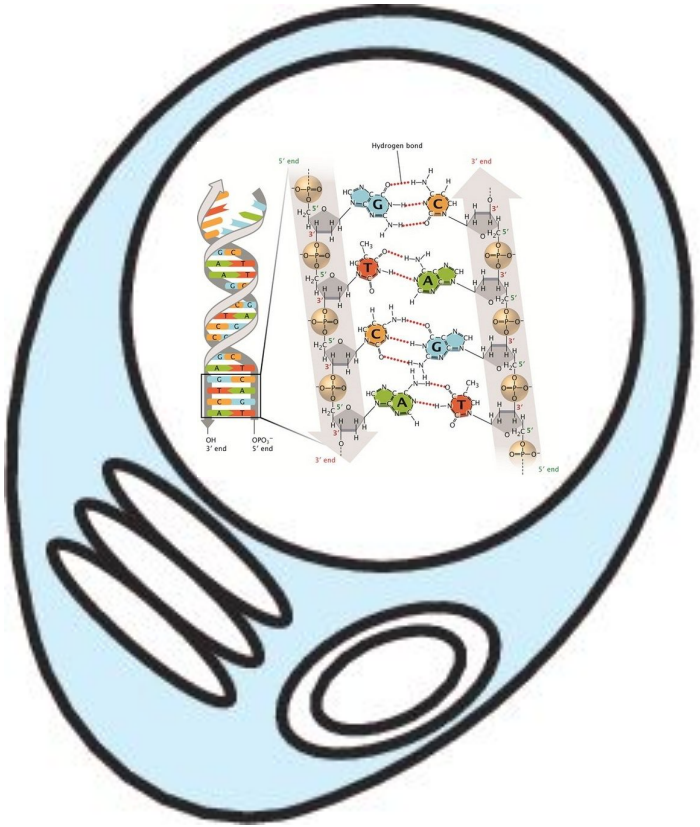
Computational approaches for integrative cancer genomics

