# **Supplementary Information**

Ordino: a visual analysis tool for ranking and exploring genes, cell lines, and tissue samples

Marc Streit<sup>1,2</sup>, Samuel Gratzl<sup>1,2</sup>, Holger Stitz<sup>1</sup>, Andreas Wernitznig<sup>3</sup>, Thomas Zichner<sup>3,\*</sup> & Christian Haslinger<sup>3,\*</sup>

<sup>1</sup> Institute of Computer Graphics, Johannes Kepler University Linz, Linz, Austria.
 <sup>2</sup> datavisyn GmbH, Linz, Austria.
 <sup>3</sup> Department of Pharmacology and Translational Research, Boehringer Ingelheim RCV GmbH & Co KG, Vienna, Austria.

\* The last two authors should be regarded as joint last authors. Email: <u>marc.streit@jku.at</u>, <u>christian.haslinger@boehringer-ingelheim.com</u>

# **Table of Contents**

Supplementary Figures	2
Supplementary Tables	12
Supplementary Notes	13
Supplementary References	25

# **Supplementary Figures**



**Supplementary Figure S1. Ordino workflow.** The workflow comprises three steps. (1) Select or define a list of items consisting of genes, cell lines, or tissue samples. (2) Rank, filter, and select items. (3) Obtain detailed information for selected items.



**Supplementary Figure S2. Case study 1: Definition of gene list.** In the start menu the user can choose predefined sets or previously saved public and private sets as a starting point for the analysis. Additionally, one can upload a custom dataset or continue a previous analysis session.

Link to Ordino state shown in this figure: <u>http://vistories.org/ordino-supplementary-figure-2</u>

🎸 Or	dino									🛔 elegant_lampoi	t O Tempor	ary Session 1 🧯 🏦
#	Genes											
Showin	n 20415	of 20415 Genes: 1	selected								×	General
	9											Database Info
A F	lank :	S Symbol	Ensembl	Name	Chromosome	Biotype	Relative Copy Number	Weighted Sum	avg TPM boxplot Relative C	opy Relative Copy Number	+	Sample Overview
		- 2.12 =	12 亩 …	12 11	12日首…	↓ <u>2</u> 日 亩 …	14:日亩…	DRIVE RSA (ER TPM of HCC1954	13日首… 111首…	↓, ⊟ 🖬 …		Copy Number
								注日首··· 注日首···			1	Expression
•	C	1			lauth series	1	1	0.00 40.87 1.00 7403.97	1	1	B	Mutation
					The second se		0.43 8.00		1.00 5009.88	0.00 0.25	*	Combined View
•	1 6	ERBB2	ENSG00000141736 ENSG00000141738	v-erb-b2 erythroblastic leukemia	17	protein_coding	8.00	5324.86	37	0.25	Î	Combined view
1	3 0	GSDMA	ENSG00000167914	gasdermin A (Source:HGNC	17	protein_coding		-	•		- 9	OncoPrint
1 -	4 C	) GSDMB	ENSG0000073605	gasdermin B [Source:HGNC	17	protein_coding		-	I III III		Q	
1	5 C	) IKZF3	ENSG00000161405	IKAROS family zinc finger 3	17	protein_coding			· · · · · · · · · · · · · · · · · · ·		=	Visualization
	6 0	) LRRC3C	ENSG00000204913	leucine rich repeat containing 3C	17	protein_coding						Co-Expression
	1 0	J MIENT	ENSG00000141741	migration and invasion enhancer i	17	protein_coding				10 mm		2 3 2 20 20
1	9 0	0 OBMDL3	ENS600000171532	ORM1-like 3 (S. cerevisiae)	17	protein_coding		_				Expression vs. Copy Numbe
1 - I	10 0	PGAP3	ENSG00000161395	post-GPI attachment to proteins 3	17	protein coding			i in'			Enternal Deservoires
1	11 C	) PNMT	ENSG00000141744	phenylethanolamine N-	17	protein_coding		-	HE HE			External nesources
1	12 0	) PPP1R1B	ENSG00000131771	protein phosphatase 1, regulatory	17	protein_coding		-				canSAR
	13 C	) STARD3	ENSG00000131748	StAR-related lipid transfer (START	17	protein_coding						Encombl
4	14 C	) TCAP	ENSG00000173991	titin-cap [Source:HGNC	17	protein_coding		_				Lindemon
-	15 0	J ZPBP2	ENSG00000186075	zona pellucida binding protein 2	17	protein_coding		-				Human Protein Atlas
1	17 0	BEDG	ENSG00000107238	bromodomain containing 9	5	protein_coding			1.0	10		
1	18 0	NKD2	ENSG00000145506	naked cuticle homolog 2	5	protein coding						Open Targets
1 -	19 C	SLC12A7	ENSG00000113504	solute carrier family 12	5	protein_coding		ľ.	L H	i		
1	20 C	) SLC6A18	ENSG00000164363	solute carrier family 6, member 18	5	protein_coding			HB			PubMed
	21 [	) SLC6A19	ENSG00000174358	solute carrier family 6 (neutral	5	protein_coding			HD			UniDeat
	22 C	) TRIP13	ENSG0000071539	thyroid hormone receptor	5	protein_coding		<b>•</b> >				OniPiot
	23 C	]	ENSG00000250173	Uncharacterized protein	5	protein_coding		_				
1	24 L	J CEP72	ENSG00000112877	centrosomal protein 72kDa	5	protein_coding		_				
	25 0	) 1999	ENSG00000171308	cubain polymenzation promoting	5	protein_coding						
1	27 0	CEPTMU	ENSG00000200055	CLPTM1-like [Source:HGNC	5	protein_coding	-		1			
1 - I	28 0	SLC6A3	ENSG00000142319	solute carrier family 6	5	protein coding		-	· · · · · · · · · · · · · · · · · · ·			
1	29 0	OXB1	ENSG00000164830	oxidation resistance 1	8	protein coding		_	1 Herei			
1 - I	30 C	ANGPT1	ENSG00000154188	angiopoietin 1 [Source:HGNC	8	protein_coding	÷	-	HII-I			
1	31 C	UIMC1	ENSG0000087205	ubiquitin interaction motif	5	protein_coding		1	H			
1	32 [	) ZNF346	ENSG00000113761	zinc finger protein 346	5	protein_coding	§	-	HB			
	33 C	) ABRA	ENSG00000174429	actin-binding Rho activating	8	protein_coding		-	H <b>B</b> H			
	34 C	) BTNL9	ENSG00000165810	butyrophilin-like 9 (Source:HGNC	5	protein_coding						
4	35 C	GNB2L1	ENSG00000204628	guanine nucleotide binding protein	5	protein_coding						
	30 27 0	000201	ENSCOODD185372	offactory receptor, family 2,	0	protein_coding		-				
	38 0	TRIM41	ENSG00000182013	tripartite motif containing 41	5	protein_coding		_				
1	39 0	TBIM7	ENSG00000146054	tripartite motif containing 7	5	protein coding		_				
1 - I	40 0	TEBT	ENSG00000164362	telomerase reverse transcriptase	5	protein coding			1			
í –	41 C	C5orf38	ENSG00000186493	chromosome 5 open reading	5	protein_coding		_	i iii			
1 -	42 C	IRX2	ENSG00000170561	iroquois homeobox 2	5	protein_coding			i i i i i i i i i i i i i i i i i i i			
1	43 C	EMC2	ENSG00000104412	ER membrane protein complex	8	protein_coding		-	E HE			
- i		) FIFOF	ENICO0000104400	automotic terrelation initiation		anabala andlan					-	

**Supplementary Figure S3. Case study 1: Ranking of genes.** Gene list with additional score columns, which are calculated on the fly. The ranking shows the highest amplification in the cell line HCC1954, located on chromosome 17, affecting about 15 genes, with *ERBB2 (HER2)* having the highest expression level (orange column) and the lowest sensitivity score (green column). Therefore, it is probably the most relevant gene of this amplicon. The two aggregated score columns (in red and brown) show that *ERBB2* is amplified in almost 25% of all assessed breast cancer cell lines. Further, it is highly expressed.

