

FROM INNOVATION TO APPLICATION

# EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment

João Carneiro, Filipe Pereira \*

Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Porto, Portugal

\* [fpereirapt@gmail.com](mailto:fpereirapt@gmail.com)



## OPEN ACCESS

**Citation:** Carneiro J, Pereira F (2016) EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment. *PLoS Negl Trop Dis* 10(7): e0004757. doi:10.1371/journal.pntd.0004757

**Editor:** Claudia Munoz-Zanzi, University of Minnesota, UNITED STATES

**Published:** July 21, 2016

**Copyright:** © 2016 Carneiro, Pereira. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Portuguese Foundation for Science and Technology (FCT), European Regional Development Fund (ERDF) and Programa Operacional Potencial Humano (Investigator FCT program to FP and research grant SFRH/BPD/100912/2014 to JC) and Direção-Geral da Saúde and Coordenação Nacional para a Infecção VIH/SIDA (VIH/SAU/0011/2011). CIIMAR was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and ERDF in the framework of the programme PT2020. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

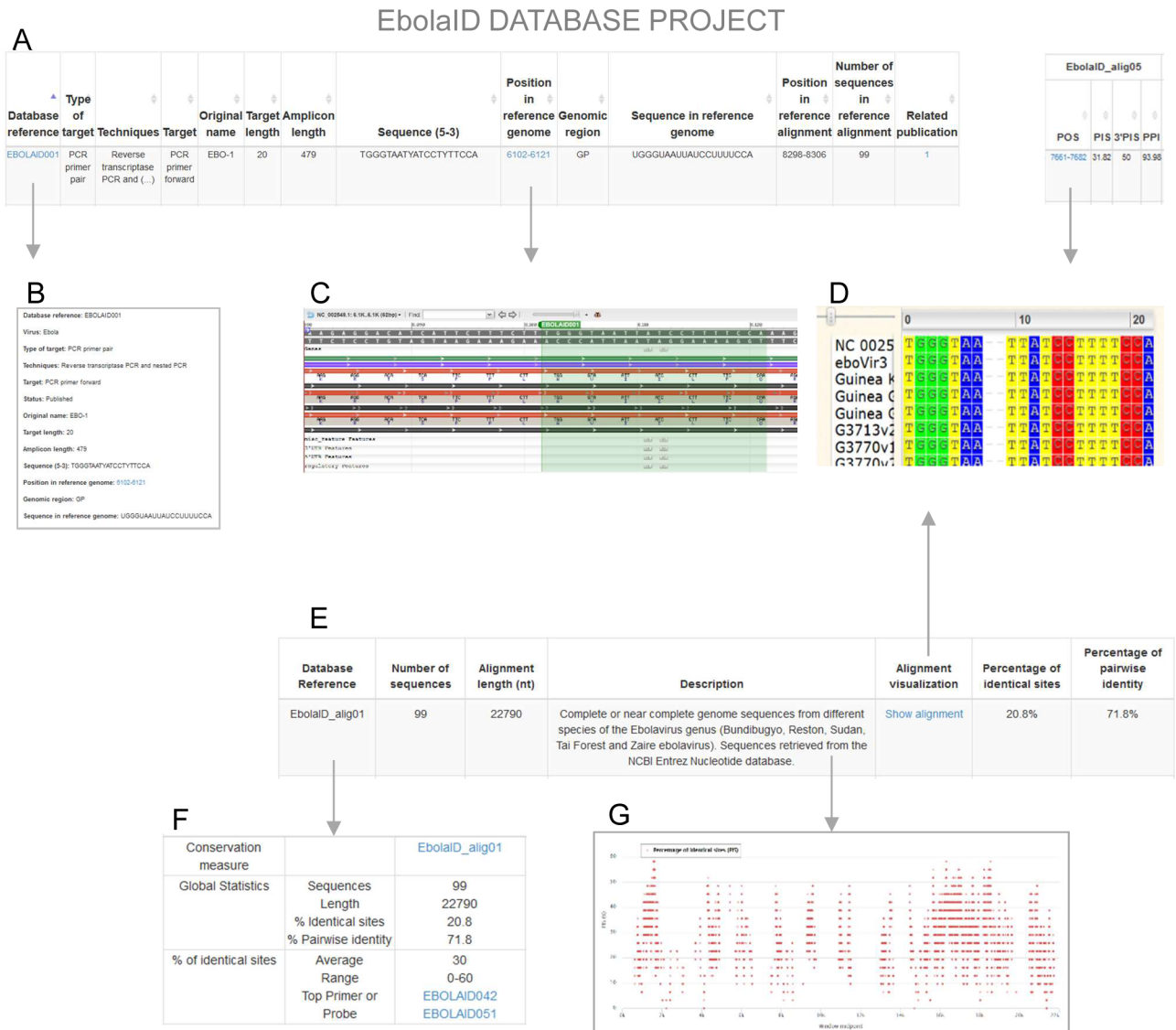
**Competing Interests:** The authors have declared that no competing interests exist.

