Pathway enrichment analysis of cancer mutations + visualization as enrichment maps

Jüri Reimand
Ontario Institute for Cancer Research, Toronto, ON, Canada
Juri.Reimand@utoronto.ca

Software requirements:

1. Cytoscape software (version 3.7.0 or above),
   see https://cytoscape.org/download.html
   (https://cytoscape.org/download-platforms.html for other OSs, including Windows)
2. EnrichmentMap app of Cytoscape (version 3.1.0 or above),
   see menu *Apps>App manager…* or http://apps.cytoscape.org/apps/enrichmentmap

PATHWAY ENRICHMENT ANALYSIS OF CANCER DRIVER MUTATIONS

Let's get a few gene lists for analysis

ARTICLE

OPEN
doi:10.1038/nature12634

Mutational landscape and significance across 12 major cancer types

Cyriac Kandoth[1]*, Michael D. McLellan[1]*, Fabio Vandin[2], Kai Ye[1,3], Beifang Niu[1], Charles Lu[1], Mingchao Xie[1], Qunyuan Zhang[1,3], Joshua F. McMichael[1], Matthew A. Wyczalkowski[1], Mark D. M. Leiserson[2], Christopher A. Miller[1], John S. Welch[4,5], Matthew J. Walter[4,5], Michael C. Wendl[1,3,6], Timothy J. Ley[1,3,4,5], Richard K. Wilson[1,3,5], Benjamin J. Raphael[2] & Li Ding[1,3,4,5]

The Cancer Genome Atlas (TCGA) has used the latest sequencing and analysis methods to identify somatic variants across thousands of tumours. Here we present data and analytical results for point mutations and small insertions/deletions from 3,281 tumours across 12 tumour types as part of the TCGA Pan-Cancer effort. We illustrate the distributions of mutation frequencies, types and contexts across tumour types, and establish their links to tissues of origin, environmental/carcinogen influences, and DNA repair defects. Using the integrated data sets, we identified 127 significantly mutated genes from well-known (for example, mitogen-activated protein kinase, phosphatidylinositol-3-OH kinase, Wnt/β-catenin and receptor tyrosine kinase signalling pathways, and cell cycle control) and emerging (for example, histone, histone modification, splicing, metabolism and proteolysis) cellular processes in cancer. The average number of mutations in these significantly mutated genes varies across tumour types; most tumours have two to six, indicating that the number of driver mutations required during oncogenesis is relatively small. Mutations in transcriptional factors/regulators show tissue specificity, whereas histone modifiers are often mutated across several cancer types. Clinical association analysis identifies genes having a significant effect on survival, and investigations of mutations with respect to clonal/subclonal architecture delineate their temporal orders during tumorigenesis. Taken together, these results lay the groundwork for developing new diagnostics and individualizing cancer treatment.

Highlights:

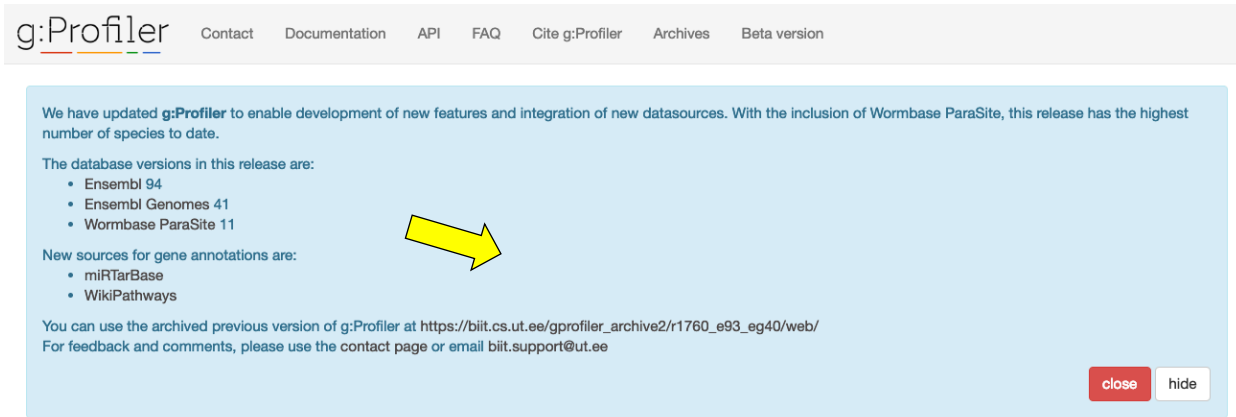
- Using the integrated data sets, the authors identified 127 significantly mutated genes as candidate cancer driver genes
- Genes under positive selection, either in individual or multiple tumour types,
- tend to display higher mutation frequencies above background.
- The statistical analysis identified 127 such genes
- The mutational significance in cancer (MuSiC) package was used to identify
- significant genes for both individual tumour types and the Pan-Cancer sample cohort. [Dees et al. MuSiC: Identifying mutational significance in cancer genomes. Genome Res. 2012]
- These significantly mutated genes are involved in a wide range of cellular processes, including transcription factors/regulators, histone modifiers, genome integrity, receptor tyrosine kinase signalling, cell cycle, mitogen- activated protein kinases (MAPK) signalling, phosphatidylinositol-3-OH kinase (PI(3)K) signalling, Wnt/b-catenin signalling, histones, ubiquitin- mediated proteolysis, and splicing (Fig. 2).

