Data Analysis Projects
MoBi 4. FS - SS2019

Concept

- **Project-oriented teaching**: provide hands-on experience with data analysis and programming

- **Goals**
  - experience data analysis challenges on real datasets related to research question
  - experience team work, also outside of your team!
  - learn to use modern data analysis tools: R / Python / markdown / notebooks / github

Research topics / projects

- **5 research projects** have been defined

- Each topic has up to **5 sub-projects**

- Each project will be worked out by **groups of 4 students**

- **one supervisor and master tutor per project**

- **Role of the master tutors:**
  - ◉ weekly meetings with groups working on project
    (Wednesday 10am-1pm)

- Meeting rooms: BioQuant SR42/43 + IPMB meeting room 5th floor
  2 other meeting places must be found (lounge corner at BioQuant?…)

- Tutors:
  Valentina Giunchiglia; Julia Rühle; David Schwarzenbacher; Nicolas
  Peschke; Alexander Mattausch

Timeline

We do…   *You do…*

17/04   Presentation of the projects

                                            selection of projects
                                            and teams;
24/04   Presentation of R markdown   registration
            and github

15/05   Presentation of
                                            project proposal
                                            (10 + 10 min)

24/07   Final presentation
(25/07 for Project 03)   (15+10 minutes)

Carl Herrmann   Health Data Science Unit - Medizinische Fakultät Heidelberg

Project proposal (15/05)

- During the project proposal presentation, you should
  - review some of the references given in the project description
  - explain what the questions / challenges are
  - describe which of these questions you want to adress in your project
  - indicate a approximate timeline
    - milestones = important steps in the analysis
    - when these milestones should be achieved

- Presentation in front of the project supervisors
  - 10 minutes presentation
  - 10 minutes discussion / questions

- *All team members are expected to contribute!*

Projects / sub-projects

- **Project 01**: *Genetic interactions in cancer*

  (Ashwini Sharma / Carl Herrmann)

  - ◉ Data types: gene expression / gene mutations / gene knockdown / CNV

- **Project 02**: *Cellular response to drug perturbation*

  (Nicolas Palacio / Javier Perales)

  - ◉ Data types: gene expression treated / untreated / mutations / metadata

- **Project 03**: *Biomedical image analysis*

  (Karl Rohr / Christian Ritter)

  - ◉ Data types: MNIST images / cell nuclei images

- **Project 04**: *Programming k-means*

  (Thorsten Beier)

  - ◉ Data types: scRNA-seq

- **Project 05**: *Cancer DNA Methylation*

  (Matthias Schlesner / Christian Heyer)