Link to Ordino state shown in this figure: <u>http://vistories.org/ordino-supplementary-figure-3</u>

😩 Ordino								🛎 elegant_la	mport O Temporar	y Session 1 🚺 🕸
🖨 Genes Copy Nun	nber									
		× General	Data Source Cell Line T Data Si	btype Relative Conv. I	Number •	Filter: (1013/1013)	<ul> <li>Showing 58 of 1013 Cell Lines 2</li> </ul>	selected	×	General
		Database Info								Database Info
Rank S Symbol	Ensembl Relative C	Sample Overview	A Rank S Name	Tumor Type	Organ	Gender	ERBB2 AA Mutated of BRCA1	AA Mutated of BRCA2	+	Gene Overview
······································	拉面… 11	E Copy Number	······································	12日首…	12日首…	12日亩… 11	日首… 12日首…	12 日 首 …		Copy Number
	0.0	- Expression		-	_	1.10	9.90		+	Expression
1 🗹 ERBB2	ENSG00000141736 8.00		▼ 1  HCC1559	breast carcinoma	breast	female 8.62	Non Mutated	Mutated	<u> </u>	
2 🗆 GRB7	ENSG00000141738	Mutation	2 UACC-893	breast carcinoma	breast	female	Non Mutated	Non Mutated	8	Mutation
3 GSDMA	ENSG00000167914	Combined View	3 ₩ HCC1954	breast carcinoma	breast	female 8.00	Mutated	Non Mutated	*	Combined View
5 D IKZE3	ENSG00000161405	Complete Herr	5 D EEM-192A	breast carcinoma	breast	female	Non Mutated	Non Mutated		combined them
6 🗆 LRRC3C	ENSG00000204913	OncoPrint	6 🗆 HCC1419	breast carcinoma	breast	female	Non Mutated	Non Mutated	Q	External Resources
7 🗆 MIEN1	ENSG00000141741		7 🗆 ZR-75-30	breast carcinoma	breast	female	Non Mutated	Non Mutated	Q	COSMIC
8 D NEUROD2	ENSG00000171532	Visualization	8 🗆 SK-BR-3	breast carcinoma	breast	female	Non Mutated	Unknown		
9 🗆 ORMDL3	ENSG00000172057	Co-Expression	9 🗆 AU565	breast carcinoma	breast	female	Non Mutated	Non Mutated		
10 D PGAP3	ENSG00000161395		10 U JIMI-1	breast carcinoma	breast	temale	Unknown	Non Mutated		
12 D DDD1B1B	ENS00000131771	Expression vs. Copy Number	11 L BP474	breast carcinoma	breast	female	Non Mutated	Linknown		
13 D STARD3	ENSG00000131748		13 D HCC202	breast carcinoma	breast	female	Unknown	Non Mutated		
14 🗆 TCAP	ENSG00000173991	External Resources	14 🗆 MDA-MB-361	breast carcinoma	breast	female	Non Mutated	Mutated		
15 🖂 ZPBP2	ENSG00000186075	canSAR	15 🗆 MDA-MB-453	breast carcinoma	breast	female	Non Mutated	Non Mutated		
16 🗆 CDK12	ENSG00000167258		16 🗆 EFM-19	breast carcinoma	breast	female	Non Mutated	Non Mutated		
17 🗆 BRD9	ENSG0000028310	Ensembl	17 🗆 HCC38	breast carcinoma	breast	female	Unknown	Mutated		
18 📋 NKD2	ENSG00000145506	Human Brotoin Atlan	18 1 1-470	breast carcinoma	breast	temale	Non Mutated	Non Mutated		
20 C SLC1247	ENSC00000164252	Homan Protein Adas	20 CAU-85-1	breast carcinoma	breast	female	Mutated	Linknown		
20 D SLC6A19	ENSG00000174358	Open Targets	21 D FVSA-T	breast carcinoma	breast	female	Non Mutated	Non Mutated		
22 D TRIP13	ENSG00000071539		22 D BT-483	breast carcinoma	breast	female	Unknown	Non Mutated		
23 🗆 –	ENSG00000250173	PubMed	23 🗆 HDQ-P1	breast carcinoma	breast	female	Unknown	Unknown		
24 🗆 CEP72	ENSG00000112877		24 🗆 MDA-MB-231	breast carcinoma	breast	female	Non Mutated	Non Mutated		
25 🗆 TPPP	ENSG00000171368	UniProt	25 🗆 MDA-MB-175-VII	breast carcinoma	breast	female	Non Mutated	Non Mutated		
26 🗆 -	ENSG0000268885		26 🗆 CAMA-1	breast carcinoma	breast	female	Non Mutated	Non Mutated		
27 D CLPIMIL	ENS00000049856		27 L HS 739.1	breast carcinoma	breast	female	New Michael	Mutated		
28 D 0Y81	ENSC00000164830		28 D 6/051	breast carcinoma	breast	female	Indi Mutateu	Non Mutated		
30 ANGPT1	ENSG00000154188		30 C HCC1599	breast carcinoma	breast	female	Unknown	Mutated		
31 🗆 UIMC1	ENSG0000087206		31 🗆 Hs 281 T	breast carcinoma	breast	female	Unknown	Non Mutated		
32 🔲 ZNF346	ENSG00000113761		32 🗆 CAL-148	breast carcinoma	breast	female	Non Mutated	Non Mutated		
33 🗆 ABRA	ENSG00000174429		33 🗆 Hs 274.T	breast carcinoma	breast	female	Unknown	Non Mutated		
34 🗆 BTNL9	ENSG00000165810		34 🗆 HCC1937	breast carcinoma	breast	female	Mutated	Unknown		
35 GNB2L1	ENSG00000204628		35 🔲 BT-20	breast carcinoma	breast	temale	Non Mutated	Mutated		
30 D 0H2V1	ENSC00000185372		30 DUA/MB-415	breast carcinoma	breast	female	Non Mutated	Non Mutated		
38 D TRIM41	ENSG00000146063		38 D Hs 742 T	breast carcinoma	breast	female	Unknown	Unknown		
39 🗆 TRIM7	ENSG00000146054		39 🗆 CAL-120	breast carcinoma	breast	female	Non Mutated	Non Mutated		
40 🔲 TERT	ENSG00000164362		40 🗆 Hs 606.T	breast carcinoma	breast	female	Unknown	Non Mutated		
41 🗆 C5orf38	ENSG00000186493		41 🗆 MDA-MB-157	breast carcinoma	breast	female	Non Mutated	Non Mutated		
42 🗆 IRX2	ENSG00000170561		42 🗆 Hs 343.T	breast carcinoma	breast	female	Unknown	Non Mutated		
43   EMC2	ENSG00000104412		43 🗆 HCC1806	breast carcinoma	breast	female	Non Mutated	Non Mutated		
44 U EIF3E	ENS00000105002		44 LI HS 5781	breast carcinoma	breast	female	Non Mutated	Non Mutated		
	ENSCO0000107249	-	45 C RT649	breast carcinoma	breast	female	Non Mutated	Non Mutated		
4	L14000000101249		47 0 0001107	brown carcinoma	hannal	Comple		Non Montel	-	

**Supplementary Figure S4. Case study 1: Cell lines ranked by copy number data of ERBB2.** The user selects the gene *ERBB2* and opens the Copy Number detail view as new focus view. The gene list remains open on the left as context view. The Copy Number view shows that HCC1954 has the highest *ERBB2* amplification among *BRCA1* mutated cell lines and that HCC1569 has the highest *ERBB2* amplification among *BRCA2* mutated cell lines.

Link to Ordino state shown in this figure: http://vistories.org/ordino-supplementary-figure-4

