

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/327163330>

bioSyntax: Syntax highlighting for computational biology

Article in BMC Bioinformatics · December 2018

DOI: 10.1186/s12859-018-2315-y

CITATIONS

0

READS

396

8 authors, including:



Artem Babaian

University of British Columbia - Vancouver

44 PUBLICATIONS 201 CITATIONS

[SEE PROFILE](#)



Anicet Ebou

Institut National Polytechnique Félix Houphouët-Boigny

8 PUBLICATIONS 0 CITATIONS

[SEE PROFILE](#)



Alyssa Fegen

University of British Columbia - Vancouver

5 PUBLICATIONS 0 CITATIONS

[SEE PROFILE](#)



Jeffrey Kam

University of Waterloo

6 PUBLICATIONS 0 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Loss of macpΨ ribosomal RNA modification is a major feature of cancer [View project](#)




A novel computational approach for drug repurposing using systems biology [View project](#)

SOFTWARE

Open Access



bioSyntax: syntax highlighting for computational biology

Artem Babaian^{1,2*} , Anicet Ebou³, Alyssa Fegen⁴, Ho Yin Kam⁵, German E. Novakovsky², Jasper Wong⁶, Dylan Aissi⁷ and Li Yao⁸

Abstract

Background: Computational biology requires the reading and comprehension of biological data files. Plain-text formats such as SAM, VCF, GTF, PDB and FASTA, often contain critical information which is obfuscated by the data structure complexity.

Results: *bioSyntax* (<https://biosyntax.org/>) is a freely available suite of biological syntax highlighting packages for *vim*, *gedit*, *Sublime*, *VSCoDe*, and *less*. *bioSyntax* improves the legibility of low-level biological data in the bioinformatics workspace.

Conclusion: *bioSyntax* supports computational scientists in parsing and comprehending their data efficiently and thus can accelerate research output.

Keywords: Syntax highlighting, *Vim*, *Sublime*, Command line interface, SAM, VCF, FASTA, FASTQ

Background

A major component of computational biology research involves the reading and comprehension of data in biological file-formats including FASTA [1], FASTQ [2], gene transfer format (GTF) [3], variant calling format (VCF) [4], protein database format (PDB) [5, 6], and sequence alignment map (SAM) [7] amongst others [8]. While being easy to parse computationally, these and other biological files often become illegible as their size and complexity increases, including header sections that contain critical data descriptors often required for downstream processing.

Syntax highlighting is designed to improve the interpretability of text files with well-defined structures through the application of colour, font, and formatting to differentiate content; typically a set of keywords, structures, and symbols. Originally developed for the code editor *Emacs* in 1971 [9] and later *LEXX* in the late 1980s [10], syntax highlighting is now ubiquitous in the computer sciences. There are contradictory findings whether syntax highlighting offers performance benefits

on computer programming comprehension or efficiency but syntax highlighting consistently decreases the reported difficulty of work tasks [11–15]. In terms of human perception, colour (via syntax highlighting) increases the visual saliency of data thereby decreasing search times to find a particular symbol in a visual search task [16–18].

A plethora of tools, both stand-alone and web-based, exist to help scientists visualize and process biological data [19–25]. However, there is no comprehensive set of software to assist in direct inspection and interpretation of raw biological data and their headers.

The objective of *bioSyntax* is to improve the human readability of scientific data-formats through seamlessly integrated syntax highlighting and to assist scientists in performing visual search tasks when working with low-level data.

Implementation

bioSyntax is currently ported for four common text editors, *Vim* (and *GVim*), *gedit* (and other linux editors using the *GTKSourceView* library), *Sublime-Text-3*, and *Visual Studio Code (VSCoDe)*, as well as the command-line pager program, *less*. Additionally, *bioSyntax* functions as a repository into which syntax highlighting definition files may be deposited by the community for additional scientific file formats.

* Correspondence: ababaian@bccrc.ca

¹Terry Fox Laboratory, BC Cancer Research Centre, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada

²Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Full list of author information is available at the end of the article



File specifications

At its core, bioSyntax is a set of syntax-highlighting definition files specific to each port which are themselves a programmed set of regular expressions.