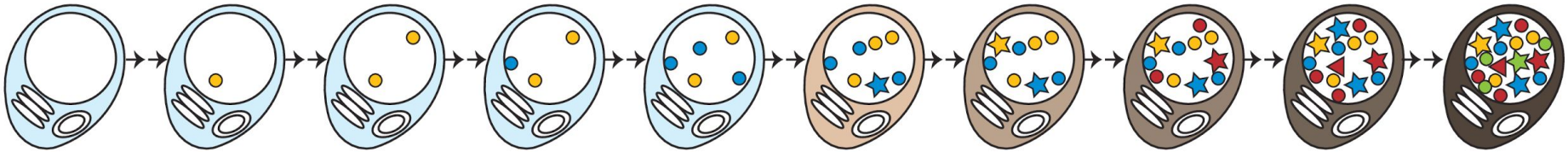
Christian Pérez Llamas

18th Dec 2015

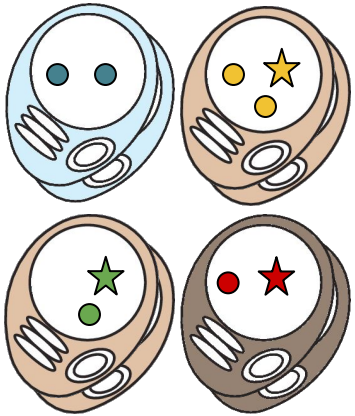
Integrative Cancer Genomics

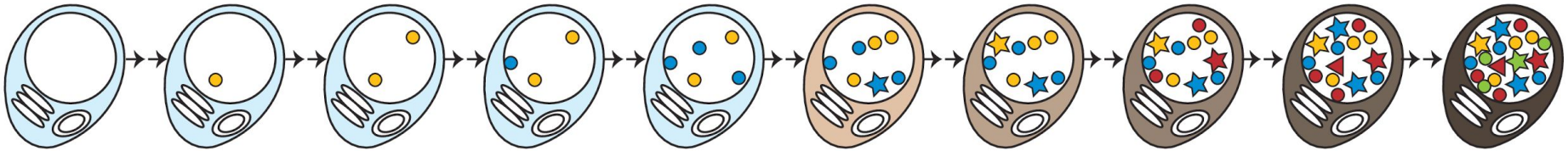


Cancer development

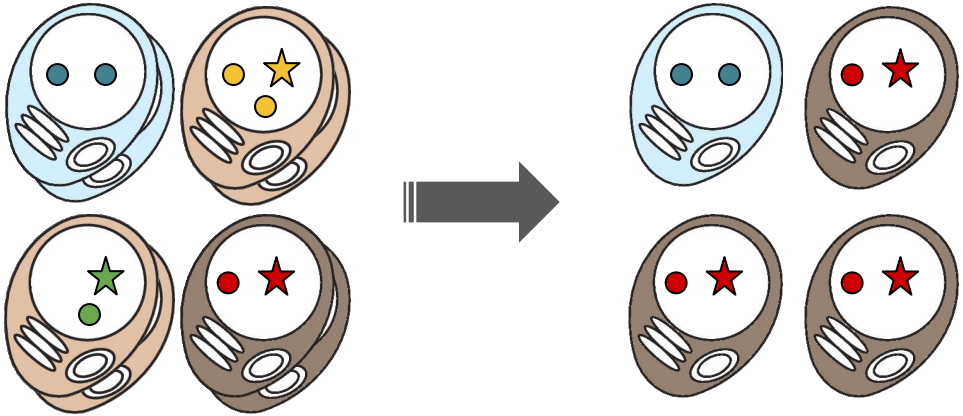


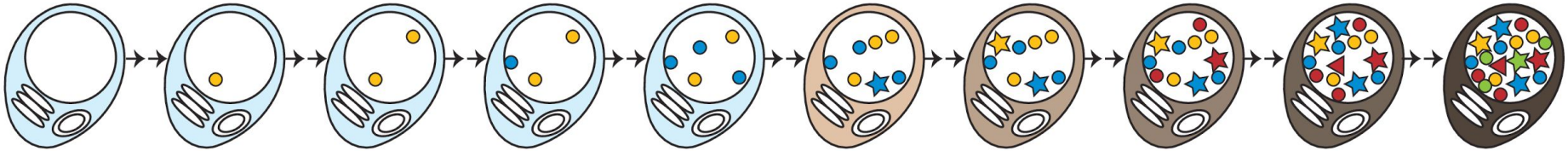
Cancer development



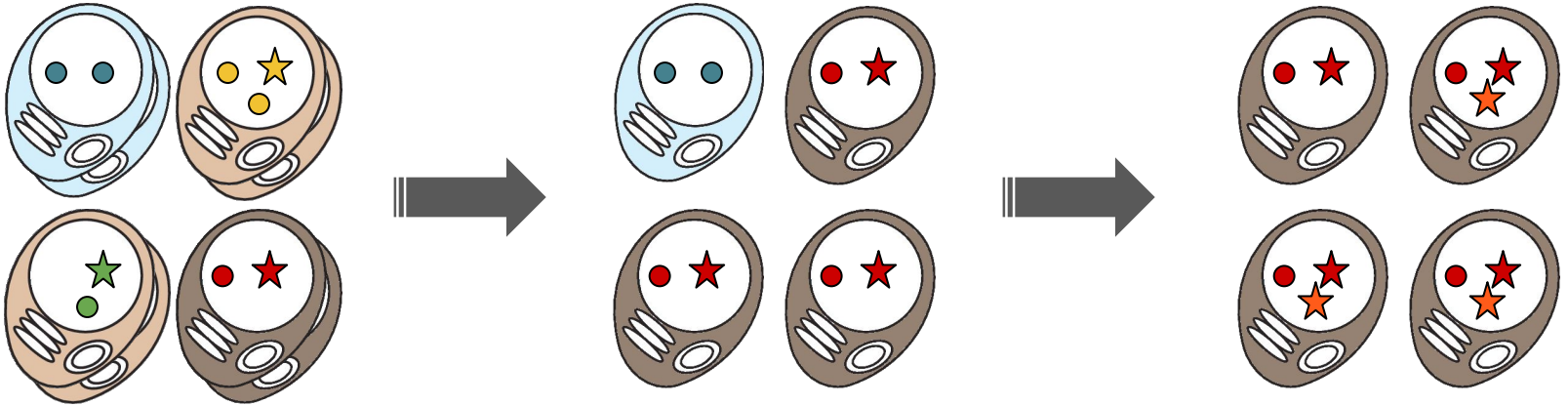


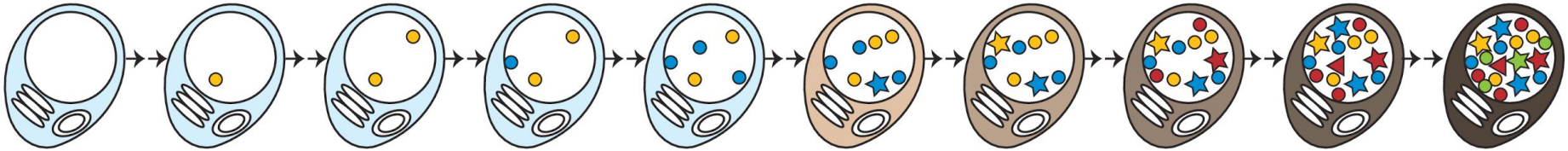
Cancer development





Cancer development



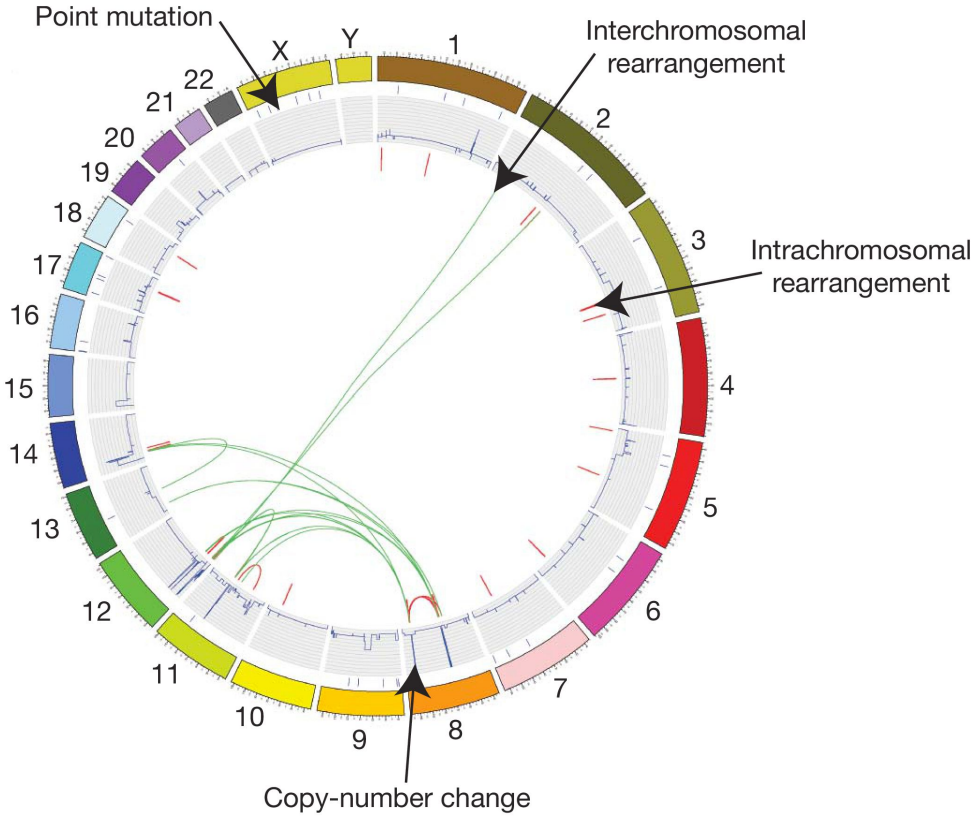


Genomic alterations

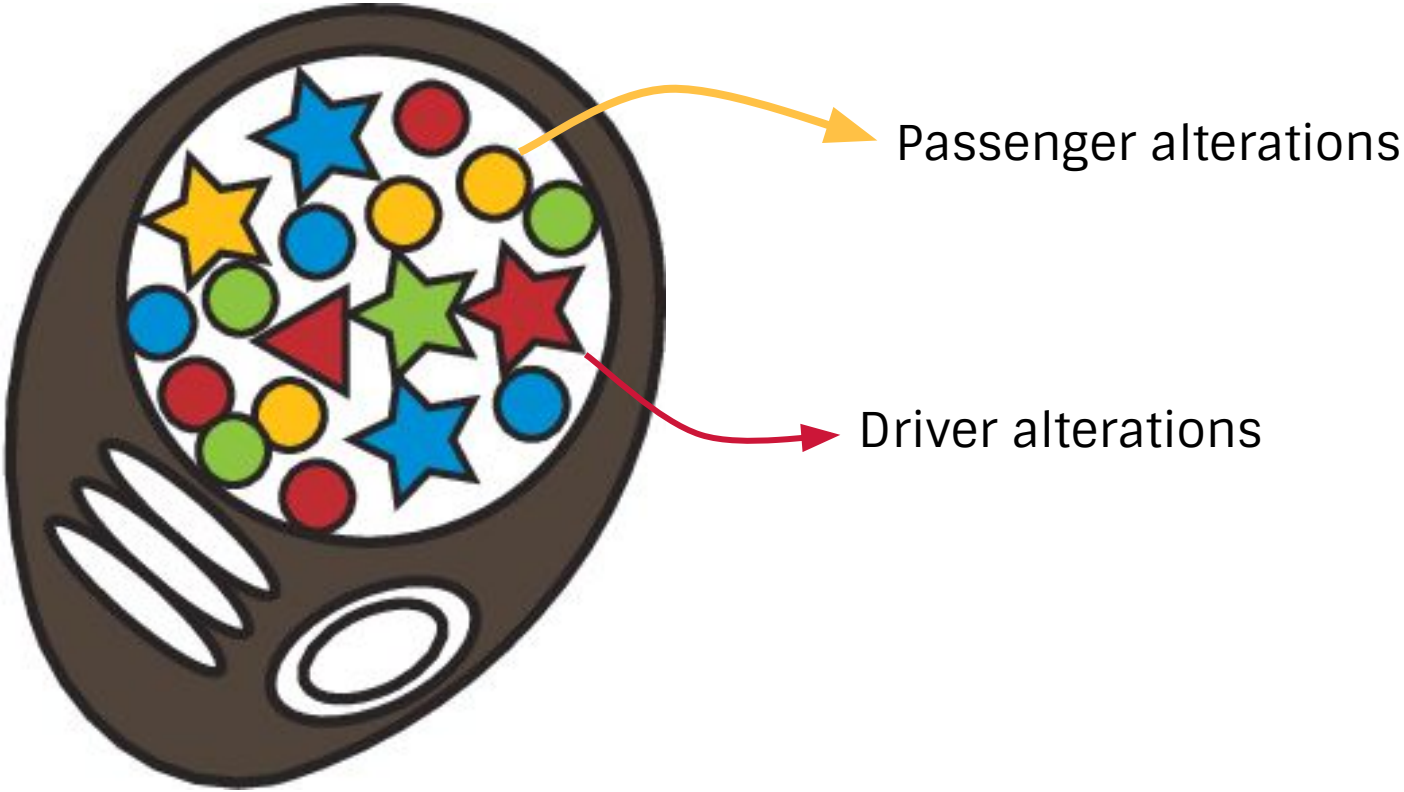
Somatic alterations

Germline alterations

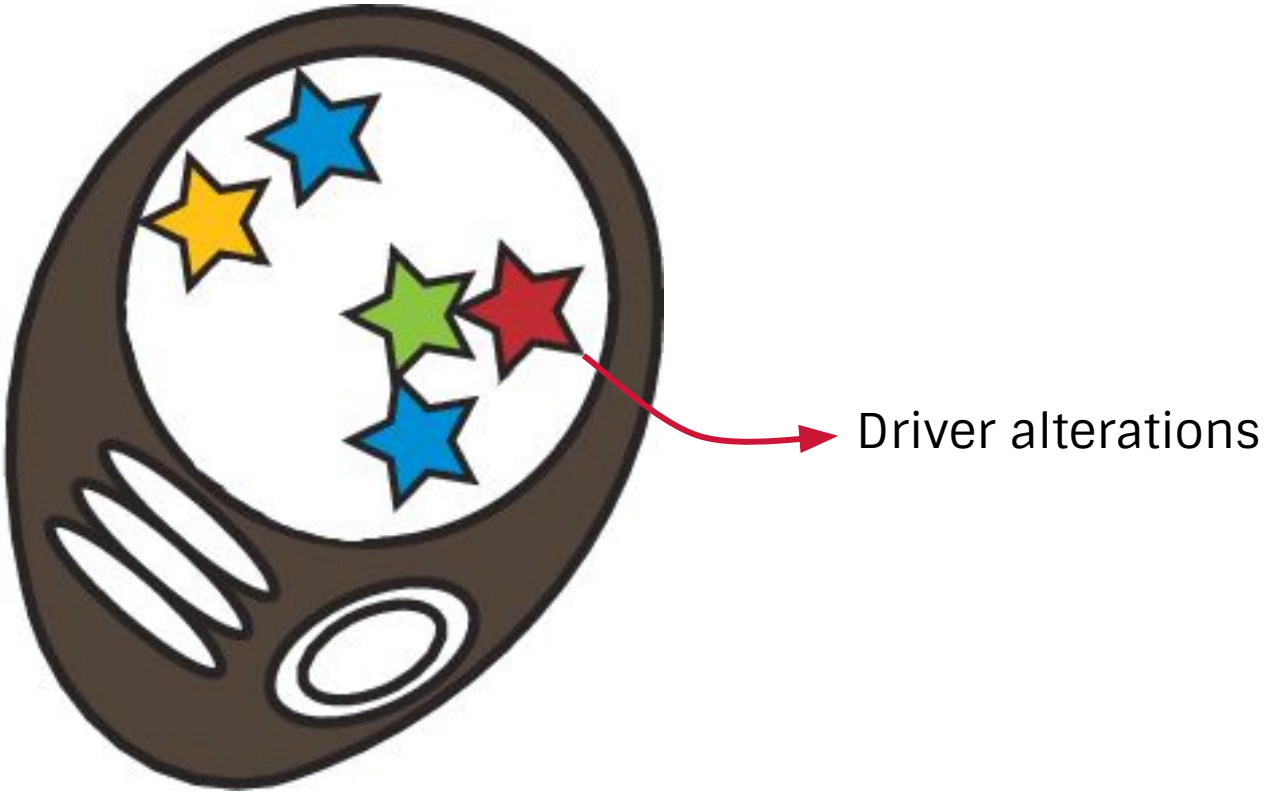
Genomic alterations



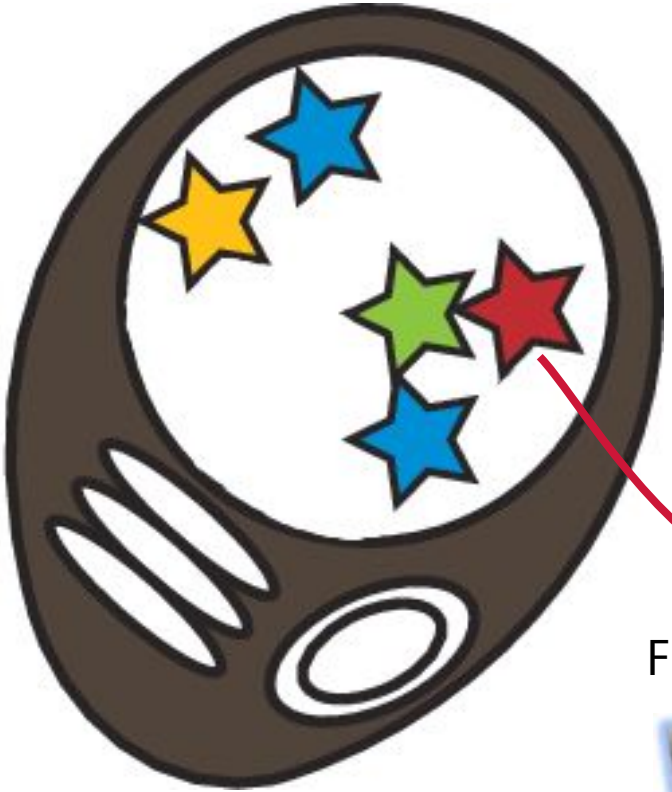
Genomic alterations



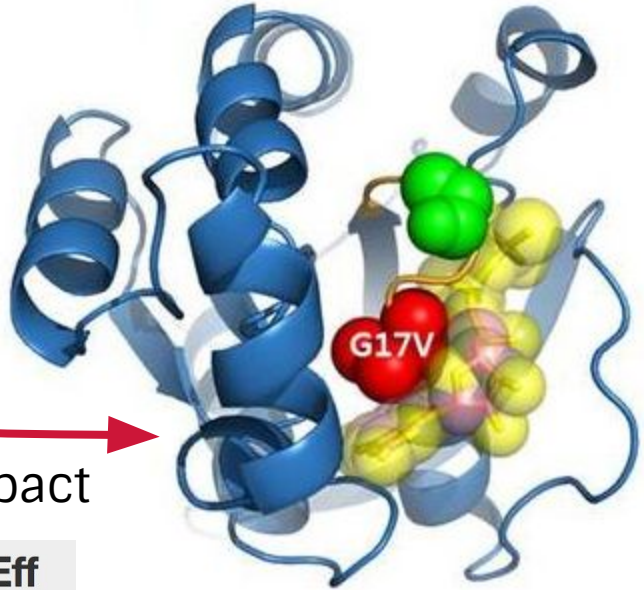
Identification of driver alterations



Identification of driver alterations



Functional Impact

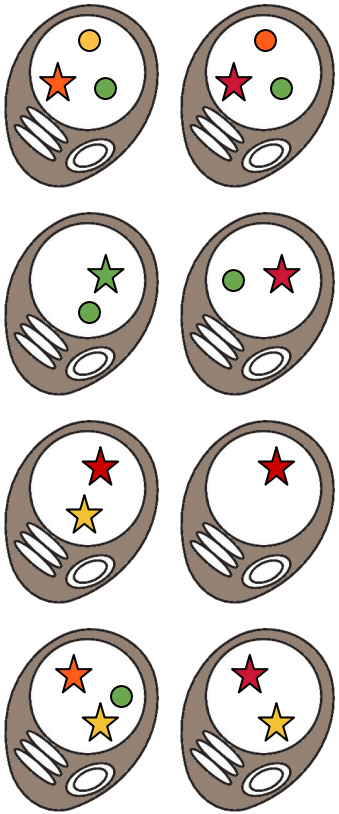


VeP

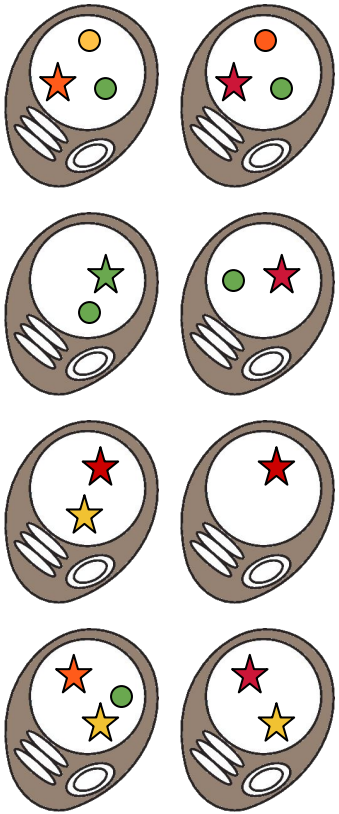
SnpEff
Genetic variant annotation and effect prediction toolbox.

Identification of driver alterations

Identification of driver alterations



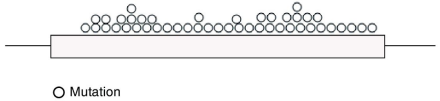
Identification of driver genes



Recurrence

R MuSiC-SMG

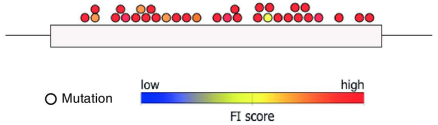
Identify genes mutated more frequently than background mutation rate



FM bias

F OncodriveFM

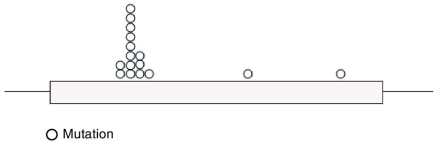
Identify genes with a bias towards high functional mutations (FM bias)



CLUST bias

C OncodriveCLUST

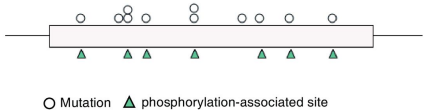
Identify genes with a significant regional clustering of mutations



ACTIVE bias

A ActiveDriver

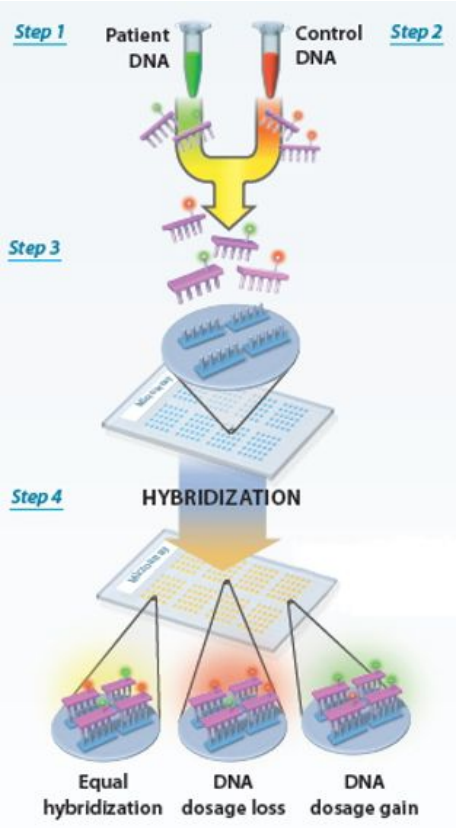
Identify genes significantly enriched in mutations affecting phosphorylation-associated sites

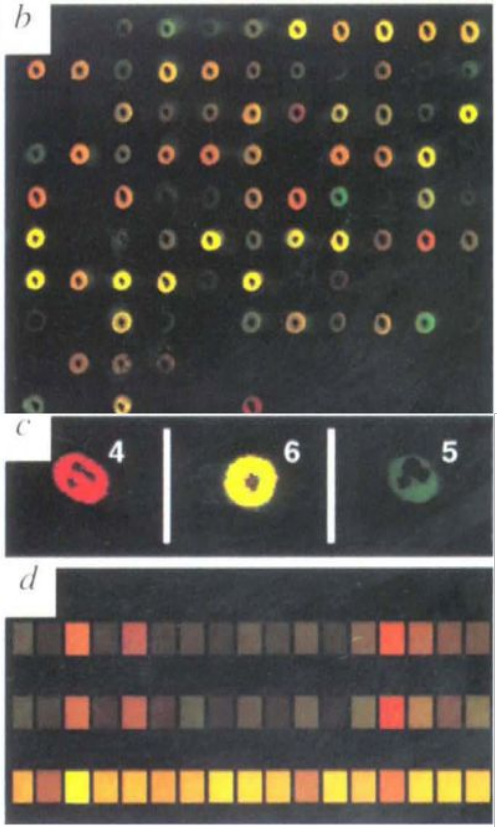


Advances in Technology

Human Genome Project	Begin	End	+10 years
Genome sequencing			
Cost to Generate a Human Genome Sequence	~$1 billion	~$10-50 million	~$3-5 thousand
Time to Generate a Human Genome Sequence	~6-8 years	~3-4 months	~1-2 days
Human Genome Sequences	0	1	Thousands
Genome Sequence Data			
Total DNA Bases in GenBank	~49 Mbases	~31 Tbases	~150 Tbases
Human Single-Nucleotide Polymorphisms	~4.4 K	~3.4 M	~53.6 M
Genomic Medicine			
Drugs with Pharmacogenomics Information on Label	4	46	106