Ordino							🛎 elegant_lamport	O Temporary Session 0
Genes PubMed								
		×	General -	The web page below is directly load	ied from 🗗//www	wincbi, nlm, nih, gov/pubmed?term=ERB82		
Rank S Symbol	Ensembl	Relati	Sample overview	S NCBI Resources	🗹 How To 🗹			Sign in to NCBI
···· 2.12	12 亩 …	14° E	Copy Number	Pub Med gov	PubMed	<ul> <li>FRBB2</li> </ul>	Search	
U		0.43	Expression	US National Library of Medicine National Institutes of Health		Create RSS Create alert Advanced		Help
1 🗹 ERBB2	ENSG00000141736	8 00 ^		Article types	Farmer	h Summary - Fast hus Mast Descent - Persona - 20	Cittan Marine Citta	
2 🗆 GRB7 3 🗆 GSDMA	ENSG00000141738 ENSG00000167914		Mutation	Clinical Trial	Forma	t Summary + Sort by, Most Recent + Per page: 20 + Send to	<ul> <li>Filters: <u>manage Filters</u></li> </ul>	
4 GSDMB	ENSG0000073605		Combined View	Review Customize	See	2555 articles about ERBB2 gene function	Results by year	
6 LRRC3C	ENSG00000161405 ENSG00000204913		OncoPrint	Text availability	See	also: ERBB2 erb-b2 receptor tyrosine kinase 2 in the Gene database		
7 🗆 MIEN1	ENSG00000141741			Abstract	See	also: 42 tests for ERBB2 in the Genetic Testing Registry		
8 D NEUROD2	ENSG00000171532 ENSG00000172057		Visualization	Free full text				
10 D PGAP3	ENSG00000161395		Co-Expression	r un text	Searc	ch results	4	Download CSV
11 D PNMT	ENSG00000141744		Expression vs. Copy Number	PubMed Commons	Items:	1 to 20 of 26049 << End < Prev Page 1 of 1303 Next > Last		
12 D PPP1R1B 13 D STARD3	ENSG00000131771 ENSG00000131748			Reader comments		THE PART OF THE PART OF THE PART	Delated second	
14 🗆 TCAP	ENSG00000173991		External resources	Trending articles	E m	B-1273g-3p promotes proliferation, migration and invasion of LoVo cells via cannabingid	related searches	
15 C ZPBP2	ENSG00000186075		canSAR	Publication dates	1. re	ceptor 1 through activation of ERBB4/PIK3R3/mTOR/S6K2 signaling pathway,	crows prease carloer	
17 BRD9	ENSG00000028310		Ensembl	10 years	Li	M, Qian X, Zhu M, Li A, Fang M, Zhu Y, Zhang J.		
18 🗆 NKD2	ENSG00000145506			Custom range	Mo	ol Med Rep. 2018 Jan 9. doi: 10.3892/mmr.2018.8397. [Epub ahead of print] diD: 29328379	PMC Images search	for ERBB2
19 U SLC12A7 20 C SLC6A18	ENSG00000113504 ENSG00000164363		Human Protein Atlas	Species	Sit	milar articles		
21 🔲 SLC6A19	ENSG00000174358		Open Targets	Humans Other Animals		amoning Magadiumst Mak appliturel or Depliturel Bath Followed by Arthurst Mark	-781344 -87.98.98	
22 🗋 TRIP13	ENSG00000071539		DubMed	Concervalinides	2. W	omparing Neoagovant Nao-paonaxer vs Haunaxer Both Followed by Anthracycline Regimens in fomen With ERBB2/HER2-Negative Breast Cancer-the Evaluating Treatment With Neoadiuvant	-367- JAG= 1	(T) L. T)
23 U	ENSG00000250173 ENSG00000112877		ruomed	Clear all	A	braxane (ETNA) Trial: A Randomized Phase 3 Clinical Trial,	-	
25 🗆 TPPP	ENSG00000171368		UniProt	Show additional filters	Gi	ianni L, Mansutti M, Anton A, Calvo L, Bisagni G, Bermejo B, Semiglazov V, Thill M, Chacon JI, Ch	an Emilia E	1.0. mm.
26 🗆 —	ENSG00000268885 ENSC00000049656				A,	, Morales S, Alvarez I, Plazaola A, Zambetti M, Redfern AD, Dittrich C, Dent RA, Magazzù D, De ato R, Valamiera P, Tuernate I		A
28 🗆 SLC6A3	ENSG00000142319				JA	MA Oncol. 2018 Jan 11. doi: 10.1001/jamaoncol.2017.4612. [Epub ahead of print]	1	2 2
29 🗆 OXR1	ENSG00000164830				PN	MID: 29327055		
30 ANGPT1 31 UMC1	ENSG00000154188 ENSG00000087206				50	11101_0105402		See more (4550)
32 🗆 ZNF346	ENSG00000113761				🗆 In	sertional mutagenesis in a HER2-positive breast cancer model reveals ERAS as a driver of		oce more (+669)
33 🗆 ABRA	ENSG00000174429				3. Ca	ancer and therapy resistance.		
34  BINL9 35  GNB2L1	ENSG00000165810 ENSG00000204628				lki Hi	Ink GJ, Boer M, Bakker ERM, Vendel-Zwaagstra A, Klijn C, Ten Hoeve J, Jonkers J, Wessels LF, Ikens I	Titles with your sear	ch terms 👘
36 🗆 OR2V1	ENSG00000185372				Or	rcogene, 2018 Jan 12, doi: 10.1038/s41388-017-0031-0. [Epub ahead of print]	ERBB2 mutation: A pro squamous cervical can	mising target in non- ce [Gynecol Oncol. 2017]
37 OR2V2	ENSG00000182613				Ph	MID: 29326437	In vivo evidence sunnor	ting a metastasis
30 TRIM41 39 TRIM7	ENSG0000146063 ENSG00000146054				St		suppress [Genes Chro	mosomes Cancer. 2017]
40 🗆 TERT	ENSG00000164362					Inical significance of targeting drug-based molecular biomarkers expression in ovarian clear ce	Designing a cancer the	apeutic peptide by
41 C5orf38	ENSG00000186493				4. <u>ca</u>	ircinoma]_	combining the mitochor	drial ta [FEBS Lett. 2018]
43 🗌 EMC2	ENSG00000170361				Li	MJ, Li HR, Cheng X, Bi R, Tu XY, Liu F, Chen LH. Internal El Chan Ke Ze Zhi 2017 Dec 25-52(12):835-843 doi: 10.3750/cma.i.issn.0520-557-2017.12.008. Chinese		See more
44 🗆 EIF3E	ENSG00000104408				PN	AllD: 29325268	50	
45 CDC37L1	ENSG00000106993				Sit	milar articles	Find related data	
40 U 0000	CH3300000107249					as Via and Vana of VV1 in Tumor Crowth and Suppression	rind related data	

**Supplementary Figure S5. Case study 1: PubMed information for ERBB2.** Users can obtain further information for selected items in detail views that load the content from external websites, such as PubMed and Open Targets.

Link to Ordino state shown in this figure: <u>http://vistories.org/ordino-supplementary-figure-5</u>

*	Tissues						
Showing	152 of 11253 Tissues						×
A Ra	nk S Name	Tumor Type	Organ	Gender	AA Mutated of TP53	avg TPM	+
	• … 212 首 …	12日首…	12日首…	拉日首…	↓ <u></u> 日 首 …	112日首…	
•			_				+
			-	-	-	0.11 105.14	1.
-	1 D TCGA-AA-A029-01	colon adenocarcinoma	colon	male	Non Mutated		î D
	2 D TCGA-AA-3590-01	colon adenocarcinoma	colon	female	Non Mutated		
	A  TCCA-AA-2921-01	colon adenocarcinoma	colon	male	Non Mutated		*
	5 D TCCA-46-2809-01	colon adenocarcinoma	colon	male	Non Mutated		
	6 □ TCGA-AA-3516-01	colon adenocarcinoma	colon	female	Non Mutated		Q
1	7 D TCGA-AA-A024-01	colon adenocarcinoma	colon	male	Non Mutated		Q
	8 T TCGA-AA-3818-01	colon adenocarcinoma	colon	female	Non Mutated		1.00
1 i	9 D TCGA-AA-A02Y-01	colon adenocarcinoma	colon	male	Non Mutated		
- i	10 🔲 TCGA-AA-3989-01	colon adenocarcinoma	colon	male	Non Mutated		
1 i	11 D TCGA-AA-3530-01	colon adenocarcinoma	colon	male	Non Mutated		
1 I	12 🗆 TCGA-AA-3845-01	colon adenocarcinoma	colon	female	Non Mutated		
1	13 🗆 TCGA-AA-3821-01	colon adenocarcinoma	colon	female	Non Mutated		
1	14 🗆 TCGA-AA-3939-01	colon adenocarcinoma	colon	male	Non Mutated		
	15 🗆 TCGA-AA-A00K-01	colon adenocarcinoma	colon	male	Non Mutated		
	16 🗆 TCGA-AA-3850-01	colon adenocarcinoma	colon	male	Non Mutated		
	17 D TCGA-AA-A01V-01	colon adenocarcinoma	colon	male	Non Mutated		
	18 🗆 TCGA-AA-3555-01	colon adenocarcinoma	colon	female	Non Mutated		
	19 🗆 TCGA-AA-3522-01	colon adenocarcinoma	colon	male	Non Mutated		
	20 🔲 TCGA-AA-3680-01	colon adenocarcinoma	colon	female	Non Mutated		
	21 U TCGA-AA-3986-01	colon adenocarcinoma	colon	male	Non Mutated		
	22 TCGA-AA-A00Q-01	colon adenocarcinoma	colon	temale	Non Mutated		
	23 D TCCA-AA-3819-01	colon adenocarcinoma	colon	remaie	Non Mutated		
	24 D TCGA-04-3715-01	colon adenocarcinoma	colon	famala	Non Mutated		
	25 D TCCA-04-3529-01	colon adenocarcinoma	colon	female	Non Mutated		
	27 D TCGA-46-2672-01	colon adenocarcinoma	colon	female	Non Mutated		
	28 T TCGA-AA-3678-01	colon adenocarcinoma	colon	female	Non Mutated		
1	29 D TCGA-AA-3684-01	colon adenocarcinoma	colon	female	Non Mutated		
1	30 🗆 TCGA-AA-3855-01	colon adenocarcinoma	colon	male	Non Mutated		
- i	31 D TCGA-AA-3833-01	colon adenocarcinoma	colon	female	Non Mutated		
	32 🔲 TCGA-AA-3864-01	colon adenocarcinoma	colon	male	Non Mutated		
	33 🔲 TCGA-AA-3994-01	colon adenocarcinoma	colon	male	Non Mutated		
1	34 🗆 TCGA-AA-3514-01	colon adenocarcinoma	colon	female	Non Mutated		
1	35 🗆 TCGA-AA-3554-01	colon adenocarcinoma	colon	female	Non Mutated		
	36 🗆 TCGA-AA-3984-01	colon adenocarcinoma	colon	female	Non Mutated		
	37 🔲 TCGA-AA-A00E-01	colon adenocarcinoma	colon	male	Mutated		
	3B 📋 TCGA-AA-3B37-01	colon adenocarcinoma	colon	male	Non Mutated		
	39 U TCGA-AA-A01X-01	colon adenocarcinoma	colon	temale	Non Mutated		
	40 U TCGA-AA-A011-01	colon adenocarcinoma	colon	maie	Non Mutated		
	41 D TUGA-AA-3947-01	colon adenocarcinoma	colon	temale	Non Mutated		
	42 D TUGRAAA020-01	colon adenocarcinoma	colon	male	Non Mutated		
	44 D TCCA-AA-2710-01	colon adenocarcinoma	colon	famala	Non Mutated		
	44 C TCGRAMSTICUT	colon adenocarcinoma	colon	female	Non Mutated		
	45 D TCCA.AA.A00N-01	colon adenocarcinoma	colon	male	Non Mutated		
	40 0 10003003004001400	Colon agenocal cinoma		mend	- mon mutated		

**Supplementary Figure S6. Case study 2: TCGA tumor samples ranked by average expression.** The list of TCGA tumor samples shows a clear correlation between the gene expression signature (orange bars) and the mutation status of *TP53*: Of the 50 samples with the highest expression only 3 are *TP53* mutated, whereas of the 50 samples with the lowest expression 35 are TP53 mutated (not visible in this figure).