Where available, syntax-files are designed using a combination of; official file specifications for SAM v1.5 (<https://samtools.github.io/hts-specs/SAMv1.pdf>), VCF v4.2 (<https://samtools.github.io/hts-specs/VCFv4.2.pdf>)

PDB v3.30 (ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_Letter.pdf), BED6 (<https://genome.ucsc.edu/FAQ/FAQformat.html>), and GTF v2.2 (<http://mblab.wustl.edu/GTF22.html>); example files from databases (NCBI Nucleotide/Protein [26], Sequence Read Archive [27], dbSNP [28], RefSeq [29], RCSB [5] and UCSC Genome Browser [23]); publicly available consortium data (1000 genomes project [30], ENCODE [31]);

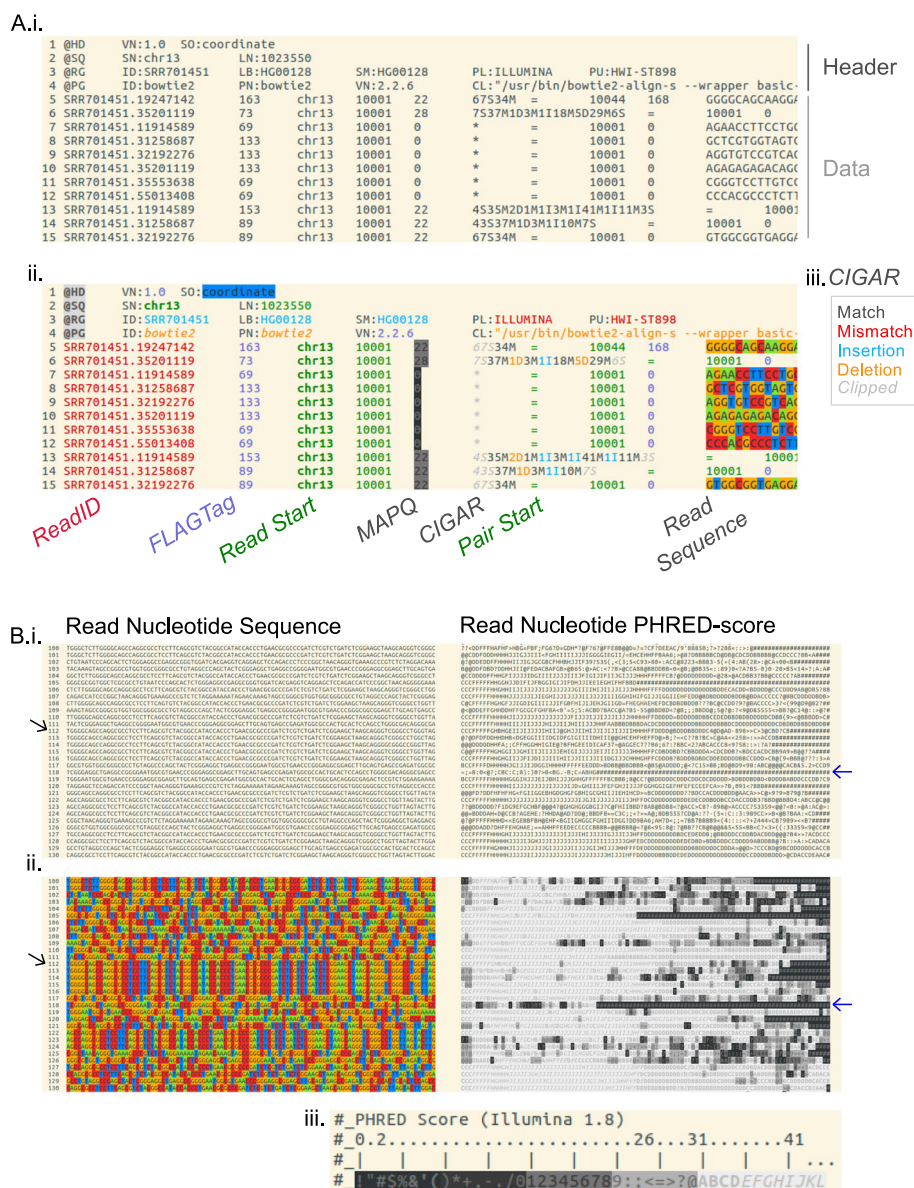


Fig. 1 Syntax highlighting for sequence alignment map (.sam) file format. **a** Terminal screenshot of the 'less HG00128_hgr1.sam' command run i. normally or ii. with bioSyntax. Related information in the header and data sections are grouped by colours (genomic coordinates, green; sample information, pale blue ...) to improve legibility. Each data-row is an individual sequencing read. Iii. CIGAR alignment strings in particular can be highlighted such that they become substantially easier to read. **b** A broad view of the nucleotide and PHRED-score for 30 reads i. before, and ii. after syntax highlighting. Underlying information of about the data becomes intuitively visible such as PCR-duplicates (black arrow) and poor quality areas and reads (blue arrow) based on iii. PHRED score

and standard outputs from commonly used software (*samtools* [7], *GATK* [32], *bowtie2* [33], *cufflinks* [34] and *ClustalX* [35]).

Using bioSyntax

Within each program, highlighting of FASTA, FASTQ, CLUSTAL, BED, GTF, PDB, VCF and SAM files is automatically detected by file extension and can also be manually set within text-editors for files with non-standard file-extensions. In *gedit*, *sublime* and *vim*, amino acid

FASTA files can be coloured using CLUSTAL [35], Taylor [36], Zappo [19] or hydrophobicity [20] colour schemes.

Large and compressed data can be manipulated on a unix command pipe with the output directly ported into *less* using explicit bioSyntax format commands: `'sam-less', 'vcf-less', etc.` (Fig. 1). For example `'samtools view -h NA12878.bam | sam-less -',` or `'gzip -dc gencode.v26.gtf.gz | grep 'MYC' - | gtf-less -x 10'.`

