PeerJ Computer Science

AliTV—interactive visualization of whole genome comparisons

Markus J. Ankenbrand[1,*], Sonja Hohlfeld[1,2,*], Thomas Hackl[2,3] and Frank Förster[2,4]

[1] Department of Animal Ecology and Tropical Biology, Julius Maximilian University, Würzburg, Germany
[2] Department for Bioinformatics, Julius Maximilian University, Würzburg, Germany
[3] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[4] Center for Computational and Theoretical Biology, Julius Maximilian University, Würzburg, Germany
[*] These authors contributed equally to this work.

ABSTRACT

Whole genome alignments and comparative analysis are key methods in the quest of unraveling the dynamics of genome evolution. Interactive visualization and exploration of the generated alignments, annotations, and phylogenetic data are important steps in the interpretation of the initial results. Limitations of existing software inspired us to develop our new tool AliTV, which provides interactive visualization of whole genome alignments. AliTV reads multiple whole genome alignments or automatically generates alignments from the provided data. Optional feature annotations and phylogenetic information are supported. The user-friendly, web-browser based and highly customizable interface allows rapid exploration and manipulation of the visualized data as well as the export of publication-ready high-quality figures. AliTV is freely available at https://github.com/AliTVTeam/AliTV.

Subjects Bioinformatics, Computational Biology
Keywords Comparative genomics, Alignment, Visualization

Submitted 22 April 2017
Accepted 25 April 2017
Published 12 June 2017

Corresponding author
Frank Förster,
frank.foerster@uni-wuerzburg.de,
foersterfrank@gmx.de

Academic editor
Shawn Gomez

Additional Information and Declarations can be found on page 7

DOI **10.7717/peerj-cs.116**

INTRODUCTION

Advances in short- and long-read sequencing and assembly over the last decade (*Salzberg et al., 2011*; *Chin et al., 2013*; *Hackl et al., 2014*) have made whole genome sequencing a routine task for biologists in various fields. Public sequence databases already contain several thousand of draft and finished genomes (*Benson et al., 2013*), with many more on the way (*Pagani et al., 2012*). In particular, high throughput sequencing projects of pathogen strains related to recent outbreaks (*Rasko et al., 2011*), and large-scale ecological studies targeting microbial communities and pan genomes of populations using metagenome and single cell sequencing approaches contribute in this process (*Turnbaugh et al., 2007*; *Kashtan et al., 2014*). These rich data sets can be explored for large-scale evolutionary processes using comparative genomics and whole genome alignments, revealing genomic recombinations (*Didelot, Méric & Falush, 2012*; *Namouchi et al., 2012*; *Yahara et al., 2014*), islands and horizontal gene transfer (*Avrani et al., 2011*; *Coleman et al., 2006*; *Langille, Hsiao & Brinkman, 2008*) as well as the often related dynamics of mobile or endogenous viral elements (*Fischer, 2015*; *Touchon & Rocha, 2007*). Other applications of whole genome

comparisons include the analysis of paleopolyploidization events (*Vanneste et al., 2014*) and quantitative measurements of intra-tumour heterogeneity (*Schwarz et al., 2015*).

However, to facilitate proper interpretation of the obtained whole genome comparisons, visualization is key. One of the first tools to provide an interactive graphical representation of aligned genomes is the multiple whole genome alignment program `Mauve` (*Darling et al., 2004*). `Mauve` represents genomes in a co-linear layout with homologous syntenic blocks indicated by colors and connecting lines. The interactive stand-alone viewer `ACT` (*Carver et al., 2008*), in addition to alignment blocks, supports the representation of genomic annotations, such as genes. The `R` library `genoPlotR` (*Guy, Kultima & Andersson, 2010*) and the `Python` based application `EasyFig` (*Sullivan, Petty & Beatson, 2011*), both also based on a co-linear layout and supporting feature annotations, lack interactive analysis features as they are designed to generate static figures.