Link to Ordino state shown in this figure: <u>http://vistories.org/ordino-supplementary-figure-6</u>

< Ordino							🛎 elegant_lamport 🛛 O Temporary Session 1 🛛 i	ŵ
Cell lines								C
Showing 326 of 1013 Cell Lines								×
A Rank S Name	Tumor Type	Organ	Gender	AA Mutated of TP53	AA Mutation of TP53	Weighted Sum		+
···· ··· 2.1% 🗇 ···	拉日首…	1:日前…	14日前…	↓ <u>2</u> 日 亩 …	12 亩 …	14天 白 …		
						TP53 Predictor Score DRIVE RSA (ER Mc		+
								±.
•	ashe at beau	and a large				1294 20910 001 714		
		- Add to the last						±.
▼[ 1 [] JL-1	mesothelioma	lung	male	Non Mutated	wt			
2 🗆 SJSA-1	bone sarcoma	bone	male	Non Mutated	wt			Q
3 📋 UACC-62	melanoma	skin	-	Non Mutated	wt			Q
1 4 U WM-115 5 □ SNGM	interus carcinoma	sign	female	Non Mutated	wt			-
6 C UACC 257	melanoma	ekin	iemaie	Non Mutated	wit			
7 0 632	melanoma	skin	male	Non Mutated	wt			
8 🗆 WM1799	melanoma	skin	-	Non Mutated	wt			
9 🖸 HT-1197	bladder carcinoma	bladder	male	Non Mutated	wt			
10 D HT-1080	sarcoma/soft tissue	soft tissue	male	Non Mutated	wt			
11 🗆 MEL-HO	melanoma	skin	female	Non Mutated	wt			
12 🗆 0S-RC-2	-	kidney	male	Non Mutated	wt			
13 🗆 KS-1	astrocytoma/glioblasto	brain	female	Non Mutated	wt			
14 🗆 Hs 939.T	melanoma	skin	female	Non Mutated	wt			
15 O TUHRATKB	renal carcinoma	kidney	-	Non Mutated	wt			
10 U HS 852.1	-	skin	male	Non Mutated	wt			
18 C) C010-678	colon carcinoma	colon	male	Non Mutated	wit			
19 C Caki-2	renal carcinoma	kidney	male	Non Mutated	wt			
20 COL0 829	melanoma	skin	male	Non Mutated	wt			
21  LNCap.FGC	prostate carcinoma	prostate	male	Non Mutated	wt			
22 🗆 CAL-54	renal carcinoma	kidney	male	Non Mutated	wt			
23 🗆 Caki-1	renal carcinoma	kidney	male	Non Mutated	wt			
24 🗆 K029AX	melanoma	skin	-	Non Mutated	wt			
25 🖸 SK-MEL-5	melanoma	skin	female	Non Mutated	wt			
26 SK-HEP-1	liver carcinoma	liver	male	Non Mutated	wt			
27 U COR-LIUS	NSCLU	lung	male	Non Mutated	wt			
28 O ACHIN	melanoma	clain	famala	Non Mutated	wi			
30 0 465	gastric carcinoma	stomach	female	Non Mutated	wt			
31 C CW-2	colon carcinoma	colon	female	Non Mutated	wt			
32 🗆 JHUEM-2	uterus carcinoma	uterus	female	Non Mutated	wt			
33 🗆 U-2 OS	bone sarcoma	bone	female	Non Mutated	wt			
34 🖸 MEL-JUSO	melanoma	skin	female	Non Mutated	wt			
35 🖸 SNU-719	gastric carcinoma	stomach	male	Non Mutated	wt			
36 🗆 COLO-679	melanoma	skin	female	Non Mutated	wt			
37 🖸 Panc 05.04	pancreas carcinoma	pancreas	female	Non Mutated	wt			
38 U SW 1573	NSCLU	lung	temale	Non Mutated	wt			
39 U WM-793	melanoma	skin	mala	Non Mutated	wit			
40 C LOAMVI	melanoma	skin	male	Non Mutated	wt			
42 □ 18513	colon carcinoma	colon	male	Non Mutated	wt			
43 🗆 769-P	renal carcinoma	kidney	female	Non Mutated	wt			
A4 () 11au 40	aunian anniaana		- famala	alan Mutatad			•	

**Supplementary Figure S7. Case study 2: Cell lines ranked by weighted sum of two scores.** The analyst has added mutation and score columns to the list of cell lines. The combined score of DRIVE RSA values for the gene *MDM2* (red bars) and the gene expression signature (*TP53* Predictor Score; purple bars) correlates with the *TP53* mutation status (column AA Mutated).

Link to Ordino state shown in this figure: http://vistories.org/ordino-supplementary-figure-7

Cell lines								
								General
ng 48 of 1013 Cell Lines; 2 sele	ected						<u></u>	Database Info
Rank S Name	Tumor Type	Organ	Gender	AA Mutated of TP53	AA Mutation of TP53	Weighted Sum	+	Gene Overview
··· ·· 2.1%	12日首…	12日首…	拉日首…	12日首…	12 亩 …	TP53 Predictor Score DRIVE BSA (ER Mo		Copy Number
							<b>T</b>	Expression
	1			-				Mutation
0						12.94 209.10 -0.01 -7.14	*	Midtation
1 🐼 MDST8	colon carcinoma	colon	-	Unknown	_	170 08 397	<u></u>	Combined View
2 🗋 OS-RC-2	-	kidney	male	Non Mutated	wt		Q	
3 🗋 Hs 852.T	_	skin	male	Non Mutated	wt			External Resources
4 🗌 COLO-678	colon carcinoma	colon	male	Non Mutated	wt		Q	COSMIC
5 CW-2	colon carcinoma	colon	female	Non Mutated	wt			
6 🗋 LS513	colon carcinoma	colon	male	Non Mutated	wt			
7 U SN0-1079	-	billary tract	-	Non Mutated	wt			
8 🗋 SW48	colon carcinoma	colon	temale	Non Mutated	WT COMES			
9 🗋 SNU-04	colon carcinoma	colon	male	Mutated	p.62455			
11 () 86001	colon carcinoma	colon	male	Non Mutated	wit			
12 1 15 180	colon carcinoma	colon	female	Non Mutated	wit			
13 C HCT 116	colon carcinoma	colon	male	Non Mutated	wt			
14 C BK0	colon carcinoma	colon	-	Non Mutated	wt			
15 C SNU-81	colon carcinoma	colon	male	Mutated	n B213* n K132T			
16 C SNU-407	colon carcinoma	colon	male	Mutated	p A88fs p P89fs			
17 🗆 LS411N	colon carcinoma	colon	male	Mutated	p.Y126*			
18 🗋 CL-34	colon carcinoma	colon	female	Mutated	p.S127P.p.K382fs			
19 🖸 CCK-81	colon carcinoma	colon	-	Mutated	p.P278H			
20 🖸 CL-11	colon carcinoma	colon	-	Mutated	p.S215N			
21 🗍 JHH-6	-	liver		Mutated	p.336_340del			
22 🗍 KM12	colon carcinoma	colon		Mutated	p.A70fs;p.P72fs;p.R72			
23 🗋 HT115	colon carcinoma	colon		Mutated	p.R213*			
24 🗆 SW1417	colon carcinoma	colon	female	Mutated	p.238_242del;p.C238f			
25 🗋 HCT-15	colon carcinoma	colon	male	Mutated	p.S241F.p.?			
26 🗋 NCI-H747	colon carcinoma	colon	male	Mutated	p.R158L			
27 C COLO 201	colon carcinoma	colon	male	Mutated	p. Y 103F			
28 [] HT-29	colon carcinoma	colon	temale	Non Mutated	WI			
29 📋 SW403	colon carcinoma	colon	temale	Mutated	p.E51*			
30 U SW948	colon carcinoma	colon	temale	Mutated	p.011/15 + D019L + D019-1			
31 U H155	colon carcinoma	colon	mala	Mutated	p.nz13L[p.HZ13>L + D200P			
33 () 0010 741	colon carcinoma	colon	famala	Mutated	o K321fe o D322fe o K			
34 () 0010-320	colon carcinoma	colon	female	Mutotad	p.102110,p.F02210,p.K			
35 C HGC-27	-	stomach	-	Mutated	n T150fs n P151fe n P			
36 C NCI-H716	colon carcinoma	colon	male	Mutated	p.E224D			
37 🗆 SNU-738	-	brain	male	Mutated	p.0192*			
38 🗋 SNU-685	-	uterus	female	Mutated	p.?			
39 🐼 NCI-H508	colon carcinoma	colon	male	Unknown				
40 🗆 SNU-C2A	colon carcinoma	colon	female	Mutated	p.R273C			
41 🗋 HuH-6	-	liver	-	Mutated	p.N239D;p.A159D			
42 🖸 SW1463	colon carcinoma	colon	female	Mutated	p.R248Q			
42 CT CW/490	colon carcinoma	colon	male	Mutated	0.02000			

**Supplementary Figure S8. Case study 2: Filtering list of cell lines ranked by weighted sum of two scores.** Filtering the list of cell lines for tumor type *colon carcinoma* reveals that the cell lines MDST8 and NCI-H508 lack *TP53* mutation status, but gene expression and DRIVE data are available. MDST8 has a very high combined score and is therefore unlikely to be *TP53* mutated. NCI-H508, however, has a very low combined score and is therefore probably *TP53* mutated.

