

VISMapper: ultra-fast exhaustive cartography of viral insertion sites for gene therapy

José M. Juanes^{1,2,+}, Asunción Gallego^{3,4,+}, Joaquín Tárraga^{2,5}, Felipe J. Chaves^{6,7}, Pablo Marín-García^{6,8}, Ignacio Medina⁵, Vicente Arnau^{1,2,8} and Joaquín Dopazo^{3,9,10,*}

¹ Departamento de Informática, Escuela Técnica Superior de Ingeniería (ETSE), Universidad de Valencia, 46100, Burjassot (Valencia), Spain.

² Computational Genomics Department, Prince Felipe Research Center, 46012, Valencia, Spain.

³ Clinical Bioinformatics Research Area, Fundación Progreso y Salud, Hospital Virgen del Rocío, 41013, Sevilla, Spain;

⁴ Bioinformatics in Rare Diseases (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Hospital Virgen del Rocío, 41013, Sevilla, Spain;

⁵ HPC Service, University Information Services, University of Cambridge, UK.

⁶ Genotyping and Genetic Diagnosis Unit, Health Research Institute, INCLIVA, Valencia, Spain

⁷ CIBERDem, Health Institute Carlos III, Madrid, Spain

⁸ Institute for Integrative Systems Biology (I2SysBio). Universidad de Valencia-CSIC. 46980, Paterna (Valencia), Spain.

⁹ Bioinformatics and Data Analysis Unit, Genomic Medicine Institute, Imegen, Valencia, Spain.

¹⁰ Functional Genomics Node, INB-ELIXIR-es, Hospital Virgen del Rocío, 42013, Sevilla, Spain.

+These authors have contributed equally to the work

*Corresponding author

ABSTRACT

Background: The possibility of integrating viral vectors to become a persistent part of the host genome makes them a crucial element of clinical gene therapy. However, viral integration has associated risks, such as the unintentional activation of oncogenes that can result in cancer. Therefore, the analysis of integration sites of retroviral vectors is a crucial step in developing safer vectors for therapeutic use.

Results: Here we present VISMMapper, a vector integration site analysis web server, to analyze next-generation sequencing data for retroviral vector integration sites. VISMMapper can be found at: <http://vismapper.babelomics.org>.

Conclusions: Because it uses novel mapping algorithms VISMMapper is remarkably faster than previous available programs. It also provides a useful graphical interface to analyze the integration sites found in the genomic context.

Keywords: gene therapy, viral insertion, viral integration, sequence mapping, genome viewer

BACKGROUND

The stable, long-term correction of diseases by integrating viral vectors carrying healthy copies defective genes in the patient's genome has become mainstream procedure in clinical gene therapy [1, 2]. However, despite its successful application, viral integration based therapies are not exempt of risks, such as the accidental activation of oncogenes that can cause malignant transformation of the cells [3, 4]. Vector locations in the host genome constitute molecular markers that help monitoring the fate of affected cells. Analysis of vector insertion sites (ISs) is carried out by the amplification (currently using Next Generation Sequencing –NGS- technologies) of sequences from retroviral vectors with a long terminal repeat (LTR). Primers mapping LTRs produce sequence reads with LTR-chromosome junctions, which can be used to accurately determine the chromosomal region of insertion of the viral vector [4]. Such monitoring is required because it is known that distinct gene transfer vectors can have preferences to target gene coding regions, CpG islands, or transcriptional start sites [5-7].

Here we present a new web server, VISMMapper, a web tool to manage sequencing data for the detection of viral vector insertion sites in gene therapy experiments. VISMMapper is much faster than other alternative software available and provides a comprehensive graphic interface that allows interactive visualization of the viral ISs in the genomic context.

IMPLEMENTATION

VISMMapper is written in Node.js (a JavaScript runtime) and uses GenomeMaps [8] for the visual representation of the results in the context of the genome. Thus the resulting viral insertion sites of an experiment can be visualized along with the genomic features they have around, including reads mapped,

genes and other type of genomic elements. Supported assemblies for the human genome are GRCh37 and GRCh38.

Cancer genes were taken from the COSMIC [9] database through the CellBase [10] webservices.

FEATURES

Data upload and workspace

VISMapper reads standard FASTQ or FASTA files containing reads corresponding to the insertion sites of the virus. If FASTA files are provided, they are converted to FASTQ format. Since FASTA files lack the quality parameter, this is set to 20 by default for the FASTQ file generated. A value of 20 minimizes the false positive rate when the original sequences are **of a reasonable** quality. In any case, the use of FASTQ containing quality values is obviously preferable. Files can be ZIP compressed. During the upload, user can optionally provide an email to be notified of the end of the data processing (given the speed of data processing it is usually unnecessary).