In addition to co-linear layouts, tools using circular representations of genomes have been developed. `BLASTatlas` (*Hallin, Binnewies & Ussery, 2008*) and `BRIG` (*Alikhan et al., 2011*) use multiple concentric rings to represent data of individual genomes, with `BRIG` also providing an interactive graphical interface. `GenomeRing` (*Herbig et al., 2012*) uses a circular representation as well, however, places all genomes on the same ring and syntenic blocks are connected with arcs extending into the center of the ring.

The web-based comparative genomics software `Sybil` (*Riley et al., 2012*) provides interactive co-linear visualization of multiple whole genome alignments with feature annotations and also supports a phylogenetic tree alongside the alignments. The software builds on a relational `Chado` database schema and, therefore, requires upload and import of custom data sets prior to analysis.

During our analysis of existing software, we found that interactive tools are useful for data exploration, but offer limited support for the figure export and at low qualities. Scripting-based tools provide higher levels of customization and figure quality, however, require familiarity with the respective language, thus often rendering the generation of figures time-consuming. For web- and database-based suites, such as `Sybil`, the upload and import procedure complicate utilization and limit applicability.

Here we present our stand-alone application `AliTV` (Alignment Toolbox and visualization) designed for interactive visualization of multiple whole genome alignments. `AliTV` aims to enable researches to either directly read or automatically generate new whole genome alignments, rapidly explore the results, manipulate and customize the visualization and, at the end of the day, export appealing, publication-grade figures. `AliTV` reads sequence and annotation or alignment data in common formats (FASTA, GenBank, GFF, MAF, Newick, and so on), and internally computes alignments using `lastz` (*Harris, 2007*). The user-friendly interface is built on the state-of-the-art `D3.js` JavaScript framework and can be utilized in a platform independent manner with common web browsers. Genomes are represented in a highly customizable co-linear layout including annotations and an optional phylogenetic tree. The tree is not computed by `AliTV` but has to be provided during data generation. Also, the order of genomes is not automatically optimized to minimize rearrangements. Customizations to the figure by the user can be saved, reloaded, and exported to high quality `SVG` files.

METHODS

Our tool AliTV is divided into two parts. The first non-interactive part is required for the generation of the input files for our interactive viewer. The second part represents that interactive viewer in the form of a SVG file embedded in a HTML5 website. The latest version of our code can be obtained from GitHub (https://github.com/AliTVTeam/AliTV). It is planned to adjust AliTV in order to integrate it into the BioJS registry (https://biojsnet.herokuapp.com/, *Corpas et al. (2014)*). The general design of AliTV assures, that AliTV runs on different hard- and software platforms, e.g., Linux, MacOSX, and Windows. The following sections describe those parts in more detail.

Data preparation

The data preparation is performed by a single Perl script named alitv.pl. This script uses a set of different Perl modules to import incoming data and generate valid JSON input data for our visualization engine described in the next paragraph. One of our aims is to support as many different input formats for sequence and annotation information as possible. Therefore, we used the well tested and broadly accepted BioPerl as basis for our modules (*Stajich et al., 2002*).

The script alitv.pl uses a YAML file to specify the different input files. Moreover, an easy-to-use-mode is available which requires only a couple of input files and generates the required YAML file on the fly. This generated YAML settings file might be used to reproduce AliTV results or can be used as starting point to alter configuration parameters.

During the preparation step, AliTV requires all-vs-all alignments of the complete sequence set. Those alignments are generated or user provided. The current version of alitv.pl requires lastz to generate all alignments in MAF format (*Harris, 2007*). Nevertheless, BioPerl supports a broad range of alignment formats. Therefore, other programs can easily be added to the list of supported alignment programs. Moreover, the ability to use existing alignments allows a huge time benefit, when AliTV parameters are changed to optimize the visualization via YAML settings file in a non-interactive manner. Thus future versions of alitv.pl will support caching of alignments based on checksums to avoid unnecessary recalculations.