Link to Ordino state shown in this figure: <u>http://vistories.org/ordino-supplementary-figure-8</u>

💲 Ordino										& elegant_lamport	O Temporary Session 1	i	ŵ
H Cell lines Exp	ression												6
	×	General		(		10 (10 0							×
		Database Info	Data Subtype	TPM •	Filter: 137 55532	Snowing 13 of 13 Genes							
Rank S Name	Weighted Sum	Gene Overview	A Rank	S Symbol	Ensembl	Name	Chromosome	Biotype	NCI-H508	MDST8			÷
···· ··· 2.1% ····	TP53 Predictor Score DRIVE RSA	Copy Number	-	···· 212 ····	12 🗰 🚥	12 11 ***		12 日 首 …	1;日日…	11 H m			+
	148 <b>0</b> ~ 148	Expression			C100000000000000				6.97 105.13	9.06 633.72			1.
0	12.94 209.10 -0.01	Mutation	2	CDKN1A	ENSG00000124762	cyclin-dependent kinase inhibitor 1A (p21,	6	protein_coding protein_coding					
1 🐼 MDST8	170.08 3.97	Combined View	3	RPS27L     TNFRSF10B	ENSG00000185088 ENSG00000120889	ribosomal protein S27-like [Source:HGNC tumor necrosis factor receptor	15 8	protein_coding protein_coding					*
2 C 0S-RC-2		External Resources	5	DDB2	ENSG00000134574 ENSG00000181026	damage-specific DNA binding protein 2, apoptosis enhancing nuclease	11	protein_coding protein_coding					Q
4 🗆 COLO-678		COSMIC	7	CCNG1	ENSG00000113328	cyclin G1 [Source:HGNC ferredenin reductees  Pennes HCNC	5	protein_coding					Q
5 🖸 CW-2 6 🗋 LS513			9	MDM2	ENSG00000135679	Mdm2, p53 E3 ubiquitin protein ligase	12	protein_coding	i (				
7 🖸 SNU-1079			10	XPC     RBM2B	ENSG00000154767 ENSG00000048392	xeroderma pigmentosum, ribonucleotide reductase M2 B (TP53	3	protein_coding protein_coding	-				
9 🖸 SNU-C4			12	ZMAT3	ENSG00000172667	zinc finger, matrin-type 3 [Source:HGNC	3	protein_coding	_				
10 C LoVo			L 13	L SESNI	ENSG0000080546	sestrin 1 [Source:HGNC	6	protein_coding					
12 🗆 LS 180													
13 🗌 HCI 116 14 🗋 RKO													
15 C SNU-81													
16 🗆 SNU-407 17 🗆 LS411N													
18 🗆 CL-34													
19 🖸 CCK-81 20 🗍 CL-11													
21 🖸 JHH-6													
22 C KM12													
23 C H1115 24 C SW1417													
25 🗋 HCT-15													
26 🗌 NCI-H747													
27 C COLO 201													
29 🖸 SW403													
30 🗋 SW948													
31 C HT55 32 C SW620													
33 C COLO 741													
34 🗌 COLO-320													
35 🗋 HGC-27													
30 O NCI-H/16 37 O SNU-738													
38 🖸 SNU-685													
39 🕑 NCI-H508	20												
40 SNU-C2A													
41 U HuH-6 42 C SW1463													
42 C CWARD													

**Supplementary Figure S9. Case study 2: Gene expression detail view showing expression of two genes.** Opening the Gene Expression detail view shows that the high gene signature expression in the cell line MDST8 is caused mainly by expression of the genes *BAX, CDKN1A,* and *RPS27L*.

Link to Ordino state shown in this figure: http://vistories.org/ordino-supplementary-figure-9



**Supplementary Figure S10. Case study 2: COSMIC detail view.** The analyst can browse the information available for selected cell lines in a detail view, in this case the corresponding COSMIC page for MDST8.

Link to Ordino state shown in this figure: <u>http://vistories.org/ordino-supplementary-figure-10</u>

# **Supplementary Tables**

**Supplementary Table S1.** Data available for cell lines and tissue samples, with mRNA expression, copy number, and mutation status data as well as number of available genes on canonical chromosomes. The gene annotation is based on Ensembl Version 70.

Dataset	Assay	Samples	Genes	Туре
tissues		12,334		
tissues	mRNA-seq expression	10,623	19,959	continuous
tissues	SNP6-array based copy number data	12,035	56,595	categorical and continuous
tissues	DNA sequencing based mutation data	6,795	20,136	categorical
cell lines		1,013		
cell lines	mRNA-seq expression	932	56,495	continuous
cell lines	SNP6-array based copy number data	974	56,595	categorical and continuous
cell lines	DNA sequencing based mutation data	810	up to 20,197	categorical and continuous
cell lines	RNAi depletion screen data (Project DRIVE [4])	385	7,267	continuous
cell lines	CRISPR-Cas9 depletion screen data (Avana CERES [5])	330	17,586	continuous

genes

56,632

# **Supplementary Notes**

# Background

To address the task of prioritizing a collection of items based on a rich set of experimental data and metadata, domain experts use either programming languages, such as R and Python, or tools, such as Microsoft Excel, Spotfire, and Tableau.

Programming languages provide the full flexibility needed for analysis. However, they require substantial programming knowledge, and data integration and preparation are time-consuming; even minor changes in the analysis can lead to major changes in the code. Furthermore, the output of scripts based on languages such as R and Python is usually static, which severely limits the exploratory nature of the analysis.

Powerful general-purpose tools such as Tableau and Spotfire can be too difficult to use for a non-expert and have limitations with respect to the aggregation and visualization of genomics data.

Spreadsheet tools, such as Microsoft Excel and Google Spreadsheets, are not designed for handling larger genomics datasets. Further, neither of these allows detailed information about individual cell lines or genes to be retrieved in an interactive manner.

# Visual Analysis Workflow

The Ordino analysis workflow consists of three steps, as outlined in **Supplementary Fig. S1**.

# Step 1: Define List of Items

The user starts the analysis by defining a set of items. The item set can be determined by manually entering a list of identifiers (e.g., a list of gene symbols), by selecting a previously saved or predefined list of items, or by uploading a comma-separated file (**Supplementary Fig. S2**).

#### Step 2: Rank, Filter, and Select Items

A core component of the Ordino system is the interactive visualization technique LineUp (http://lineup.caleydo.org) [1], which allows users to flexibly create and explore rankings of items based on a set of heterogeneous attributes. The exploration is supplemented with filtering features, such as setting cutoff values for numerical attributes, specifying a string or regular expression for textual columns, and specifying one or more categories in categorical attributes. In addition, users can change the visual representation of columns on demand. Numerical attributes, for instance, can be visualized using bars, varying brightness, or as circles whose sizes are proportional to the data values.

As a starting point, Ordino presents the list of items selected in Step 1 as a table containing metadata attributes specific to the item type. For genes, the default columns are gene symbol, Ensembl ID, chromosome, and biotype (**Supplementary Figure S3**). For cell lines and tissue samples, the default columns are name, tumor type, organ, and gender (**Supplementary Figure S6**). Initially, gene lists are sorted alphabetically by gene symbol, and lists of cell lines and tissue samples by their name. Further columns can be added by clicking on the plus icon shown on the right-hand side of the interface (**Supplementary Fig. S3**).

Ordino supports the following column types:

- **Database columns** contain metadata about genes (such as biotype, chromosome, Ensembl ID, name, sequence region start & end, strand, and gene symbol), cell lines (age at surgery, gender, growth type, histology type, metastatic site, morphology, name, organ, and tumor type), and tissue samples (age, body mass index (BMI), days to death, days to last follow up, ethnicity, gender, height, name, organ, race, tumor type, tumor type adjacent, vendor name, vital status, and weight).
- **Parameterized Score columns.** Depending on the item type of the main table, users can add single score columns by specifying a single item (gene, cell line, or tissue sample) together with the data attribute of interest (e.g., expression, copy number, or mutation). The following single scores are available:
  - For **genes**: single tissue-sample score, single cell-line score, and single depletion-screen score.
  - For **cell lines**: single gene score and single depletion-screen score.
  - For **tissue samples**: single gene score.

