

RNA-Seq workflow development for a clinical use case

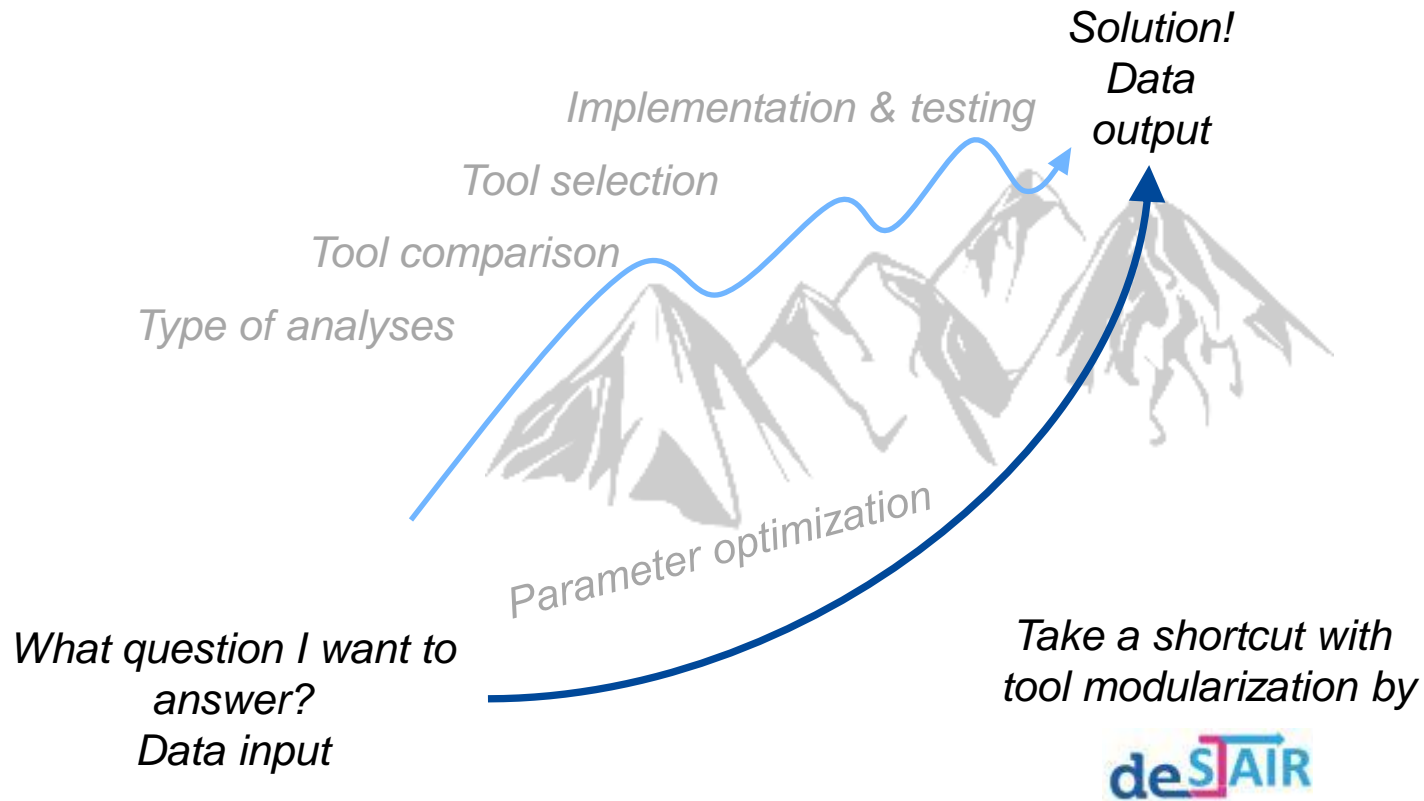
Markus Wolfien and Andrea Bagnacani

de.NBI Training – 28th June 2018 Jena

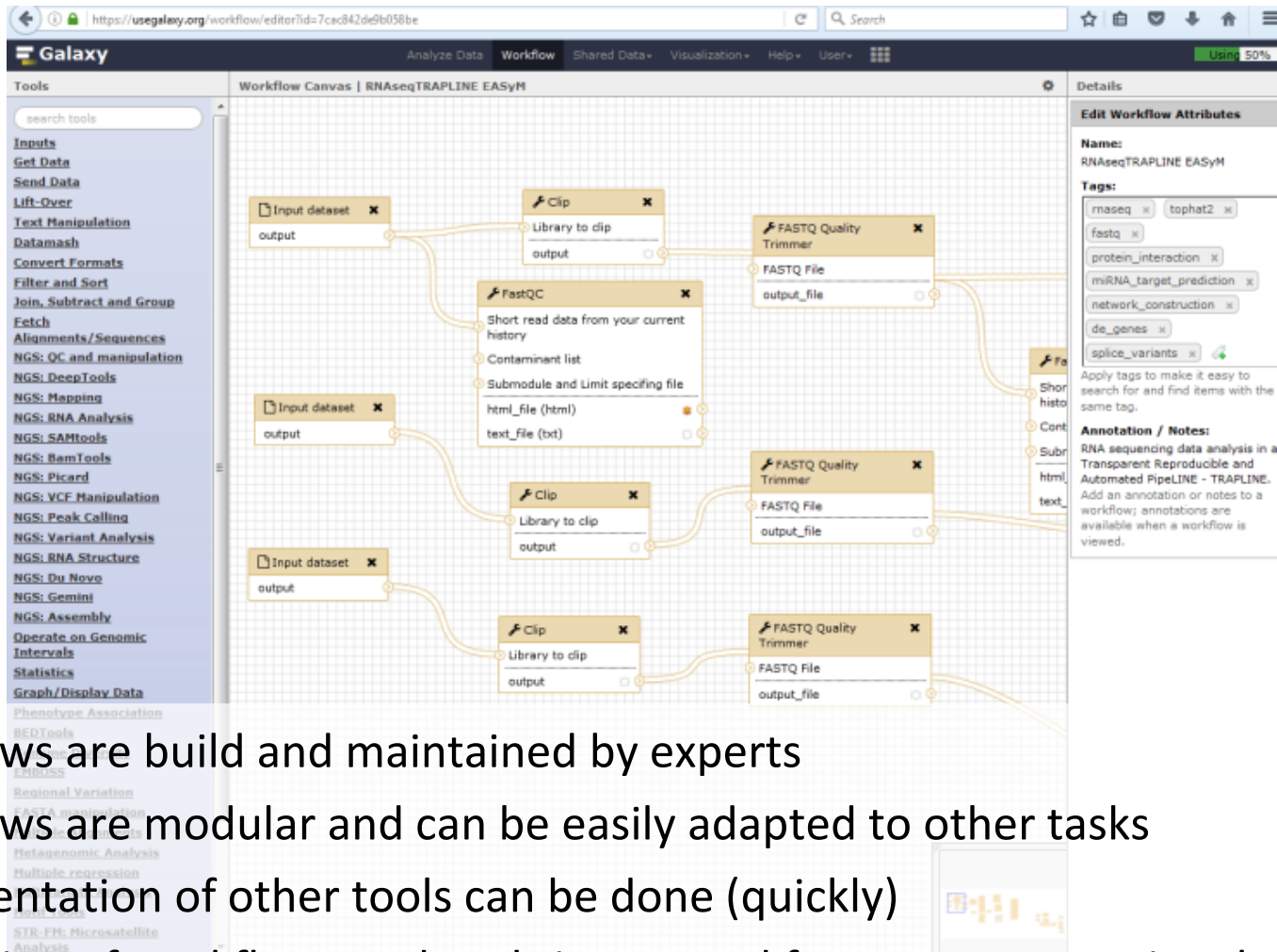
www.sbi.uni-rostock.de



Why using workflows?



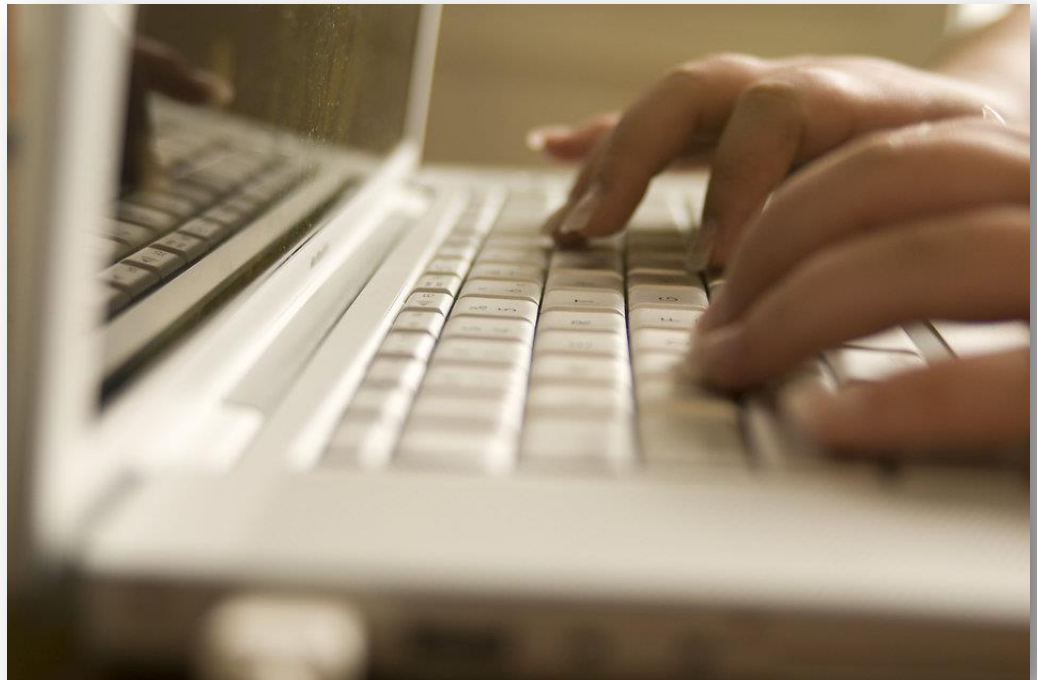
Lott, Wolfien, Riege, Bagnacani, et al., *J.Biotech*, 2017