Supplementary Data, Table 4

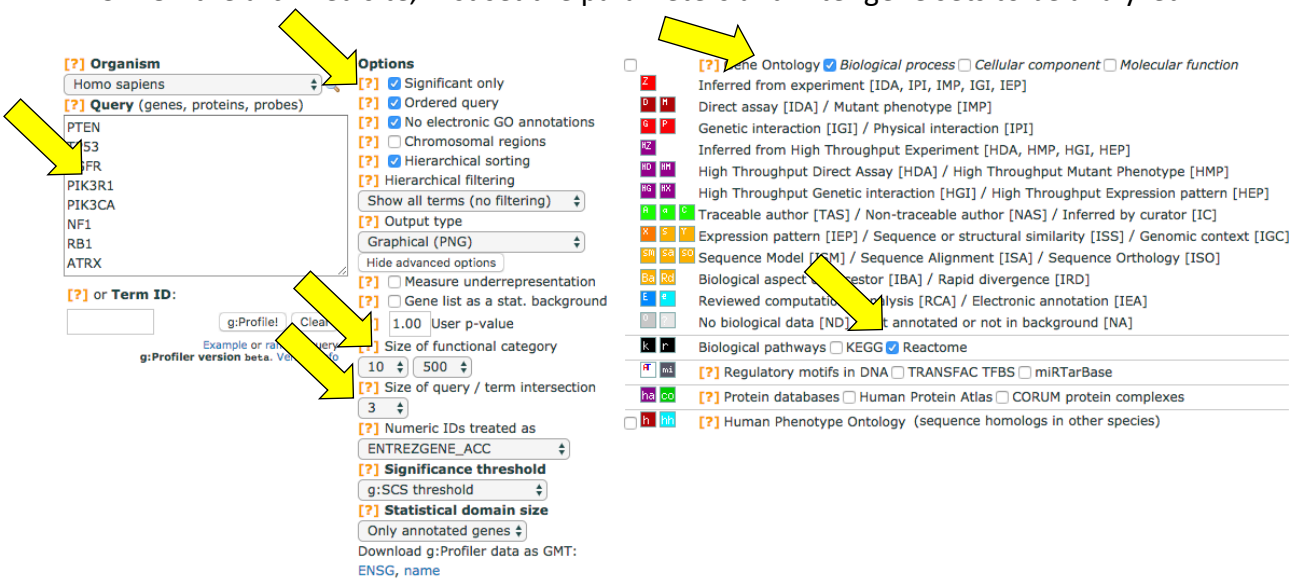
- globally significant, frequency >= 1% for glioblastoma multiforme (GBM): 46

Use g:Profiler to obtain pathway enrichment results for GBM driver genes

1. Go to g:Profiler website at http://biit.cs.ut.ee/gprofiler/
2. You need to use the archived version as the recently updated g:Profiler interface does not include some important parameters we use.