Microarrays





Gene Expression Profiling

Comparative Genomic Hybridization

Chromatin Immunoprecipitation

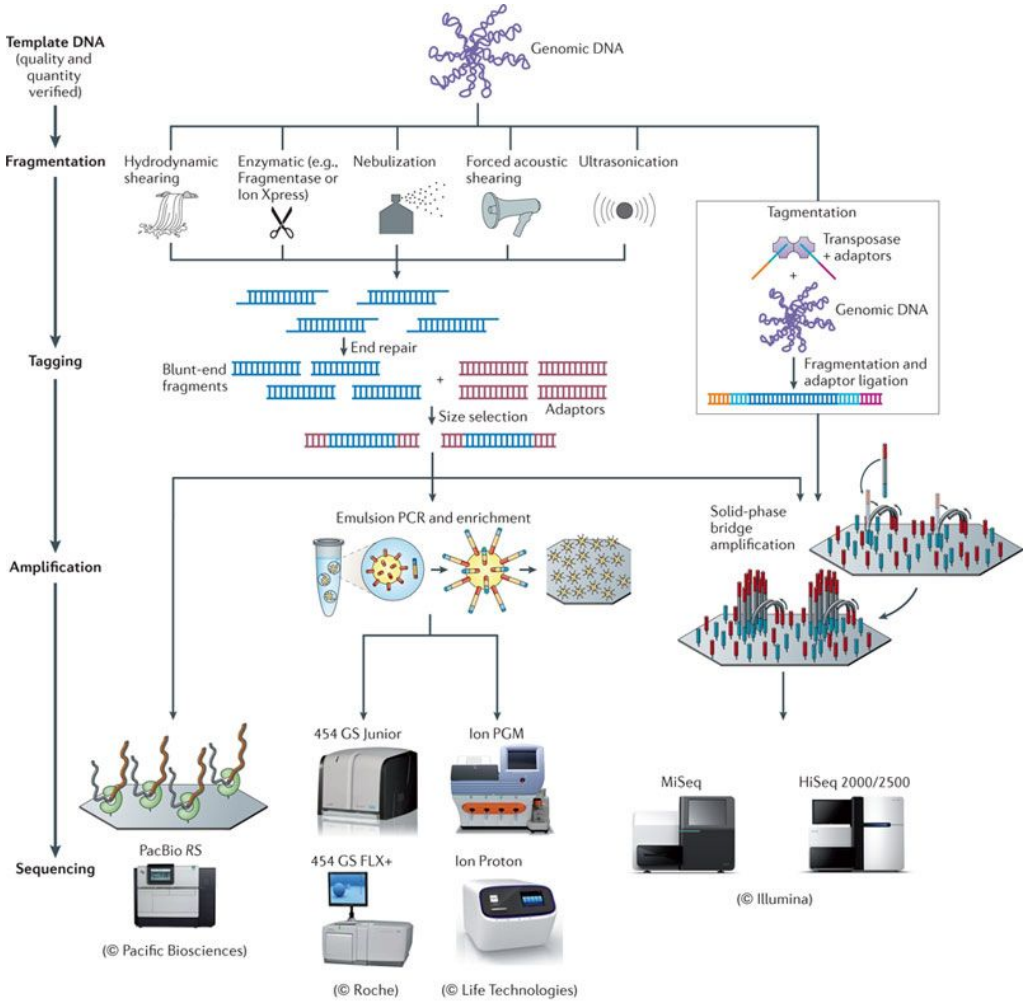
SNP detection

Next Generation Sequencing

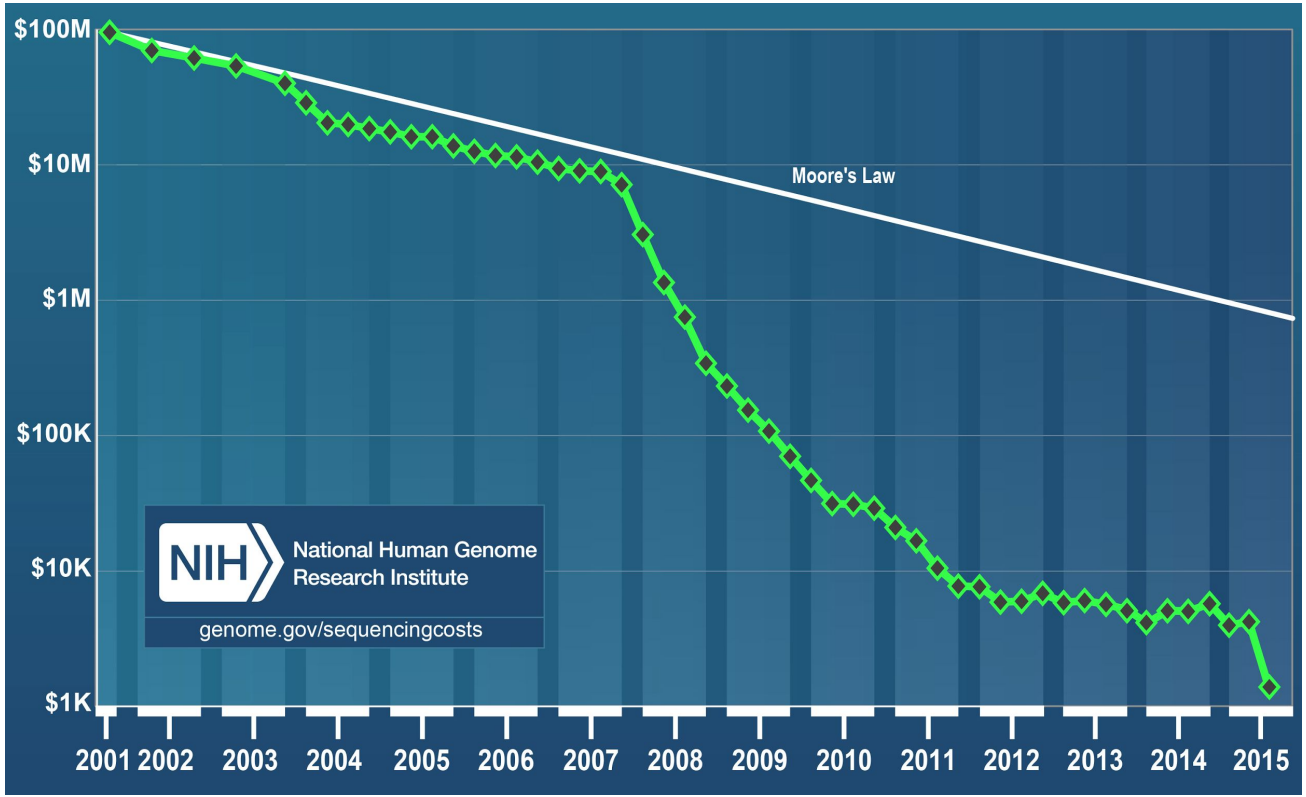
Whole Genome Sequencing

Whole Exome Sequencing

Whole Transcriptome Sequencing

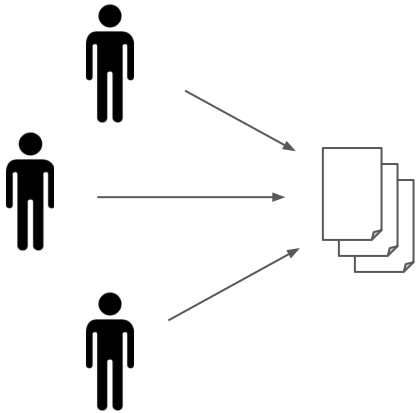


Next Generation Sequencing

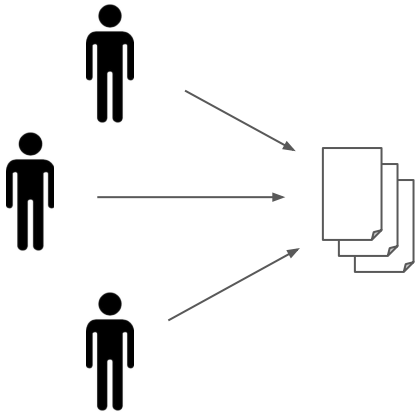


Integration of data

Integration of data



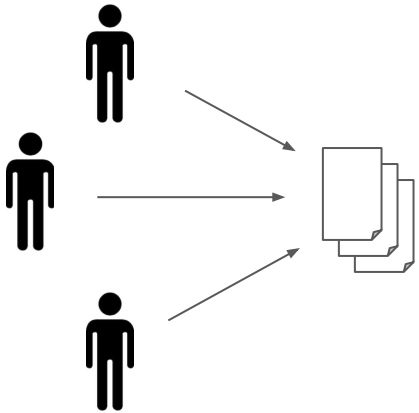
Integration of data



Gene Expression Omnibus

ARRAYEXPRESS

Integration of data

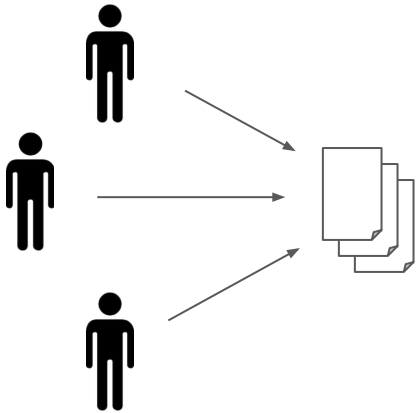


COSMIC
Catalogue of somatic mutations in cancer

GEO
Gene Expression Omnibus

ARRAYEXPRESS

Integration of data



COSMIC
Catalogue of somatic mutations in cancer

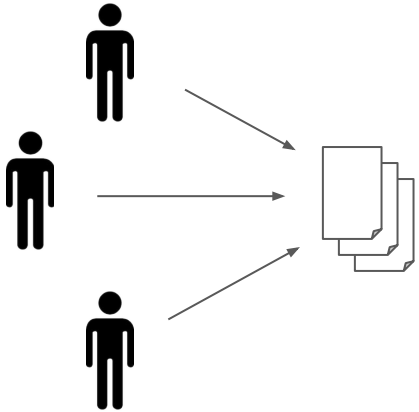
GENEONTOLOGY
Unifying Biology

SO The Sequence Ontology Project

GEO
Gene Expression Omnibus

ARRAYEXPRESS

Integration of data



COSMIC
Catalogue of somatic mutations in cancer

GENEONTOLOGY
Unifying Biology

SO The Sequence Ontology Project

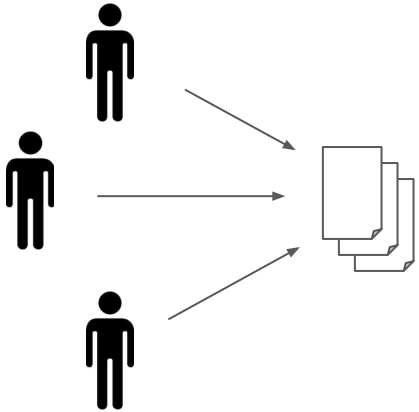
GEO
Gene Expression Omnibus

UCSC Genome Bioinformatics

ARRAYEXPRESS

e! Ensembl

Integration of data



COSMIC
Catalogue of somatic mutations in cancer

GENEONTOLOGY
Unifying Biology

SO The Sequence Ontology Project

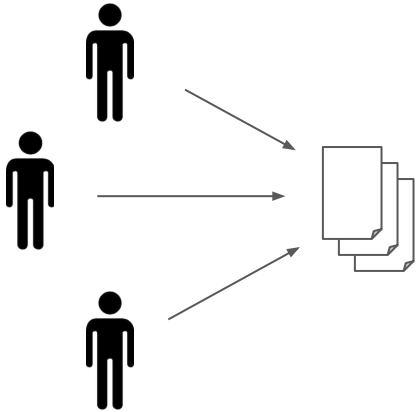
GEO
Gene Expression Omnibus

UCSC Genome Bioinformatics

ARRAYEXPRESS

Ensembl

Integration of data



COSMIC
Catalogue of somatic mutations in cancer

GENEONTOLOGY
Unifying Biology

SO The Sequence Ontology Project

GEO
Gene Expression Omnibus

UCSC Genome Bioinformatics

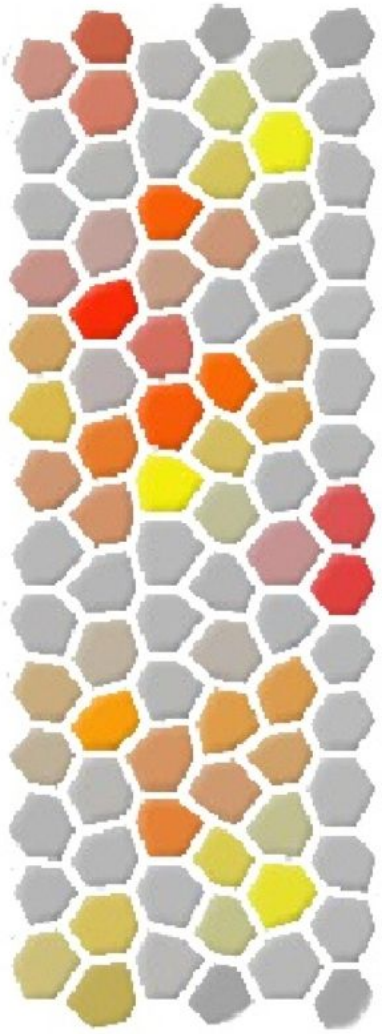
ARRAYEXPRESS

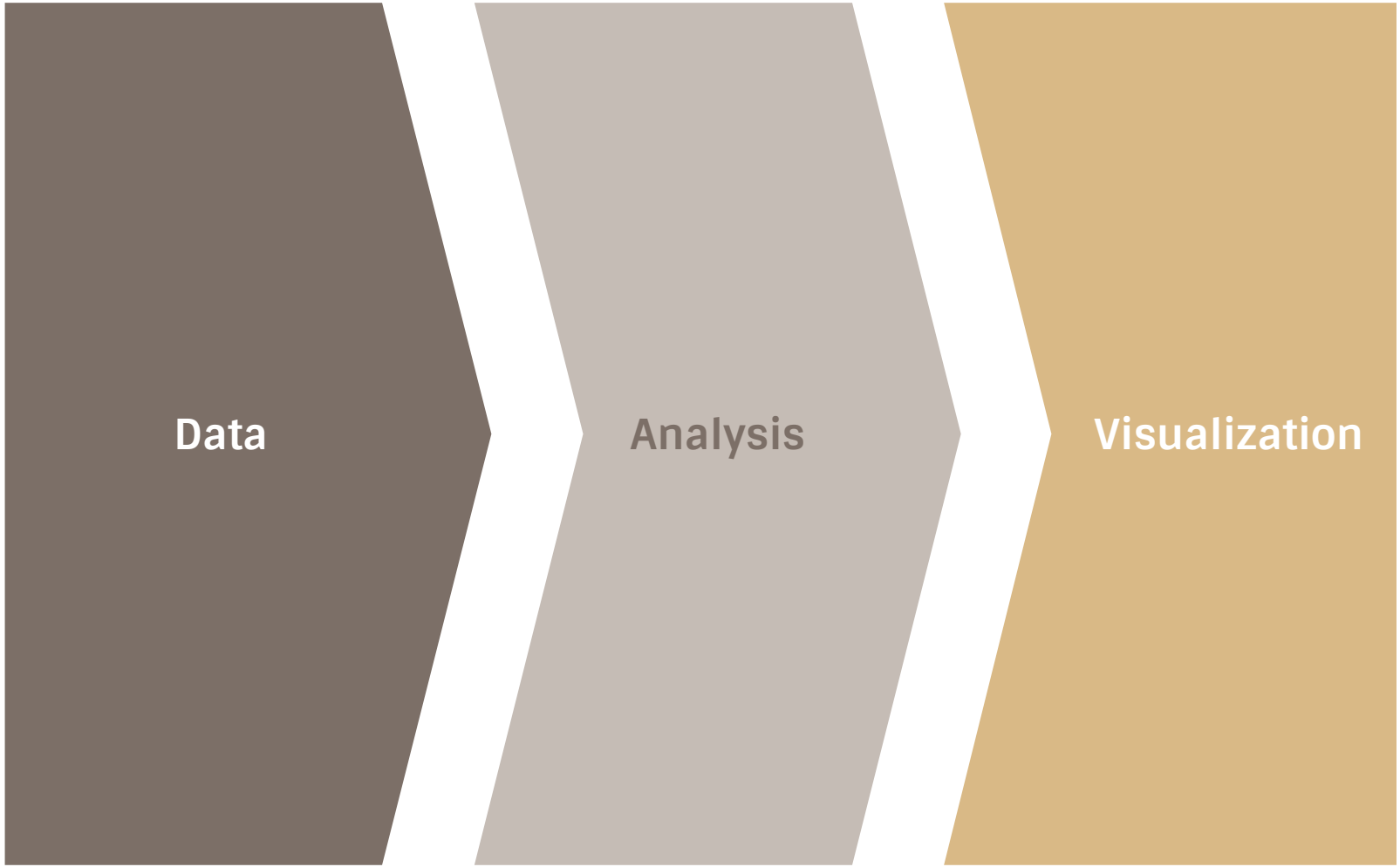
e! Ensembl

TCGA

International Cancer Genome Consortium

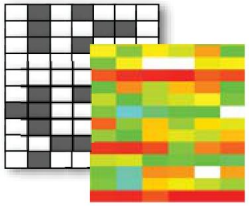

gitools




Data
Analysis
Visualization

Data | Analysis | Visualization

Matrices



Modules

Tables

Binary Data Matrix (BDM)

	S1	S2	S3	S4
ENSG00000000003	0	1	0	1
ENSG00000000005	-	1	0	0
ENSG00000000419	1	0	1	0
ENSG00000000457	0	1	0	0

Continuous Data Matrix (CDM)

	S1	S2	S3	S4
ENSG00000000003	0,01	1,00	0,40	0,94
ENSG00000000005	-	1,00	1,00	1,00
ENSG00000000419	1,00	0,95	0,98	1,00
ENSG00000000457	1,00	1,00	1,00	1,00

Tab Separated Values (TSV)

id	symbol	chromosome	band
ENSG00000000003	TSPAN6	X	q22.1
ENSG00000000005	TNMD	X	q22.1
ENSG00000000419	DPM1	20	q13.13
ENSG00000000457	SCYL3	1	q24.2

Two Columns Mapping (TCM)

ENSG00000115204	GO:0000002
ENSG00000025708	GO:0000002
ENSG00000151729	GO:0000002
ENSG00000137074	GO:0000012
ENSG00000174405	GO:0000012
ENSG00000042088	GO:0000012
ENSG00000169621	GO:0000012
ENSG00000073050	GO:0000012
ENSG00000118245	GO:0000012

Gene Matrix (GMX)

bound_0h	bound_27h	bound_96h
genes bound at 0h	genes bound at 27h	genes bound at 96h
ENSG00000000005	ENSG00000000971	ENSG00000001036
ENSG00000000419	ENSG00000001167	ENSG00000003096
ENSG00000000457		ENSG00000004455

Indexed Mapping (IXM)

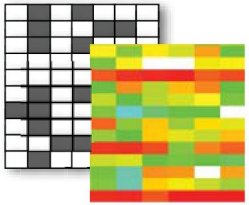
This format is designed to use less space but it can not be easily edited.
There is a command line conversor that can generate it from other formats.

Gene Matrix Transposed (GMT)

bound_0h	bound at 0h	ENSG00000000005	ENSG00000000419	ENSG00000000457
bound_27h	bound at 27h	ENSG00000000971	ENSG00000001167	
bound_96h	bound at 96h	ENSG00000001036	ENSG00000003096	ENSG00000004455

Data | Analysis | Visualization

Matrices



Modules

Tables

biomart

BioMart Central Portal	CyanoBase	HTGT
Ensembl	Wormbase	Cildb
Ensembl Bacteria	DroSpeGe	Pancreatic Expression Database
Ensembl Metazoa	ArrayExpress DW	Reactome
Ensembl Protists	Eurexpress	EU Rat Mart
Ensembl Plants	HapMap	Paramecium DB
Ensembl Fungi	Dictybase	International Potato Center (CIP)
Phytozome	COSMIC	Mouse Genome Informatics (MGI)
Gramene	Rat Genome Database	
Europhenome	GermOnLine	
UniProt	PRIDE	
InterPro	PepSeeker	International Cancer Genome Consorcium
HGNC	VectorBase	

intOgen Integrative Onco Genomics

Genes significantly altered
Modules of genes significantly altered

levels → Experiments, Tumour types

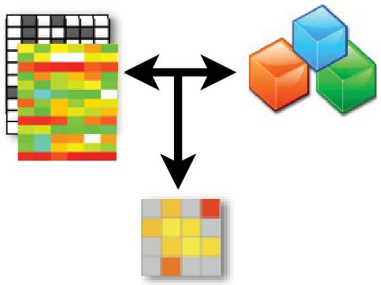
alterations → Up-regulation, Down-regulation, Gain, Loss

KEGG Kyoto Encyclopedia of Genes and Genomes

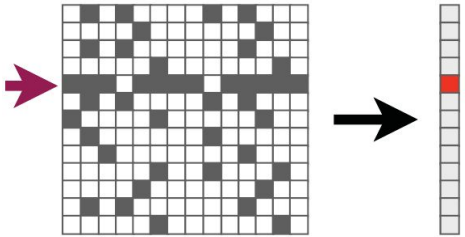
Pathways

Data | Analysis | Visualization

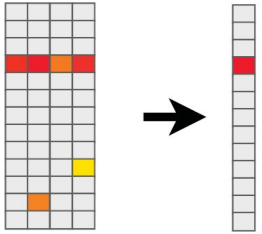
Enrichment Analysis



Oncodrive

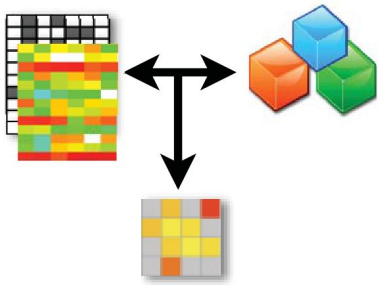


Combination

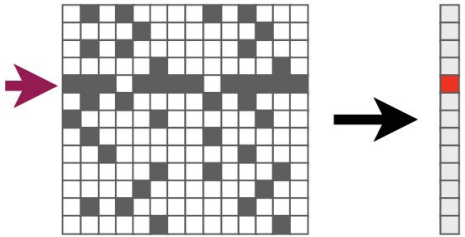


Data | Analysis | Visualization

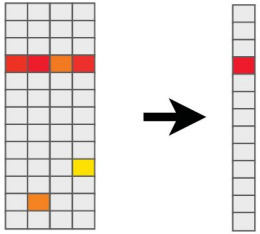
Enrichment Analysis



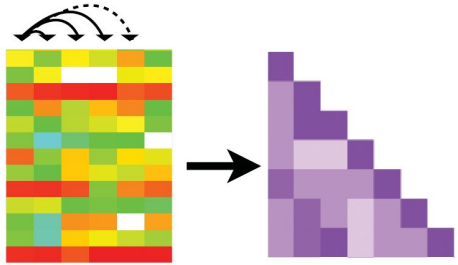
Oncodrive



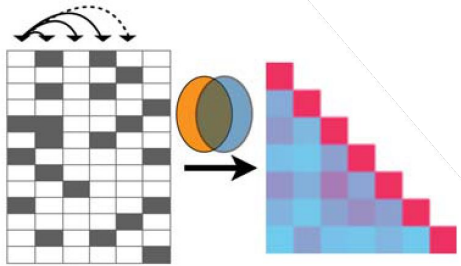
Combination



Correlations



Overlapping



→ Scale

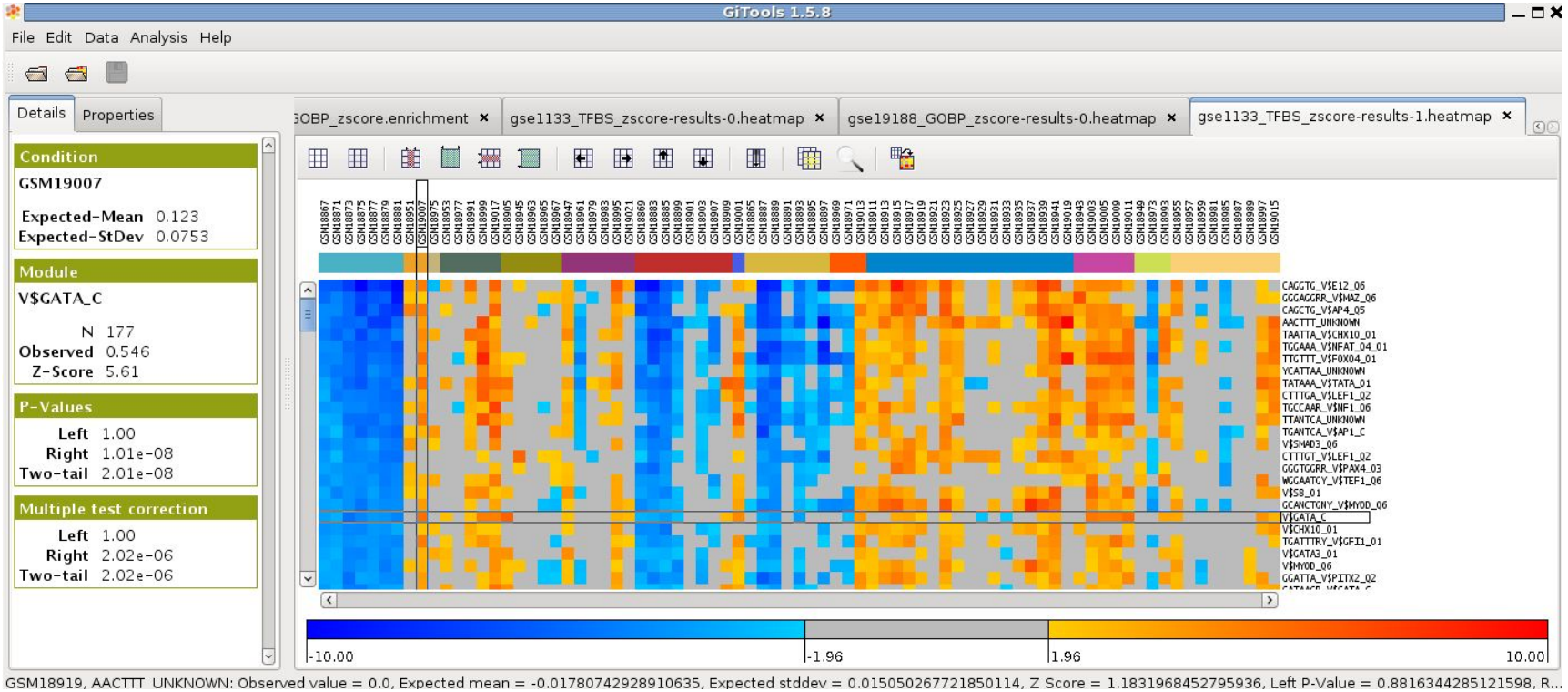
→ Grid

→ Sizes

→ Font

→ Labels

Data → Analysis → Visualization



Explore the data by searching, filtering, sorting, and reordering columns and rows

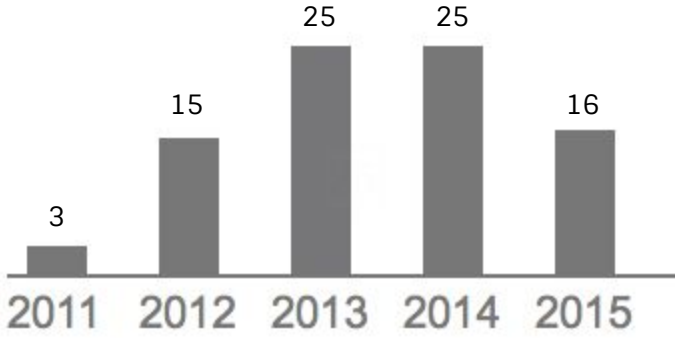
Gitools: Analysis and Visualisation of Genomic Data Using Interactive Heat-Maps.
Perez-Llamas, C. & Lopez-Bigas, N. PLoS ONE 6, e19541 (2011)

DIANA miRPath v. 2.0: investigating the combinatorial effect of microRNAs in pathways.

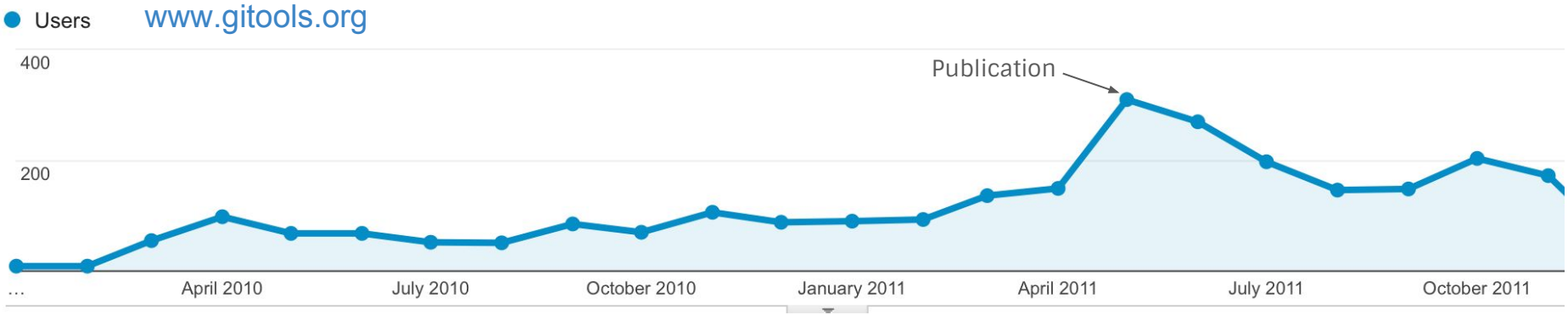
A polycomb group protein is retained at specific sites on chromatin in mitosis.

Transcriptomic characterization of temperature stress responses in larval zebrafish.

温度刺激对斑马鱼仔鱼基因转录表达的影响.



84 total citations



Gitools: Analysis and Visualisation of Genomic Data Using Interactive Heat-Maps.