Read mapping

Reads in the FASTQ file are mapped onto the reference human genome using BWA [11] or HPG-Align [12]. Typically mapping runtimes are in the range of seconds, which makes of VISMapper a truly interactive and accurate tool for exploring the result of retroviral insertion experiments. IS locations are detected by identified reads partially mapped. We use the CIGAR information for this. When the CIGAR of a mapping contains soft or hard clippings it indicates that the corresponding read have part of the genome sequence and part of the viral sequence. The reads are arranged by chromosome using SAMTools [13] and are inserted in a MySQL database for facilitating a faster access to them.

Dashboard

The Dashboard is a graphical working environment composed by three panels: the karyotype viewer, the genome viewer and the control panel (See Figure 1). The karyotype viewer provides a general perspective of all the ISs along the chromosomes. Clicking with the left mouse button magnifies the chromosome, with ISs marked as red lines. Exact details on the IS location are provided by setting the cursor over them. A vertical panel on its left (See Figure 1) allows filtering IS by the number of reads supporting them. It

also allows searching those reads which are closer to oncogenes or genes related to specific tumor types. When the mouse hovers the chromosome in the karyotype a detailed view of the selected chromosome with the IS is displayed. Setting the mouse over the ISs pops up information on its exact location and the number of reads supporting it.

A more detailed view of the region in which the ISs occur (that can be selected by clicking in the karyotype viewer) can be obtained with the genome viewer, which implements GenomeMaps [8]. Several tracks are available at different detail level depending on the zoom level in the genome viewer: a) the surrounding genomic region, b) oncogenes located in the neighborhood (the cursor over them displays information on the genes) and c) reads mapped around the IS (again, information on the read, such as strand, mapping quality, etc. is provided by hovering the mouse on them)

Finally, the control panel allows setting a threshold based on the number of reads that support ISs and allows finding specific cancer genes or genes of specific cancer types (see Figure 1, left part). Specifically, a box allows setting a threshold with the minimum number of reads to consider a IS (5 by default). The second box allows selecting a specific oncogene (can be searched by name or selected from a list). The list of oncogenes has been extracted from COSMIC. Another box allows displaying only the genes known to be associated with a given tumor.

Report

The control panel allows generating a comprehensive tabular report of the results found. The button report directs to another page with a table containing all the ISs found that can be arranged by all the criteria shown in the header of the columns (chromosome, position, quality, etc.) Different filters (number of reads that support the IS and distance to a cancer gene) can be applied to expand or reduce the number of ISs to consider. This list can be downloaded in tab delimited format and a BAM file with the alignments found by the mapper can also be downloaded.

For any IS considered with the filtering schema used, the report contains the following items:

- Chromosome
- Position
- Number of reads mapped in this position
- Average quality of all the reads mapped in the position

- Closest oncogene
- Distance to the oncogene (0 means that the IS maps within the oncogene)
- Position of the oncogene with respect to the IS
- Entrez entry of the oncogene
- URL to the Entrez entry of the oncogene

COMPARISON TO OTHER WEB SERVERS FOR VIRAL IS MAPPING

There are a few web servers for viral vector insertion site analysis, such as, HISAP [14], SeqMap (requires user registration) or QuickMap [15], or the recently published VISA [16]. However, all of them use BLAST [17] or BLAT [18] for read mapping that involve comparatively much longer runtimes. Figure 2 shows a comparative of runtimes where the increase in speed gained by the use of more sophisticated mapping algorithms in VISMMapper is obvious. The data used in the comparison were taken from the VISA website and can also be downloaded at the VISMMapper documentation site (<https://github.com/jmjuanes/vismapper/tree/master/ismapper-test>).

In addition, a more detailed comparison was made with the VISA program by generating 4 datasets with known number of IS using the IS generator program from the VISA website (https://visa.pharmacy.wsu.edu/bioinformatics/random_site_generator.html). Table 1 shows the results of the comparison. Relative runtimes are similar to the ones shown in Figure 2. While both methods give a very small number of false positives, in general VISMMapper was able to map a higher percentage of sequences and found more IS sites than VISA.

In addition, QuickMap does not process more than 50,000 sequences and VISA limits are between 50,000 and 100,000. HISAP could manage up to 100,000 in about 50 minutes, but **cannot** arrive to 250,000 sequences. Moreover, none of the other programs provide a graphic interface to analyze the results. Furthermore, QuickMap and HISAP do not support GRCh38.

CONCLUSIONS

Because of its speed and sensitivity, VISMMapper constitutes an attractive alternative to the options available for viral insertion site analysis. VISMMapper offers a unique, interactive graphical working

environment that allows a detailed and exhaustive exploration of the consequences and potential risks of the viral vectors inserted in the analyzed genome.

LIST OF ABBREVIATIONS

BAM: Binary Alignment Map

BWA: Burrows–Wheeler Algorithm

IS: Insertion Site

LTR: long terminal repeat

NGS: Next Generation Sequencing

DECLARATIONS

Ethics and consent to participate

Not applicable

Consent to publish

Not applicable

Availability of data and materials

VISMapper can be found at: <http://vismapper.babelomics.org>.