3. On the archived site, first set the parameters and filter gene sets to be analyzed:



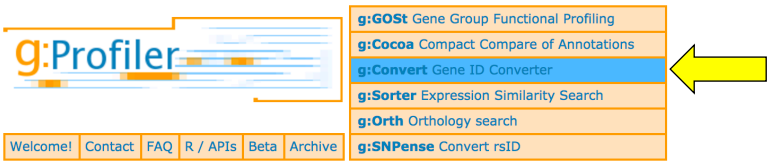
4. Use *Ordered Query* option because the input genes are ordered according to p-value.
5. Paste brain cancer gene list (glioblastoma, GBM) into *Query* box (***Genelist_GBM.txt***).
6. Press *g:Profile* to start the analysis.
7. Scroll down to see significantly enriched pathways and processes. Scroll right to see gene annotations of GO processes (colored) and Reactome pathways (black; scroll further down).
8. Repeat the analysis for kidney cancer genes (KIRC) (***Genelist_KIRC.txt***) (optional).

source	term name Gene Ontology (Biological process)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value	PTEN	TP53	EGFR	PIK3R1	PIK3CA	NF1	RB1	ATRX	IDH1	STAG2	PDGFRA	EPPK1	RPL5	TSH2	PIK3CG	MAP2K1	BRAF	FLT3	CHEK2
BP	cell cycle checkpoint	GO:0000075	196	37	6	2.53e-03		M-					A												M-
BP	DNA integrity checkpoint	GO:0031570	146	37	5	1.10e-02		M-																	M-
BP	DNA damage checkpoint	GO:0000077	136	37	5	7.80e-03		M-																	M-
BP	signal transduction involved in cell cycle checkpoint	GO:0072395	73	37	4	1.44e-02		M-																	M-
BP	signal transduction involved in DNA integrity checkpoint	GO:0072401	72	37	4	1.36e-02		M-																	M-
BP	signal transduction by p53 class mediator	GO:0072331	200	37	8	5.26e-06		D-						S					M						A
BP	regulation of signal transduction by p53 class mediator	GO:1901796	115	30	6	2.77e-05		A											M						A
BP	signal transduction in response to DNA damage	GO:0042770	124	37	6	1.72e-04		D-						S											D-
BP	DNA damage response, signal transduction by p53 class mediator	GO:0030330	101	37	6	5.04e-05		D-						S											A
BP	signal transduction involved in DNA damage checkpoint	GO:0072422	72	37	4	1.36e-02		M-																	M-
BP	cellular response to external stimulus	GO:0071496	196	38	6	2.99e-03		D														X			

Browse results:

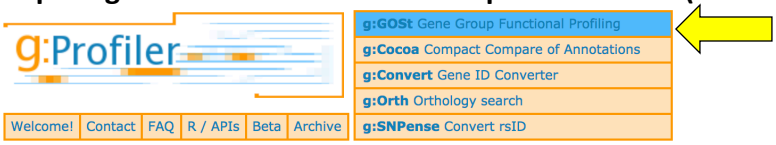
1. Click on numbers in the column *n. of common genes* to find genes that are part of a given process.
2. Uncheck checkbox *Hierarchical sorting* and run g:Profiler again to reveal ranking of results by corrected p-value. Note that many top results are very similar to each other.

g:Convert – gene ID conversion:



1. Paste the same gene list from ***Genelist_GBM.txt*** into the *Query* box.
2. In the *Target Database* list, select the desired type of gene/protein identifiers, for example *UNIPROTSWISSPROT*. Click *Convert IDs* to continue.
3. This tool helps convert many types of gene and protein IDs. Note that many types of IDs can be mixed in g:Profiler and conversion is usually not needed.