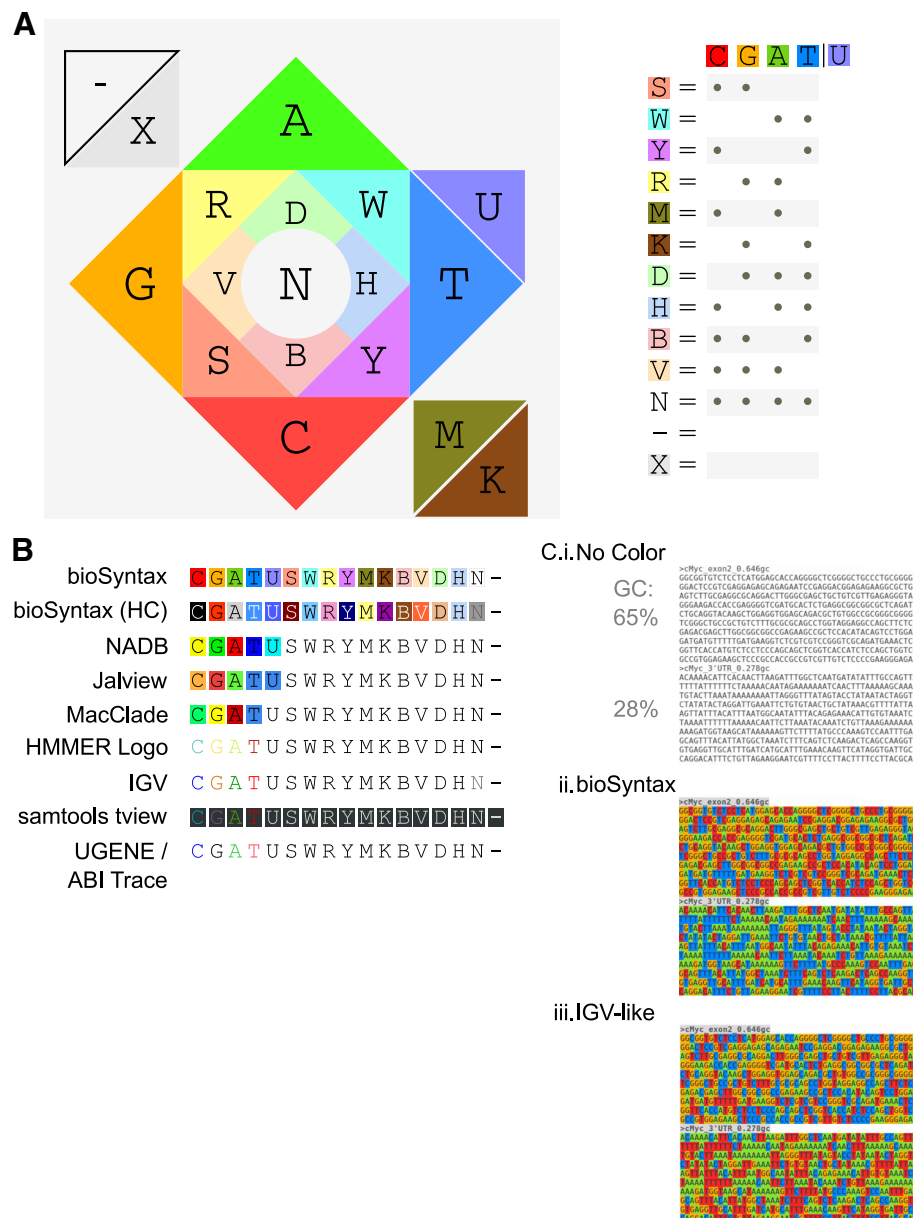


Fig. 2 bioSyntax nucleotide colour scheme. **a** The four primary bases are coloured in two pairs of contrasting colours. IUPAC ambiguous bases are then coloured in increasingly lighter tones of the approximately mixed colours. To accommodate 4-dimensional bases in 3-dimensional colours, aMino (A or C) and Keto (G or T) bases are darker. **b** A comparison of nucleotide colour-schemes in the literature. **c** bioSyntax colouring allows for approximation of a sequences GC-content by how warm (high GC) or cool (high AT) it appears

Results and discussion

Syntax highlighting for computational biology file formats

bioSyntax currently recognizes FASTA, FASTQ, CLUSTAL, BED, GTF, PDB, SAM and VCF formats across all four text editors and *less*. Upon installation, *bioSyntax* automatically recognizes file-extensions and seamlessly assigns syntax highlighting to these data files.

The main benefit of syntax highlighting is immediately apparent through its increased legibility (Fig. 1, Additional file 1: Figures S1 and S2), especially in the deconstruction of verbose content such as plain-text CIGAR strings (Fig. 1a). In each file-format data is organized using contrasting colours to accentuate keywords or data fields. Nucleotides and amino acids are highlighted with distinct colours allowing for users to read sequences and interpret patterns in the alignment. Data fields containing scores such as PHRED base quality or mapping scores are gradient coloured.

The overall system of highlighting is also designed to group biological classes, even across file formats (Additional file 2: Table S1). For instance, dark-green is reserved for genomic coordinates in BED, GTF, SAM and VCF, so even if a user is unfamiliar with the SAM format, previous experience associating dark-green in BED or GTF will inform them of the meaning of those fields when presented in a SAM file (Additional file 1: Figure S2).