The final result of our alitv.pl is a JSON file, which can be load into our interactive visualization page.

Interactive visualization

AliTV is implemented in JavaScript. Our code is documented using JSDoc 3 (version 3.4.0 http://usejsdoc.org/, 02.06.2016). AliTV generates a SVG which is presented within a browser using HTML5. A tutorial is available at https://alitv.readthedocs.io/en/latest/index.html.

To gain advanced application possibilities we use different libraries. The JavaScript library D3.js 3.5.17 (http://d3js.org/, 06.06.2016) provides a wide range of pre-built functions for calculating and drawing the interactive figure. In addition, AliTV employes JQuery 2.2.4 (https://jquery.com/, 06.06.2016) to ease access to several parts of the figure. This is helpful for hiding selected sequences, genes or links. JQueryUI 1.11.4 (https://jqueryui.com/, 06.06.2016) gives us the possibilities to add user-friendly

Table 1 Chloroplast genomes of the parasitic and non-parasitic plants used in the case study.

Species	Accession	Life-style	Reference
Olea europaea	NC_013707	Non-parasitic	*Messina (2010)*
Lindenbergia philippensis	NC_022859	Non-parasitic	*Wicke et al. (2013)*
Cistanche phelypaea	NC_025642	Holo-parasitic	*Wicke et al. (2013)*
Epifagus virginiana	NC_001568	Holo-parasitic	*Wolfe, Morden & Palmer (1992)*
Orobanche gracilis	NC_023464	Holo-parasitic	*Wicke et al. (2013)*
Schwalbea americana	NC_023115	Hemi-parasitic	*Wicke et al. (2013)*
Nicotiana tabacum	NC_001879	Non-parasitic	*Kunnimalaiyaan & Nielsen (1997)*

interactions to `AliTV`. With sliders the user has the chance to specify values for link length and link identity. Context menus offer direct and native interactions with the figure.

To guarantee correct code functionality we engineer `AliTV` according to the Test Driven Development. First we write an automated test case that defines a new function. Then we add the minimum amount of code to make the test pass. Finally we refactor the code to accepted standards. We use `Jasmine 2.3` (http://jasmine.github.io/, 06.06.2016), as framework for testing our JavaScript code. The tests can run either via the SpecRunner or the command line using the taskrunner `grunt 1.0.0` (http://gruntjs.com/, 06.06.2016).

RESULTS AND DISCUSSION

To demonstrate the capabilities of `AliTV` we describe a short case study using seven published chloroplast genomes (Table 1). Four of the chloroplasts belong to parasitic plant species and three to non-parasitic ones. Parasitic plants rely much less or not at all on photosynthetic activity, a trait that should be reflected in the genomic structure of their chloroplast genomes. To assess this hypothesis the chloroplast genomes were downloaded from NCBI and processed with `alitv.pl`. For demonstration purposes, the chloroplast genome of *Nicotiana tabacum* was split in two pieces to represent an unfinished genome with more than one contig, and the genome sequence of Schwalbea americana was reverse-complemented (flipped). The pair-wise whole genome alignments are visualized by `AliTV` (Fig. 1A). The left-hand side of the display panel shows the phylogenetic tree for the seven species with species names as tip labels (parasitic plants are highlighted with an asterisk). The tree has been created provided in accordance to NCBI taxonomy (*Sayers et al., 2009*). Next to the tip labels, each genome is drawn as a scaled and annotated horizontal bar. The orientation of the *S. americana* genome was swapped back to match the orientation of the other genomes, indicated by the tick coordinates in reverse order (0 on the right side). *N. tabacum* is represented by two bars as the sequence has been split into two parts. On those bars features (e.g., genes or (IRs)) are shown as either rectangles or arrows. Alignments between adjacent genomes are represented as colored ribbons. The bottom legend shows the default color scale from red to green corresponding to low and high identity respectively.