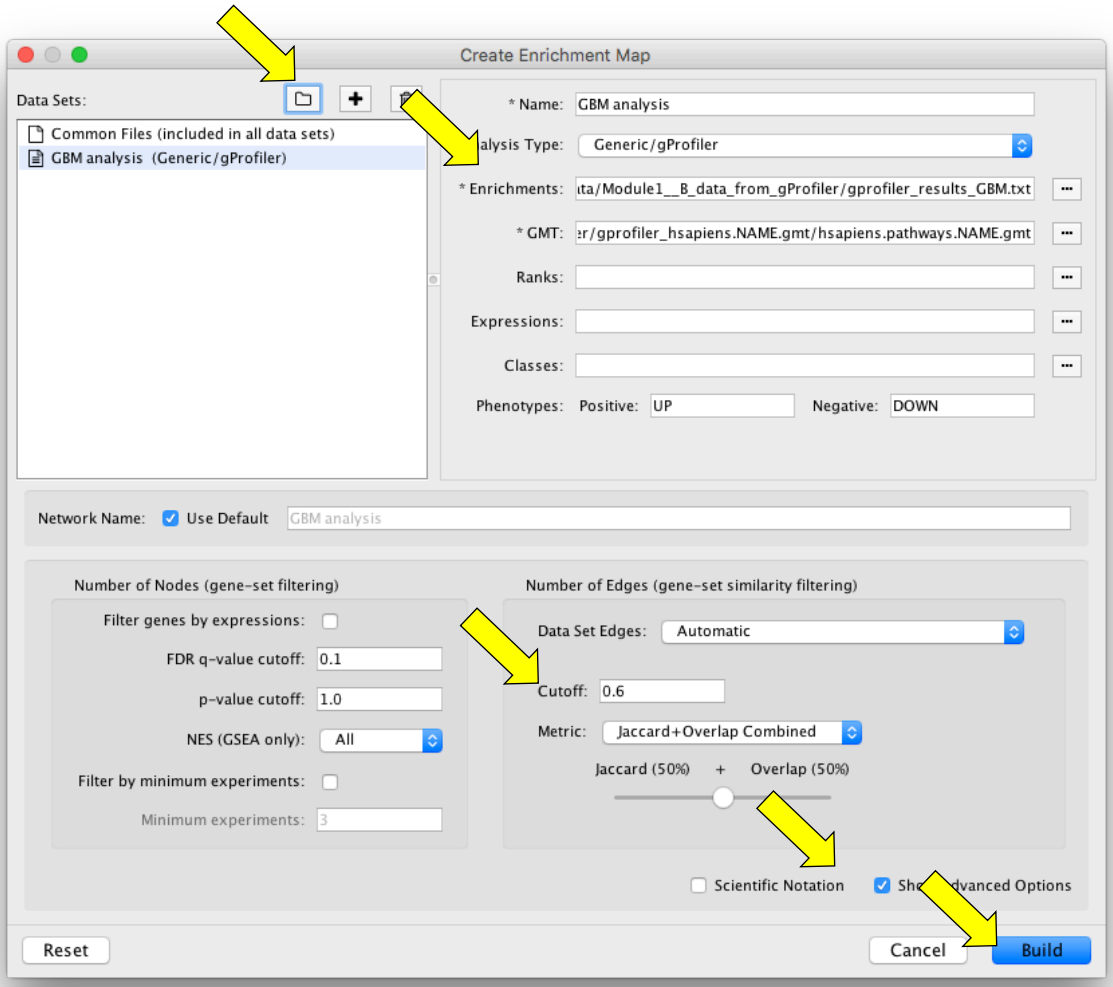
Steps in g:Profiler for Enrichment Map construction (# Click on the g:GOSt tab):



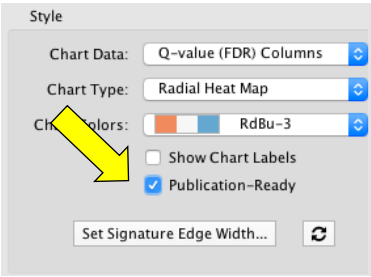
1. Set *Output Type* to *Generic Enrichment Map (TAB).*
2. Click on *g:Profile* to run GBM analysis again (steps 2-5 on previous page).
3. Right-click on *Download data in Generic Enrichment Map (GEM) format* to save the file.
4. Browse the downloaded file in a text editor. Note lists of genes in the rightmost column. These genes are part of the input list and also the pathway. These genes are responsible for the given pathway enrichment.
5. At the bottom of *Advanced Options,* find *Download g:Profiler data as GMT* and right-click the link *name* to save the zip file with gene-set annotations.
6. From the zip file, you will need the file ***hsapiens.pathways.NAME.gmt***
   (# Since we already have this file in the original workshop dataset at ***hsapiens.pathways.NAME.gmt***, we don't need this download step; it is still worth checking where the download link is located in g:Profiler).

