*. PLoS One. 2011; 6:e23014. doi: 10.1371/journal.pone.0023014 PMID: 21876735

30. Burrus V, Pavlovic G, Decaris B, Guedon G. The ICESt1 element of *Streptococcus thermophilus* belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. Plasmid. 2002; 48:77–97. doi: 10.1016/S0147-619X(02)00102-6 PMID: 12383726

31. Farrow KA, Lyras D, Rood JI. Genomic analysis of the erythromycin resistance element Tn*5398* from *Clostridium difficile*. Microbiology. 2001; 147:2717–2728. Available: http://mic.sgmjournals.org/content/147/10/2717.full.pdf+html PMID: 11577151

32. Eichler EE, Clark RA, She X. An assessment of the sequence gaps: unfinished business in a finished human genome. Nature Rev Genet 2004; 5:345–354. doi: 10.1038/nrg1322 PMID: 15143317

33. Garber M, Zody MC, Arachchi HM, Berlin A, Gnerre S, Green LM, et al. Closing gaps in the human genome using sequencing by synthesis. Genome Biol. 2009; 10:R60. doi: 10.1186/gb-2009-10-6-r60 PMID: 19490611

34. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27:573–580. doi: 10.1093/nar/27.2.573 PMID: 9862982

35. Butler JM. Genetics and genomics of core short tandem repeat loci used in human identity testing. J Forensic Sci. 2006; 51:253–265. doi: 10.1111/j.1556-4029.2006.00046.x PMID: 16566758

36. National Library of Medicine. Basic Local Alignment Search Tool [Internet]. Available: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=DeveloperInfo

37. NET Bio [Internet]. Available: https://github.com/dotnetbio/bio

38. Microsoft takes .NET open source and cross-platform, adds new development capabilities with Visual Studio 2015, .NET 2015 and Visual Studio Online [Internet]. 2014. Microsoft News Center. Available: http://news.microsoft.com/2014/11/12/microsoft-takes-net-open-source-and-cross-platform-adds-new-development-capabilities-with-visual-studio-2015-net-2015-and-visual-studio-online/

39. Martinez K, Cupitt J. VIPS: a highly tuned image processing software architecture. IEEE International Conference on Image Processing. IEEE. 2005;574–577. doi: 10.1109/ICIP.2005.1530120