  - ◉ Data types: DNA methylation WGBS
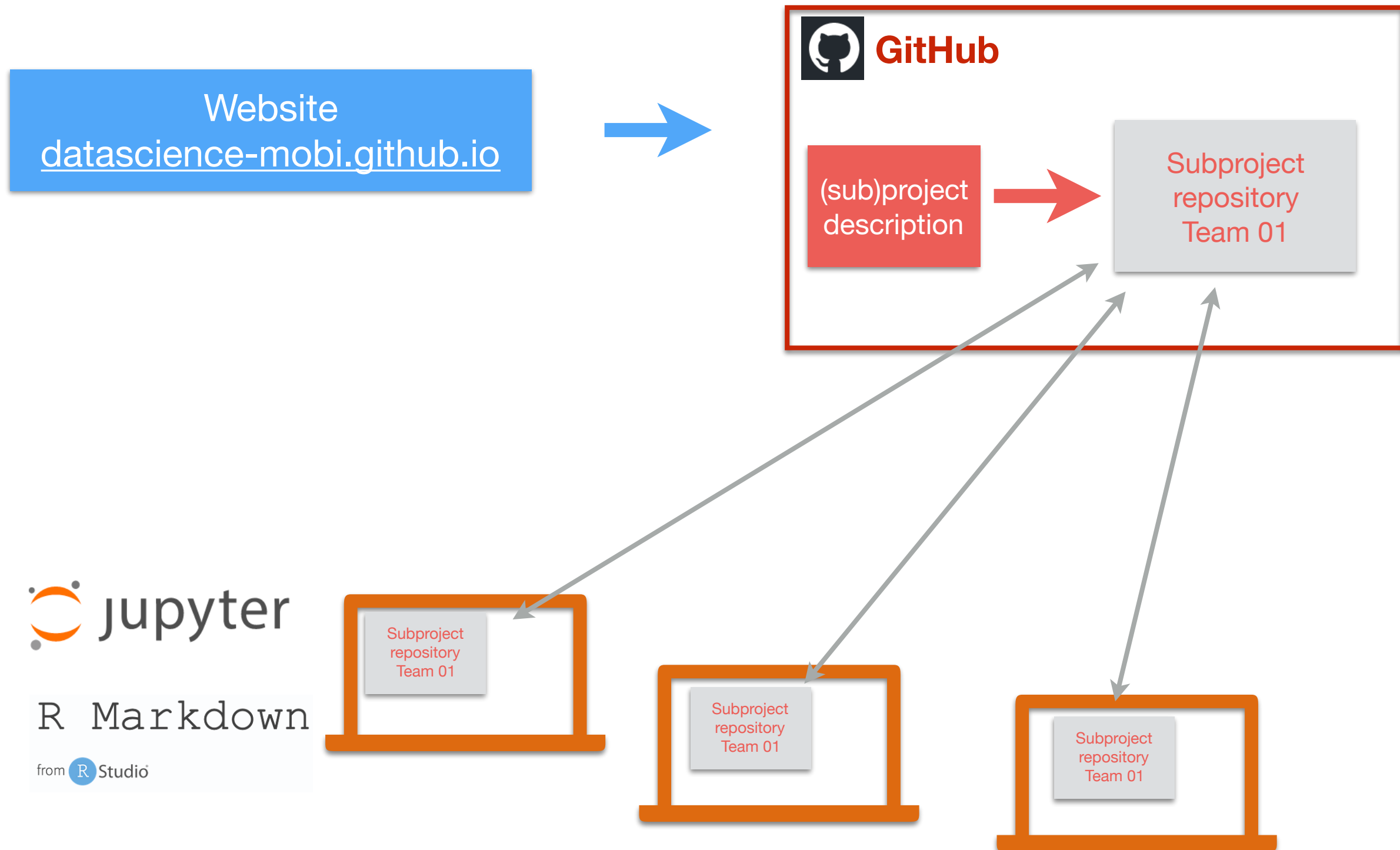
Project selection / registration

- Listen to the description of the projects / sub-projects

- Check this webpage
  https://datascience-mobi.github.io/

- Once you have selected your team and project, **register your team** in the Google sheet
  https://docs.google.com/spreadsheets/d/1LEQLH2LaDulMq3Qepx-7-5KWjgZjK4dfWDDIbm1vu0Q/edit?usp=sharing

- ***Selection of the projects should be done by 24/04 10am !***

Organizing your work

Website
datascience-mobi.github.io

GitHub

(sub)project description

Subproject repository Team 01

Jupyter

R Markdown
from R Studio

Subproject repository Team 01

Subproject repository Team 01

Subproject repository Team 01

Project 01: genetic interactions in Cancer cells

- Find **synthetic lethal interactions** in cancer cell lines for driver mutations