## Challenge

The Ebola virus disease (EVD) is a rare and deadly disease affecting humans and other primates caused by infection with a virus of the family *Filoviridae*, genus *Ebolavirus* [1]. The March 2014 Ebola epidemic is the largest in history, affecting multiple countries in West Africa, with more than 11,301 deaths reported by the World Health Organization by March 2016. Methods based on polymerase chain reaction (PCR) are commonly used for virus detection with the advantage of requiring low amounts of viral samples, minimal manipulation, and minimal equipment. In particular, quantitative PCR (qPCR) has been used for the sensitive and fast identification of patients with EVD [2–7]. However, the high genetic diversity of RNA viruses poses a challenge for the design of efficient nucleic acid-based assays, as suggested by the high substitution rate observed in Ebola virus from the 2014 outbreak [5,8]. Recent studies suggest that false-negative diagnosis or inefficient therapeutics can occur due to sequence variation at binding sites of PCR primers, probes, small interfering RNAs (siRNAs), or monoclonal antibodies targeting the Ebola virus [5,9]. For this reason, the selection of reliable oligonucleotides and target genomic regions for use in nucleic acid-based assays is crucial for the correct diagnosis and treatment of EVD.

## The EbolaID Database

The EbolaID database (<http://ebolaid.portugene.com>) is a free, web-accessible database built to facilitate the design of accurate molecular methods for detection and identification of the Ebola virus (Fig 1). The database provides an interface for browsing, filtering, and downloading data from published oligonucleotide sequences (PCR primers, real-time PCR probes, etc.) annotated according to a reference genome. The user can find varied information for each oligonucleotide: the type of technique where it was originally used, the sequence, location in the reference genome, references, etc. The database provides different measures of sequence conservation for each target region obtained from multiple sequence alignments, allowing the selection of the most conserved oligonucleotides. The sequences can be visualized, edited, and exported using the Wasabi (<http://wasabi2.biocenter.helsinki.fi/>) and NCBI (<https://www.ncbi.nlm.nih.gov/tools/sviewer/>) sequence viewers. The database also describes the most conserved and variable regions of the Ebola genome for those researchers interested in designing new oligonucleotides. [Box 1](#) discusses the advantages and disadvantages of the database.



**Fig 1. Screenshot of the data and tools included in the EbolaID database.** A) Section with the list of oligonucleotides. B) Section with detailed information for each oligonucleotide. C) NCBI sequence viewer of reference genome with oligonucleotide annotation and feature tracks. D) Wasabi alignment viewer of a genomic region. E) Table with description of a multiple sequence alignment. F) Summary of different conservation measures for an alignment. G) Scatter plot describing a sliding window analysis of diversity measures across the Ebola genome.

doi:10.1371/journal.pntd.0004757.g001

## Database Coverage

The EbolaID database was constructed using information from 17 peer-reviewed works published from 1999 to 2014. Currently, the database includes 57 oligonucleotides (48 primers and 9 probes) retrieved from publications describing assays for virus detection or from epidemiological and phylogenetic studies [8,10–12]. The oligonucleotides included in the database are located in the NP (33.3%), GP (21.0%), L (38.6%), VP24 (3.5%), and 3'UTR (3.5%) regions of the Ebola genome. The database includes multiple sequence alignments of complete or nearly complete genomic sequences of ebolaviruses and marburgviruses used to estimate the level of sequence conservation in oligonucleotides (a description of the alignments can be found in the “Sequence alignments” section of the website). For example, the database includes an

## Box 1. Advantages and Disadvantages of the EbolaID Database

### Advantages

- The use of different multiple sequence alignments and measures of sequence conservation
- Detailed information on each oligonucleotide with links to references
- Query tools for selecting and evaluating the oligonucleotides

### Disadvantages

- Restricted to oligonucleotides described in peer-reviewed publications
- Large tables with numeric values can be difficult to visualize
- Data from different publications may have different levels of reliability

alignment with 99 sequences from different species of the *Ebolavirus* genus (Bundibugyo, Reston, Sudan, Tai Forest, and Zaire ebolavirus) retrieved from the NCBI Entrez Nucleotide database (<http://www.ncbi.nlm.nih.gov>). The recent 2014 outbreak is represented by an alignment with 81 sequences from 2014 plus 20 sequences from earlier outbreaks (from 1976 to 2008) built by Gire et al. [8].