The most striking observation is that three of the chloroplast genomes have drastically reduced sizes. All of those are parasitic (Table 1). Interestingly the chloroplast genome

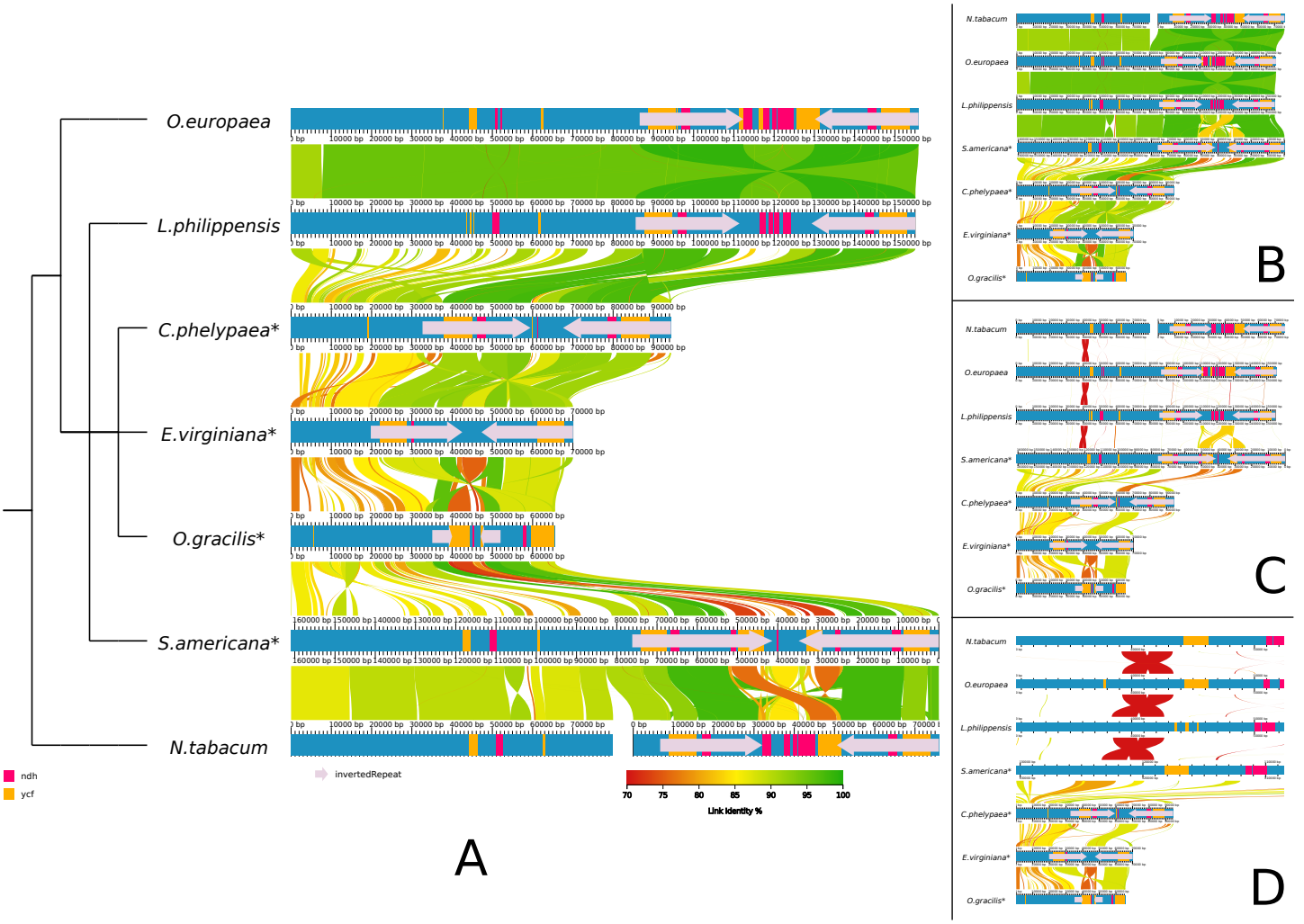

Figure 1 Whole genome alignment of seven chloroplasts visualized by `AliTV`. Species names were italicized and parasites marked with asterisks ex post. (A) Default layout with a phylogenetic tree on the left-hand side and genomes represented by co-linear horizontal bars on the right; genes and inverted repeats are displayed as rectangles and arrows, respectively; colored ribbons connect corresponding regions in the alignment. (B–D) Customized layouts: (B) reordered genomes, non-parasitic plants at the top and holo-parasitic plants at the bottom; (C) links filtered by identity (only those with 50%–90% identity are drawn); (D) zoom in on a potential segmental duplication (red 'X'-shaped links) in the top four genomes.

size of *S. americana* is similar to that of the non-parasitic plants. This can be explained by the life style of *S. americana* which is hemi-parasitic in contrast to the other parasitic plants which are holo-parasites. The features shown are the IR regions as arrows, the hypothetical chloroplast open reading frames as orange and the genes of the ndh family as pink rectangles. First, it can be seen that there is a big variation in size of the inverted repeats. While the IR of Orobanche gracilis is the shortest with roughly 5,000 bp, that of *S. americana* is the largest with roughly 35,000 bp. Second, there are less genes of the ndh family on *Cistanche phelypaea, Epifagus virginiana, O. gracilis*, and *S. americana*. Members of the ndh gene family encode subunits of the NADH dehydrogenase-like complex, which is involved in chlororespiration (*Martín & Sabater, 2010*). However, they are not required for plant growth under optimal conditions (*Burrows, 1998*). The absence of ndh genes

in chloroplasts of parasitic plants has been studied in detail in *Wicke et al. (2013)*. Loss of ndh genes has also been reported for photosynthetic plants such as some conifers and orchids (*Wakasugi et al., 1994*; *Kim et al., 2015*). Looking at the pairwise similarities of adjacent genomes, it is apparent that the non-parasitic plants (e.g., *Olea europaea and Lindenbergia philippensis*) have high overall sequence identity. In contrast, the sequence