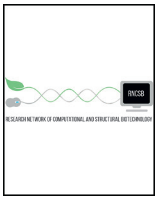
Computational and Structural Biotechnology Journal 14 (2016) 87–90

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj



Short communication

# Common integration sites of published datasets identified using a graph-based framework

Alessandro Vasciaveo [a,b], Ivana Velevska [a], Gianfranco Politano [b], Alessandro Savino [b], Manfred Schmidt [a], Raffaele Fronza [a,*]

[a] Department of Translational Oncology, National Center for Tumor Diseases and German Cancer Research Center, Im Neuenheimer Feld 581, 69120 Heidelberg, Germany
[b] Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Available online 29 November 2015

ABSTRACT

With next-generation sequencing, the genomic data available for the characterization of integration sites (IS) has dramatically increased. At present, in a single experiment, several thousand viral integration genome targets can be investigated to define genomic hot spots. In a previous article, we renovated a formal CIS analysis based on a rigid fixed window demarcation into a more stretchy definition grounded on graphs. Here, we present a selection of supporting data related to the graph-based framework (GBF) from our previous article, in which a collection of common integration sites (CIS) was identified on six published datasets. In this work, we will focus on two datasets, $IS_{RTCGD}$ and $IS_{HIV}$, which have been previously discussed. Moreover, we show in more detail the workflow design that originates the datasets.

## Specifications Table

| | |
|---|---|
| Subject area | Computational biology, systems biology |
| More specific subject area | Gene therapy, integrational mutagenesis analysis |
| Type of data | Table, image, dataset |
| How data was acquired | In silico experiments |
| Data format | Analyzed datasets, analyzed Excel tables, PNG files |
| Experimental factors | Integration sites datasets were analyzed with a new computational method for common integration sites identification |
| Experimental features | A proposed set of common integration sites from two published integration sites datasets (see [1]) |
| A pathway enrichment analysis is also reported | |
| Data source location | Heidelberg, Germany |
| Data accessibility | Data is with this article and in ref. [1] |

## Value of the data

- The analyzed dataset here provided can be used as benchmark to compare the results of the graph modeling approach for CIS identification and analysis implemented in software tools.
- Graph modeling approach to the identification of common integration sites.
- Validation of the graph-based framework (GBF) against well-known datasets.
- Detailed illustrated procedure for the identification of CIS via GBF.

## 1. Data

The dataset containing the identified CIS from the Retroviral Tagged Cancer Gene Database (RTCGD) [6] is provided in Table 1 Appendix A and it is obtained by using a Cytoscape 2.8 plugin, which implements some of the features of the GBF method (see how to retrieve the code in [1]). The other datasets are collected using a normal Internet browser. Fig. 1 shows a Venn diagram in which two datasets are compared. The first dataset is the collection of all the genes found with the GBF method, while the second dataset is the list of genes provided by RTCGD which uses the standard window method (SWM) to identify CIS and the next gene approach (NGA) to discover and associate an annotated

[*] Corresponding author.
E-mail address: raffaele.fronza@nct-heidelberg.de (R. Fronza).

http://dx.doi.org/10.1016/j.csbj.2015.11.004

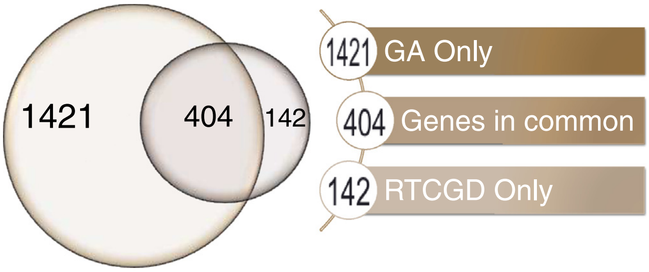


**Fig. 1.** Venn diagram of the gene atmosphere of all identified CIS from the RTCGD dataset using the GBF (graph-based framework) [1] and using the SWM (standard window method) [2].

gene to the identified CIS. For further details about the two approaches, see [1]. With the GBF method, it is possible to discover 1421 genes which are not present in the RTCGD dataset. Only 142 genes were not discovered by the GBF method while they are present in the RTCGD gene list, and 404 of the genes can be found by both methods.