Hands on part

11:30 – 14:45

“Clinical use case for RNA-Seq, combining all previous processing steps and linking results to further resources“





www.pinkribbon-deutschland.de

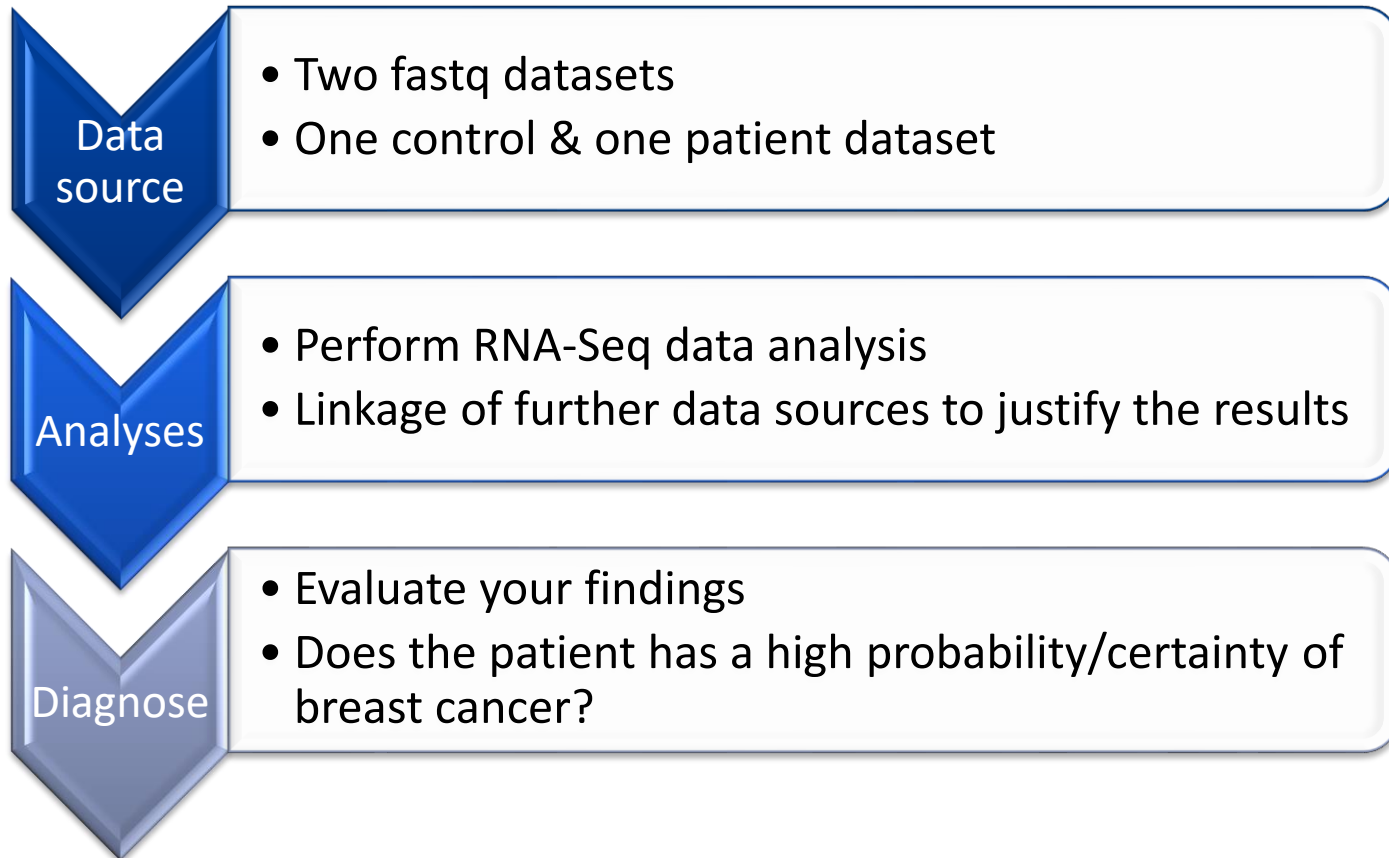
Most common cancer in women worldwide

Leading cause of death from cancer in women worldwide

Predictive factors that identify a benefit

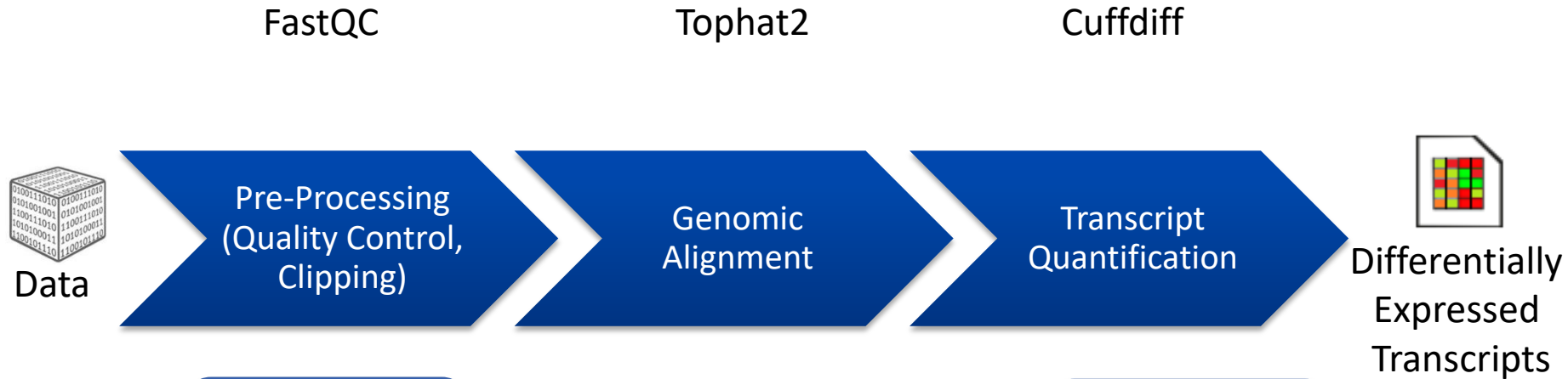
Many different variations and subtypes

Many different therapeutically approaches



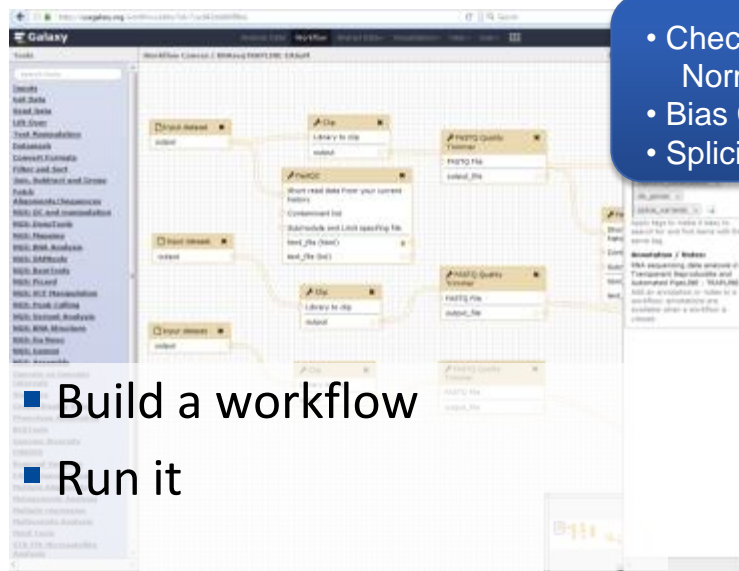


<https://usegalaxy.eu/u/mwolfien/h/rna-seq-training-jena>



- Evaluate Reads (e.g. Sequence Quality, GC Content, Read length)