### Selecting the Best Oligonucleotides

The EbolaID database provides different measures of sequence conservation for each oligonucleotide (details at [http://ebolaid.portugene.com/EbolaID\\_definitions.html](http://ebolaid.portugene.com/EbolaID_definitions.html)): a) percentage of identical sites (PIS), calculated by dividing the number of equal positions in the alignment for an oligonucleotide by its length; b) percentage of identical sites in the last five nucleotides at the 3' end of oligonucleotides (3'PIS); and c) percentage of pairwise identity (PPI), calculated by counting the average number of pairwise matches across the positions of the alignment. The oligonucleotides can be selected by considering a ranking score ("EbolaID score") with the mean value of the three different measures (PIS, 3'PIS, and PPI). In general, the oligonucleotides with the highest values for these measures have a higher probability of binding to the target genomic region, increasing the probability of a positive detection. The three oligonucleotides with the highest EbolaID score (above 88) are located in the L gene (EBOLAID027, EBOLAID051, and EBOLAID036) and are part of the most conserved PCR primer pairs. All data can be accessed through the "Search" section of the database.

### Example of Use

Here we describe a hypothetical situation where a researcher needs an oligonucleotide for the detection of the Ebola viruses of the recent 2014 outbreak. The user can start by accessing the "Sequence alignments" tab at the top of the EbolaID homepage, which opens a table describing all sequence alignments included in the database. The user can visualize each alignment and progress to the table with the list of all oligonucleotides ordered according to the degree of conservation. Alternatively, the researcher can go to the "Search" tab and select "Oligonucleotides in specific alignments", which describes different measures of sequence conservation for each oligonucleotide and alignment. The list of oligonucleotides can be ordered by the PIS or PPI

measures. For example, the user will find that the most conserved oligonucleotide currently in our database for the “EbolaID\_Alig03” alignment is EBOLAID039. The most conserved oligonucleotides can also be found in the “The top alignment” section of the search page. All details on the selected oligonucleotide, including the sequence, genomic location, and references, can be found by clicking on the primer name (Fig 1). The researcher can also access the list with the most conserved genomic regions in each alignment in the top menu named “Genome variation” or in the table at the “Sequence alignments” section.

## Conclusion

Although several online resources include information on the *Ebolavirus* genus, the EbolaID database is by far the largest set of oligonucleotides currently available for this group of viruses. For example, the Hemorrhagic Fever Viruses (HFV) database [13] has curated nucleotide and protein sequence alignments describing the species and outbreak associated with each Ebola virus record but lacks information on target regions for molecular methods. The Virus Pathogen Database and Analysis Resource (ViPR) and the NCBI *Ebolavirus* databases [14,15] are very useful resources with multiple sequence alignments and phylogenetic trees but were not designed to guide the design of molecular methods. The EbolaID database can help researchers interested in designing accurate assays for the identification of the Ebola virus or to select conserved genomic regions for sequence-based therapeutics. The annotated reference genome and multiple sequence alignments can also be useful for epidemiological and phylogenetic studies.