- Which mutation / copy-number alteration does lead together with a driver mutation to a significant phenotype in cancer cells?
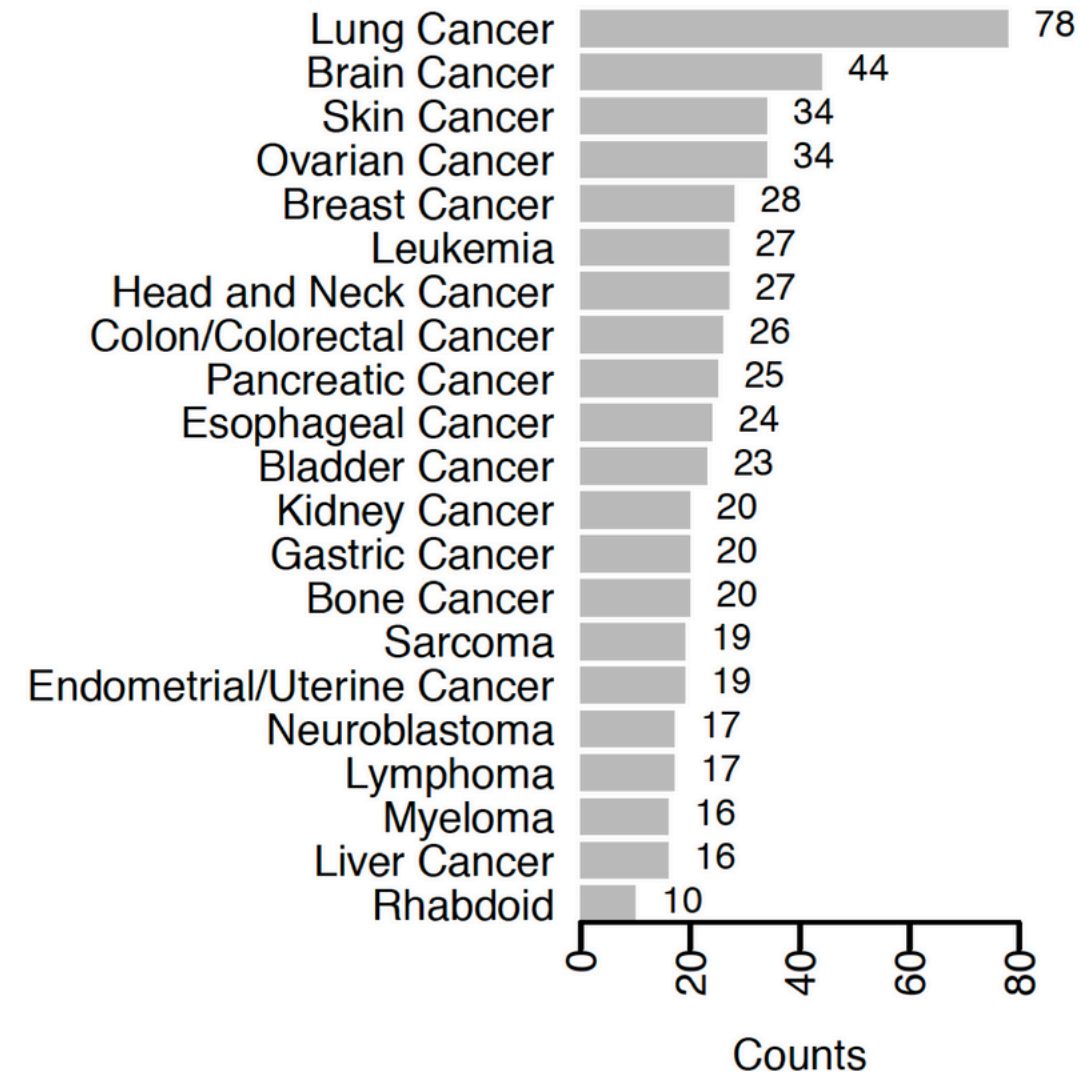