In addition to single score columns, the values of which are loaded directly from the the Ordino database, users can define aggregations of multiple items that are calculated on the fly by the Ordino server. To define an aggregation, users must specify (1) the set of items on whose basis the aggregation will be calculated (either by selecting previously stored named sets, by entering lists of items, or by selecting categorical attributes, such as the tumor type of cell lines), (2) the data type (expression, copy number, mutation, and depletion screen), and (3) the aggregation function (average, median, min, max, box plot, frequency, and count). The following aggregated scores are available:

- For **genes**: aggregated tissue-sample score, aggregated cell-line score, and aggregated depletion-screen score.
- For **cell lines**: aggregated gene score and aggregated depletion-screen score.
- For **tissues samples**: aggregated gene score.
- Combining columns allow users to combine the content of multiple columns by dragging the header of single columns onto the combined column header. Users can create weighted sum columns, which are visualized as stacked bars highlighting the contribution of individual attributes to the total score, min/mean/max combination columns, which show only the minimum, mean or maximum of all combined columns, scripted columns, for which users can define how individual columns are to be combined using JavaScript, nested columns for semantically grouping multiple columns, and imposition columns, which color numerical columns by a categorical attribute.
- **Uploaded columns** allow users to fuse external data to the currently shown table. The data can be loaded from a comma-separated file in which the first column contains the unique identifier of the primary identifier in the table, followed by one or multiple columns holding the data to be integrated. The system automatically detects common annotations, such as gene symbols and Ensembl IDs.

### Step 3: Obtain Detailed Information

Users can select one or more items in a ranking table for exploration using a collection of detail views. The detail views offered to the user depend on the type of items selected. The publicly deployed version of Ordino includes the following detail views:

- **Database Info view** for showing metadata stored in the database for all selected items. The information is represented as a table containing a row for each database attribute and a column for each selected item.
- Expression view, Copy Number view, and Mutation view visualizing experimental data for the currently selected items with the ranking visualization technique described in Step 2 of the analysis workflow.
- **Combined view** is a specialized ranking view that is able to show copy number, expression, and mutation data in combination.
- Expression vs. Copy Number view showing a scatterplot for each selected gene with copy number mapped to the x-axis and expression to the y-axis (cf. Step 3 in Supplementary Fig. S1). The analyst can determine via a drop-down list whether the scatterplot shows cell lines or tissue samples. Dots in the plots can be colored by preloaded categorical attributes, such as tumor type, gender and organ.
- **Co-Expression view** for comparing the expression of multiple selected genes. If multiple genes are selected, one plot is shown for each combination. Analogously to the Expression vs. Copy Number view, the dots represent either cell lines or tissue samples and can be colored by categorical attributes.
- **OncoPrint view** showing a horizontal series of colored blocks (glyphs) for each gene selected. Depending on the chosen data subset, each block represents a cell line or tissue sample. The background color of the blocks indicates the copy number status (pink=amplification, blue=deep deletion, gray=normal, white=unknown), while the small block contained visualizes the mutation status (green=mutated, gray=non mutated, white=unknown) of a cell line or tissue sample.
- **External resource views** loading the content of external websites. For selected genes, the user can look at the information available on canSAR, Ensembl, Human Protein Atlas, Open Targets, PubMed, and UniProt. For cell lines, the analyst can load the information available on COSMIC (Catalogue Of Somatic Mutations In Cancer).

# Implementation and Availability

Ordino is publicly available at <u>https://ordino.caleydoapp.org</u> and best viewed in Google Chrome browsers. The source code is freely available at <u>https://github.com/Caleydo/ordino</u>.

Ordino is based on the extensible Phovea platform (<u>http://phovea.caleydo.org</u>). The web client is implemented in TypeScript and the server in Python. The source code is open source under the Mozilla Public License (MPL) and hosted on Github. The public version of the system is deployed on AWS (Amazon Web Services) infrastructure using docker images (<u>https://www.docker.com</u>).

# **Data Processing & Integration**

The publicly deployed Ordino instance contains the following data sets (**Supplementary Table S1**):

- The Cancer Genome Atlas (TCGA) <u>https://cancergenome.nih.gov</u>:
   Gene expression, mutation, and copy number data
- Cancer Cell Line Encyclopedia (CCLE) <u>https://portals.broadinstitute.org/ccle</u> [3]: Gene expression, mutation, and copy number data
- Project DRIVE [4]: RNAi depletion-screen data (RSA and ATARiS)
- Avana CERES [5]: CRISPR-Cas9 depletion-screen data

#### **TCGA Sample Selection**

All TCGA samples from cBioPortal stage files (downloaded on Sept 29th, 2016) and all samples from the Firehose release 2016-01-28 were included, except for the sample cohorts COADREAD, FPPP, GBMLGG, KIPAN, STES, since these are combined or experimental cohorts.

#### TCGA Metadata

The R package TCGAbiolinks (Version 2.5.9) [6] was used to extract sample and patient information for TCGA samples by using a custom-made R script.

#### TCGA Gene Expression Data

TCGA data was downloaded from Firehose (https://gdac.broadinstitute.org) using the version 2016 01 28 stddata between March 11th and March 16th, 2016. Expression data computed by RSEM and represented in tau values were extracted from the archives with the base name \*CANCER.Merge\* illumina\* rnaseqv2 \*RSEM genes data.Level 3\*tar.gz, where CANCER resolves to all cancer types. Gene expression values were converted to TPM values with the equation TPM = tau \* 1e6. Gene names were split into gene identifier and gene symbols. All gene information was mapped to Ensembl 70 gene identifiers using the following procedure: (i) assign all genes for which there is a unique 1:1 correspondence between gene id and gene ids reported by Ensembl, (ii) assign remaining genes if their gene ids and gene symbols map 1:1, and (iii) map remaining genes if their gene ids and symbols map 1:1 after excluding the following gene types: pseudogene, lincRNA, sense. The mapping rate was 19,959/20,531 (97.2%) with respect to mapping genes from the original source. With respect to the samples, only samples of types 1 (Primary solid Tumor), 3 (Primary Blood Derived Cancer - Peripheral Blood), or 11 (Solid Tissue Normal) were kept in the final output. For about 20 samples, more than one aliquot or sequencing result (Illumina HiSeq or GA) was available. In these cases, either the HiSeq sequencing result was taken or the first occurrence in the file in case of a tie situation.

#### **TCGA Mutation Data**

cBioPortal stage files were downloaded on Sept 29th, 2016, and subsequently parsed using a custom-made R script. Additional parsing was necessary, as the transcript for which TCGA reports

mutations (column Transcript\_ID) did not necessarily match the canonical Ensembl transcript. Column "all\_effects" of the stage files contains the Variant Effect Predictor (VEP) annotation for additional Ensembl transcripts, which we split on the default delimiter character (";") and inserted into the database. This additional step allowed us to align the mapping between gene symbol and canonical Ensembl transcript ID between CCLE and TCGA.

#### TCGA Copy Number Data

TCGA SNP6 copy number segmentation data was downloaded from Firehose (https://gdac.broadinstitute.org) using version 2016\_01\_28 stddata between March 11th and March 16th 2016. The segmentation information was obtained from the files \*snp\*seg.txt stored in the archives \*.Merge snp \*cnv hg19\*.Level 3\*gz.

Gene-wise copy numbers were determined by overlapping the segmentation information with Ensembl 70 gene annotation. If a gene was covered by a single segment, the copy number of the segment was assigned to the gene. If a gene was covered by multiple segments, a weighted average copy number was computed based on the size of the overlap between the gene and each segment.

Relative copy numbers <= 1.0 were considered as "deep deletion", and relative copy numbers >= 3.5 were considered as "amplification".

#### **CCLE** Metadata

Cell line names and descriptions (organ of origin, metastatic site, histology type, morphology, growth type, gender, and age at surgery) were taken from the provider's cell-line data sheet. If a cell line was available from various vendors, the cell-line name was taken from the top rank in a hierarchy of vendors in the following order: atcc, dsmz, ecacc, jcrb, iclc, riken, kclb. A controlled vocabulary of 30 tumor types was derived from the cell-line annotation and assigned to the cell lines with the help of a pathologist.

#### **CCLE Gene Expression Data**

CCLE BAM files were downloaded from <u>https://cghub.ucsc.edu/</u> using gtdownload in March 2014. BAM files were converted to FASTQ files using SAMtools 0.1.19, bedtools 2.20.1, and the FASTX-Toolkit 0.0.14 as follows: secondary and vendor failed alignments were excluded from the input BAM files, and alignments were shuffled. BAM files were converted to paired-end FASTQ files length-trimmed to 100 bp per read.