similarity within parasitic plants is lower. This observation can help framing a hypothesis about the evolutionary pressure on chloroplasts of parasitic plants. Another interesting observation is the distribution of missing regions of *C. phelypaea* in comparison to *L. philippensis*. Missing regions are distributed all over the genome and the order of the remaining parts remains stable. *Wicke et al. (2013)* describe an inversion in the large single copy region of *S. americana* compared to non-parasitic plants which is clearly visible by the link to *N. tabacum* around the 115 kbp position. All these observations can be made by simply looking at the raw figure created by `alitv.pl` and visualized by `AliTV`. However the figure can be analyzed interactively in more detail. One shortcoming of the linear representation of whole genome alignments is the limited comparability of non-adjacent sequences. Therefore, `AliTV` provides a way for the user to re-order the genomes on the figure (Fig. 1B). If reordering causes inconsistencies with the phylogenetic tree, the tree is hidden and a warning message is displayed. Furthermore, the links can be filtered by their alignment identity. The default setting is to display only links with minimal identity of 70%. But sometimes it might be interesting to look at regions with less similarity. To see these regions it is also important to hide large regions with high similarity. This can be achieved by changing the identity via a slider (Fig. 1C). After setting the identity range to 50%–90% red 'X'-shaped links between *N. tabacum, O. europaea, L. philippensis*, and *S. americana* become apparent. For detailed inspection of regions of interest, `AliTV` provides a zoom function (Fig. 1D). This way the exact location of the alignments can be traced to the locations of psaA and psaB. Moreover `AliTV` provides functions like alignment length filtering, selective hiding of sequences, links and features, change of orientation (reverse complement) and rotation of circular chromosomes. Finally, it is possible to tweak many graphical parameters, such as colors, labels or spacing, directly via the interface to produce a publication quality figure which can be saved in `SVG` format. Furthermore, the current state can be saved in `JSON` format in order to share it with collaborators or continue the work with `AliTV` at a later time.