- Check RPKM Normalization
- Bias Correction
- Splicing detection

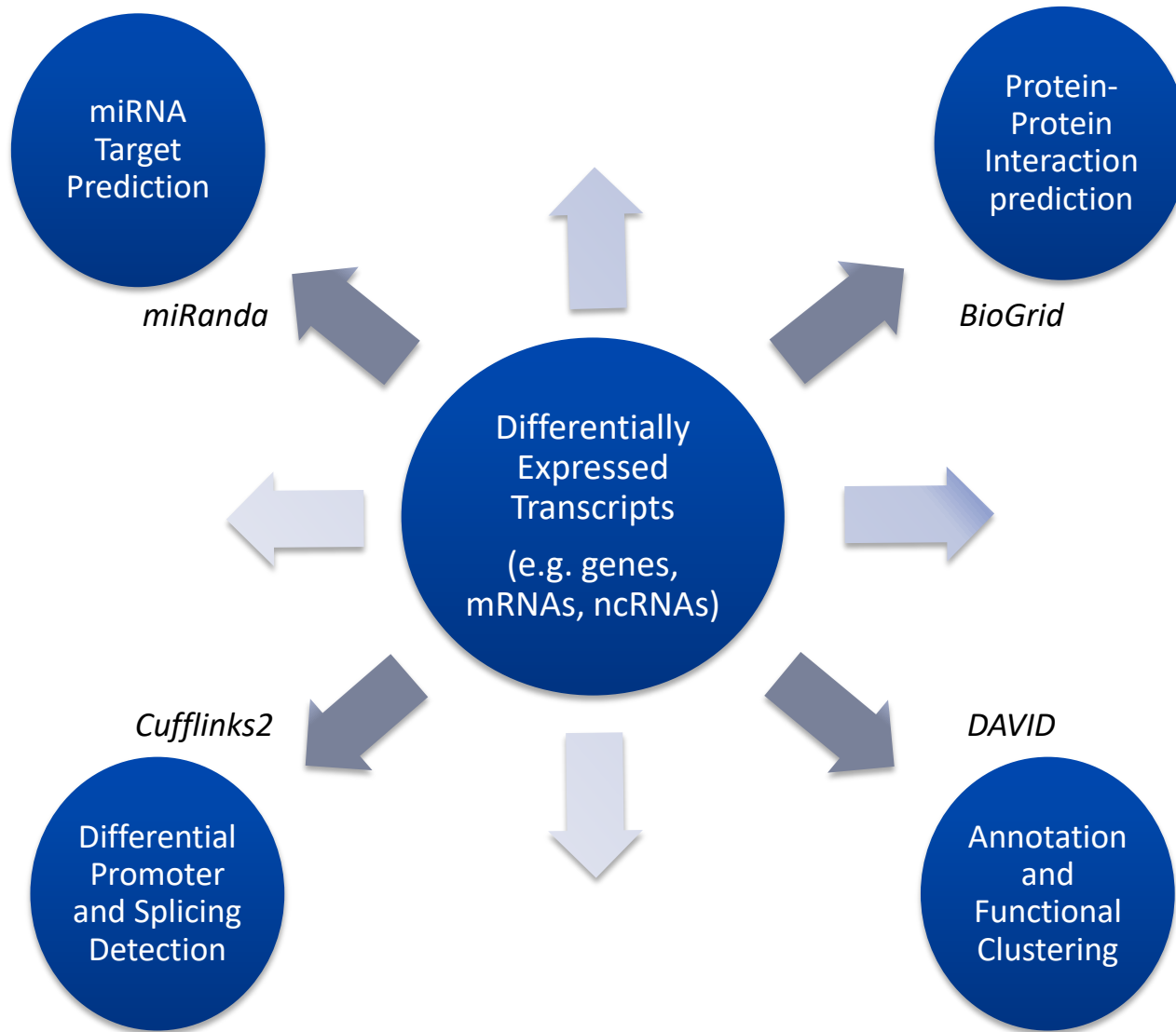


- Build a workflow
- Run it

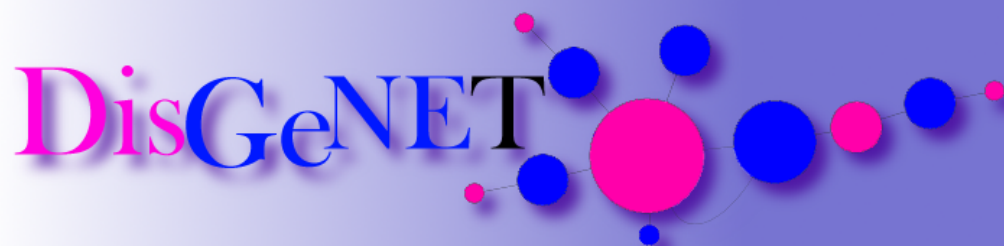
Explore your data!

gene_exp - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_ch	test_stat	p_value	q_value	significant						
1	ADAMTS4	ADAMTS4	ADAMTS4	chr1:1611595	patient	control	OK	118.001	48.454	-128.411	-43.619	5,00E-05	0.0140016	yes						
2	AOC3	AOC3	AOC3	chr17:410032	patient	control	OK	11.335	457.446	-130.911	-43.654	5,00E-05	0.0140016	yes						
3	APOD	APOD	APOD	chr3:1952955	patient	control	OK	207.113	540.507	13.839	471.016	5,00E-05	0.0140016	yes						
4	ARHGAP40	ARHGAP40	ARHGAP40	chr20:372305	patient	control	OK	643.756	188.355	154.887	530.044	5,00E-05	0.0140016	yes						
5	ARHGEF19	ARHGEF19	ARHGEF19	chr1:1652455	patient	control	OK	357.156	112.397	-166.795	-60.984	5,00E-05	0.0140016	yes						
6	ARRDC1	ARRDC1	ARRDC1	chr9:1405000	patient	control	OK	699.613	139.438	0.994996	340.275	0.0003	0.0493798	yes						
7	ATL1	ATL1	ATL1	chr14:509997	patient	control	OK	76.407	249.703	-161.349	-441.121	5,00E-05	0.0140016	yes						
8	ATP6V0D2	ATP6V0D2	ATP6V0D2	chr8:8711113	patient	control	OK	141.496	602.768	-123.109	-398.314	0.0002	0.0383333	yes						
9	BCAT1	BCAT1	BCAT1	chr12:249625	patient	control	OK	191.027	586.341	-170.397	-622.628	5,00E-05	0.0140016	yes						
10	BMP2	BMP2	BMP2	chr20:674874	patient	control	OK	961.681	336.814	-151.361	-466.973	5,00E-05	0.0140016	yes						
11	BP1FB1	BP1FB1	BP1FB1	chr20:318705	patient	control	OK	36.651	761.925	10.558	395.806	0.00015	0.03125	yes						
12	C9orf152	C9orf152	C9orf152	chr9:1229618	patient	control	OK	126.281	293.493	121.669	444.641	5,00E-05	0.0140016	yes						
13	CCL5	CCL5	CCL5	chr17:341984	patient	control	OK	263.041	571.128	111.852	400.647	5,00E-05	0.0140016	yes						
14	CD109	CD109	CD109	chr6:7440362	patient	control	OK	245.725	920.931	-141.588	-522.319	5,00E-05	0.0140016	yes						
15	CEMIP	CEMIP	CEMIP	chr15:810717	patient	control	OK	838.152	219.043	-1.936	-651.489	5,00E-05	0.0140016	yes						
16	CHI3L1	CHI3L1	CHI3L1	chr1:2031480	patient	control	OK	521.762	786.714	-272.948	-907.519	5,00E-05	0.0140016	yes						
17	CITED4	CITED4	CITED4	chr1:4132672	patient	control	OK	550.335	145.476	14.024	520.669	5,00E-05	0.0140016	yes						
18	CNN1	CNN1	CNN1	chr19:116495	patient	control	OK	237.007	108.075	-11.329	-374.136	0.0003	0.0493798	yes						
19	COL10A1	COL10A1	COL10A1	chr6:1164215	patient	control	OK	168.116	608.885	-146.522	-479.166	5,00E-05	0.0140016	yes						
20	CRYAB	CRYAB	CRYAB	chr11:111775	patient	control	OK	741.219	291.357	-134.711	-423.311	5,00E-05	0.0140016	yes						
21	CYP4B1	CYP4B1	CYP4B1	chr1:4726466	patient	control	OK	382.688	134.253	181.071	526.254	5,00E-05	0.0140016	yes						
22	CYP4X1	CYP4X1	CYP4X1	chr1:4748923	patient	control	OK	430.477	106.728	130.993	392.501	0.00015	0.03125	yes						
23	DEGS2	DEGS2	DEGS2	chr14:100612	patient	control	OK	123.647	369.609	157.977	528.636	5,00E-05	0.0140016	yes						
24	DKK3	DKK3	DKK3	chr11:119845	patient	control	OK	48.43	250.027	-0.953814	-359.497	0.00025	0.043125	yes						
25	EDIL3	EDIL3	EDIL3	chr5:8323641	patient	control	OK	138.012	653.211	-107.917	-392.637	5,00E-05	0.0140016	yes						
26	EDNRA	EDNRA	EDNRA	chr4:1484020	patient	control	OK	525.935	269.993	-0.961961	-353.657	0.00015	0.03125	yes						
27	ELL2	ELL2	ELL2	chr5:9522080	patient	control	OK	312.428	149.467	-10.637	-395.955	5,00E-05	0.0140016	yes						
28	ERBB2	ERBB2	ERBB2	chr17:378443	patient	control	OK	379.353	15.602	-20.451	195.473	0.0004655	0.00632554	yes						
29	ERMN	ERMN	ERMN	chr2:1581751	patient	control	OK	605.981	119.265	-23.451	-60.247	5,00E-05	0.0140016	yes						
30	FBXL16	FBXL16	FBXL16	chr16:742495	patient	control	OK	160.236	341.278	109.075	408.836	0.0001	0.0250727	yes						
31	FGFR2	FGFR2	FGFR2	chr10:123237	patient	control	OK	116.387	364.171	164.569	40.751	5,00E-05	0.0140016	yes						
32	FXD6	FXD6	FXD6	chr11:117690	patient	control	OK	40.807	190.584	-109.839	-398.591	0.0001	0.0250727	yes						
33	GALNT15	GALNT15	GALNT15	chr3:1621618	patient	control	OK	233.808	821.554	-15.089	-556.106	5,00E-05	0.0140016	yes						
34	GALNT5	GALNT5	GALNT5	chr2:1581143	patient	control	OK	586.511	189.127	-163.281	-438.311	0.00015	0.03125	yes						
35	GFPT2	GFPT2	GFPT2	chr5:1797276	patient	control	OK	344.814	16.149	-109.437	-408.532	0.0001	0.0250727	yes						
36	GJB2	GJB2	GJB2	chr13:207616	patient	control	OK	309.823	147.739	-10.684	-390.409	0.0002	0.0383333	yes						
37	GOLIM4	GOLIM4	GOLIM4	chr3:1677276	patient	control	OK	463.025	227.216	-102.703	-384.509	0.00015	0.03125	yes						
38	GPR68	GPR68	GPR68	chr14:916988	patient	control	OK	227.216	950.869	-125.674	-449.161	5,00E-05	0.0140016	yes						
39	GPR68	GPR68	GPR68	chr14:916988	patient	control	OK	227.216	950.869	-125.674	-449.161	5,00E-05	0.0140016	yes						