All 935 pairs of FASTQ files were then individually aligned to the human genome (hs37d5) using GSNAP version 2012-12-20 together with splice-site information as well as SNP data from the 1000 Genomes Project Phase I (parameters: gsnap \$readgroup -A sam --ordered -n 50 -N 1 -t \$threads --gunzip -D \$gsnapDB -d \$gsnapName -s \$spliceiit -V \$snpDB -v \$snpName --show-refdiff --sam-use-OM \$r1 \$r2). Gene and transcript quantification was performed using Cufflinks version 2.0.2 (parameters: cufflinks -u -p \$threads -0 \$fpkmout --max-bundle-frags 999999999 --no-effective-length-correction --compatible-hits-norm --max-frag-multihits 1 -G \$gtf \$output.bam) and the human gene annotation from Ensembl 70. The resulting BAM file with mapped reads was sorted by read names using the SAMtools version 1.0 and used as input to HTSeq version 0.5.3p9 to summarize read counts over gene models (parameters: samtools view -h -f 0x2 \$output.namesorted.bam | htseq-count --quiet --idattr gene\_id --type exon -a 20 --stranded no --mode intersection-strict - \$GTF > \$output.counts). Various quality control measurements were performed using the following tools: Picard tools version 1.64, RSeQC version 2.3.5, R version 3.0, and FastQC version 0.10.0.

#### **CCLE Mutation Data**

Variant calling of cell lines followed community best practices. The reads were aligned using BWA version 0.5.8 (parameters: -q 10) against the reference genome hg19 including decoy sequences (hs37d5), followed by base recalibration and indel realignment and subsequent variant calling using GATK version 1.6. and filtering for artefacts such as polymerase slippage in homopolymer regions, strandness of detected variants, and overall quality. Variants were annotated based on Ensembl 70, and putative germline variation was flagged using external data sets such as dbSNP v135 and data from the 1000 genomes consortium. Putative alignment artefacts were filtered out using a mutation blacklist derived from the Sanger COSMIC Cell line Project VCF files (v70), for which putative artefacts/germline variation is flagged in the VCF files. We computed coverage statistics for each gene in each sample: In the absence of a mutation, we called a gene wild-type if and only if at least 80% of bases of the gene body (excluding the first exon) were sufficiently covered, and NA otherwise.

#### **CCLE Copy Number Data**

SNP6 CEL files were downloaded from <a href="https://cghub.ucsc.edu/">https://cghub.ucsc.edu/</a> in October 2012. Relative copy number segments were computed using the R package aroma.affymetrix version 2.13.0 [7-9]: the SNP6 data was processed with the method CRMA v2 followed by CBS segmentation. Afterwards, the copy number segments were overlapped with Ensembl 70 gene annotation analogously to the TCGA processing in order to obtain gene-wise relative copy number values. "Amplification" and "deep deletion" status were also assigned as in the TCGA processing.

Absolute copy number segments were computed using PICNIC version c\_release 2010-10-29 [10] with reference files adapted for reference genome hg19 and default parameters. The resulting segments were overlapped with Ensembl 70 gene annotation as in the TCGA processing in order to obtain gene-wise absolute copy number values.

#### **DRIVE Data**

DRIVE (deep RNAi interrogation of viability effects in cancer) is a large shRNA screen of ~8000 genes and ~400 cancer cell lines [4]. Raw data and processed RSA and ATARIS scores were transferred via email by the authors. siRNAs targeting multiple genes were discarded. Gene symbols were translated into Ensembl stable identifiers for genes by using the official gene symbol provided by the Ensembl database Version 70. Cell-line names are identical to CCLE cell-line names and were translated to the Boehringer Ingelheim cell-line nomenclature.

#### Avana Data

The Avana single guide RNA (sgRNA) library was used in a large CRISPR/Cas9 loss-of-function screen [5] of ~340 cell lines and ~17,500 genes. Processed CERES scores (representing the estimated gene knockout effects) were taken from the supplement. Gene symbols were translated into Ensembl gene identifiers using the official gene symbol provided by the Ensembl database Version 70. As for the DRIVE data set, CCLE cell-line names were used to translate cell-line identifiers into Boehringer Ingelheim's cell-line names.

### **Case Study 1: Assessment and Selection of Breast Cancer Cell Lines**

This case study summarizes an analysis session carried out by a scientist working in a drug discovery team at a pharmaceutical company. In order to identify potential drug targets in a set of tumor types, the analyst performs experiments with cancer cell lines—cultured cells that are derived from tumors and that can proliferate indefinitely in the laboratory. These cell lines are characterized by various properties, such as tumor type (lung cancer, prostate cancer, etc.) and the set of genes that are mutated, deleted, or amplified.

In this case study, the analyst first wants to identify the most relevant amplified genes in the breast cancer cell line HCC1954. Based on these results, the analyst then wants to study a larger set of breast cancer cell lines.

The scientist starts by loading the list of all protein coding genes and adding a column with the relative copy number information for the cell line HCC1954. After sorting by this column, the analyst observes that about 15 genes on chromosome 17 are affected by a large genomic amplification. In order to identify the most relevant gene of these, the analyst adds a column with the gene expression (a measure of activity) in HCC1954 as well as a gene sensitivity score (a measure of importance for cell survival) for HCC1954 (RSA scores obtained from DRIVE data set [4]). The assumption is that amplified cancer genes are highly expressed and that cell lines are sensitive to their removal. Of the highly amplified genes, *ERBB2* (also known as *HER2*) has the highest expression and the most significant sensitivity score, which becomes even more obvious when combining the two columns as stacked bars. It is therefore probably the most relevant gene within this amplified genomic region.

This finding leads the scientist to the question whether *ERBB2* is also highly expressed and frequently amplified in other breast cancer cell lines. To investigate this, the analyst adds a column with the average gene expression, a column with the gene copy number distribution, and a column with the gene amplification frequency across all breast cancer cell lines. He observes that *ERBB2* is amplified in almost 25% of all assessed breast cancer cell lines and that it is often highly expressed.

In order to obtain additional information about this gene, the analyst selects it and opens a series of detail views. Based on the the Expression vs. Copy Number detail view, the analyst notices that there is a clear correlation between copy number and expression of *ERBB2* (the higher the copy number, the higher the expression). The Open Targets detail view provides further information about the gene, for instance, the corresponding protein structure and a list of drugs that target *ERBB2*.

Next, the scientist is interested in obtaining a list of breast cancer cell lines that have both an *ERBB2* amplification and either a *BRCA1* or *BRCA2* mutation (*BRCA1* and *BRCA2* are highly relevant genes in the context of breast cancer). To this end, the analyst opens the Copy Number detail view, filters the cell

lines for the tumor type "breast cancer" and sorts the remaining cell lines by their *ERBB2* copy number. After adding columns for *BRCA1* and *BRCA2* mutations, the scientist observes that HCC1954 has the highest *ERBB2* amplification among *BRCA1* mutated cell lines and that HCC1569 has the highest ERBB2 amplification among *BRCA2* mutated cell lines. Finally, in order to obtain further information about these two cell lines, the scientist selects them, opens the COSMIC detail view, and browses the available data.

# Analysis Steps and Observations

Question: What are the most amplified genes in breast cancer cell line HCC1954?

- Open list of all protein-coding genes
- Add single cell-line score
  - Cell line: HCC1954
  - Data type: Relative Copy Number
- Sort by copy number column
- Observe: Highest amplification on chromosome 17, affecting about 15 genes

### Question: Which of these genes is the most relevant?

- Add single cell-line score
  - Cell line: HCC1954
  - Data type: Expression (TPM)
- Add single depletion-screen score
  - Cell line: HCC1954
  - Data type: DRIVE RSA (NB: the lower this value, the more sensitive a cell line is to the depletion of a specific gene)
  - Note that his data is only available for a subset of genes.
- Invert depletion screen score (large bars represent small (very negative) values)
- Observe: Of the highly amplified genes, *ERBB2* (*HER2*) has the highest expression and the lowest sensitivity score. Therefore, it is probably the most relevant gene of this amplicon.
- Combine both score columns to obtain stacked bars
- Observe: Combining the columns highlights the importance of *ERBB2*

#### Question: Is ERBB2 also highly expressed and frequently amplified in other breast cancer cell lines?

- Add aggregated cell-line score
  - Tumor type: breast carcinoma
  - Data type: Expression (TPM)
  - Aggregation: Average
- Add aggregated cell-line score
  - Tumor type: breast carcinoma
  - Data type: Relative Copy Number
  - Aggregation: Boxplot
- Add aggregated cell-line score
  - Tumor type: breast carcinoma
  - Data type: Relative Copy Number
  - Aggregation: Frequency (> 4)

• Observe: *ERBB2* is amplified in almost 25% of all assessed breast cancer cell lines. Further, it is highly expressed.

Aim: Get more information about ERBB2

- Select ERBB2
- Open Expression vs. Copy Number detail view
- Observe: direct correlation between copy number and expression of *ERBB2*
- Open Open Targets detail view
- Open Pubmed detail view

Question: In what cell lines is ERBB2 amplified? Select cell lines with ERBB2 amplification that have mutation in BRCA1 or BRCA2.