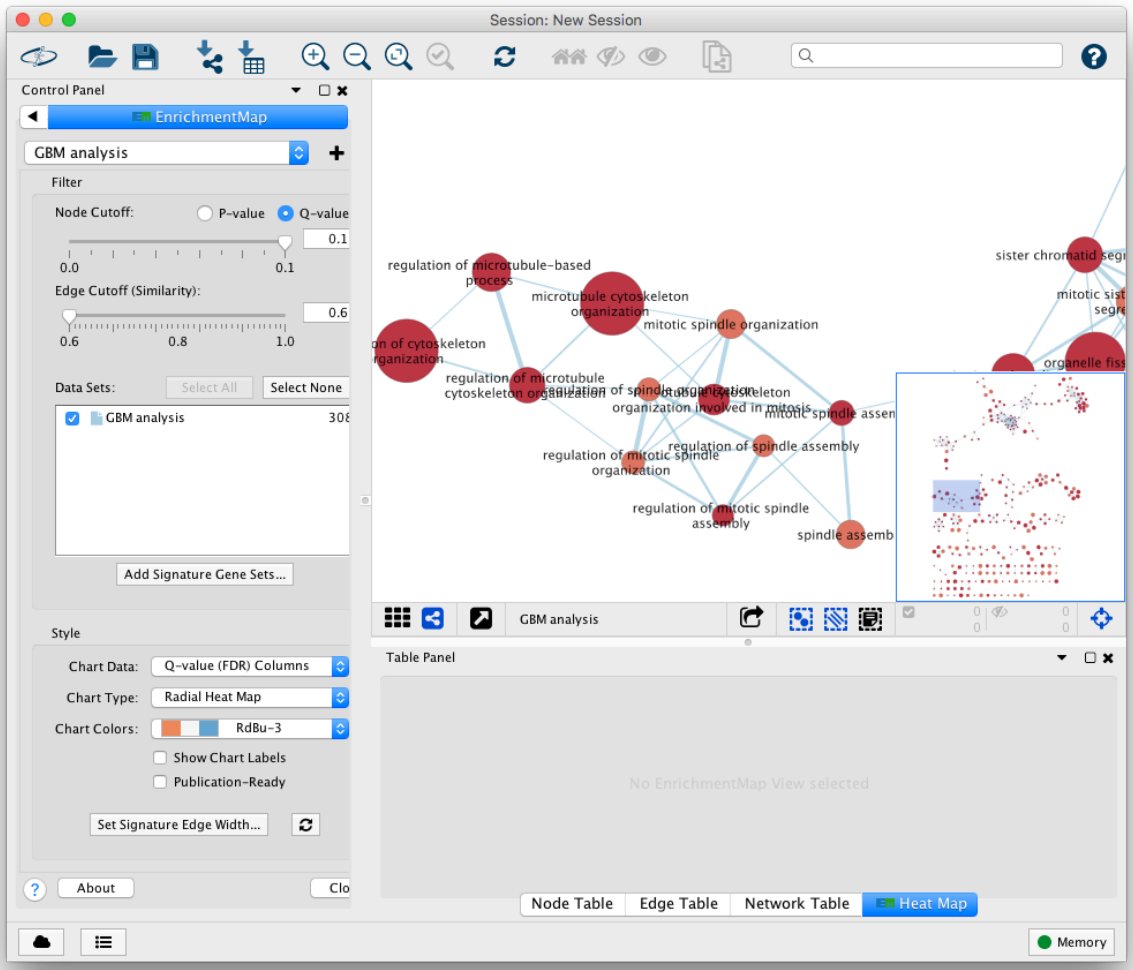
Building an Enrichment Map visualization in Cytoscape

1. Start Cytoscape
2. From the main menu, select Apps>EnrichmentMap.
    a. if you are the first-time Cytoscape user, click on the App manager tab. Choose EnrichmentMap from the second column and click on install button. Then select Apps>EnrichmentMap from main menu once you have done that.
3. Click '+' to start new analysis.
4. Set up GBM analysis by filling the form:
    a. *Generic/gProfiler* for *Analysis Type*,
    b. g:Profiler pathway enrichment results for *Enrichments (e.g.,* ***gprofiler_results_GBM.txt****)*,
    c. pathways file for *GMT (e.g.,* ***hsapiens.pathways.NAME.gmt****).*
    d. Set *Name* to *GBM analysis*.
5. Then click checkbox *Show Advanced Options* and set *Cutoff* to 0.6. Cutoff determines how dense the network is and ranges between 0 and 1. Higher numbers mean that more edges between similar pathways are removed (two pathways are 'similar' if they share many genes).
    a. Optionally, try building enrichment maps with varying parameters (0.2, 0.4, 0.6).
6. Click *Build*.