Ultimately, *bioSyntax* aims to help computational biologists comprehend data using graphical highlighting rather than simple syntactic highlighting. When the data does not have to be read per character or per word, but can be viewed as graphical patterns, underlying information in the data becomes salient, similar to alternative nucleotide representations [37, 38]. This is best seen in complex files such as SAM in which PCR-duplicate reads form block patterns and read density can be approximated by the diagonal similarity of reads at a locus (Fig. 1b). In a user-experience survey of bioinformaticians (Additional file 3: Text 1), 98.6% of users ($N = 72$) selected *bioSyntax* highlighted alignments as being easier to identify nucleotide variants compared to a standard monochrome text. Future research on how syntax highlighting can be refined to optimize for user performance is necessary.

bioSyntax nucleotide representation

bioSyntax implements a novel nucleotide colouring scheme for the complete IUPAC ambiguous base set [39], unlike other colour-sets which are designed for four or five bases (Fig. 2). The four primary base colours are chosen such that additive colour mixing also represents complete base ambiguity. For instance, thymine (blue) and cytosine (red) are pyrimidines (magenta), and the “any base”, N, is white. This colour-set visually distinguishes the strong bases (G,C) and weak bases (A,T)

as warm and cool colours respectively, allowing for an intuitive approximation of the sequence GC-content (Fig. 2c). Additionally, a high-contrast colour-scheme is available to aid visually impaired or colour-blind users (Additional file 1: Figure S3).