- DisGeNET (<http://www.disgenet.org/>)



[Home](#) [About](#) [Search](#) [Browser](#) [Downloads](#) [Cytoscape](#) [RDF](#) [Help](#)

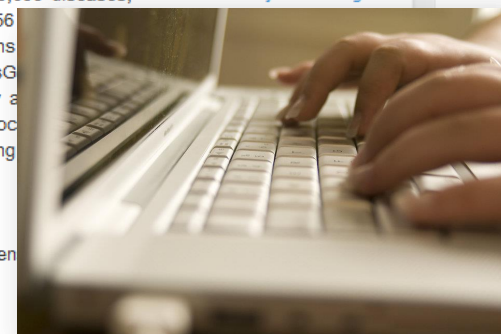
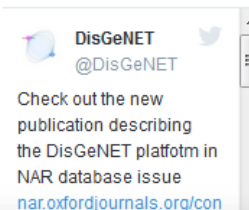
One of the most challenging problems in biomedical research is to understand the underlying mechanisms of complex diseases. Great effort has been spent on finding the genes associated to diseases (Botstein and Risch, 2003; Kann, 2009). However, more and more evidences indicate that most human diseases cannot be attributed to a single gene but arise due to complex interactions among multiple genetic variants and environmental risk factors (Hirschhorn and Daly, 2005). Several databases have been developed storing associations between genes and diseases such as CTDTM (Davis, *et al.*, 2014), OMIM[®] (Hamosh *et al.*, 2005) and the NHGRI-EBI GWAS catalog (Welter *et al.*, 2014). Each of these databases focuses on different aspects of the phenotype-genotype relationship, and due to the nature of the database curation process, they are not complete. Hence, integration of different databases with information extracted from the literature is needed to allow a comprehensive view of the state of the art knowledge within this research field. With this need in mind, we have created DisGeNET.

DisGeNET is a discovery platform integrating information on gene-disease associations (GDAs) from several public data sources and the literature (Piñero *et al.*, 2015). The current version contains (DisGeNET v4.0) contains 429,036 associations, between 17,381 genes and 15,093 diseases, disorders and clinical or abnormal human phenotypes, and 72,870 variant-disease associations (VDAs), between 46,589 SNPs and 6,356 phenotypes. Given the large number of GDAs compiled in DisGeNET, we have also developed a *score* in order to rank the associations supporting evidence. Importantly, useful tools have also been created to explore and analyze the data contained in DisGeNET. DisGeNET can be queried through [Search](#) and [Browse](#) functionalities available from this web interface, or by a plugin created for Cytoscape to query a network representation of the data. Moreover, DisGeNET data can be queried by downloading the SQLite [database](#) to your local machine. Furthermore, an RDF (Resource Description Framework) representation of DisGeNET database is also available. It can be queried using a SPARQL endpoint and a Faceted Browser. Follow the [link](#) for more information.

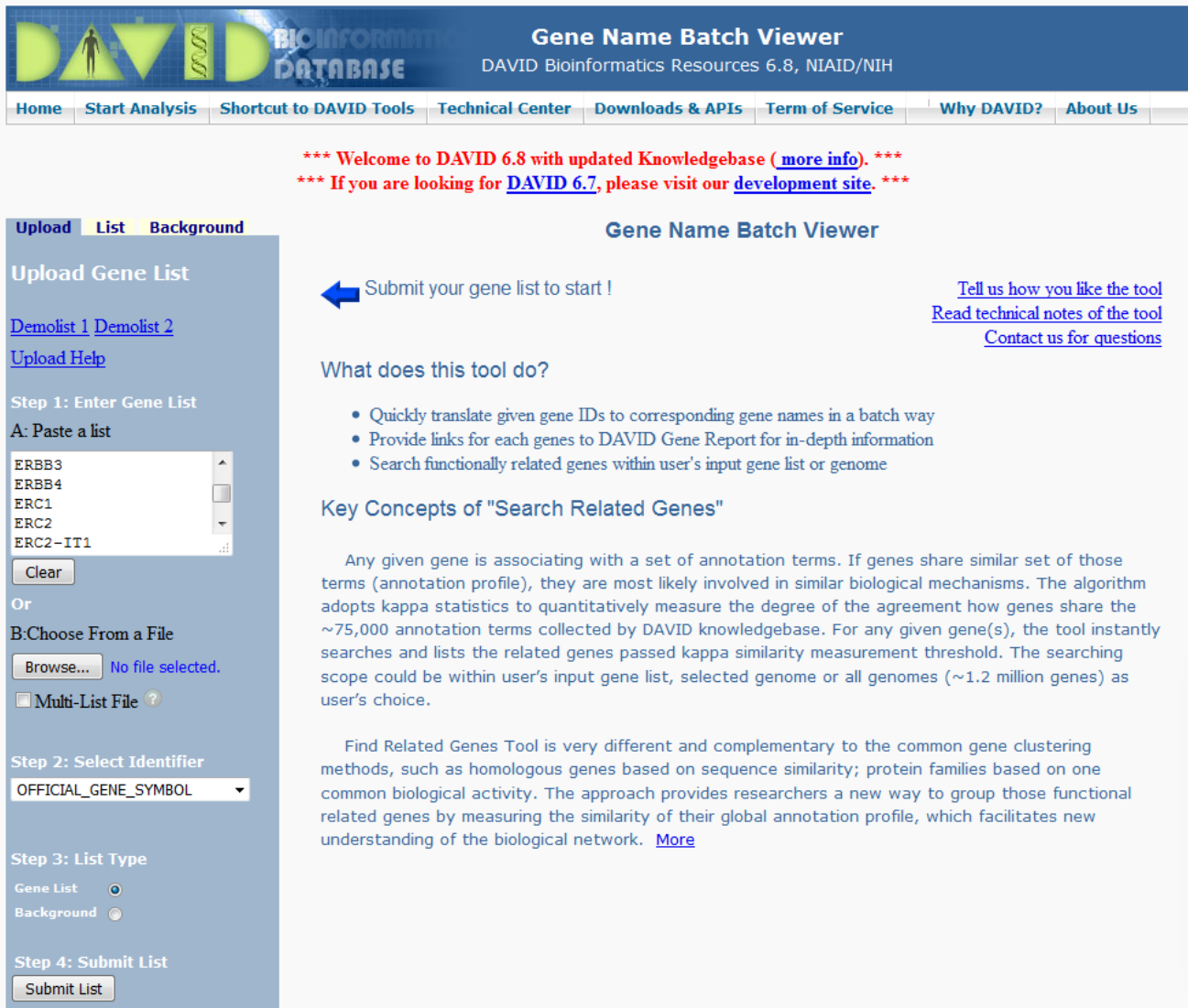
DisGeNET database has been cited by several papers. Some of them can be reviewed [here](#).

The DisGeNET database is made available under the [Open Database License](#). Any rights in individual contents of the database are licensed under the [Database Contents License](#).

Tweets by @DisGeNET



■ David (<https://david.ncifcrf.gov/list.jsp>)



The screenshot shows the DAVID Bioinformatics Resources 6.8 web interface. The header includes the DAVID logo and navigation links: Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About Us. A welcome message states: "*** Welcome to DAVID 6.8 with updated Knowledgebase ([more info](#)). ***" and "*** If you are looking for [DAVID 6.7](#), please visit our [development site](#). ***". The main section is titled "Gene Name Batch Viewer" and contains a left sidebar with tabs for Upload, List, and Background. The Upload tab is active, showing "Upload Gene List" with links for Demolist 1, Demolist 2, and Upload Help. It includes a "Step 1: Enter Gene List" section with a text area containing gene IDs (ERBB3, ERBB4, ERC1, ERC2, ERC2-IT1) and a "Clear" button. Below this is a "B: Choose From a File" section with a "Browse..." button (showing "No file selected.") and a "Multi-List File" checkbox. The "Step 2: Select Identifier" section has a dropdown menu set to "OFFICIAL_GENE_SYMBOL". The "Step 3: List Type" section has radio buttons for "Gene List" (selected) and "Background". The "Step 4: Submit List" section has a "Submit List" button. The main content area on the right says "Submit your gene list to start !" with a left arrow icon, and includes links: "Tell us how you like the tool", "Read technical notes of the tool", and "Contact us for questions". It also has a section "What does this tool do?" with a bulleted list: "Quickly translate given gene IDs to corresponding gene names in a batch way", "Provide links for each genes to DAVID Gene Report for in-depth information", and "Search functionally related genes within user's input gene list or genome". Below this is a section "Key Concepts of 'Search Related Genes'" with a paragraph explaining the kappa statistics algorithm and a "Find Related Genes Tool" paragraph explaining its complementary nature to other methods, with a "More" link.

Gene Name Batch Viewer

Submit your gene list to start !