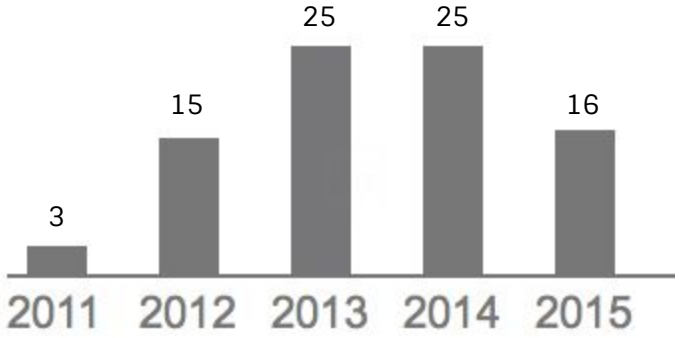
Perez-Llamas, C. & Lopez-Bigas, N. PLoS ONE 6, e19541 (2011)

DIANA miRPath v. 2.0: investigating the combinatorial effect of microRNAs in pathways.

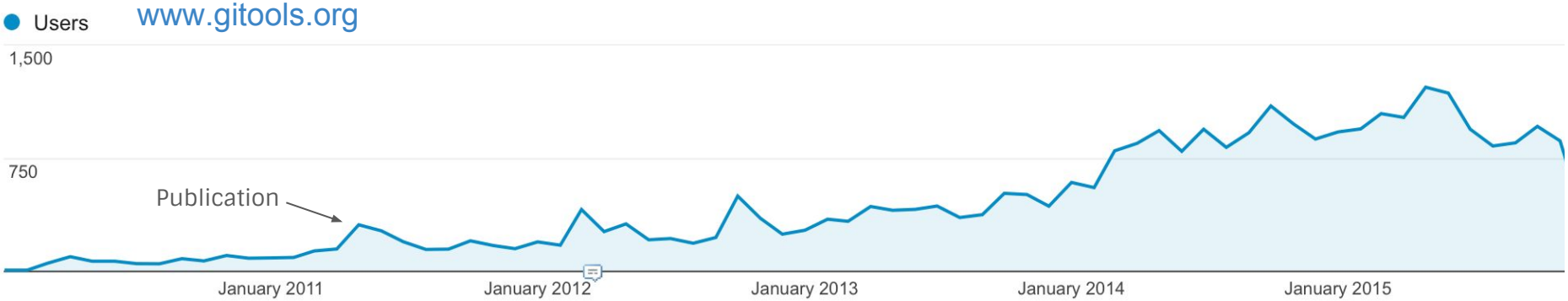
A polycomb group protein is retained at specific sites on chromatin in mitosis.

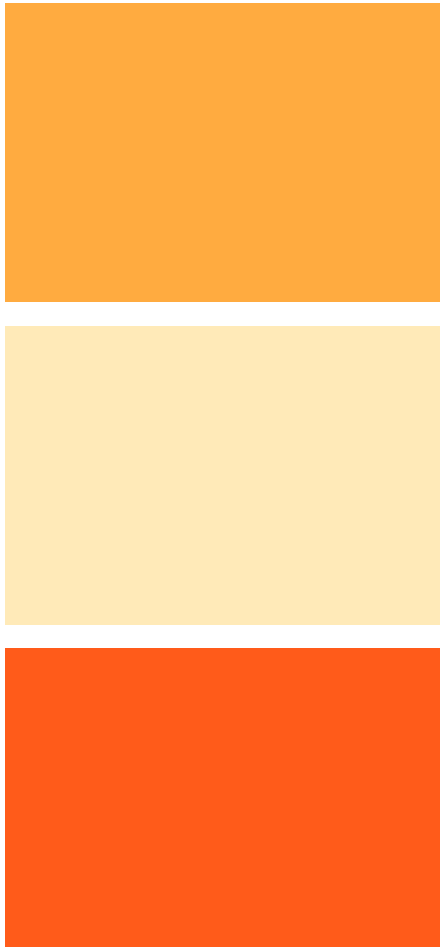
Transcriptomic characterization of temperature stress responses in larval zebrafish.

温度刺激对斑马鱼仔鱼基因转录表达的影响.



84 total citations






intOgen
Integrative Onco Genomics


intOgen
Integrative Onco Genomics


Data
Analysis
Visualization

Cancer Data

Tumour Alterations

Multiple Views

Nuria López-Bigas

Abul Islam

Gunes Gundem

Alba Jene-Sanz

Jordi Deu-Pons

Abel Gonzalez-Perez

Carlota Rubio-Perez

David Tamborero

Simon J Furney

Michael P Schroeder

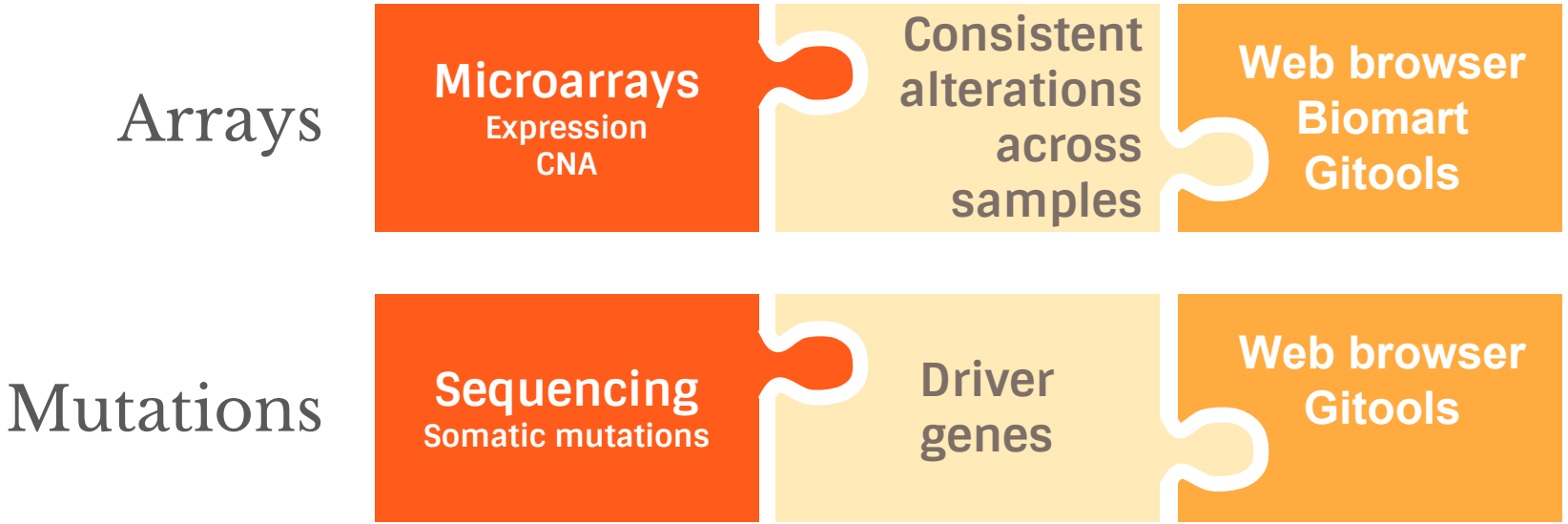
Alberto Santos

Anna Kedzierska

Christian Perez-Llamas

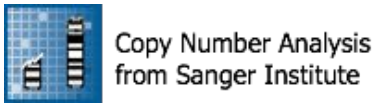
intOgen

Integrative Onco Genomics



Arrays intOgen
Integrative Onco Genomics

Data



Collect → Organize → Annotate

Arrays intOgen
Integrative Onco Genomics

Data

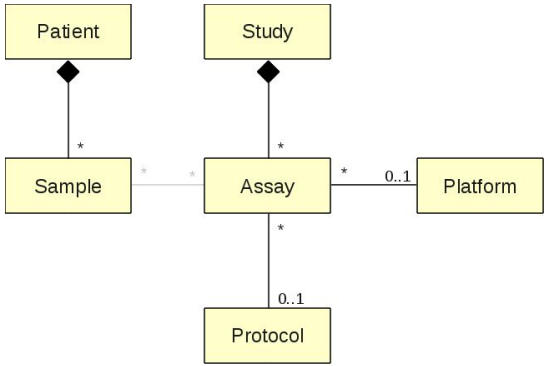
Gene Expression Omnibus

ARRAYEXPRESS

progenetix

Copy Number Analysis from Sanger Institute

COSMIC
Catalogue of somatic mutations in cancer



Collect | Organize | Annotate

Arrays intOgen
Integrative Onco Genomics

Data

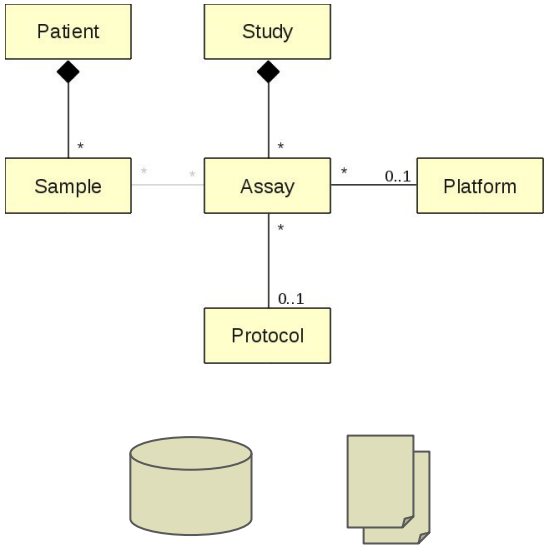
Gene Expression Omnibus

ARRAYEXPRESS

progenetix

Copy Number Analysis from Sanger Institute

COSMIC
Catalogue of somatic mutations in cancer



GENEONTOLOGY
Unifying Biology

KEGG
Kyoto Encyclopedia of Genes and Genomes

Ensembl

Collect | Organize | Annotate

Arrays intOgen
Integrative Onco Genomics

Data

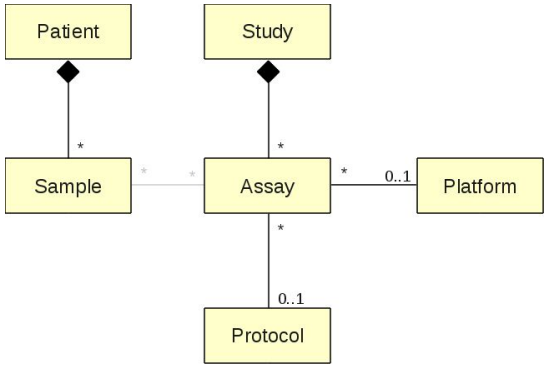
Gene Expression Omnibus

ARRAYEXPRESS

progenetix

Copy Number Analysis from Sanger Institute

COSMIC
Catalogue of somatic mutations in cancer



GENEONTOLOGY
Unifying Biology

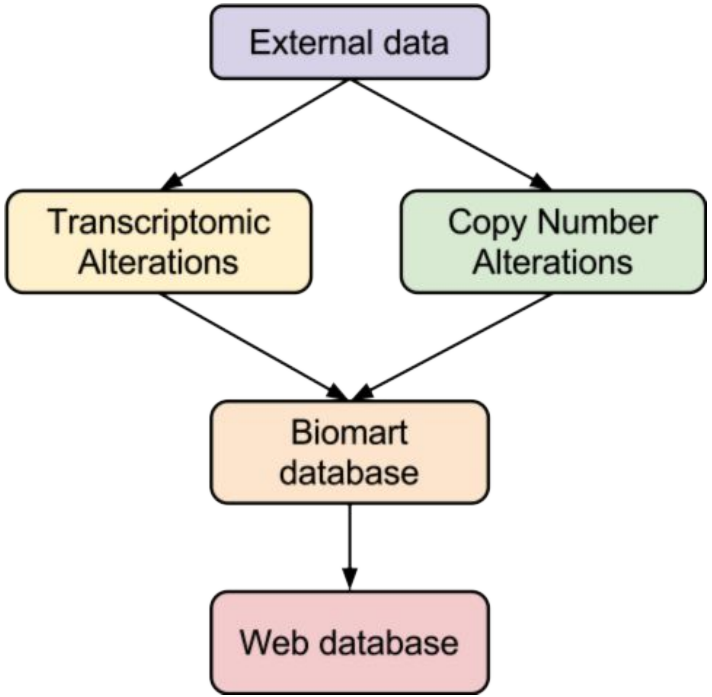
KEGG
Kyoto Encyclopedia of Genes and Genomes

e!Ensembl

Collect → Organize → Annotate

Arrays intOgen
Integrative Onco Genomics

Analysis



Arrays intOgen
Integrative Onco Genomics

Analysis

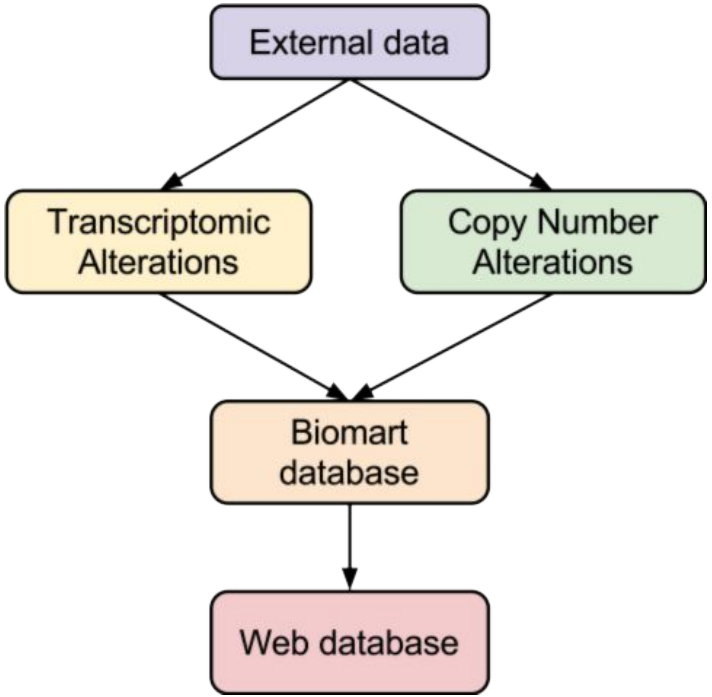
Makefiles

↓

Python Tasks

↓

Wok



Arrays intOgen
Integrative Onco Genomics

Analysis

Kepler

Ruffus

BPEL

Taverna

hadoop

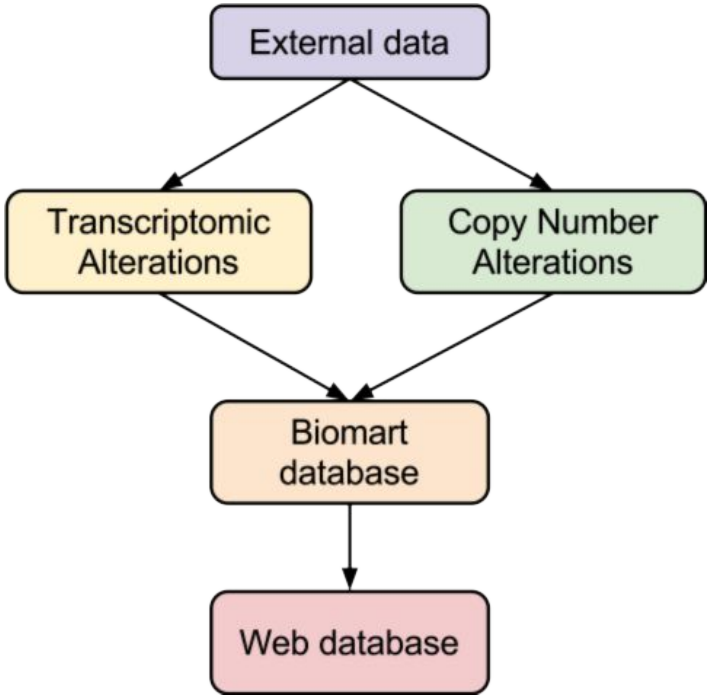
Makefiles

↓

Python Tasks

↓

Wok



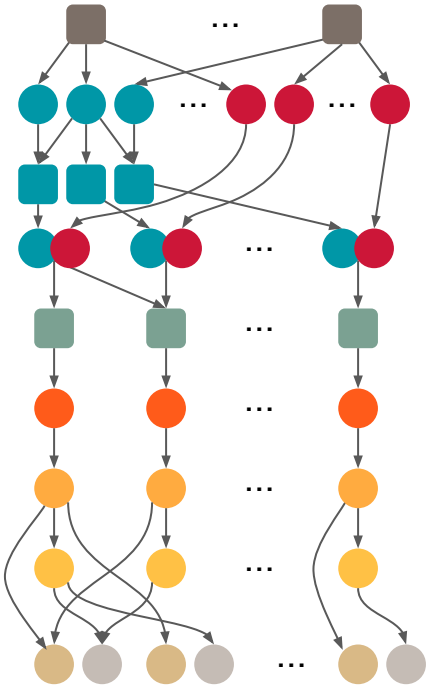
Arrays intOgen
Integrative Onco Genomics