CONCLUSION

The case study demonstrates the suitability of `AliTV` as a tool for visualizing and analyzing whole genome comparisons. `AliTV` can be used to easily create a figure that show cases many genomic features at once. Furthermore, the rich interactive features enable the exploratory analysis and discovery of previously unknown features. Thus, novel hypotheses can be generated that can then be validated with experimental methods. Therefore, `AliTV` is a useful tool that will help scientists to find biologically meaningful information in the vast amount of genomic data.

ACKNOWLEDGEMENTS

We would like to thank Felix Bemm for fruitful discussions about file formats and must-have-features during the development of `AliTV` and for supervising MJA during his bachelor thesis.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. The publication fee was funded by the MIT Libraries. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg.
MIT Libraries.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Markus J. Ankenbrand and Frank Förster conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Sonja Hohlfeld performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Thomas Hackl conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, performed the computation work, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:
Github: https://github.com/AliTVTeam/AliTV.

REFERENCES

Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**:402 DOI 10.1186/1471-2164-12-402.

Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. 2011. Genomic island variability facilitates Prochlorococcus—virus coexistence. *Nature* **474(7353)**:604–608 DOI 10.1038/nature10172.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Research* **41(Database issue)**:D36–D42 DOI 10.1093/nar/gks1195.

Burrows PA. 1998. Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid ndh genes. *The EMBO Journal* **17(4)**:868–876 DOI 10.1093/emboj/17.4.868.

Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream M-A. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24(23)**:2672–2676 DOI 10.1093/bioinformatics/btn529.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**:563–569 DOI 10.1038/nmeth.2474.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. 2006. Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311(5768)**:1768–1770 DOI 10.1126/science.1122050.

Corpas M, Jimenez R, Carbon SJ, García A, Garcia L, Goldberg T, Gomez J, Kalderimis A, Lewis SE, Mulvany I, Pawlik A, Rowland F, Salazar G, Schreiber F, Sillitoe I, Spooner WH, Thanki A, Villaveces JM, Yachdav G, Hermjakob H. 2014. BioJS: an open source standard for biological visualisation—its status in 2014. *F1000 Research* **3**:55 DOI 10.12688/f1000research.3-55.v1.

Darling A. CE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14(7)**:1394–1403 DOI 10.1101/gr.2289704.

Didelot X, Méric G, Falush D. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* **13**:256 DOI 10.1186/1471-2164-13-256.

Fischer MG. 2015. Virophages go nuclear in the marine alga Bigelowiella natans. *Proceedings of the National Academy of Sciences of the United States of America* **112(38)**:11750–11751 DOI 10.1073/pnas.1515142112.

Guy L, Kultima JR, Andersson S. GE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26(18)**:2334–2335 DOI 10.1093/bioinformatics/btq413.

Hackl T, Hedrich R, Schultz J, Förster F. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30(21)**:3004–3011 DOI 10.1093/bioinformatics/btu392.

Hallin PF, Binnewies TT, Ussery DW. 2008. The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology. *Molecular BioSystems* **4(5)**:363–371 DOI 10.1039/b717118h.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. PhD thesis, Pennsylvania State University.

Herbig A, Jäger G, Battke F, Nieselt K. 2012. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics* **28(12)**:i7–i15 DOI 10.1093/bioinformatics/bts217.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. *Science* **344(6182)**:416–420 DOI 10.1126/science.1248575.

Kim HT, Kim JS, Moore MJ, Neubig KM, Williams NH, Whitten WM, Kim J-H. 2015. Seven new complete plastome sequences reveal rampant independent loss of the ndh gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. *PLOS ONE* **10(11)**:e0142215 DOI 10.1371/journal.pone.0142215.