[Tell us how you like the tool](#)
[Read technical notes of the tool](#)
[Contact us for questions](#)

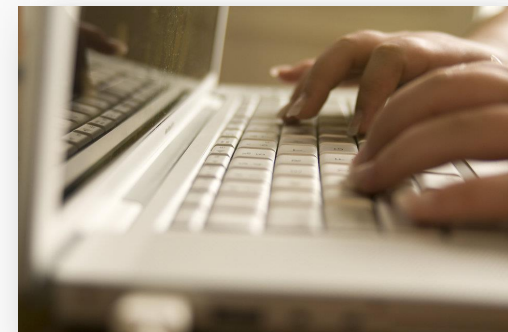
What does this tool do?

- Quickly translate given gene IDs to corresponding gene names in a batch way
- Provide links for each genes to DAVID Gene Report for in-depth information
- Search functionally related genes within user's input gene list or genome

Key Concepts of "Search Related Genes"

Any given gene is associating with a set of annotation terms. If genes share similar set of those terms (annotation profile), they are most likely involved in similar biological mechanisms. The algorithm adopts kappa statistics to quantitatively measure the degree of the agreement how genes share the ~75,000 annotation terms collected by DAVID knowledgebase. For any given gene(s), the tool instantly searches and lists the related genes passed kappa similarity measurement threshold. The searching scope could be within user's input gene list, selected genome or all genomes (~1.2 million genes) as user's choice.

Find Related Genes Tool is very different and complementary to the common gene clustering methods, such as homologous genes based on sequence similarity; protein families based on one common biological activity. The approach provides researchers a new way to group those functional related genes by measuring the similarity of their global annotation profile, which facilitates new understanding of the biological network. [More](#)



- Enrichr (<http://amp.pharm.mssm.edu/Enrichr/>)



[Login](#) | [Register](#)

9,941,912 lists analyzed

245,575 terms

132 libraries

Analyze

[What's New?](#)

[Libraries](#)

[Find a Gene](#)

[About](#)

[Help](#)

Input data

Choose an input file to upload. Either in BED format or a list of genes. For a quantitative set, add a comma and the level of membership of that gene. The membership level is a number between 0.0 and 1.0 to represent a weight for each gene, where the weight of 0.0 will completely discard the gene from the enrichment analysis and the weight of 1.0 is the maximum.

Try an example [BED file](#).

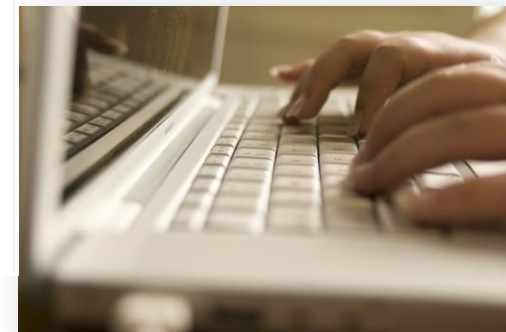
Keine ausgewählt

Or paste in a list of gene symbols optionally followed by a comma and levels of membership. Try two examples: [crisp set example](#), [fuzzy set example](#)

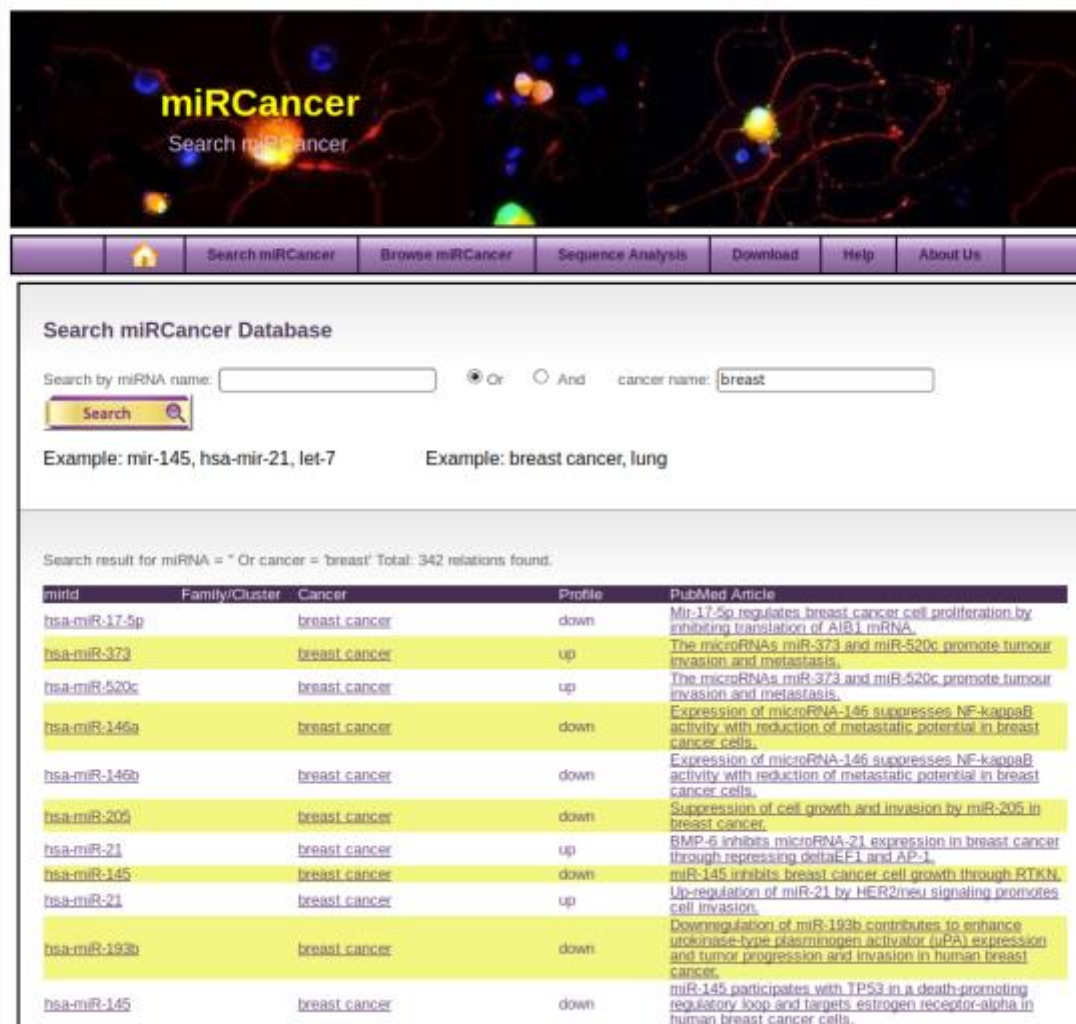
0 gene(s) entered

Enter a brief description for the list in case you want to share it. (Optional)

☐ Contribute



- miRCancer db - Find up regulated miRNAs (<http://mircancer.ecu.edu/index.jsp>)



miRCancer
Search miRCancer

Search miRCancer | Browse miRCancer | Sequence Analysis | Download | Help | About Us

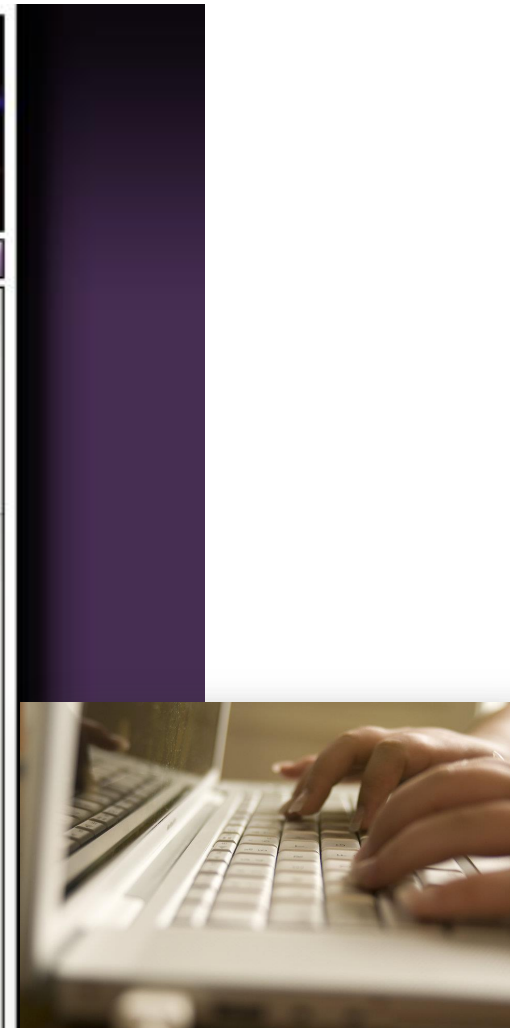
Search miRCancer Database

Search by miRNA name: Or ☐ And cancer name:

Example: mir-145, hsa-mir-21, let-7 Example: breast cancer, lung

Search result for miRNA = " Or cancer = 'breast' Total: 342 relations found.

mirid	Family/Cluster	Cancer	Profile	PubMed Article
hsa-miR-17-5p		breast cancer	down	Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA.
hsa-miR-373		breast cancer	up	The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis.
hsa-miR-520c		breast cancer	up	The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis.
hsa-miR-146a		breast cancer	down	Expression of microRNA-146 suppresses NF-kappaB activity with reduction of metastatic potential in breast cancer cells.
hsa-miR-146b		breast cancer	down	Expression of microRNA-146 suppresses NF-kappaB activity with reduction of metastatic potential in breast cancer cells.
hsa-miR-205		breast cancer	down	Suppression of cell growth and invasion by miR-205 in breast cancer.
hsa-miR-21		breast cancer	up	BMP-6 inhibits microRNA-21 expression in breast cancer through repressing deltaEF1 and AP-1.
hsa-miR-145		breast cancer	down	miR-145 inhibits breast cancer cell growth through RTKH.
hsa-miR-21		breast cancer	up	Up-regulation of miR-21 by HER2neu signaling promotes cell invasion.
hsa-miR-193b		breast cancer	down	Downregulation of miR-193b contributes to enhance urokinase-type plasminogen activator (uPA) expression and tumor progression and invasion in human breast cancer.
hsa-miR-145		breast cancer	down	miR-145 participates with TP53 in a death-promoting regulatory loop and targets estrogen receptor-alpha in human breast cancer cells.



miRNA	Regulation	Target
140-5b	up	
148b	Down	
150	Up	
106b	Up	
143	Down	
19b	Up	
21	up	
...