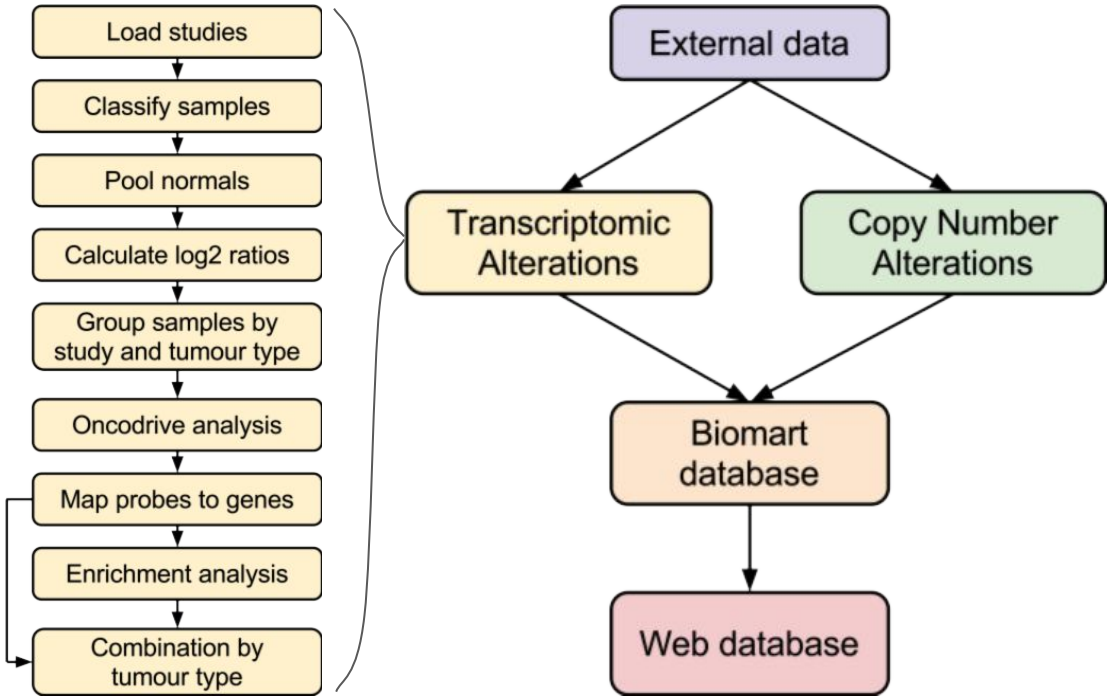
Analysis

Wok

gitools

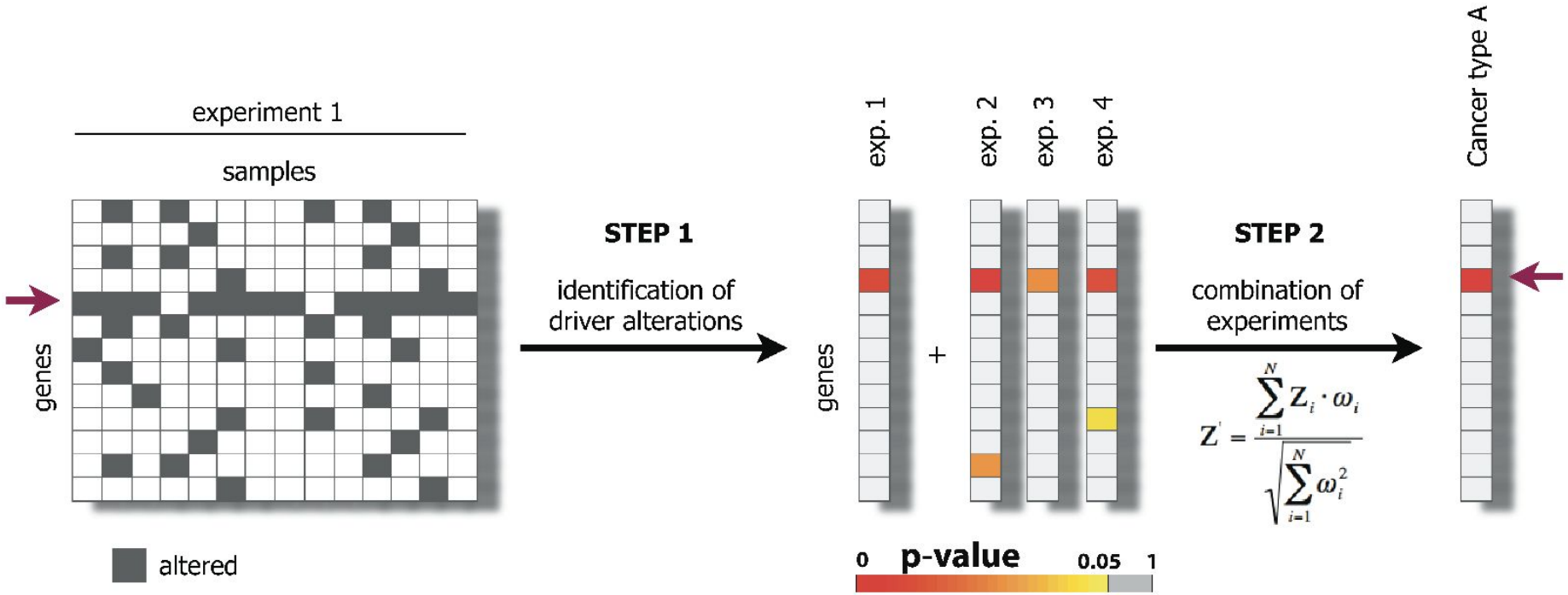
GRID ENGINE





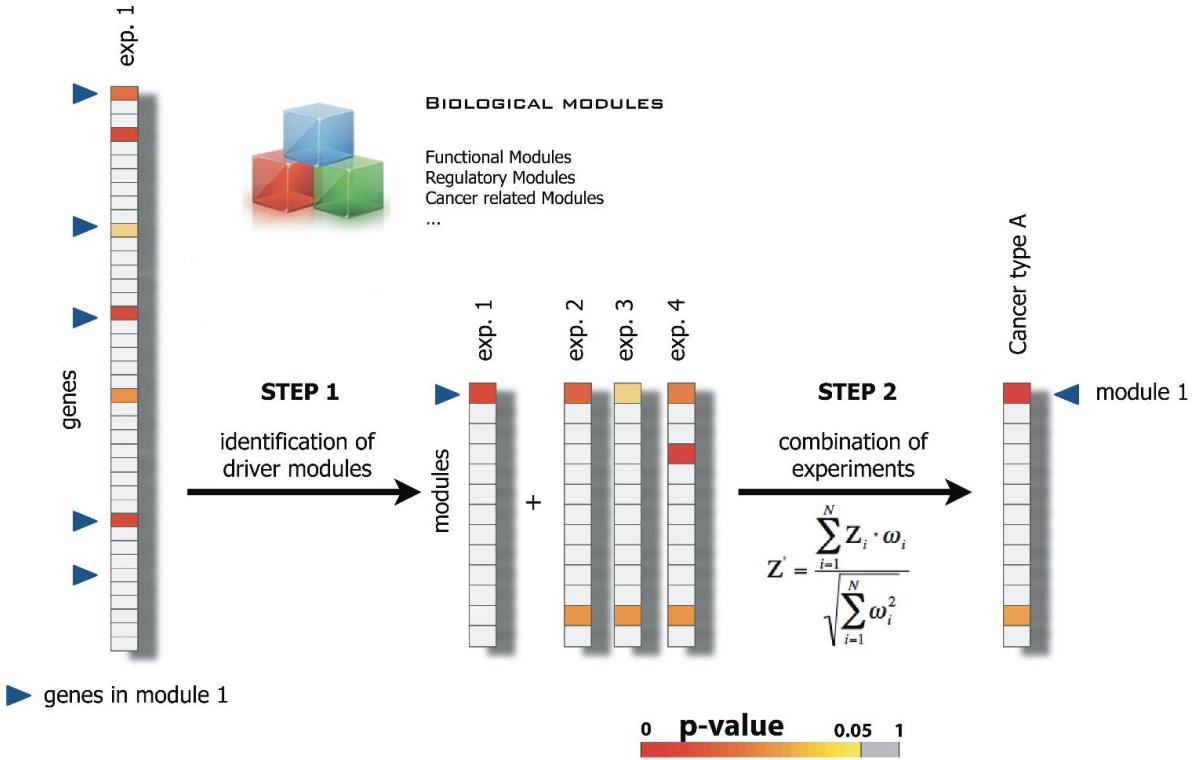
Arrays intOgen
Integrative Onco Genomics

Analysis



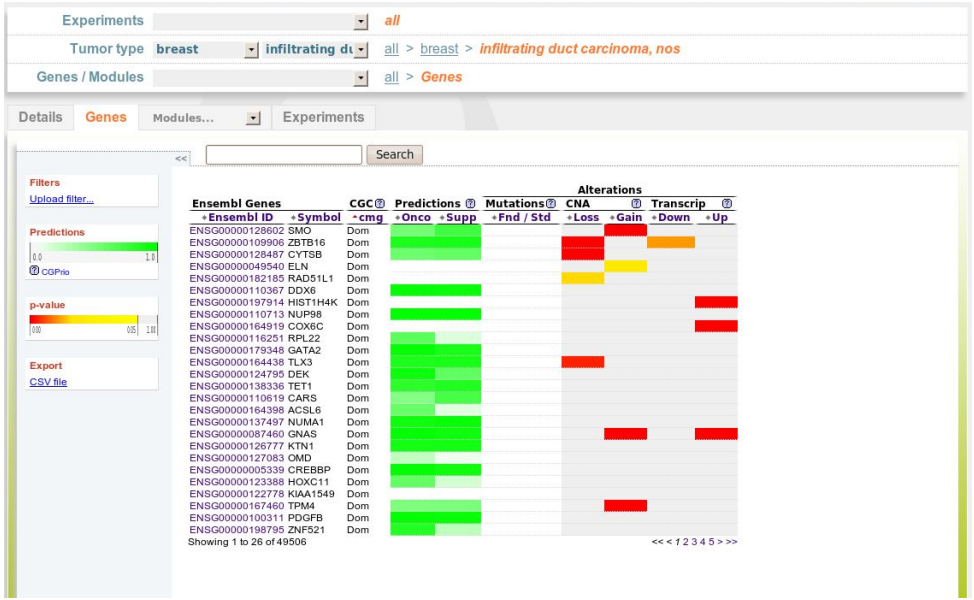
Arrays intOgen
Integrative Onco Genomics

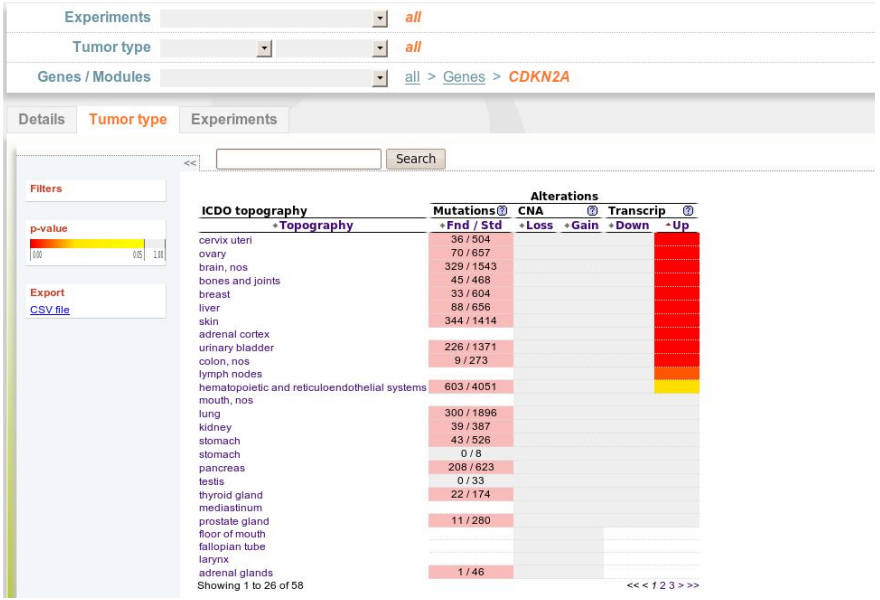
Analysis



Arrays intOgen
Integrative Onco Genomics

Visualization





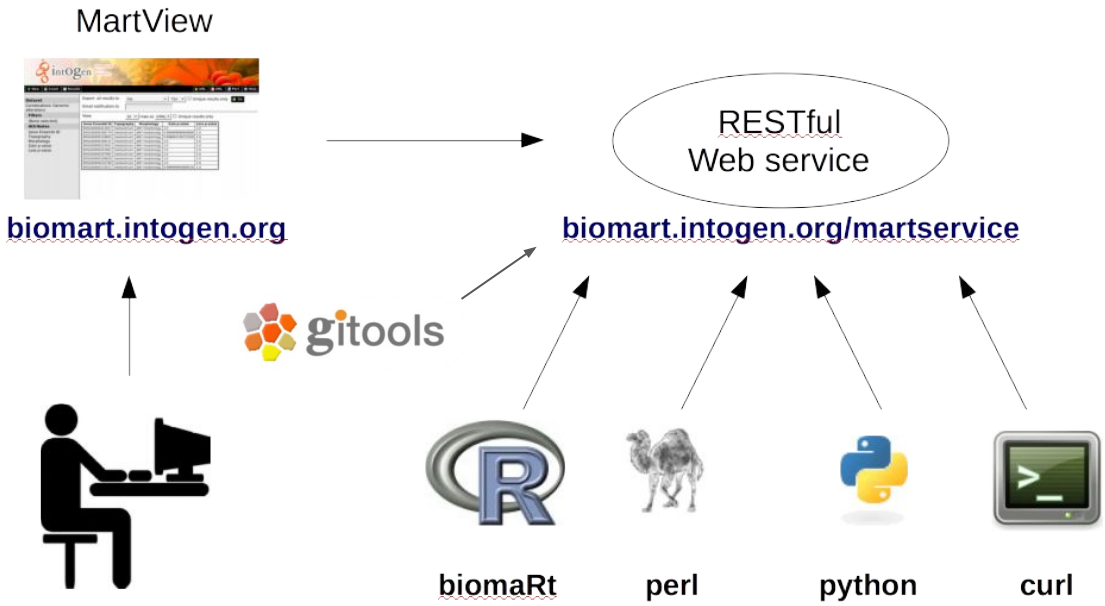
Web browser by Jordi Deu-Pons

Arrays intOGen
Integrative Onco Genomics

Visualization

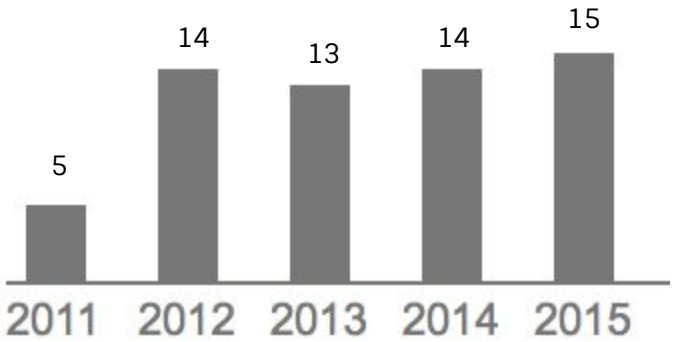
Perez-Llamas, C., Gundem, G., & Lopez-Bigas, N. (2011).
Integrative cancer genomics (IntOGen) in Biomart. Database
The Journal of Biological Databases and Curation, 2011, bar039. http://doi.org/10.1093/database/bar039

Guberman, J. M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., ... Kasprzyk, A. (2011).
BioMart Central Portal: An open database network for the biological community.
Database, 2011.



Arrays intOGen
Integrative Onco Genomics

Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., ... Lopez-Bigas, N. (2010).
IntOGen: integration and data mining of multidimensional oncogenomic data.
Nature Methods. http://doi.org/10.1038/nmeth0210-92



62 total citations

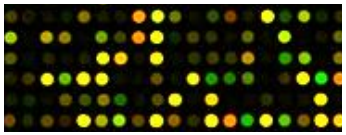
Mutations

intOgen

Integrative Onco Genomics

Sequencing
Somatic mutations

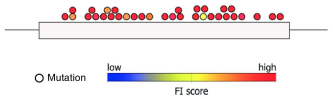
Driver genes



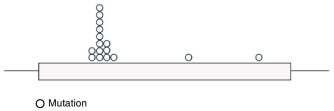
TCGA

International Cancer Genome Consortium

OncodriveFM



OncodriveCLUST



Mutations intOgen
Integrative Onco Genomics

Data

Independent projects

TCGA

International Cancer Genome Consortium

Collect

Organize

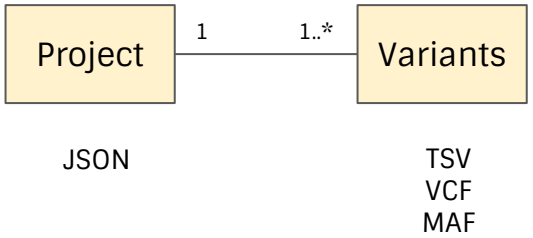
Mutations intOgen Integrative Onco Genomics

Data

Independent projects

TCGA

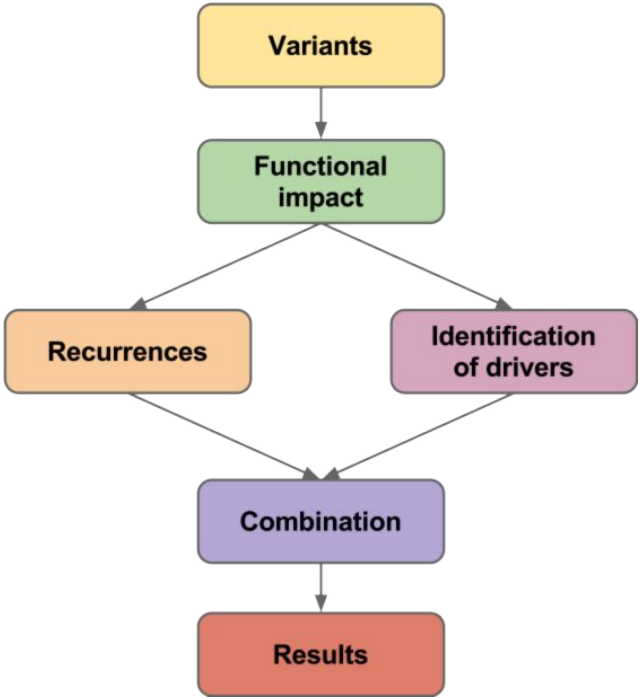
International Cancer Genome Consortium



Collect → Organize

Mutations intOgen
Integrative Onco Genomics

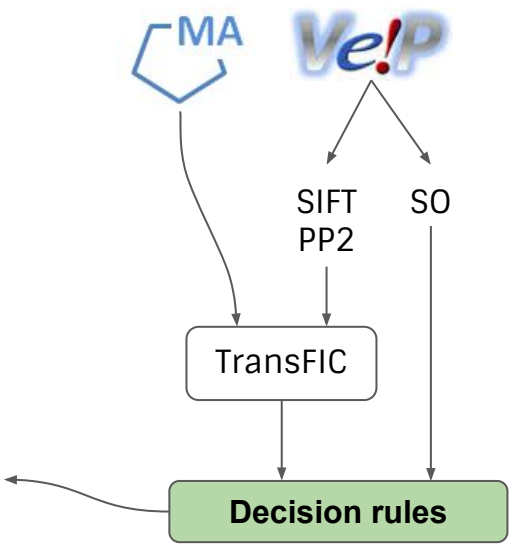
Analysis

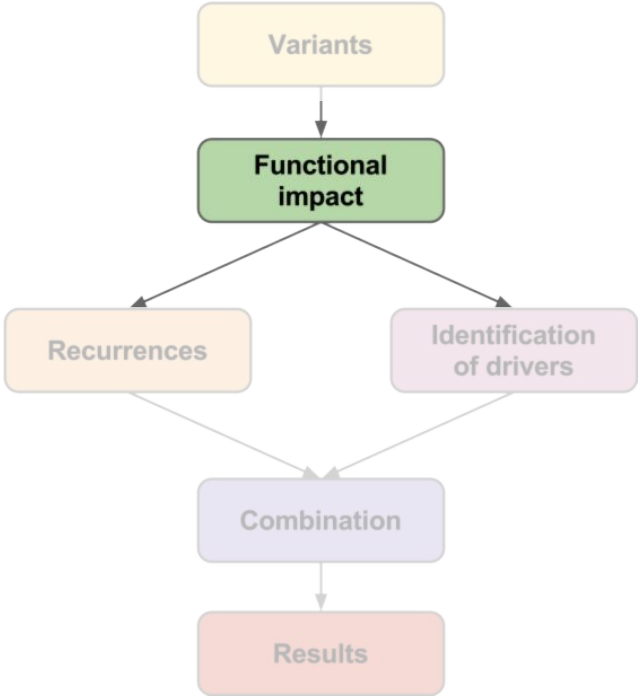


Mutations intOgen
Integrative Onco Genomics

Analysis

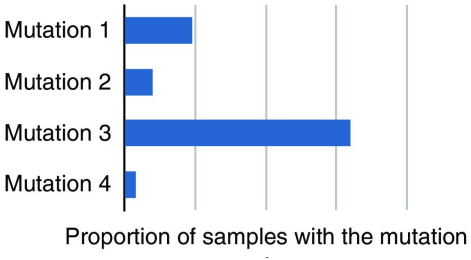
Mutation	Gene	Consequence	Functional impact
Mutation 1	Gene 1	Missense	Medium
Mutation 2	Gene 2	Synonymous	None
Mutation 3	Gene 2	Missense	Low
Mutation 4	Gene 3	STOP gain	High
Mutation 5	Gene 3	Missense	High
Mutation 6	Gene 3	Frameshift	High
Mutation 7	Gene 4	Synonymous	None
...			
Mutation *n*	Gene *n*	Missense	High