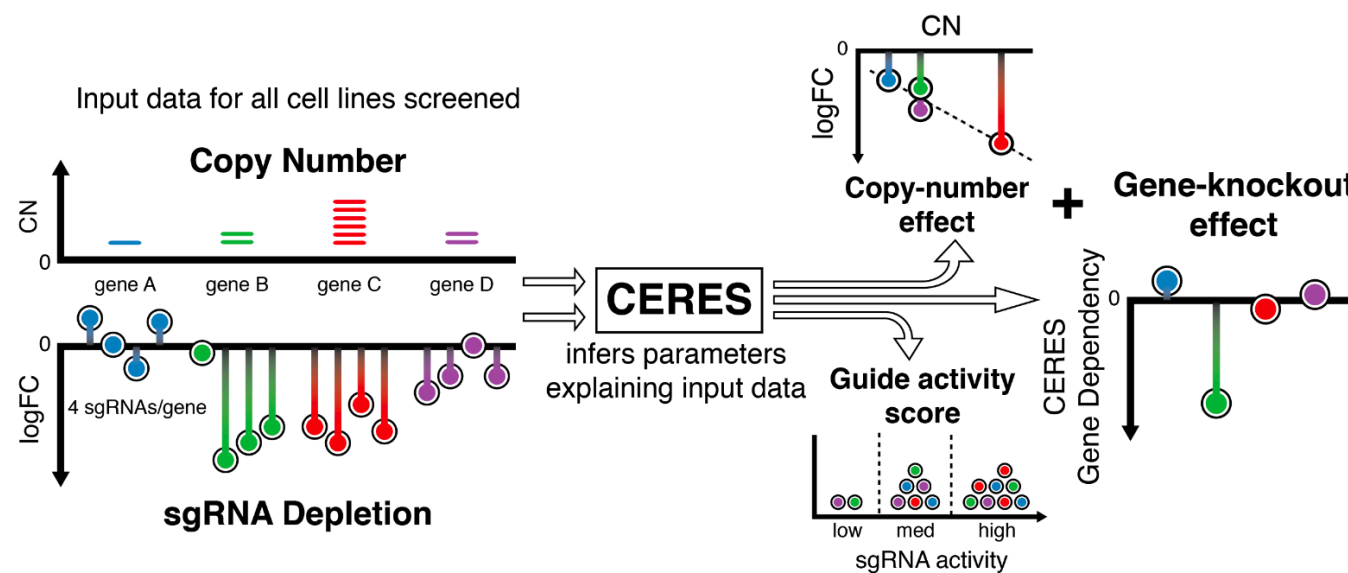 → identify potential therapeutic targets
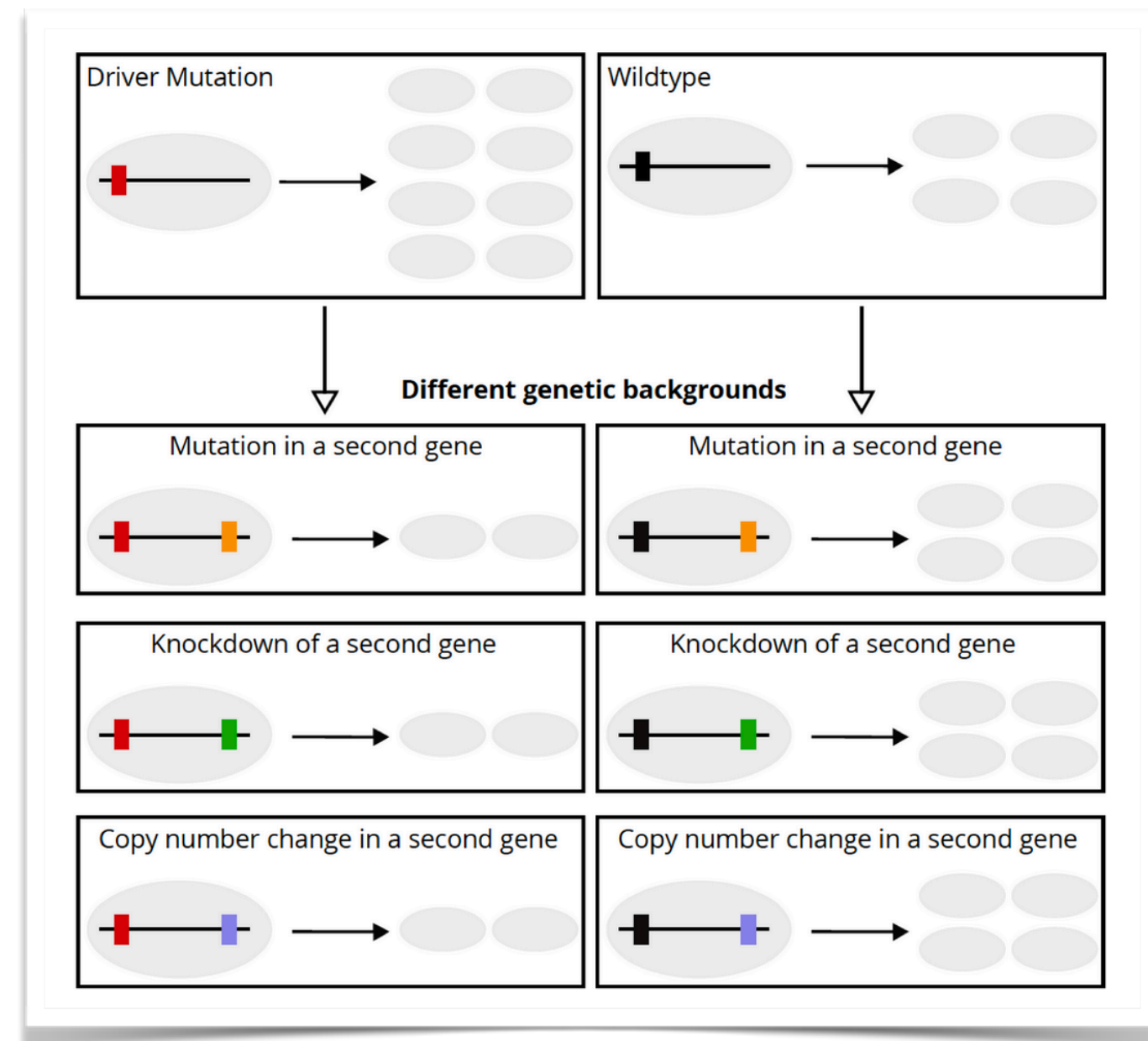
Project 01: genetic interactions in Cancer cells