- TriplexRNA database (<https://www.sbi.uni-rostock.de/triplexrna/>)



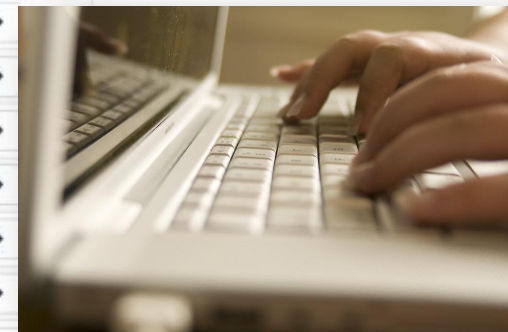
Search targets of synergistic microRNA regulation

Search in **Human** for **miRNA ID** **hsa-miR-140-5p**

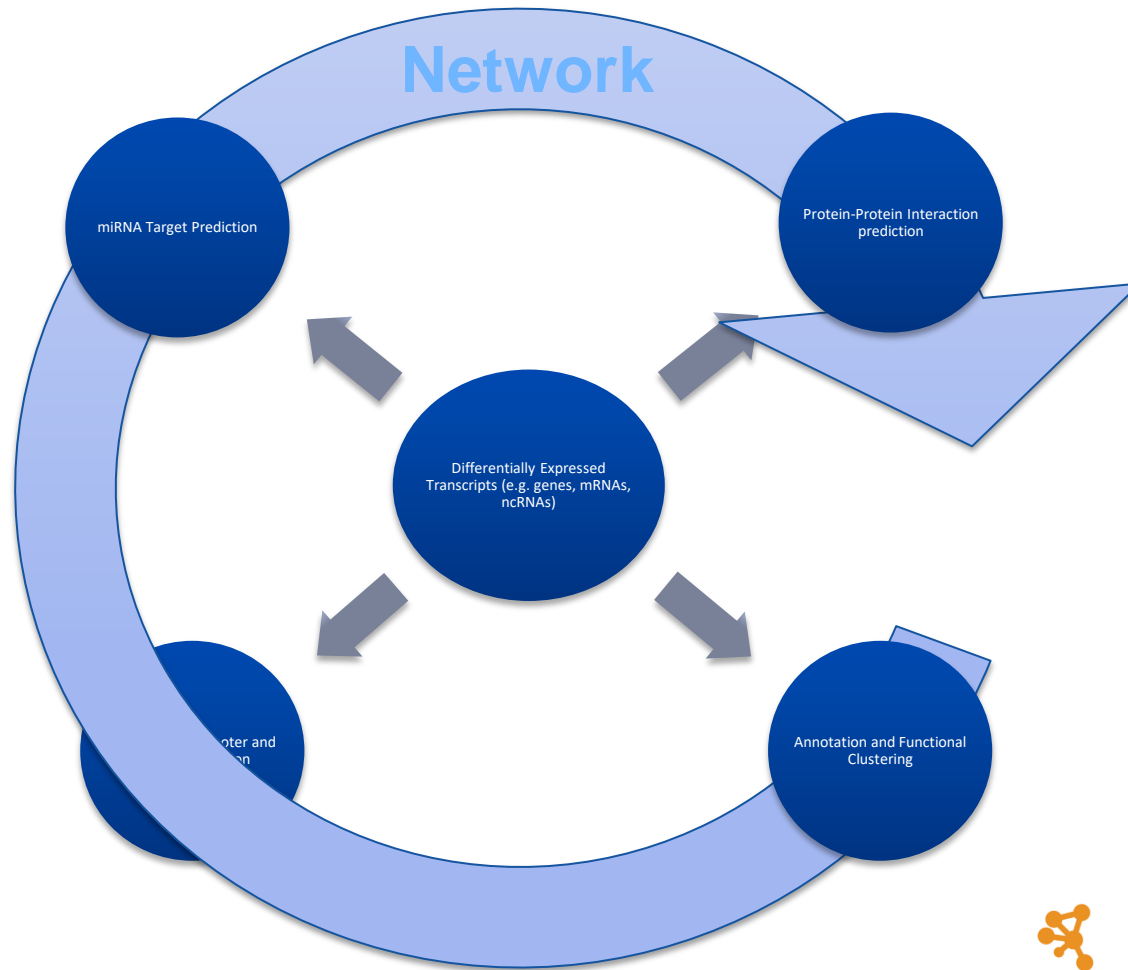


results 

Gene ID	RefSeq ID	miRNA1 ID	miRNA2 ID	Seed distance (nt)	Free energy (Kcal/mol)	Energy gain (Kcal/mol)	Triplex details
ADCY6	NM_015270	hsa-miR-197	hsa-miR-140-5p	23	-48.66	-14.38	more >
ATG4B	NM_178326	hsa-miR-140-5p	hsa-miR-346	28	-47.36	-15.58	more >
ZNF705A	NM_001004328	hsa-miR-140-5p	hsa-miR-296-3p	17	-43.76	-14.28	more >
FGR	NM_005248	hsa-miR-140-5p	hsa-miR-326	33	-43.56	-11.58	more >
PTCD1	NM_015545	hsa-miR-140-5p	hsa-miR-339-5p	34	-43.26	-12.98	more >
AARS	NM_001605	hsa-miR-24	hsa-miR-140-5p	32	-43.16	-17.18	more >
WEE1	NM_003390	hsa-miR-15b	hsa-miR-140-5p	16	-42.86	-16.28	more >
WNT1	NM_005430	hsa-miR-31	hsa-miR-140-5p	28	-42.56	-12.78	more >
ZBTB9	NM_152735	hsa-miR-140-5p	hsa-miR-296-3p	29	-41.96	-11.68	more >
ADRA1A	AY491776	hsa-miR-140-5p	hsa-miR-150	21	-41.96	-12.18	more >