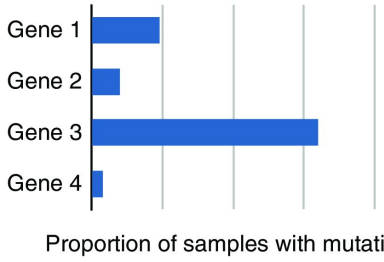


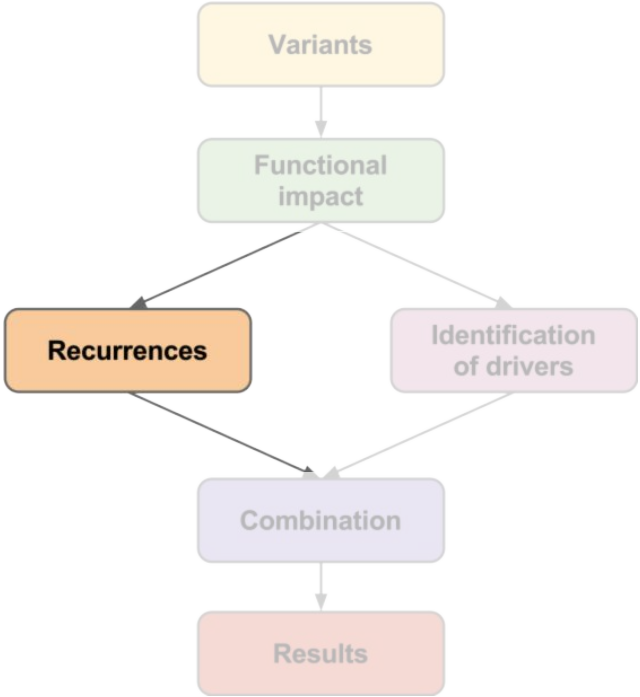


Mutations intOgen Integrative Onco Genomics

Analysis

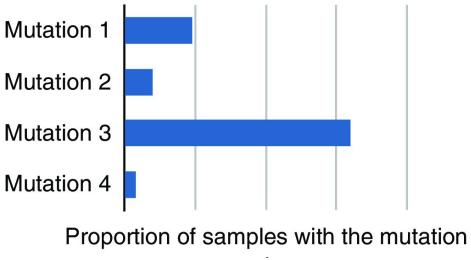


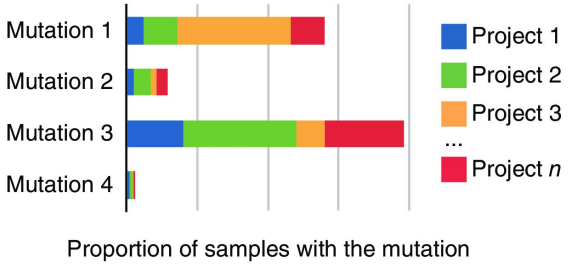


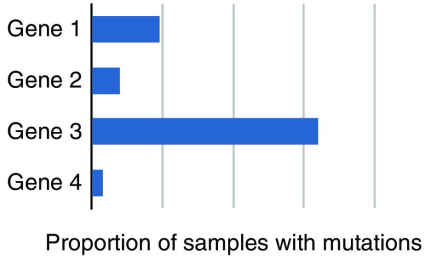


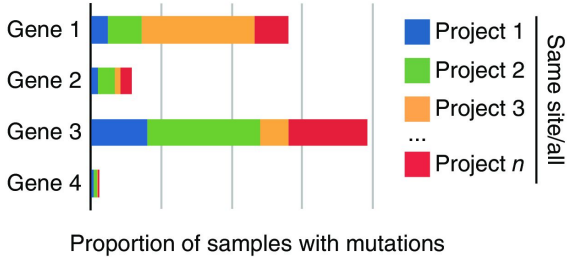
Mutations intOgen
Integrative Onco Genomics

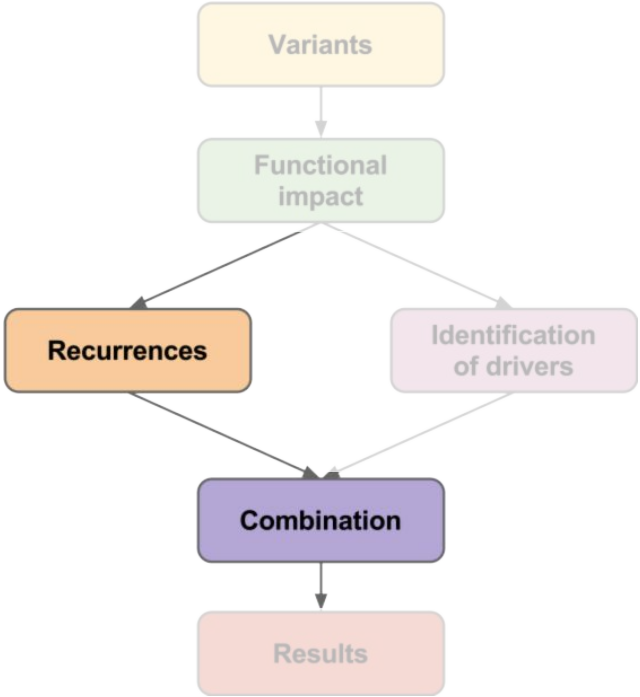
Analysis





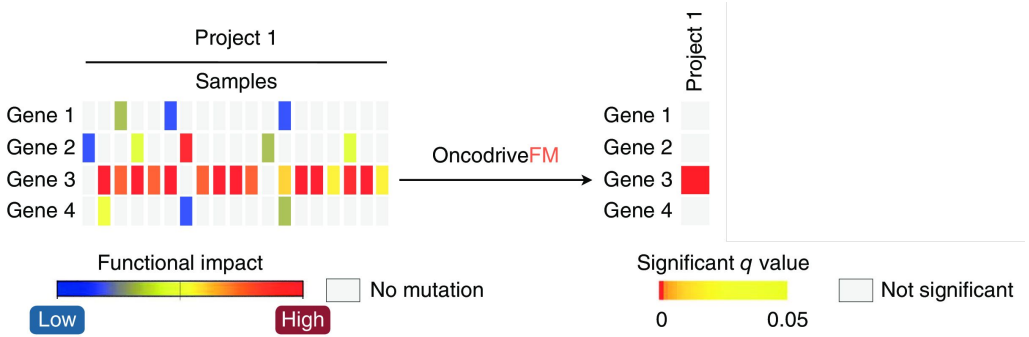


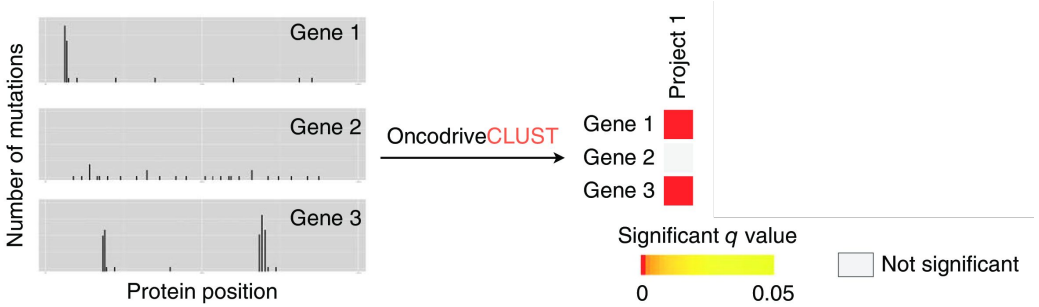


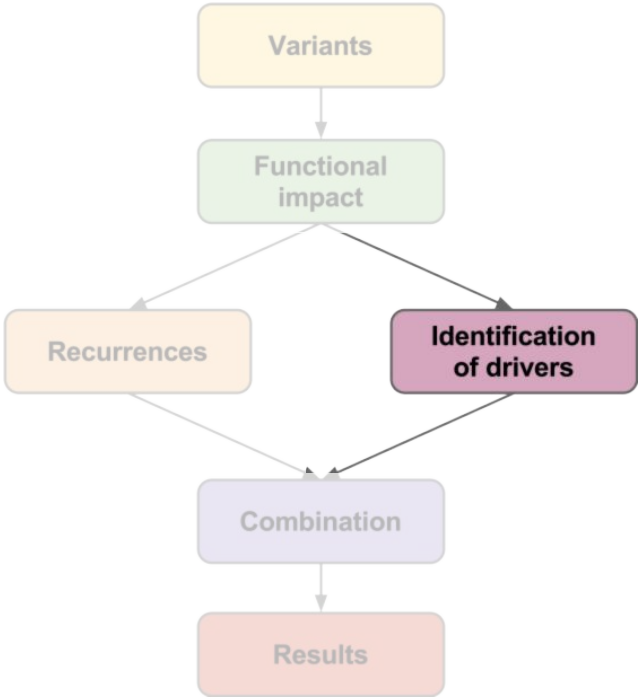


Mutations intOgen
Integrative Onco Genomics

Analysis

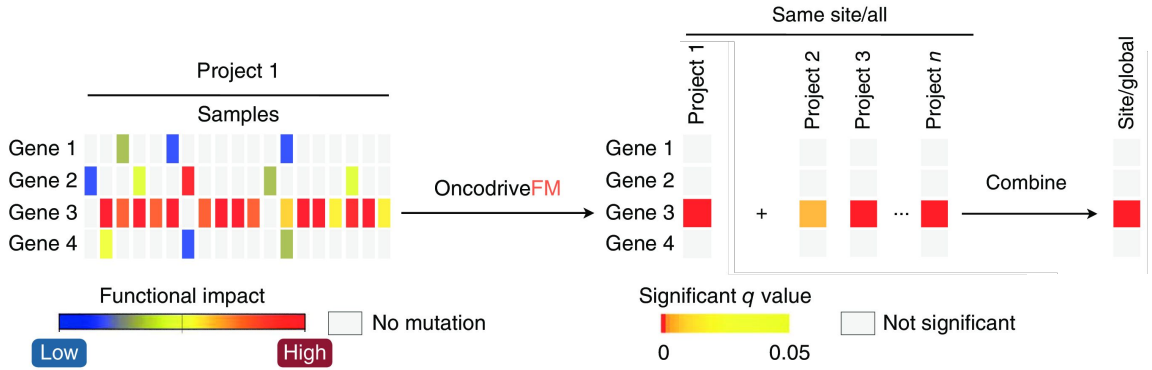


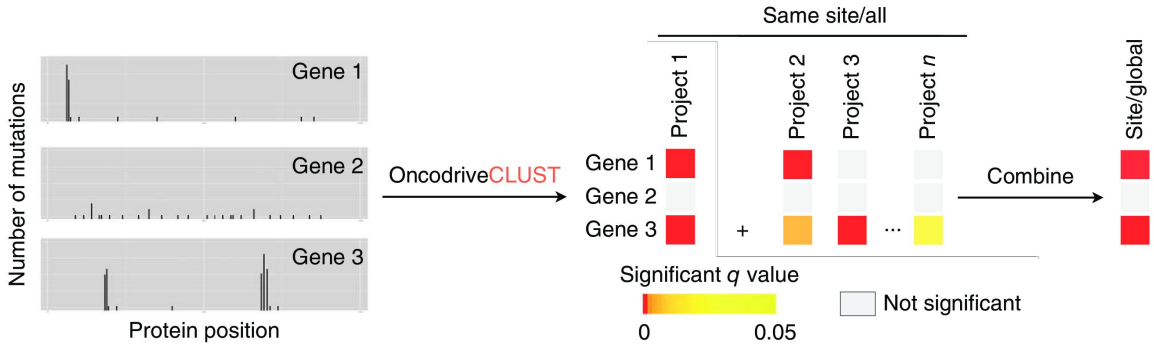


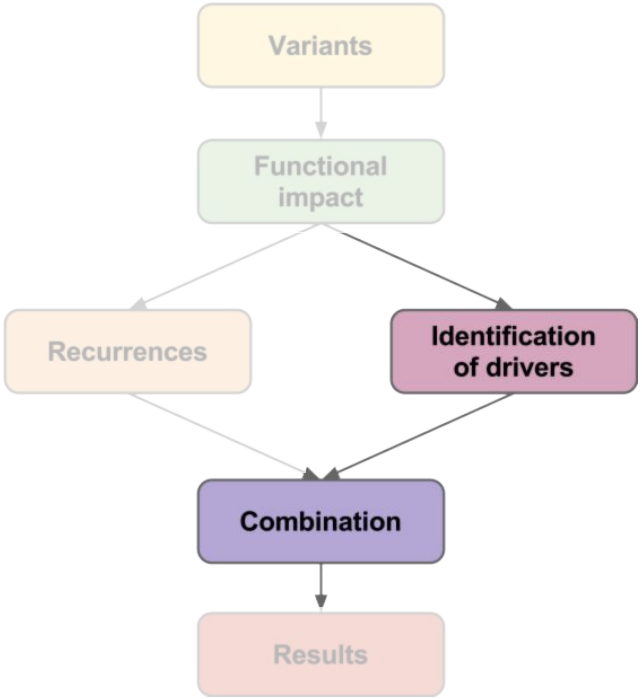


Mutations intOgen
Integrative Onco Genomics

Analysis



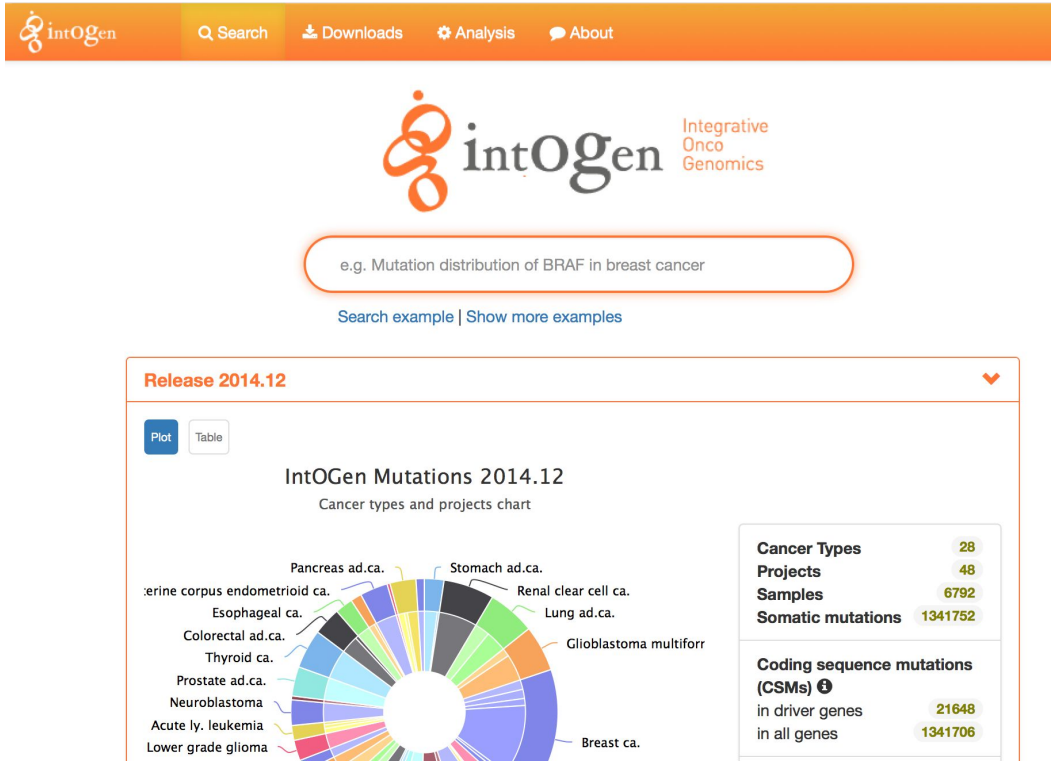


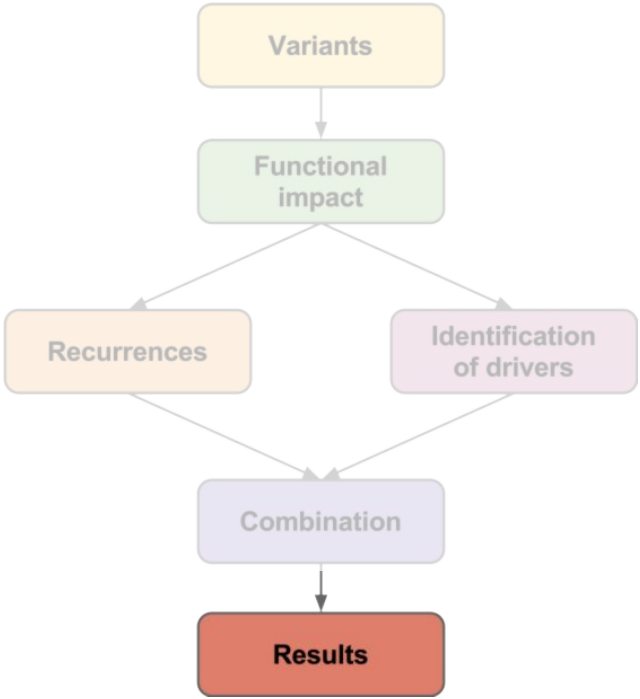


Mutations

intOGen
Integrative Onco Genomics

Visualization



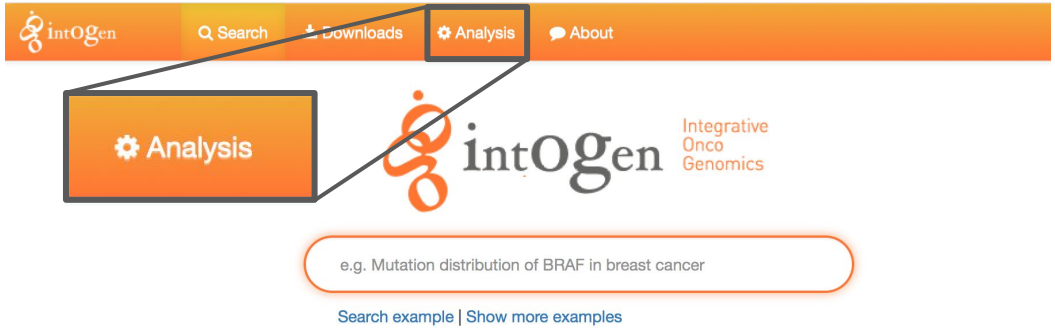


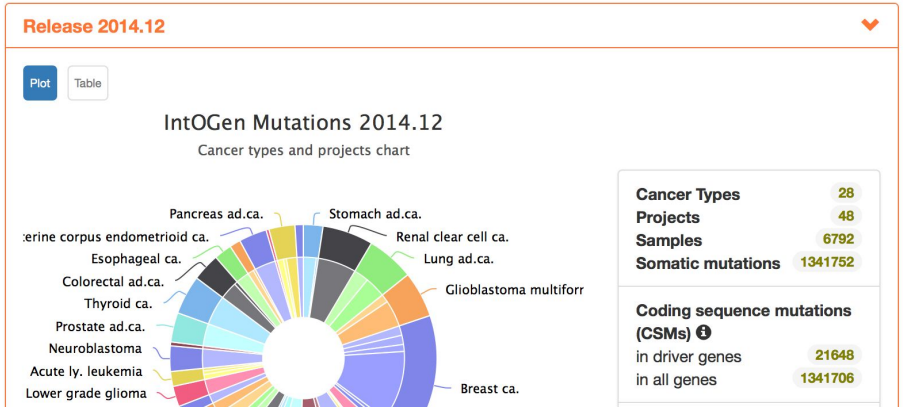
Mutations

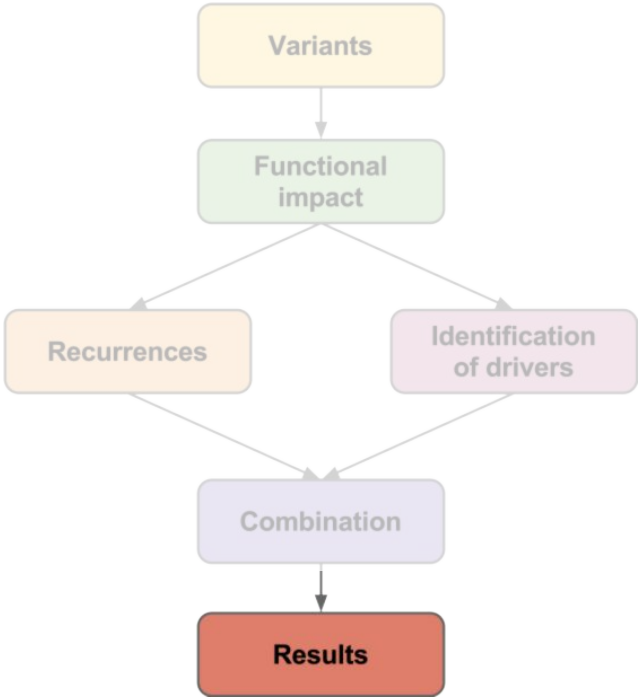
intOGen

Integrative Onco Genomics

Visualization







IntOGen Mutations Analysis 2.4.1-maintenance

Home Analysis Results Documentation Download

IntOGen Mutations Analysis

Download

To interpret catalogs of cancer somatic mutations.