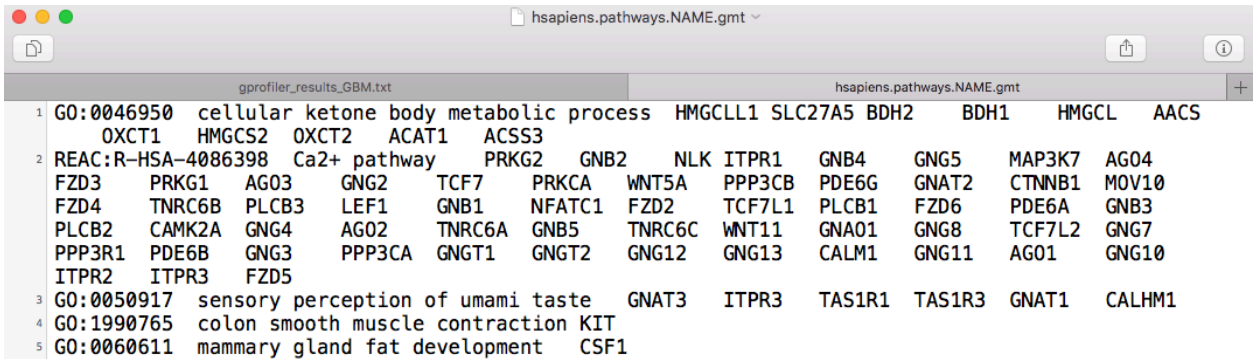


7. Enrichment map diagram will be generated. Use mouse to browse and zoom using buttons above.  .
8. Try the checkbox *Publication-Ready* on the left bottom Style submenu. This will remove node labels and allow you to annotate major functional themes for every subnetwork.

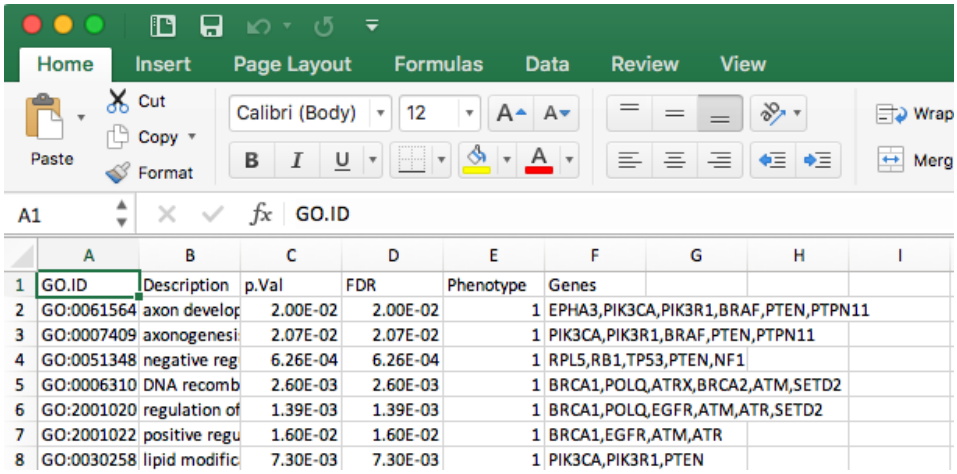




9. Export network to PDF (menu *File > Export > Network as Image*). Zoom out the network to capture it entirely in the PDF.
10. Review input files using a text editor or spreadsheet software (***hsapiens.pathways.NAME.gmt*** and ***gprofiler_results_GBM.txt***).
11. The GMT file contains one process or pathway per row. Open in a text editor.