## 2. Experimental design, materials and methods

### 2.1. Experiment workflow

The workflow of the analysis is depicted in Fig. 2. The input is a dataset composed of a list of integration sites (IS). The graph-based framework (GBF) presented in [1] is adopted to perform all the following analyses. The first step is the CIS identification and the computation of some statistics for every CIS. Further steps are optional but they have to follow the order. The second step consists of enhancing the CIS dataset with information from genomic annotated data. This step generates the gene atmosphere (GA) dataset as shown in Table 2 Appendix A. Using the GA dataset, the next step consists of the functional analysis, as shown in Table 3 Appendix A.

### 2.2. Data preparation

The dataset used for the analysis should contain few attributes in order to be properly analyzed by the GBF method. Some of these attributes are mandatory and they are shown in Table 1. The mandatory attributes for the CIS enhancing phase are shown in Table 2.

**Table 1**
Mandatory attributes of the input dataset for the identification of CIS using the GBF method.

| Attributes | Description |
|---|---|
| Chromosome number | The ordinal number of the chromosome in which the integration event was found |
| Insertion site position | The position on the genome: a very long integer number representing the base pair where the virus was integrated |
| Entropy label (e.g. Kind of tumor, virus type) | Meta-information used for the computation of the CIS entropy. It is a label that represents a factor of the experiment. For example, it could be the tumor model or type from which the IS has been associated |

### 2.3. Common integration sites identification

The method presented in [1] allows the identification of CIS on the basis of very few attributes found in the dataset under analysis (see Table 1). Fig. 3 shows the flowchart of the global method that builds the model and identifies the CIS with their statistics.

Starting from the dataset containing the integration sites (IS dataset), it is convenient to order the dataset according to the integration position to improve the algorithm efficiency. This is the data preparation part (Table 1). Afterwards, as depicted in Fig. 3, the building of the model starts creating an empty graph. For every IS present in the dataset, a node is created and added to the graph. A nested loop checks if all the vertices instantiated in the graph are at a distance below a certain threshold from the current IS previously added as a node to the graph itself. An edge connecting two nodes of the same type (i.e. two IS nodes) is created and added to the graph if the distance is lower than the threshold. When all the IS from the dataset are analyzed, the main loop terminates and the graph is ready to be analyzed by the main algorithm for CIS identification. This algorithm can be implemented in different ways (e.g. an algorithm that extracts the connected components (CC) from an undirected and disconnected graph). An efficient version of this algorithm is presented in [3].

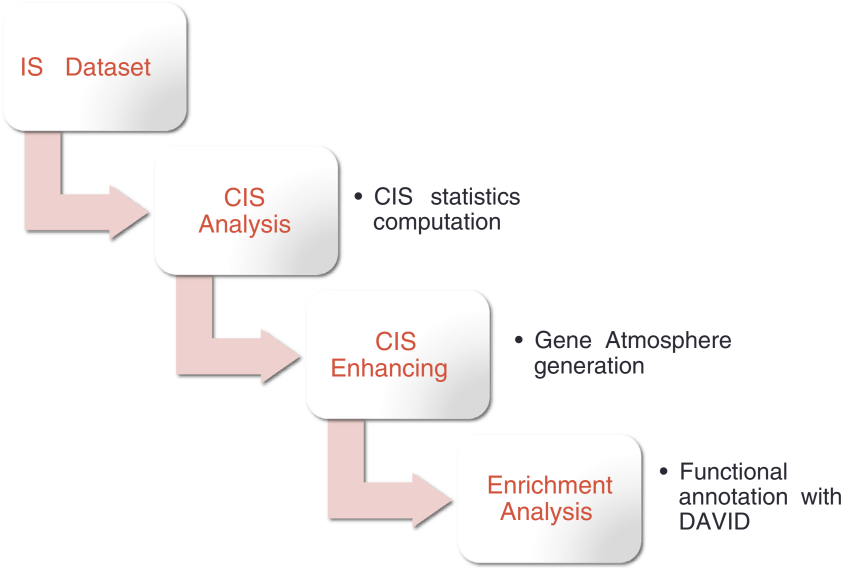

IS Dataset
CIS Analysis
• CIS statistics computation
CIS Enhancing
• Gene Atmosphere generation
Enrichment Analysis
• Functional annotation with DAVID

**Fig. 2.** Workflow of the full analysis process: starting from the raw dataset to the functional analysis.