- Knockdown assays (Project DepMap/CERES):
  CRISPR-Cas9 for ~17.000 genes across 544 cell lines

[Meyer etal., 2017]

Datasets

- **Gene expression** in various cancer cell lines

	ACH-000004	ACH-000005	ACH-000007	ACH-000009	ACH-000011	ACH-000012	ACH-000013
TSPAN6	2.61588707	3.06695024	4.06608919	6.50795317	4.57773093	5.82094479	5.53387478
TNMD	0.00000000	0.00000000	0.00000000	0.09761080	0.00000000	0.00000000	0.05658353
DPM1	5.32337007	5.76261458	5.88996020	7.98162436	5.53574194	6.00652245	7.53216108
SCYL3	2.40599236	2.99276843	3.04963077	2.24792751	2.08746284	1.91456452	1.91073266
C1orf112	3.90207358	5.35966172	3.76022095	4.49121176	2.64616266	3.75167795	3.54720302
FGR	0.92599942	0.23878686	0.02856915	0.00000000	0.00000000	0.02856915	0.02856915
CFH	4.88898672	5.70016225	0.01435529	0.02856915	0.35614381	2.68929916	1.98550043

- **Mutations** in genes in each cell line & copy-number alterations in each cell line

Hugo_Symbol	Entrez_Gene_Id	NCBI_Build	Chromosome	Start_position	End_position	Strand	Variant_Classification
RNF207	388591	37	1	6279339	6279339	+	Missense_Mutation
PLEKHG5	57449	37	1	6533165	6533165	+	Missense_Mutation
PDPN	10630	37	1	13940848	13940848	+	Missense_Mutation
CASP9	842	37	1	15819484	15819484	+	Missense_Mutation
RAP1GAP	5909	37	1	21924552	21924552	+	Missense_Mutation
C1QC	714	37	1	22974054	22974054	+	Silent
CNKSR1	10256	37	1	26514778	26514778	+	Frame_Shift_Del
AHDC1	27245	37	1	27875454	27875454	+	Missense_Mutation
COL16A1	1307	37	1	32133218	32133218	+	Silent
CSMD2	114784	37	1	32133218	34383844	+	Silent
LRRC8C	84230	37	1	90179033	90179033	+	Missense_Mutation
SPAG17	200162	37	1	118567972	118567973	+	Frame_Shift_Del
TBX15	6913	37	1	119427935	119427935	+	Missense_Mutation

- **Knock-down sensitivity** in all cell lines for 17.000 genes

	ACH-000004	ACH-000005	ACH-000007	ACH-000009
A1BG	0.1346453616	−0.212445060	0.043317923	0.0705119999
A1CF	0.0755362715	0.233123579	0.066837574	0.0084297636
A2M	−0.1402086015	0.044364933	−0.036196515	0.0271141959
A2ML1	0.0139284337	0.173837240	0.134781001	0.0559267727
A3GALT2	0.0291310328	−0.124389318	0.082995584	0.0463253889
A4GALT	−0.1472838445	−0.298849014	0.119084008	0.0159682666
A4GNT	0.2758291936	0.120259815	0.057116006	0.0535023006
AAAS	−0.3636329615	−0.339925280	−0.352541473	−0.4988600588

- **Annotations** for cell lines (tissue, …)

Project 01: genetic interactions in Cancer cells

- Select cancer type and corresponding cell lines

- Determine driver mutations from the literature (e.g. EGFR mutations in lung cancer)

- Determine potential synthetic lethal mutations/ copy- number alterations from knock-down screens by splitting **mutated**/**non-mutated** cell lines

non-mutated cell lines

mutated cell lines