There is nothing more practical than a network

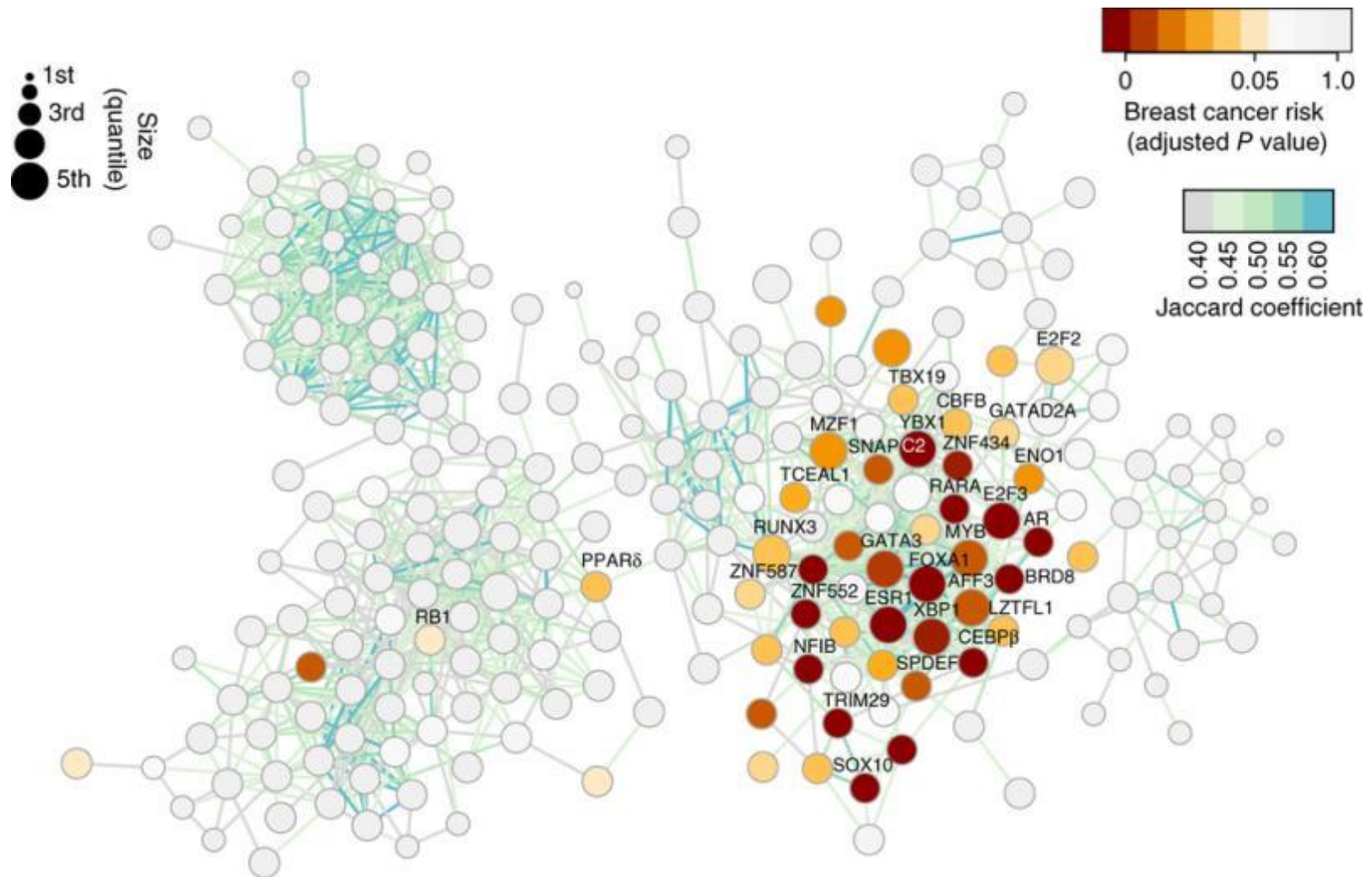


Cytoscape



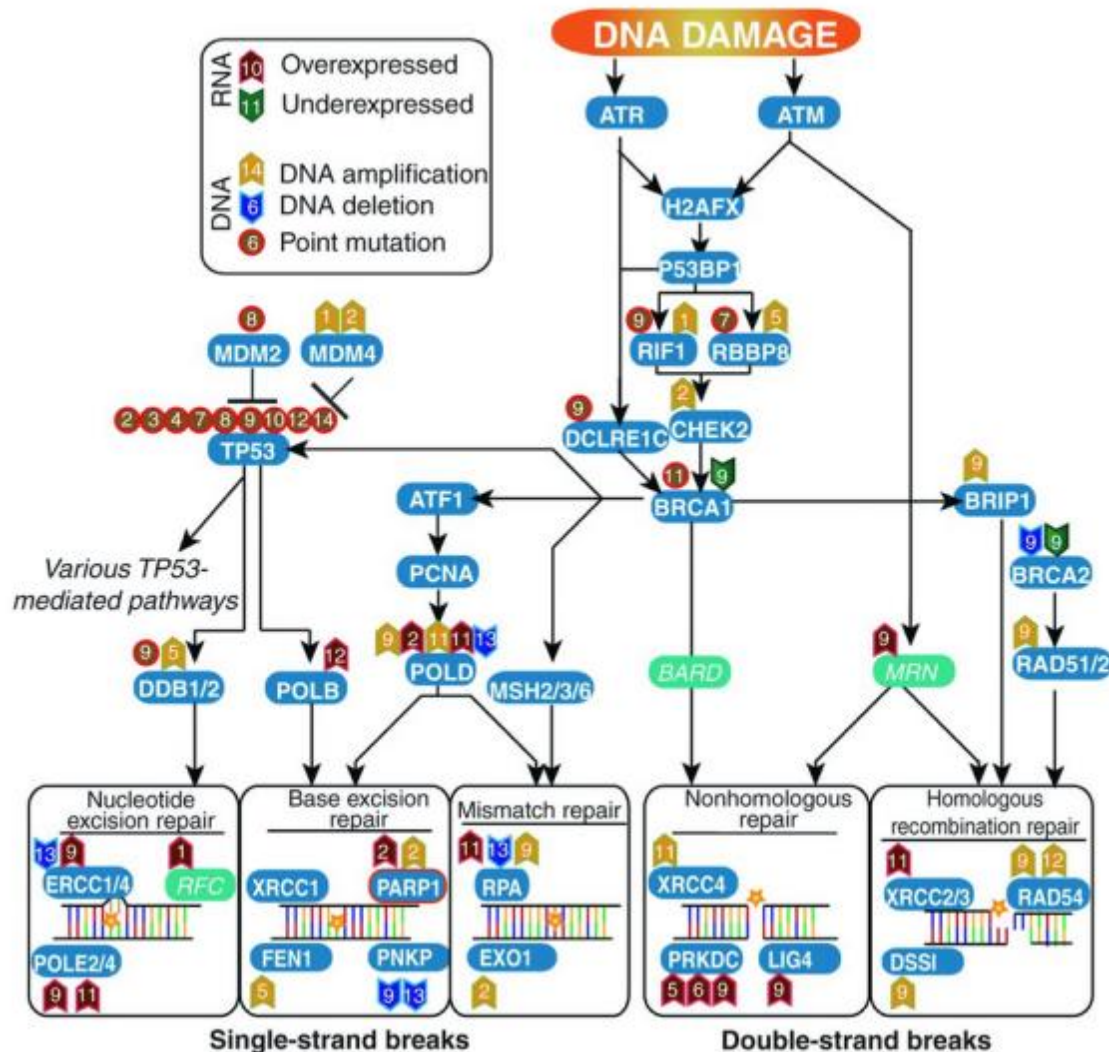
Vanted / CellDesigner

Network comparisons also reveal differences



Castro, *Nat. Gen.*, 2016

What else is done in this field with NGS?



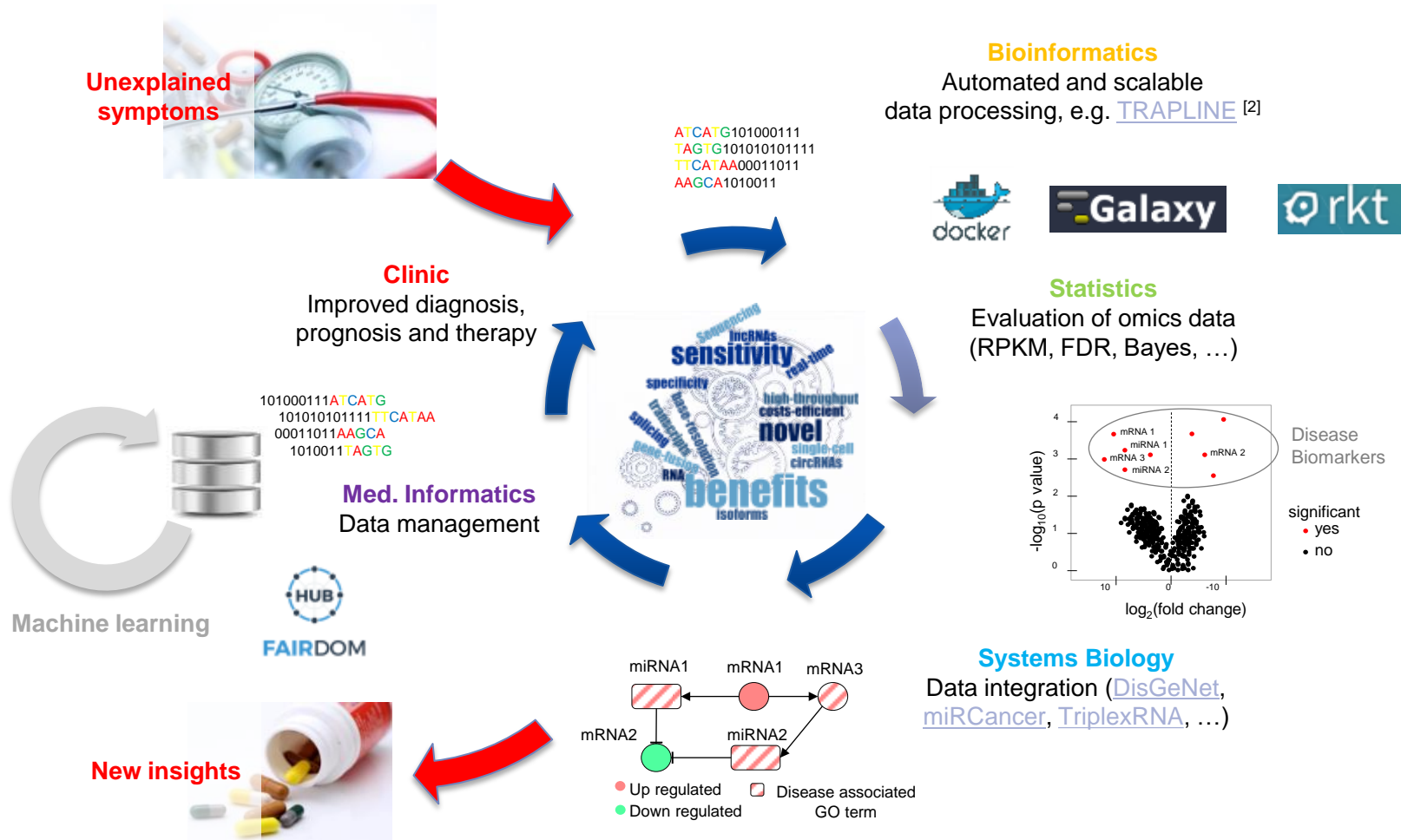
- Craig, D. W. et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. Mol. Cancer Ther. 12 , 104–116 (2013). One of the first papers investigating integration of whole-transcriptome sequencing and genome sequencing for targeted therapy selection in advanced metastatic triple-negative breast cancer