The bioSyntax repository

There are scores of biological and scientific file-formats which would benefit from syntax highlighting. To facilitate future development of syntax definition files in science, the *bioSyntax* repository (<https://bioSyntax.org>) was set-up. The repository is both a library for scientific syntax highlighting and a community-oriented resource for learning syntax highlighting development. In this manner, researchers experienced in the use-cases of a file-format can quickly develop and share new syntax definition files.

Conclusions

bioSyntax assists researchers to intuitively read and navigate biological files in the context of familiar and common text editing tools. The cross-format unifying colour theme for biological data classes aids users in the rapid and accurate parsing of data with minimal prior knowledge regarding the file-format. Altogether, *bioSyntax* offers a substantial improvement to the legibility of biological data and helps researchers to grok their data.

Availability and requirements

- **Project name:** *bioSyntax*
- **Project home page:** <https://biosyntax.org>
- **Source page:** <https://github.com/bioSyntax/bioSyntax>
- **Operating system(s):** Linux, MacOS, Windows
- **Programming language:** vim-script, xml, yaml
- **Other requirements:** *gedit* 3.18+ or *GTKSourceView3*-dependent editors; *Sublime-text-3*; *Visual Studio Code* 1.25+; *vim* or *gvim* 7.4+; *less* and *source-highlight*
- **License:** GNU General Public License v3.0
- **Any restrictions to use by non-academics:** none

Additional files

Additional file 1: Figure S1. Screenshots of *bioSyntax* in; A) *gedit* for the nucleotide sequence FASTA, an amino acid FASTA with CLUSTAL colour scheme and a FASTQ file and; B) *sublime-text-3* for a PDB file. **Figure S2.** Screenshots of *bioSyntax* in; A) *vim* for the human dbSNP (hg38 build-150) VCF file and; B) *less* for the Gencode v26 Annotation GTF and C) an example SAM file. In the GTF format, note how background colouring of “start_codon”, “stop_codon”, “CDS”, and “UTR” graphically distinguishes protein-coding transcripts from non-coding transcripts. **Figure S3.** The standard and high-contrast *bioSyntax* colour set for IUPAC nucleotides under the different forms of simulated colour-blindness. Hue and lightness variations in the standard theme allow for accessibility with

colour-blindness. The alternative high-contrast set retains higher visual distinction between bases, even at the monochrome level. (PDF 3125 kb)

Additional file 2: Table S1. Biological class definitions and colour definitions for the default bioSyntax theme in hexadecimal (*gedit*, *sublime*, *VSCode*, and *gvim*), cterm (*vim*), and 8-bit ANSI escape character (*less*) colours. (XLS 186 kb)

Additional file 3: Text 1. bioSyntax user-experience survey of bioinformaticians. (PDF 3063 kb)

Abbreviations

GTF: Gene Transfer Format; IUPAC: International Union of Pure and Applied Chemistry; PDB: Protein Database Format; SAM: Sequence Alignment Map; VCF: Variant Call Format

Acknowledgements

bioSyntax was initiated at the *hackseq17* (<https://www.hackseq.com/>) genomics hackathon [40]. We would like to thank the organizers for their dedication to making the event happen. We'd like to thank the developers who have contributed to *bioSyntax*; Ching Pan Eric Chu and Benjamin A Barad; and Joseph O'Brien for his artistic insight in developing a high-contrast palette for nucleotide colouring.

Author contributions

AB conceived and led the design/development of *bioSyntax*. AB, AE, AF, HYK, GEN and JW developed the initial release of *bioSyntax* and wrote the manuscript. DA packaged *bioSyntax* for Linux releases. LY ported *bioSyntax* for VSCode. All authors read and approved the final version of this manuscript.

Funding

AB is supported by the Roman M. Babicki Fellowship in Medical Research. GN is supported by the UBC International Doctoral Fellowship award. No funding body influenced the development of this software or writing of the manuscript.