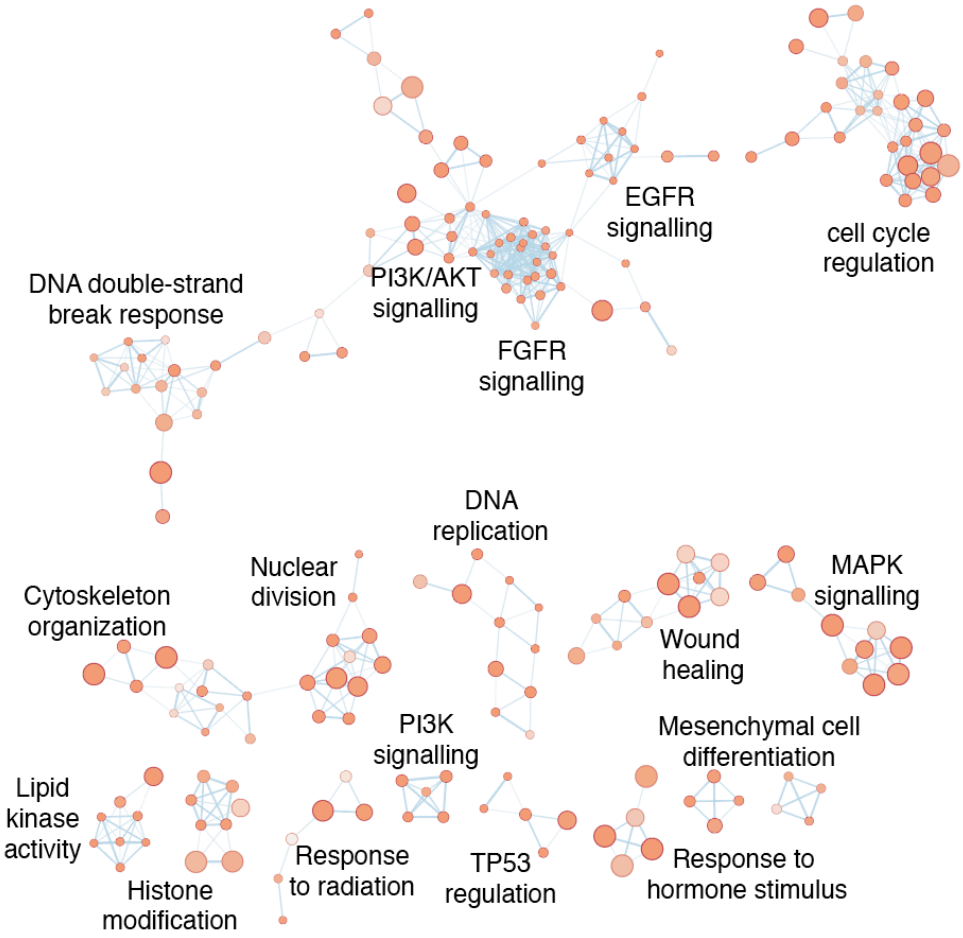


12. The enrichment text file contains info on enriched pathways. Open in a spreadsheet software like MS Excel.



	A	B	C	D	E	F
1	GO.ID	Description	p.Val	FDR	Phenotype	Genes
2	GO:0061564	axon develop	2.00E-02	2.00E-02	1	EPHA3,PIK3CA,PIK3R1,BRAF,PTEN,PTPN11
3	GO:0007409	axonogenesi	2.07E-02	2.07E-02	1	PIK3CA,PIK3R1,BRAF,PTEN,PTPN11
4	GO:0051348	negative reg	6.26E-04	6.26E-04	1	RPL5,RB1,TP53,PTEN,NF1
5	GO:0006310	DNA recomb	2.60E-03	2.60E-03	1	BRCA1,POLQ,ATRX,BRCA2,ATM,SETD2
6	GO:2001020	regulation of	1.39E-03	1.39E-03	1	BRCA1,POLQ,EGFR,ATM,ATR,SETD2
7	GO:2001022	positive regu	1.60E-02	1.60E-02	1	BRCA1,EGFR,ATM,ATR
8	GO:0030258	lipid modific	7.30E-03	7.30E-03	1	PIK3CA,PIK3R1,PTEN

13. Annotate the GBM enrichment map by highlighting biological themes of subnetworks. The full network with all pathway names annotated at nodes is usually too busy to be useful in a publication. Therefore, we recommend removing individual pathway labels and only showing labels for entire groups (subnetworks) of similar pathways. Labelling groups is best achieved manually, based on your expert knowledge of the biology and experimental details. Annotation can be also done automatically using the AutoAnnotate2 app in Cytoscape (your mileage may vary).
14. Build another enrichment map for kidney cancer genes (KIRC) (optional).


EGFR
signalling
cell cycle
regulation
DNA double-strand
break response
PI3K/AKT
signalling
FGFR
signalling
DNA
replication
Nuclear
division
MAPK
signalling
Cytoskeleton
organization
Wound
healing
PI3K
signalling
Mesenchymal cell
differentiation
Lipid
kinase
activity
Response
to radiation
TP53
regulation
Response to
hormone stimulus
Histone
modification