Kunnimalaiyaan M, Nielsen BL. 1997. Fine mapping of replication origins (ori A and ori B) in *Nicotiana tabacum* chloroplast DNA. *Nucleic Acids Research* **25(18)**:3681–3686 DOI 10.1093/nar/25.18.3681.

Langille M, Hsiao W, Brinkman F. 2008. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* **9**:329 DOI 10.1186/1471-2105-9-329.

Martín M, Sabater B. 2010. Plastid ndh genes in plant evolution. *Plant Physiology and Biochemistry* **48(8)**:636–645 DOI 10.1016/j.plaphy.2010.04.009.

Messina R. 2010. `Olea europaea chloroplast, complete genome`. *Available at http://www.ncbi.nlm.nih.gov/nuccore/NC_013707.2* (accessed on 30 June 2015).

Namouchi A, Didelot X, Schöck U, Gicquel B. 2012. After the bottleneck: genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Research* **22**:721–734 DOI 10.1101/gr.129544.111.

Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The genomes online database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **40(Database issue)**:D571–D579 DOI 10.1093/nar/gkr1100.

Rasko DA, Dale WR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin C-S, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-møller J, Struve C, Petersen AM, Krogfeld KA, Nataro JP, Schadt EE, Waldor MK. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *The New England Journal of Medicine* **365(8)**:709–717 DOI 10.1056/NEJMoa1106920.

Riley DR, Angiuoli SV, Crabtree J, Dunning Hotopp JC, Tettelin H. 2012. Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics* **28(2)**:160–166 DOI 10.1093/bioinformatics/btr652.

Salzberg SL, Phillippy AM, Zimin AV, Puiu D, Magoc T, Koren S, Treangen T, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA. 2011. GAGE: a critical

evaluation of genome assemblies and assembly algorithms. *Genome Research* **22(3)**:557–567 DOI 10.1101/gr.131383.111.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **37(Database issue)**:D5–D15 DOI 10.1093/nar/gkn741.

Schwarz RF, Ng CKY, Cooke SL, Newman S, Temple J, Piskorz AM, Gale D, Sayal K, Murtaza M, Baldwin PJ, Rosenfeld N, Earl HM, Sala E, Jimenez-Linan M, Parkinson CA, Markowetz F, Brenton JD. 2015. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLOS Medicine* **12(2)**:e1001789 DOI 10.1371/journal.pmed.1001789.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl toolkit: perl modules for the life sciences. *Genome Research* **12(10)**:1611–1618 DOI 10.1101/gr.361602.

Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* **27(7)**:1009–1010 DOI 10.1093/bioinformatics/btr039.

Touchon M, Rocha E. 2007. Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution* **24(4)**:969–981 DOI 10.1093/molbev/msm014.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature* **449(7164)**:804–810 DOI 10.1038/nature06244.

Vanneste K, Baele G, Maere S, Peer YVD. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Research* **24(8)**:1334–1347 DOI 10.1101/gr.168997.113.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. 1994. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii. *Proceedings of the National Academy of Sciences of the United States of America* **91(21)**:9794–9798 DOI 10.1073/pnas.91.21.9794.

Wicke S, Müller KF, De Pamphilis CW, Quandt D, Wickett NJ, Zhang Y, Renner SS, Schneeweiss GM. 2013. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *The Plant Cell* **25(10)**:3711–3725 DOI 10.1105/tpc.113.113373.

Wolfe KH, Morden CW, Palmer JD. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Sciences of the United States of America* **89(22)**:10648–10652 DOI 10.1073/pnas.89.22.10648.

Yahara K, Didelot X, Ansari M, Sheppard S. 2014. Efficient inference of recombination hot regions in bacterial genomes. *Molecular Biology and Evolution* **31(6)**:1593–1605 DOI 10.1093/molbev/msu082.