Cohort analysis

Use this if you have a list of somatic mutations for a cohort of tumors and want to identify driver mutations, genes and pathways.

View an example | Analyse your data

Single tumor analysis

Use this if you have a list of somatic mutations for a single tumor and want to rank them based on their implication in cancer development.

View an example | Analyse your data

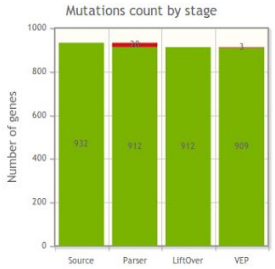
Mutations intOgen Integrative Onco Genomics

Visualization

Quality control

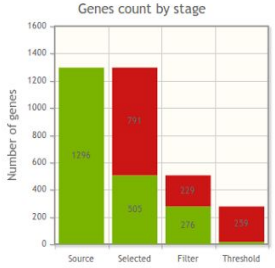
Mutations

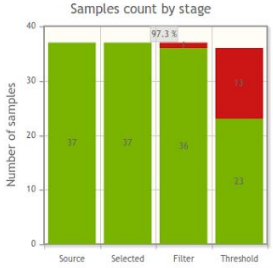
00-meduloblastoma_cohort_tier1.muts



	Passed	Discarded
Source	932	
Parser	912	20
LiftOver	912	0
VEP	909	3
Total		23

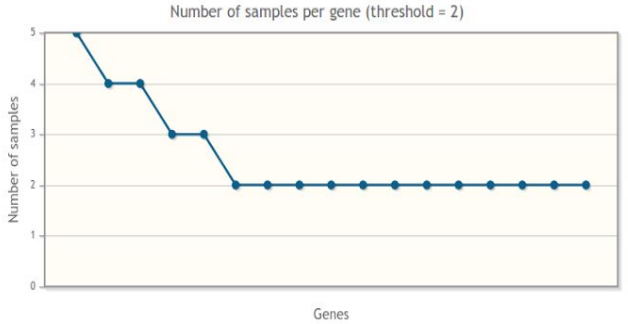
OncodriveFM

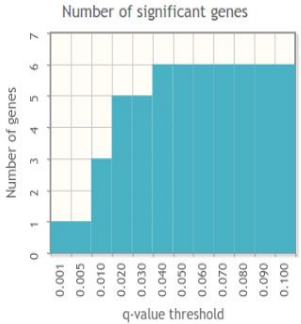




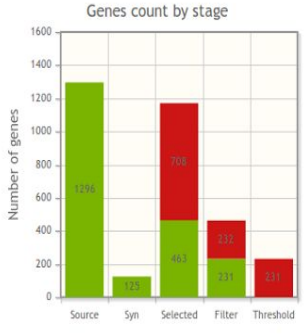
	Genes		Samples	
	Passed	Discarded	Passed	Discarded
Source	1296		37	
Selected	505	791	37	0
Filter	276	229	36	1
Threshold*	17	259	23	13
Total		1279		

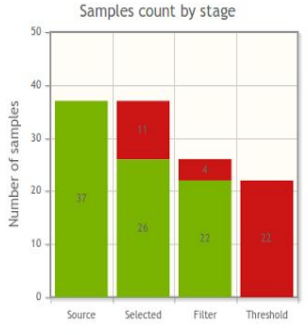
* 2 samples





OncodriveCLUST



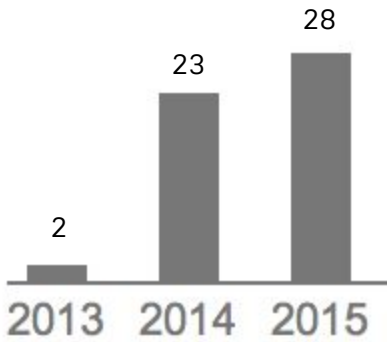


	Genes		Samples	
	Passed	Discarded	Passed	Discarded
Source	1296		37	
Selected	463	708	26	11
Filter	231	232	22	4
Threshold*	0	231	0	22
Total		1171		

* 5 samples

Mutations intOGen
Integrative Onco Genomics

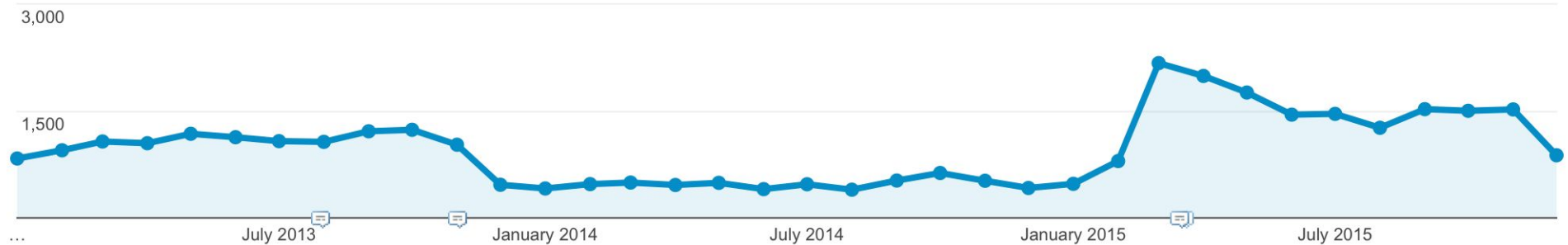
Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., ... Lopez-Bigas, N. (2013).
IntOGen-mutations identifies cancer drivers across tumor types.
Nature Methods, 10(11), 1081–2. http://doi.org/10.1038/nmeth.2642



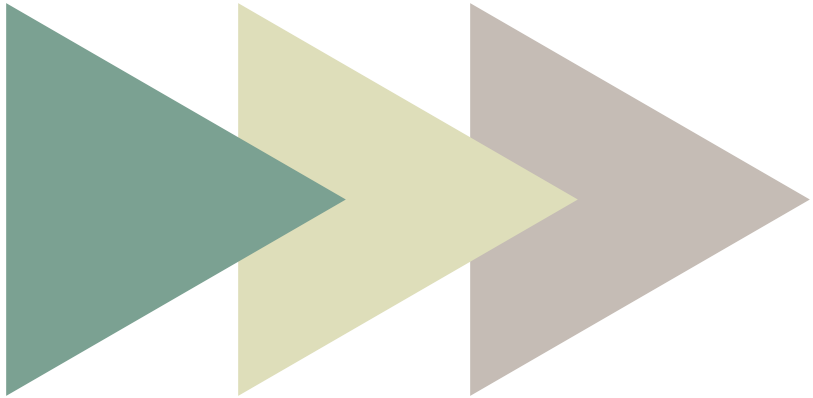
54 total citations

653 registered users (Oct 2015)

http://www.intogen.org



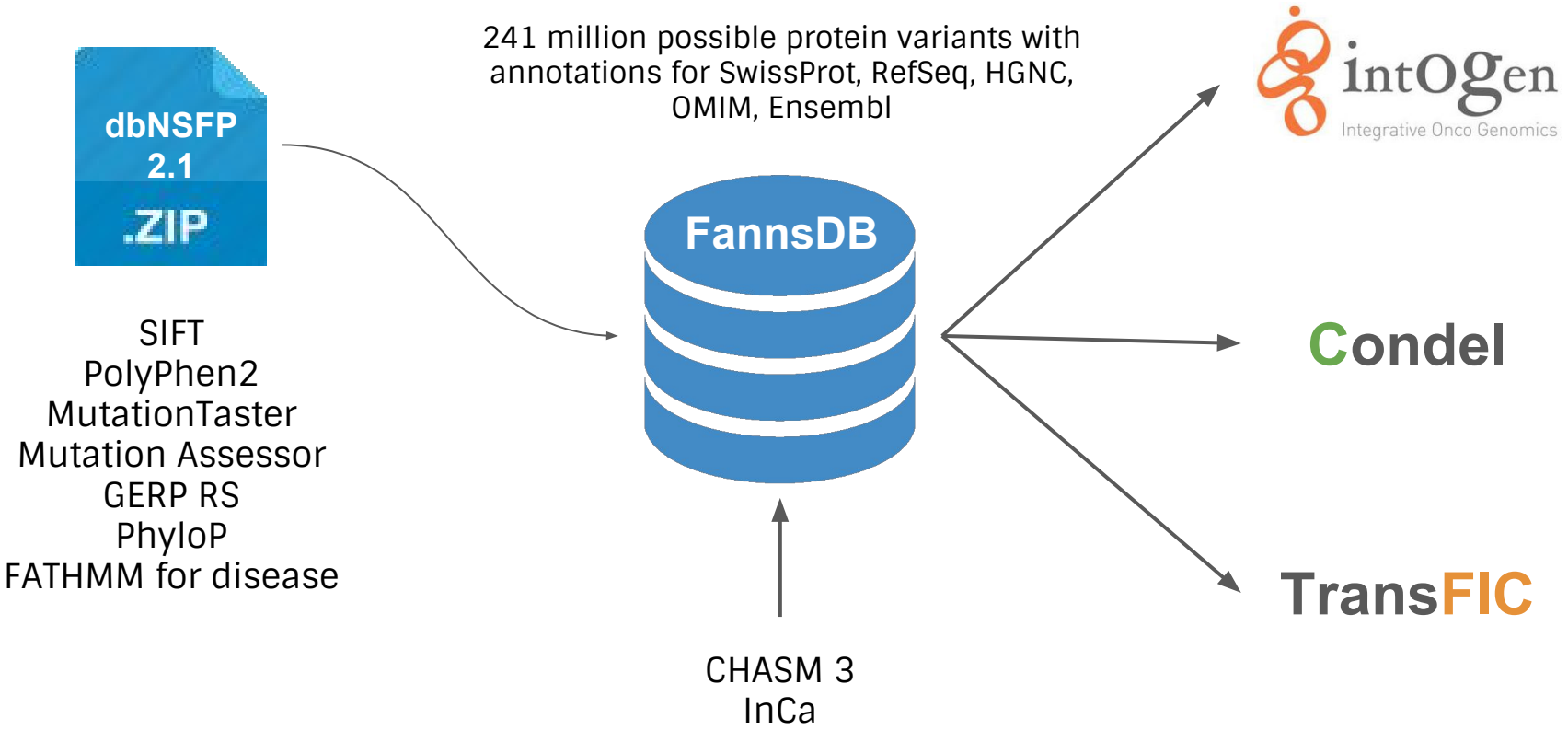
Benchmark of impact predictor tools



Prediction tools

General functional impact	SIFT, PolyPhen2, MutationTaster, FATHMM for diseases
Cancer functional impact	Mutation Assessor, FATHMM for cancer, CHASM, InCa
General conservation scores	GERP RS, PhyloP, C-Score
Combined / transformed scores	Condel, TransFIC

Precalculated scores



Proxy datasets

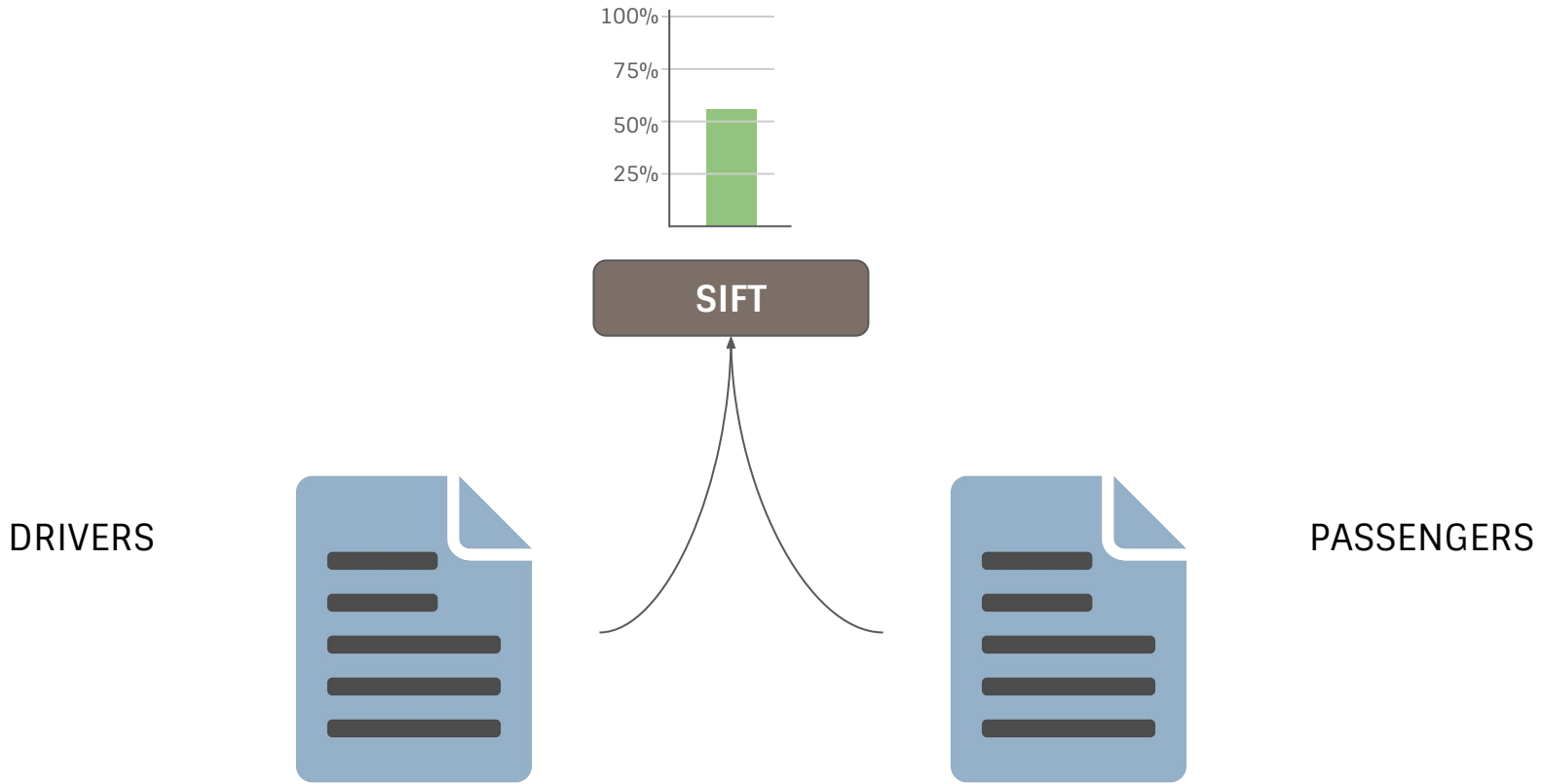
CHROM	POS	STRAND	REF	CHANGE	LABEL
1	10000	1	A	T	DRIVER
12	20000	0	C	G	PASSENGER

...