## References

1. Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, et al. Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. *J Virol*. 2013; 87: 2608–16. doi: [10.1128/JVI.03118-12](https://doi.org/10.1128/JVI.03118-12) PMID: [23255795](https://pubmed.ncbi.nlm.nih.gov/23255795/)
2. Leroy EM, Baize S, Lu CY, McCormick JB, Georges AJ, Georges-Courbot MC, et al. Diagnosis of Ebola haemorrhagic fever by RT-PCR in an epidemic setting. *J Med Virol*. 2000; 60: 463–467. doi: [10.1002/\(SICI\)1096-9071\(200004\)60:4<463::AID-JMV15>3.0.CO;2-M](https://doi.org/10.1002/(SICI)1096-9071(200004)60:4<463::AID-JMV15>3.0.CO;2-M) PMID: [10686031](https://pubmed.ncbi.nlm.nih.gov/10686031/)
3. Liu Y, Shi Z-X, Ma Y-K, Wang H-T, Wang Z-Y, Shao D-H, et al. [Development of SYBR Green I real-time RT-PCR for the detection of Ebola virus]. *Bing Du Xue Bao*. 2012; 28: 567–571. PMID: [23233935](https://pubmed.ncbi.nlm.nih.gov/23233935/)
4. Jääskeläinen AJ, Moilanen K, Aaltonen K, Putkuri N, Sironen T, Kallio-Kokko H, et al. Development and evaluation of a real-time EBOV-L-RT-qPCR for detection of Zaire ebolavirus. *J Clin Virol*. 2015; 67: 56–58. doi: [10.1016/j.jcv.2015.04.003](https://doi.org/10.1016/j.jcv.2015.04.003) PMID: [25959160](https://pubmed.ncbi.nlm.nih.gov/25959160/)
5. Sozhamannan S, Holland MY, Hall AT, Negrón DA, Ivancich M, Koehler JW, et al. Evaluation of Signature Erosion in Ebola Virus Due to Genomic Drift and Its Impact on the Performance of Diagnostic Assays. *Viruses*. 2015; 7: 3130–54. doi: [10.3390/v7062763](https://doi.org/10.3390/v7062763) PMID: [26090727](https://pubmed.ncbi.nlm.nih.gov/26090727/)
6. Cherpillod P, Schibler M, Vieille G, Cordey S, Mamin A, Vetter P, et al. Ebola virus disease diagnosis by real-time RT-PCR: A comparative study of 11 different procedures. *J Clin Virol Elsevier*; 2016; 77: 9–14. doi: [10.1016/j.jcv.2016.01.017](https://doi.org/10.1016/j.jcv.2016.01.017)
7. Pettitt J, Higgs ES, Adams RD, Jahrling PB, Hensley LE. Use of existing diagnostic RT-PCR assays for detection of Ebola virus RNA in semen. *J Infect Dis*. 2015; doi: [10.1093/infdis/jiv454](https://doi.org/10.1093/infdis/jiv454)
8. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science (80-)*. 2014; 345: 1369–1372. doi: [10.1126/science.1259657](https://doi.org/10.1126/science.1259657) PMID: [25214632](https://pubmed.ncbi.nlm.nih.gov/25214632/)
9. Kugelman JR, Sanchez-Lockhart M, Andersen KG, Gire S, Park DJ, Sealfon R, et al. Evaluation of the potential impact of ebola virus genomic drift on the efficacy of sequence-based candidate therapeutics. *MBio*. 2015; 6. doi: [10.1128/mBio.02227-14](https://doi.org/10.1128/mBio.02227-14)
10. Grolla A, Lucht A, Dick D, Strong JE, Feldmann H. Laboratory diagnosis of Ebola and Marburg hemorrhagic fever. *Bull Pathol Exot*. 2005; 98: 205.
11. Towner JS, Sealy TK, Ksiazek TG, Nichol ST. High-throughput molecular detection of hemorrhagic fever virus threats with applications for outbreak settings. *J Infect Dis*. 2007; 196: S205–S212. PMID: [17940951](https://pubmed.ncbi.nlm.nih.gov/17940951/)

12. He J, Kraft AJ, Fan J, Van Dyke M, Wang L, Bose ME, et al. Simultaneous detection of cdc category "a" DNA and rna bioterrorism agents by use of multiplex pcr & rt-pcr enzyme hybridization assays. *Viruses*. 2009; 1: 441–459. PMID: [20224751](#)
13. Kuiken C, Thurmond J, Dimitrijevic M, Yoon H. The LANL hemorrhagic fever virus database, a new platform for analyzing biothreat viruses. *Nucleic Acids Res*. 2012; 40. doi: [10.1093/nar/gkr898](#)
14. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res*. 2012; 40. doi: [10.1093/nar/gkr859](#)
15. Brister JR, Bao Y, Zhdanov S a, Ostapchuck Y, Chetvermin V, Kiryutin B, et al. Virus Variation Resource—recent updates and future directions. *Nucleic Acids Res*. 2014; 42: D660–5. doi: [10.1093/nar/gkt1268](#) PMID: [24304891](#)