- Open Copy Number detail view
- Sort by copy number
- Filter for breast cancer via column menu of column tumor type (also filter out cell lines with unknown tumor type)
- Add single gene score
  - Genes: *BRCA1*, *BRCA2*
  - Data type: AA mutated
- Observe: HCC1954 has the highest *ERBB2* amplification among *BRCA1* mutated cell lines. HCC1569 has the highest *ERBB2* amplification among *BRCA2* mutated cell lines.

Aim: Show information provided by COSMIC about these two cell lines

- Select HCC1569 and HCC1954
- Open COSMIC detail view
- Use the drop-down menu to switch between the two cell lines

#### Case Study 2: Prediction of TP53 Mutation Status

This case study summarizes another analysis session carried out by a research scientist working in a drug discovery team at a pharmaceutical company. In order to identify potential drug targets in a set of tumor types, the analyst performs experiments with cancer cell lines—cultured cells that are derived from tumors and that can proliferate indefinitely in the laboratory. These cell lines are characterized by various properties, such as tumor type (lung cancer, prostate cancer, etc.) and the set of genes that are mutated. One very important gene in the context of cancer is TP53. It encodes the p53 protein, whose presence is known to suppress the uncontrolled division of cells. However, when TP53 is mutated—which is the case in over 50% of cancer patients—it can lose its suppressing function, which results in tumor growth. Due to its important role, scientists want to know whether TP53 is mutated in a set of cell lines. However, the mutation status of TP53 is not always known. It has recently been shown that the mean expression level (expression is a measure of the activity of genes) of 13 genes that are biologically related to TP53 is correlated with its mutation status. The expression level of these genes can therefore be used to predict the mutation status of TP53 [11]. Furthermore, it has been shown that in many TP53 non-mutated cell lines, the p53 protein is downregulated through interaction with the protein MDM2. As a consequence, these cell lines are in many cases dependent on the expression of MDM2. A downregulation of MDM2 can result in re-activation of p53 and therefore in the induction of cell death. Hence, cell lines that react sensitively to the removal of MDM2 are often TP53 non-mutated, and thus this sensitivity can also be used as a predictor.

In this case study, the analyst first wants to find out how well these two predictors work for the samples contained in the database. Secondly, the analyst wants to predict the *TP53* mutation status for colon carcinoma cell lines for which this information is not available in the database, and seeks further information about the cell lines.

After creating a set containing the 13 *TP53* status prediction genes [11], the analyst loads the list of all TCGA tumor samples, filters for the tumor type colon adenocarcinoma, adds a column showing the *TP53* mutation status, and removes all samples for which the mutation status is unknown. Furthermore, the analyst loads the average expression of the 13 genes. After sorting by the gene expression level, the scientist observes that there is a clear correlation between the gene expression signature and the mutation status of *TP53*: Of the 50 samples with the highest expression only 3 are *TP53* mutated, whereas of the 50 samples with the lowest expression 35 are *TP53* mutated.

The scientist continues the analysis by assessing human cell lines. After loading the list of all available cell lines, the analyst adds columns for the *TP53* mutation status, the average expression of the 13 genes, and the *MDM2* sensitivity data (RSA score from DRIVE data set [4]; a small value indicates high sensitivity). The scientist observes that there is a clear enrichment for *TP53* non-mutated among the cell lines with highest average expression. Furthermore, the *MDM2* RSA values are clearly correlated with the average expression score and the *TP53* mutation status. Based on these observations, the analyst concludes that both predictors are working reasonably well. In order to further improve the *TP53* mutation status prediction, the scientist combines the average expression and the *MDM2* RSA score, observing that the combined score correlates even better with the *TP53* mutation status than the individual scores.

Since the scientist is particularly interested in colon carcinoma, he limits the set of cell lines to this tumor

type and observes that the joint predictor also works well for this subset.

Finally, the analyst wants to predict the *TP53* mutation status for colon carcinoma cell lines for which this information is not available in the database. The scientist includes cell lines for which no *TP53* mutation informations is available. The data set used contains two colon carcinoma cell lines that lack *TP53* mutation status and for which gene expression and DRIVE data is available: MDST8 has a very high combined score and is therefore probably *TP53* non-mutated. NCI-H508 has a very low combined score and is therefore probably *TP53* non-mutated. NCI-H508 has a very low combined score and is therefore probably *TP53* non-mutated to assess which genes contribute most to the expression-based *TP53* predictor score, the scientist opens the Expression detail view and observes that the high average expression value in MDST8 can be attributed mainly to the expression of *BAX, CDKN1A*, and *RPS27L*. Finally, the scientist opens the COSMIC detail view to obtain further information about the cell lines.

# Analysis Steps and Observations

Aim: Create gene set for the 13 genes of the expression signature

- Paste the following list of genes into the gene input field on the welcome page and click "Save": AEN, BAX, CCNG1, CDKN1A, DDB2, FDXR, MDM2, RPS27L, RRM2B, SESN1, TNFRSF10B, XPC, ZMAT3
- Name set TP53 Predictor

Aim: Test applicability of gene signature using TCGA tumor samples

- Open list of all TCGA tumors
- Filter tumor type *colon adenocarcinoma*
- Add single gene score
  - Gene: *TP53*
  - Data type: AA mutated
- Filter out samples with unknown *TP53* mutation status
- Add aggregated gene score
  - Filter: My Named Sets = TP53 Predictor
  - Data type: Expression (TPM)
  - Aggregation: Average
  - Compute score only for current sample subset
- Sort by gene expression column
- Observe: There is a clear correlation between gene expression signature and mutation status of *TP53*: Of the 50 samples with the highest expression only 3 are *TP53* mutated, whereas of the 50 samples with the lowest expression 35 are *TP53* mutated.

Aim: Test applicability of gene signature and MDM2 sensitivity using all cell lines

- Open list of all cell lines from the start menu
- Add single gene score
  - Gene: *TP53* 
    - Data types: AA Mutated and AA Mutation
- Filter out cell lines with unknown *TP53* mutation status

- Add aggregated gene score
  - Filter: My Named Sets = *TP53* Predictor
  - Data type: Expression (TPM)
  - Aggregation: Average
  - Compute scores for all cell lines, not only selected subset (i.e., uncheck option)
- Rename "avg TPM" to "TP53 Predictor Score"
- Filter out missing values
- Sort by TP53 predictor score
- Observe: There is a clear enrichment of *TP53* non-mutated among the cell lines with high score.
- Add single depletion-screen score
  - Gene: *MDM2*
  - Data type: DRIVE RSA (NB: the lower this value, the more sensitive a cell line is to the depletion of the gene of interest)
- Invert score so that large bars represent small (very negative) values
- Filter out missing values
- Sort by *MDM2* RSA score
- Observe: Small *MDM2* RSA values (large bars) are correlated to the expression score (*TP53* predictor score) and the *TP53* mutation status
- Combine the two score columns by dragging one column onto the other, which results in a weighted sum column
- Sort by the combined column
- Observe: Combined score correlates even better with *TP53* mutation status than individual scores.

#### Aim: Assess predictors in colon carcinoma cell lines

- Filter for tumor type *colon carcinoma*
- Observe: Combined score correlates well with *TP53* mutation status for colon carcinoma cell lines.
- Conclusion: The TP53 target score as well as the *MDM2* RSA score can be used to predict a *TP53* mutation status of cell lines for which it is not available.

Aim: Predict TP53 mutation status for colon carcinoma cell lines for which this information is not available in the database and assess which genes contribute most to the TP53 predictor score. Furthermore, obtain additional information from COSMIC about these cell lines.

- Include cell lines with unknown *TP53* mutation status.
- Observe: The data set used contains two colon carcinoma cell lines that lack *TP53* mutation status and for which gene expression and DRIVE data is available: MDST8 has a very high combined score and is therefore probably *TP53* non-mutated. NCI-H508, however, has a very low combined score and is therefore probably *TP53* mutated.
- Select the two cell lines and open Expression detail view
- Observe: The high average expression value in MDST8 is caused mainly by the expression of *BAX*, *CDKN1A*, and *RPS27L*
- Open COSMIC detail view and browse available information

# **Supplementary References**

[1] Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H. P., Streit, M., LineUp: Visual Analysis of Multi-Attribute Rankings. IEEE Trans Vis Comput Graph **19**, 2277–2286 (2013).

[2] Gratzl, S., Lex, A., Gehlenborg, N., Cosgrove, N., Streit, M., From Visual Exploration to Storytelling and Back Again. *Computer Graphics Forum* **35**, 491–500 (2016).

[3] Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**(7391):603–607 (2012).

[4] McDonald III, E. R. et. al. Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577-592.e10 (2017).

[5] Meyers, R. M. et. al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics* **49**, 1779–1784 (2017).

[6] Colaprico, A. et. al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71 (2016).

[7] Bengtsson, H. et. al. aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Tech Report #745, Department of Statistics, University of California, Berkeley* (2008).

[8] Bengtsson, H. et al. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759–767 (2008).

[9] Bengtsson, H. et al. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics* **25**, 2149–2156 (2009).

[10] Greenman, C. D. et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164–175 (2010).

[11] Jeay, S. et al. A distinct p53 target gene set predicts for response to the selective p53–HDM2 inhibitor NVP-CGM097. *eLife*, 4e06498 (2015).