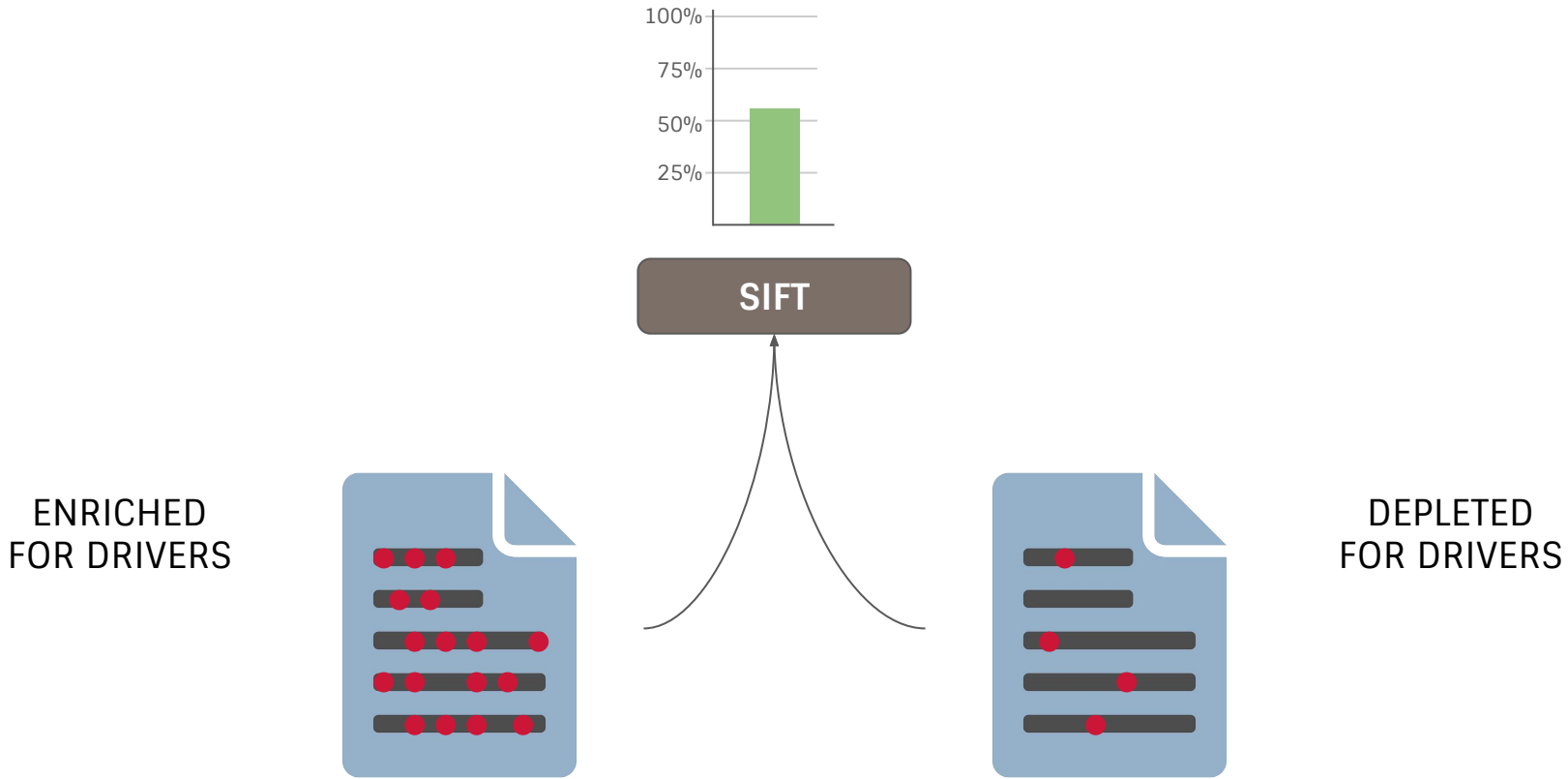
DRIVERS

PASSENGERS

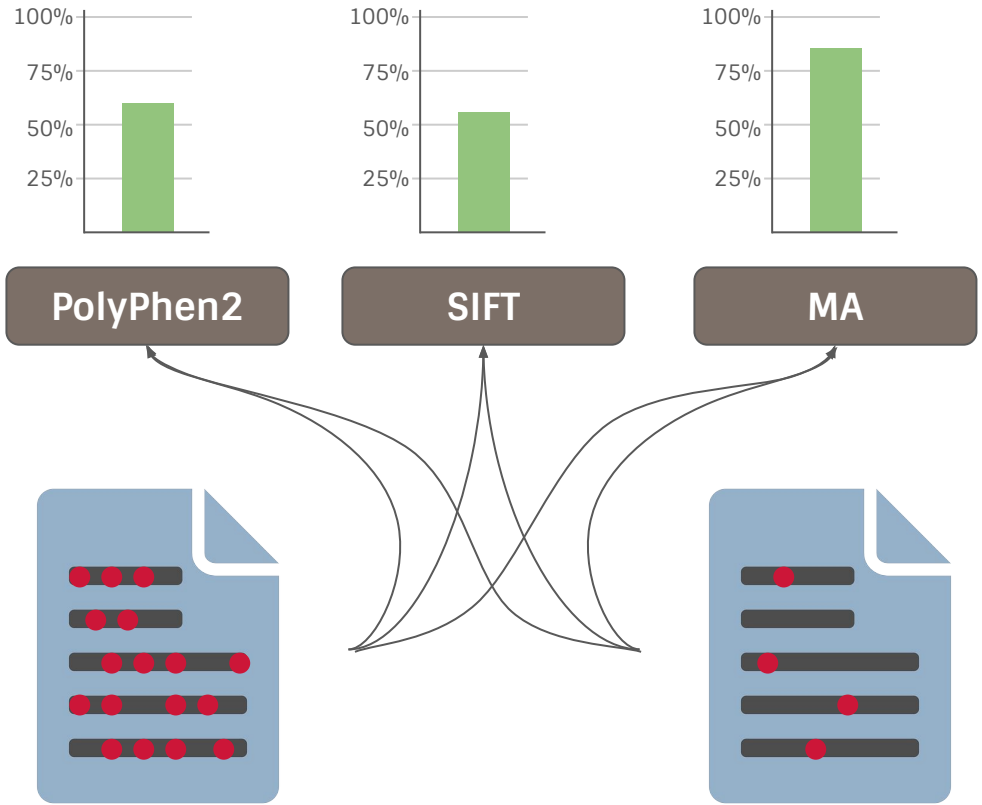
Proxy datasets



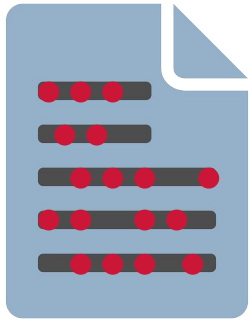
Proxy datasets



Proxy datasets



Proxy datasets



ENRICHED FOR DRIVERS

Humvar

Mutations intOGen
Integrative Onco Genomics

wg-*

Rubio-**P**erez et al **D**rivers

Tamborero et al **D**rivers

CHASM **D**rivers

Cancer **G**ene **C**ensus

COSMIC
Catalogue of somatic mutations in cancer

cm-*
cw-*

Evaluation of performance

		Proxy Dataset	
		Enriched	Reference
Predictor	Func Impact Driver	TP	FP
	Non Func Impact Non Driver	FN	TN

Evaluation of performance

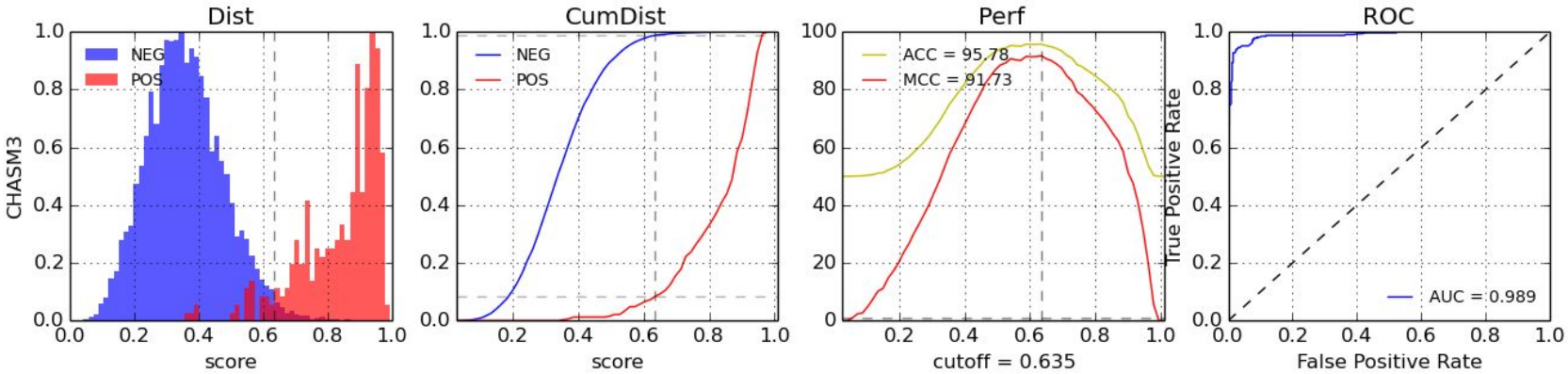
		Proxy Dataset	
		Enriched	Reference
Predictor	Func Impact Driver	TP	FP
	Non Func Impact Non Driver	FN	TN

Accuracy (ACC)	Sensitivity	Specificity
$\frac{\sum TP + \sum TN}{TOTAL}$	$\frac{\sum TP}{\sum Enriched}$	$\frac{\sum TN}{\sum Enriched}$

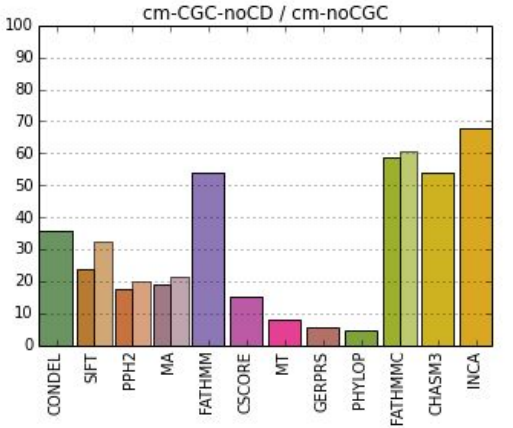
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

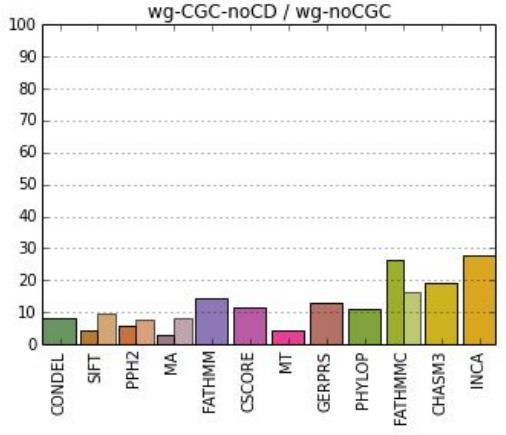
Evaluation of performance

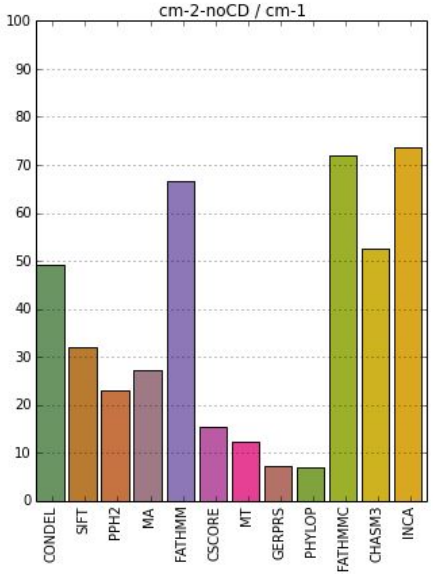
		Proxy Dataset	
		Enriched	Reference
Predictor	Func Impact Driver	TP	FP
	Non Func Impact Non Driver	FN	TN

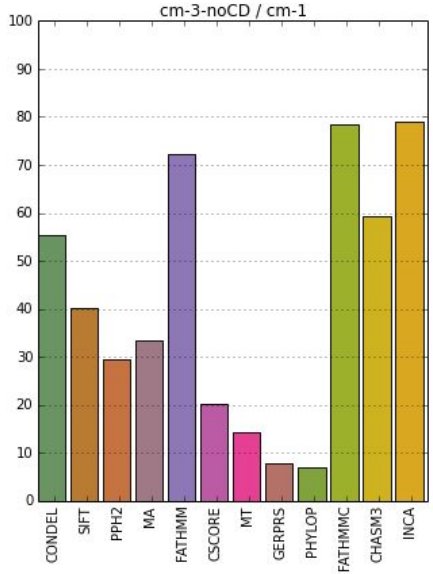


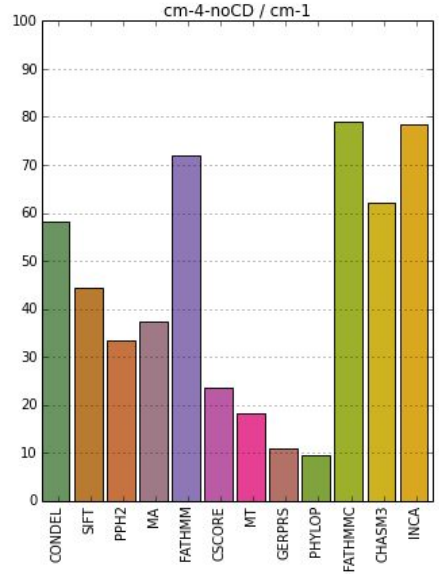
Benchmark results







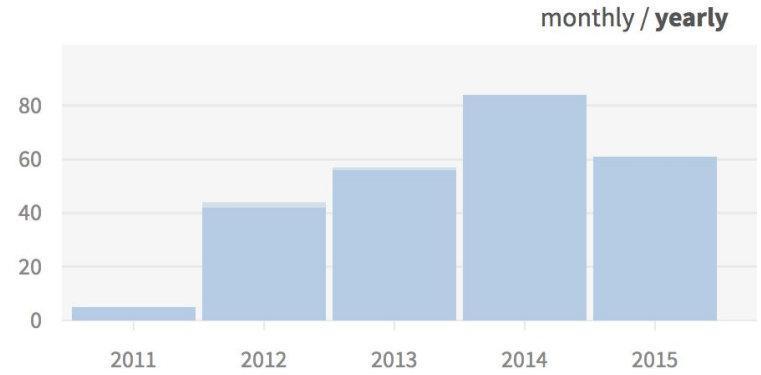


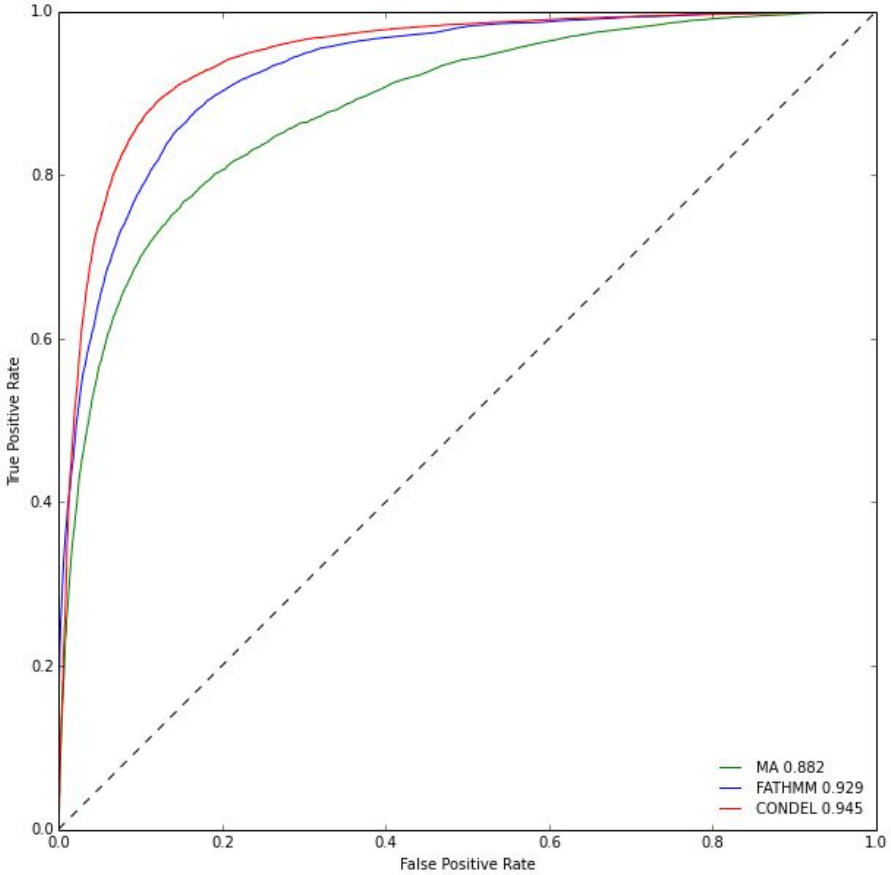


Condel

González-Pérez, A., & López-Bigas, N. (2011).
Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel.
The American Journal of Human Genetics, 88(4), 440–449. http://doi.org/10.1016/j.ajhg.2011.03.004

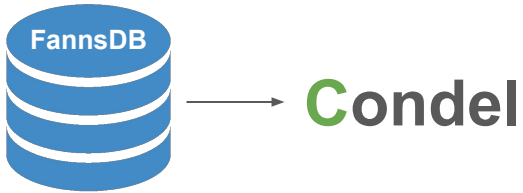
251 CITATIONS **248** EXCLUDING SELF-CITATIONS

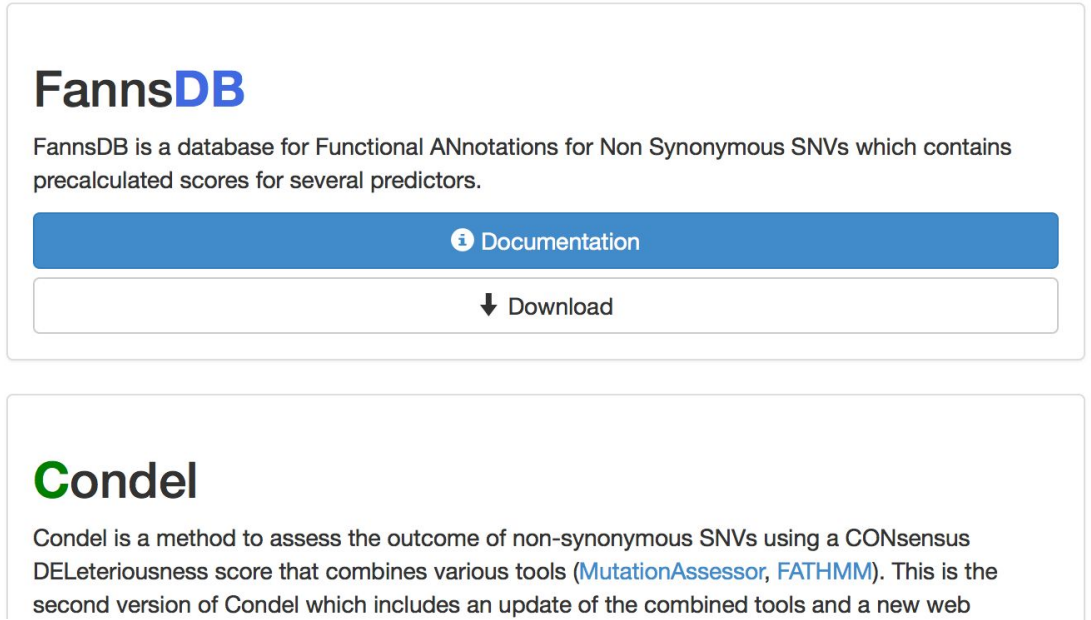




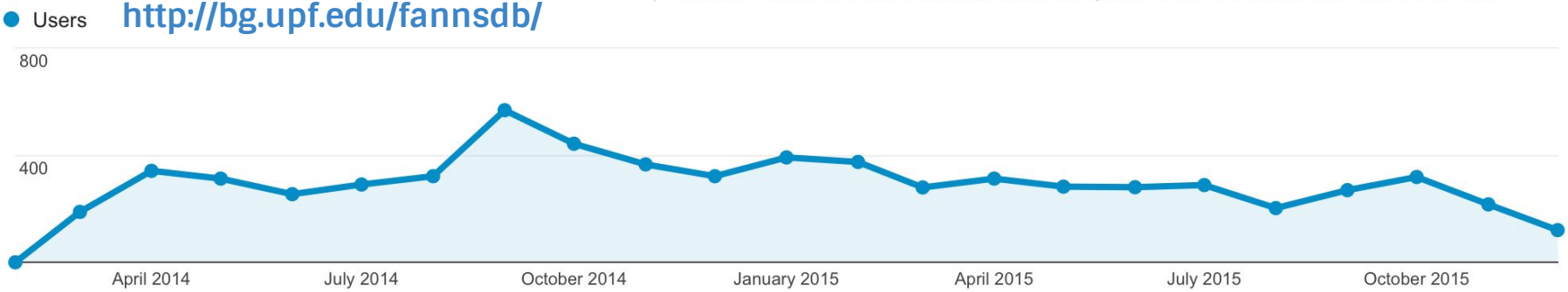
FannsDB

241 million possible protein variants with annotations for SwissProt, RefSeq, HGNC, OMIM, Ensembl





http://bg.upf.edu/fannsdb/



Conclusions

- I created **Gitools** for accessing and analysing biological data, as well as for visualizing multi-dimensional results with **interactive-heatmaps**.
- We collected and **integrated cancer data** from several repositories and large scale projects under the framework of **IntOGen**. The use of simple models and controlled vocabularies was of vital importance for the integration.
- I implemented the **IntOGen workflows** using state of the art tools and methodologies to detect consistently altered genes and identify candidate genes that drive the tumorigenesis for thousand of samples.
- **IntOGen results were easily available** to researchers in different ways. The IntOGen Mutations methods were also available for researchers to analyse their data.
- I integrated scores from several functional impact, cancer driver, and conservation predictors in a **database**, and evaluated their **performance**, and **updated Condel**. Results suggesting that **they alone are not enough for finding drivers** and need to be combined with other methodologies.

Questions

Gitools Case Study