
IntOGen FAQs

Biomedical Genomics Lab • www.intogen.org • January 28, 2010





Table of Contents

General	2
Q: <i>What is IntOGen?</i>	2
Q: <i>How do I cite IntOGen?</i>	2
Q: <i>Who has developed IntOGen?</i>	2
Q: <i>What is “an experiment” in IntOGen?</i>	2
Q: <i>What is a “module” in IntOGen?</i>	2
Q: <i>Why is IntOGen gene-centric?</i>	2
Data collection and pre-processing	3
Q: <i>How is the data in IntOGen collected and annotated?</i>	3
Q: <i>Do you process the raw data yourself or use authors processed data?</i>	3
Q: <i>Can any data set be included in IntOGen? Do you have specific data requirements?</i>	3
Statistics	4
Q: <i>How are different experiments combined?</i>	4
Q: <i>What are the results that I see when I click on a module other than genes/bands? How should I interpret them?</i>	4
Q: <i>What is CGPrio?</i>	4

GENERAL

Q: What is IntOGen?

A: IntOGen is a project to integrate multidimensional data from numerous independent experiments to identify the most important genes and modules involved in the development of different cancer types.

Q: How do I cite IntOGen?

A: Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney S and Lopez-Bigas N. IntOGen: Integration and data-mining of multidimensional oncogenomic data. [Nature Methods 7, 92-93](#) (2010).

Q: Who has developed IntOGen?

A: IntOGen is a research project developed in the [Biomedical Research Lab](#) at the [University Pompeu Fabra](#).

Q: What is “an experiment” in IntOGen?

A: An experiment is a set of samples that can be analyzed together according to some requirements. The requirements for oncogenomic experiments are that all the samples come from the same publication, they have been analyzed with the same platform and they share a particular clinical annotation.

Q. What is a “module” in IntOGen?

A. A module is a set of genes related to each other in a biologically meaningful way. For example, all genes with the Gene Ontology term “cell cycle” form a module.

Q. Why is IntOGen gene-centric?

A. We do analyze the oncogenomic data at the lowest level possible (probes, genomic coordinates, etc.). However, since most of the biological knowledge accumulates in “genes”, we map all detected alterations to the genes affected to be able to make use of biological annotations. However, the system is very generic. We model the oncogenomic data in such a way that we can focus on any biological entity when needed.

DATA COLLECTION AND PRE-PROCESSING

Q: How is the data in IntOGen collected and annotated?

A: We collect genome-wide experimental data from different public repositories (such as [GEO](#), [ArrayExpress](#), [Progenetix](#), [COSMIC](#), [TCGA](#), etc). Samples in every experiment are annotated manually with International Classification of Disease ([ICD-10](#) and [ICD-0](#)) according to original authors' information.

Q: Do you process the raw data yourself or use authors processed data?

A: For the integrative methodology in IntOGen, what we need is a measure of how much the cancerous sample is altered when compared to normal samples. When raw data is available, we preprocess it ourselves in order to calculate a measure to the comparison with a standardized pipeline. This pipeline depends on the type of platform. If raw data is not available, but such a measure is included in the analysis results of a study (for example log2 ratio), we include the results directly as they are calculated by the authors of the original publication.

Q. Can any data set be included in IntOGen? Do you have specific data requirements?

A: For statistical reliability, we require at least 20 tumor samples having the same ICD topography (tissue) annotation. In addition, for transcriptomic data, since the current implementation compares tumorous to normal samples, we need at least one normal sample of the same tissue in the same "experiment". Basically, we should be able to calculate a measure of how the tumorous sample is altered when compared to normal ones.

STATISTICS

Q: How are different experiments combined?

P-values for different experiments are combined with the weighted Z-method using the following formula:

$$Z' = \frac{\sum_{i=1}^N Z_i \cdot \omega_i}{\sqrt{\sum_{i=1}^N \omega_i^2}}$$

Where:

Z_i = Z-score corresponding to the p-value of the experiment

ω_i = Weight of the experiment determined by the number of samples analyzed

N = Number of experiments to combine

For more information on this method see:

Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol. 2005 Sep;18(5):1368-73.

[\[LINK\]](#)

Q: What are the results that I see when I click on a module other than genes/bands? How should I interpret them?

A. The p-value for a module, such as a pathway, indicates if the genes in that pathway are altered in different cancer types more than expected by chance. A significant p-value for up-regulation, for example, means that a significant number of genes annotated in that pathway is up-regulated.

Q: What is CGPrio?

A: CGprio is a resource for the prioritization of candidate cancer genes after genomic experiments. The prioritization of oncogenes and tumor suppressor genes is based on computational classifiers that use different combinations of sequence and functional data including sequence conservation, protein

domains and interactions, and regulatory data. The method assigns a prediction rank, ranging from 0 to 1, which corresponds to the likelihood of a gene being a proto-oncogene or a tumor suppressor gene according to our prediction method.

For more information see the manuscript describing the development and validation of the classifiers:

Furney SJ, Calvo B, Larranaga P, Lozano JA, Lopez-Bigas N. Prioritization of candidate cancer genes - an aid to oncogenomic studies. *Nucleic Acids Research*, 36(18):e115 (2008).

[CGprio web]

<http://bg.upf.edu/cgprio>

[Open access PDF]

<http://nar.oxfordjournals.org/cgi/reprint/gkn482v1>

[PUBMED]

<http://www.ncbi.nlm.nih.gov/pubmed/18710882>