Availability of data and materials

Source code and all development materials are available at <https://github.com/bioSyntax/bioSyntax>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Author details

¹Terry Fox Laboratory, BC Cancer Research Centre, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada. ²Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada. ³Département de Formation et de Recherches Agriculture et Ressources Animales, Institut National Polytechnique Felix Houphouët-Boigny, Yamoussoukro, Côte d'Ivoire. ⁴Faculty of Science, University of British Columbia, Vancouver, BC, Canada. ⁵Faculty of Mathematics, University of Waterloo, Waterloo, ON, Canada. ⁶Genome Science and Technology, University of British Columbia, Vancouver, BC, Canada. ⁷Department of General and Interventional Cardiology, University Heart Center Hamburg, Hamburg, Germany. ⁸THU-PKU Joint Center for Life Sciences, Tsinghua University, Beijing, China.

Received: 3 May 2018 Accepted: 14 August 2018

Published online: 22 August 2018

References

- Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science*. 1985;227:1435–41.

- Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2010;38:1767–71.
- Keibler E, Brent MR. Eval: a software package for analysis of genome annotations. *BMC Bioinformatics*. 2003;4:50.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28:235–42.
- Berman HM. The protein data Bank: a historical perspective. *Acta Crystallogr A*. 2008;64:88–95.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
- Zhang H. Overview of sequence data formats. *Methods Mol Biol Clifton NJ*. 2016;1418:3–17.
- Hansen WJ. Creation of hierarchic text with a computer display [internet]. Department of Computer Science, Stanford University; 1971 [cited 2017 Dec 11]. Available from: <https://books.google.ca/books?id=IZVFAAAIAAJ>.
- Cowlishaw MF. LEXX—A programmable structured editor. *IBM J Res Dev*. 1987;31:73–80.
- Sarkar A. The impact of syntax colouring on program comprehension. *Proc 26th Annu Conf Psychol Program Interest Group Pp*. 2015:49–58.
- Dimitri G. The impact of syntax highlighting in sonic pi. *Psychol Program Interest Group*. 2015;2015
- Hakala T, Nykyri P, Sajaniemi J. An experiment on the effects of program code highlighting on visual search for local patterns. *Psychol Program Interest Group*. 2006:38–52.
- Beelders TR, du Plessis J-PL. Syntax highlighting as an influencing factor when reading and comprehending source code. *J Eye Mov Res*. 2015;9
- Hannebauer C, Hesenius M, Gruhn V. Does syntax highlighting help programming novices? *Empir Softw Eng*. 2018:1–34.
- Duncan J, Humphreys G. Beyond the search surface: visual search and attentional engagement. *J Exp Psychol Hum Percept Perform*. 1992;18:578–88. discussion 589–593
- Ramachandran VS, Hubbard EM. Psychophysical investigations into the neural basis of synaesthesia. *Proc Biol Sci*. 2001;268:979–83.
- Wilkinson KM, Carlin M, Jagaroo V. Preschoolers' speed of locating a target symbol under different color conditions. *Augment Altern Commun Baltim Md*. 1985. 2006;22:123–33.
- Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. *Nat Methods*. 2010;7:516–25.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinforma Oxf Engl*. 2009;25:1189–91.
- Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33:1870–4.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14:178–92.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res*. 2002;12:996–1006.
- Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. 2016;44:W3–10.
- Carver T, Bleasby A. The design of Jemboss: a graphical user interface to EMBOSS. *Bioinforma Oxf Engl*. 2003;19:1837–43.
- Resource NCBI. Coordinators. Database resources of the National Center for biotechnology information. *Nucleic Acids Res*. 2017;45:D12–7.
- Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res*. 2011;39:D19–21.
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29:308–11.
- O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44:D733–45.
- 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65.
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome: Nature nature publishing group. *Nature*. 2012;489:57–74.

32. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
33. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods.* 2012;9:357–9.
34. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat Protoc.* 2012;7:562–78.
35. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinforma Oxf Engl.* 2007;23:2947–8.
36. Taylor WR. The classification of amino acid conservation. *J Theor Biol.* 1986; 119:205–18.
37. Rozak DA, Rozak AJ. Using a color-coded ambigraphic nucleic acid notation to visualize conserved palindromic motifs within and across genomes. *BMC Genomics.* 2014;15:52.
38. Jarvius J, Landegren U. DNA skyline: fonts to facilitate visual inspection of nucleic acid sequences. *BioTechniques.* 2006;40:740.
39. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Biochem J.* 1985;229:281–6.
40. hackseq Organizing Committee 2016. hackseq: Catalyzing collaboration between biological and computational scientists via hackathon. *F1000Research.* 2017;6:197.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