VISMapper code can be found in the GitHub repository <https://github.com/jmjuanes/vismapper>

Associated documentation can be found at: <https://github.com/jmjuanes/vismapper/wiki>

The data used in the general comparison can be found at:
<https://github.com/jmjuanes/vismapper/tree/master/ismapper-test>

Competing interests

The authors declare that they have no competing interests

Authors' contributions

JMJ, and AG programmed the code, JT and IM programmed and optimized the mapping of sequences, FJC and PMG helped with the programming, VA coordinated the programming work and JD conceived the work and wrote the paper. All the authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

Not applicable

FUNDING

This work is supported by grants BIO2014-57291-R from the Spanish Ministry of Economy and Competitiveness (MINECO), and Plataforma de Recursos Biomoleculares y Bioinformáticos PT13/0001/0007 from the ISCIII, both co-funded with European Regional Development Funds (ERDF); H2020-INFRADEV-1-2015-1 ELIXIR-EXCELERATE (ref. 676559). None of the funding bodies played any role in the design or conclusions of the study.

REFERENCES

1. Gaspar HB, Parsley KL, Howe S, King D, Gilmour KC, Sinclair J, Brouns G, Schmidt M, Von Kalle C, Barington T *et al*: **Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector**. *Lancet* 2004, **364**(9452):2181-2187.
2. Cartier N, Hacein-Bey-Abina S, Bartholomae CC, Veres G, Schmidt M, Kutschera I, Vidaud M, Abel U, Dal-Cortivo L, Caccavelli L *et al*: **Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy**. *Science* 2009, **326**(5954):818-823.
3. Cavazzana-Calvo M, Payen E, Negre O, Wang G, Hehir K, Fusil F, Down J, Denaro M, Brady T, Westerman K *et al*: **Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia**. *Nature* 2010, **467**(7313):318-322.
4. Paruzynski A, Arens A, Gabriel R, Bartholomae CC, Scholz S, Wang W, Wolf S, Glimm H, Schmidt M, von Kalle C: **Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing**. *Nat Protoc* 2010, **5**(8):1379-1395.
5. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F: **HIV-1 integration in the human genome favors active genes and local hotspots**. *Cell* 2002, **110**(4):521-529.
6. Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD: **Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences**. *PLoS Biol* 2004, **2**(8):E234.
7. Wu X, Li Y, Crise B, Burgess SM: **Transcription start regions in the human genome are favored targets for MLV integration**. *Science* 2003, **300**(5626):1749-1751.

8. Medina I, Salavert F, Sanchez R, de Maria A, Alonso R, Escobar P, Bleda M, Dopazo J: **Genome Maps, a new generation genome browser.** *Nucleic Acids Res* 2013, **41**(Web Server issue):W41-46.
9. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A *et al*: **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic Acids Res* 2011, **39**(Database issue):D945-950.
10. Bleda M, Tarraga J, de Maria A, Salavert F, Garcia-Alonso L, Celma M, Martin A, Dopazo J, Medina I: **CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources.** *Nucleic Acids Res* 2012, **40**(Web Server issue):W609-614.
11. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754-1760.
12. Tarraga J, Arnau V, Martinez H, Moreno R, Cazorla D, Salavert-Torres J, Blanquer-Espert I, Dopazo J, Medina I: **Acceleration of short and long DNA read mapping without loss of accuracy using suffix array.** *Bioinformatics* 2014, **30**(23):3396-3398.
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078-2079.
14. Arens A, Appelt JU, Bartholomae CC, Gabriel R, Paruzynski A, Gustafson D, Cartier N, Aubourg P, Deichmann A, Glimm H *et al*: **Bioinformatic clonality analysis of next-generation sequencing-derived viral vector integration sites.** *Human gene therapy methods* 2012, **23**(2):111-118.
15. Appelt JU, Giordano FA, Ecker M, Roeder I, Grund N, Hotz-Wagenblatt A, Opelz G, Zeller WJ, Allgayer H, Fruehauf S *et al*: **QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis.** *Gene Ther* 2009, **16**(7):885-893.
16. Hocum JD, Battrell LR, Maynard R, Adair JE, Beard BC, Rawlings DJ, Kiem HP, Miller DG, Trobridge GD: **VISA--Vector Integration Site Analysis server: a web-based server to rapidly identify retroviral integration sites from next-generation sequencing.** *BMC Bioinformatics* 2015, **16**:212.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
18. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.

FIGURE LEGENDS

Figure. 1. Screenshot showing the different graphical representations in the dashboard: the karyotype viewer and the genome viewer. Also, a table with the list of IS found is displayed.

Figure. 2. Runtimes observed for different programs QuickMap (line with diamonds), VISA (line with squares) HISAP (line with triangles) and VISMMapper (line with circles) with datasets of increasing sizes. In the case of QuickMap, VISA and HISAP, the lines are interrupted according to internal hard limits for the number to sequences that the programs can process.

TABLE LEGENDS

Table 1. Comparison of VISA and VISMMapper using four datasets generated with the IS generator program from the VISA website (https://visa.pharmacy.wsu.edu/bioinformatics/random_site_generator.html)