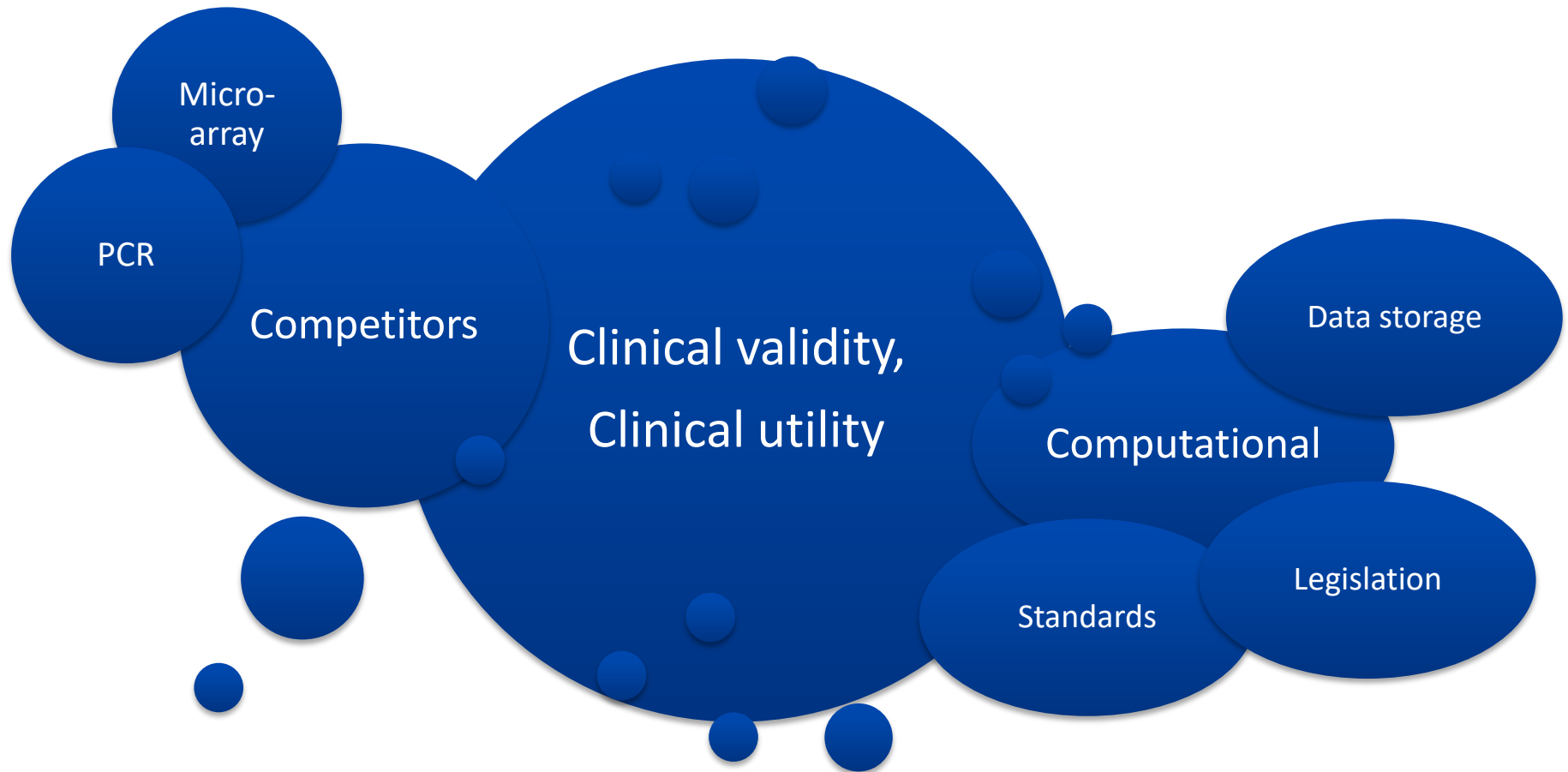
Does the patient has a high risk of cancer

Patient

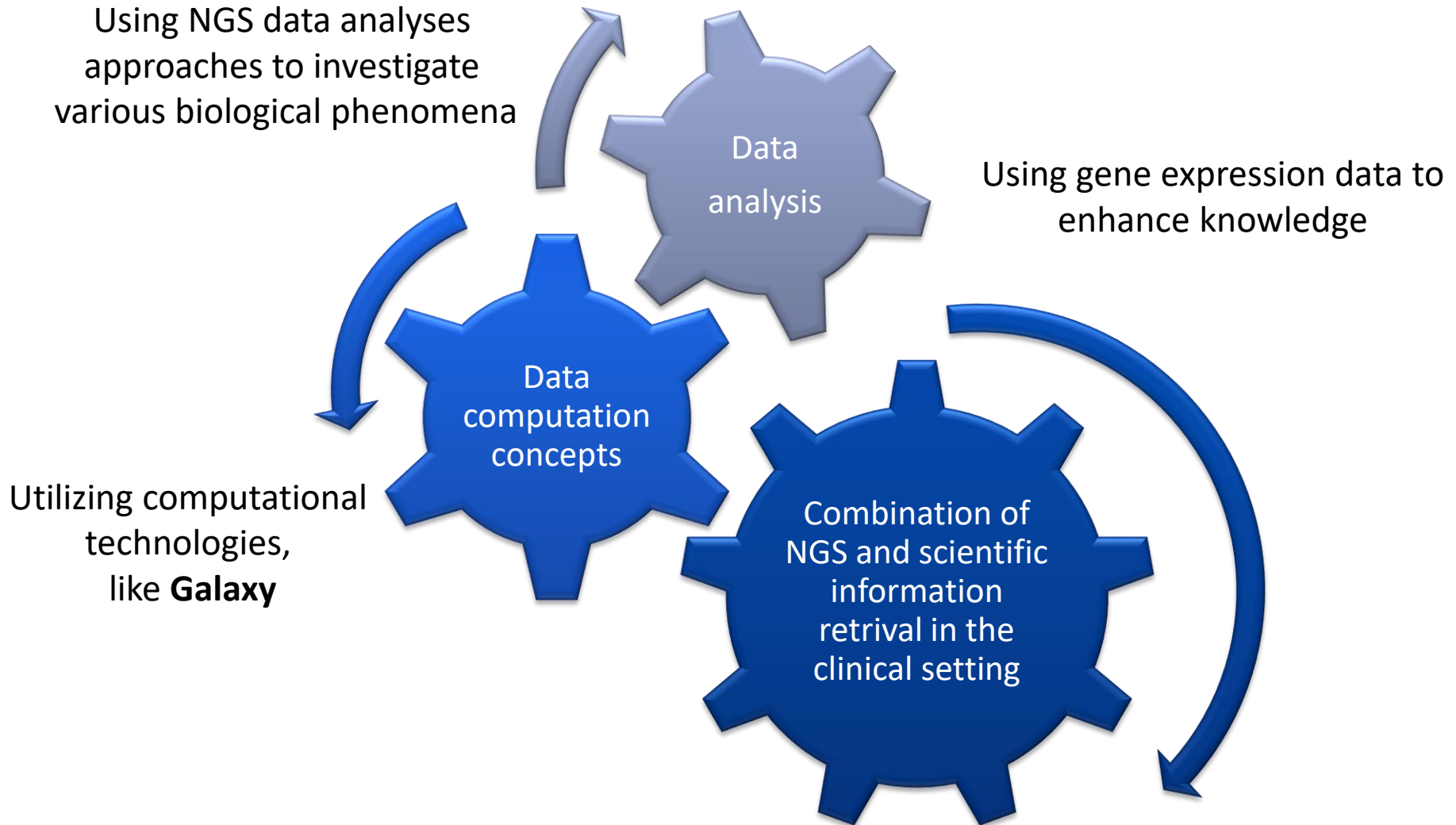


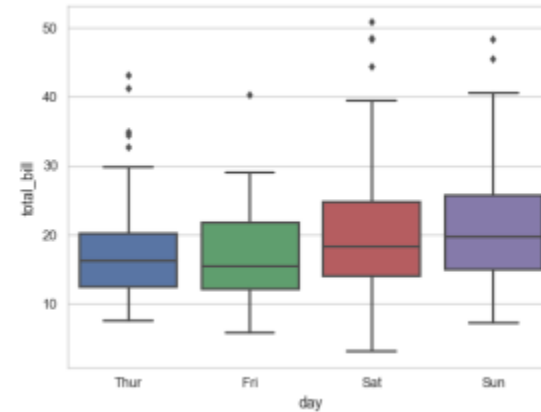
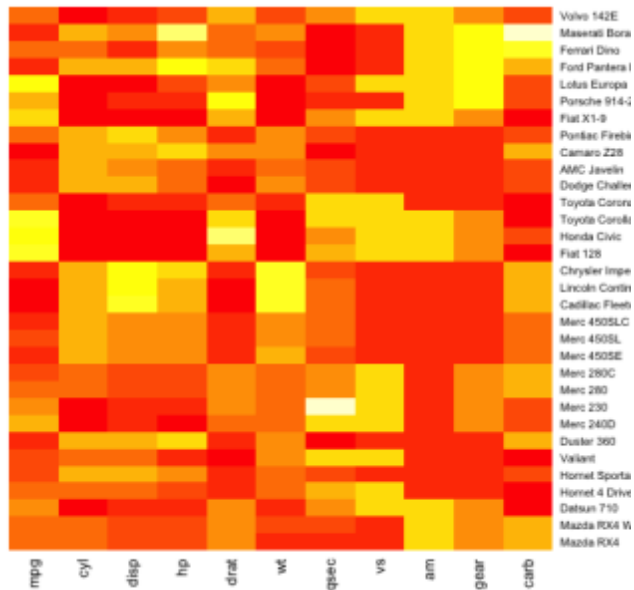
Our implementation strategy





- *“With its unprecedented ability to simultaneously detect global gene transcript levels and diverse RNA species, RNA-seq has the potential to revolutionize clinical testing for a wide range of diseases.”* Byron et al., Nat. Rev. Genet., 2016
 - *“Once the discovery phase is complete, many diagnostic tests will become targeted assays, sensitive enough to detect small numbers of rare transcripts.”*
Andersson et al., Nat. Genet., 2015
- *“Feed in latest scientific findings and analyze the same dataset over and over again [...]”*. Comment on crowdsourced research in Medicine (*Nature*)
 - *“Value of incorporating RNA sequencing (RNA-seq) with DNA sequencing to evaluate the expression of mutant alleles, to detect both known and novel gene fusions, and to detect splice variants.”* Robinson et al., Cell, 2015





Now:
Visualizations of RNA-Seq results with
Galaxy