tugHall v 3.0: USER-GUIDE-Analysis

Requirements

R version **3.6**

Libraries: **stringr, ape, ggplot2**

Operation systems: Window, Mac. The code for analysis is not tested under Linux based systems.

Note that the program comprises two different procedures in general: the first is simulation and the second is the analysis of the simulation results. This User-Guide pertains to the **analysis** alone. The programs for the analysis can be run only after the simulation is completed, and the **cloneout.txt** file exists in the **CNA/Output/** folder.

Table of Contents

1. Quick start guide

To perform the simulation, kindly see the **User-Guide-tugHall_v_3.0** file. After the simulation the file **CNA/Output/cloneout.txt** is generated, which is used to analyze the evolution of cells. Also, since the functions and objects are used after the simulation, the **.RData** file saved after the simulation must be loaded, if required.

The simplest way to analyze the results after simulation:

- open R;
- set **CNA/** as the working directory;
- load **.RData**;
- run script after a simulation in **tugHall_3.0.R** file like:

```
source("Code/Functions_clones.R")
get_flow_data(cloneoutfile, genefile )

vf = get_VAF()
VAF  =  get_rho_VAF( vf = vf, rho = c( 0.0, 0.1, 0.2, 0.5, 0.7, 0.9 ) , file_name = './Output/VAF.txt' )

source( './Code/my_plots.R' )
rdr_dysf  =  get_order_of_genes_dysfunction()
plot_order_dysfunction( rdr_dysf , pos = c(28.5,200), logscale = 'y', cex = 0.7 )

plot_average_simulation_data()

plot_clone_evolution( threshold = c(0.01, 1 ), lwd = 2.0,
                      hue = c(" ", "random", "red", "orange", "yellow",
                              "green", "blue", "purple", "pink", "monochrome")[1],
                      luminosity = c(" ", "random", "light", "bright", "dark")[4],
                      yr = NA , add_initial = TRUE, log_scale = FALSE )

plot_clone_evolution( threshold = c(0.0, 0.01), lwd = 2.0,
                      hue = c(" ", "random", "red", "orange", "yellow",
                              "green", "blue", "purple", "pink", "monochrome")[1],
                      luminosity = c(" ", "random", "light", "bright", "dark")[4],
                      yr = NA , add_initial = FALSE, log_scale = TRUE )
```

The code has initial input parameters and input files in the **/Input/** folder to define the names of the genes. In the dialogue box, the user can see the results of the simulation, which will be saved to the **/Output/** and **/Figures/** folders (if file name is defined in functions).

2. Inputs

To analyze the output data, the user has to obtain the results of the simulation in the **CNA/Output/cloneout.txt** file and the functions and objects of simulations should be present in the R environment. That is why the **cloneout.txt** file is the input file for the analysis. For detailed information, kindly see the "Outputs" section in **User-Guide-tugHall_v3.0**.

3. Outputs

Output data contain several files and figures:

- **order_genes_dysfunction.txt** has information about the order of gene dysfunction during evolution.
- **VAF.txt** and **VAF_data.txt** files have information about the variant allele frequencies (VAFs) for each gene and each site in the genes.
- the folder **CNA/Figures/** has many plots (see Figures).

order_genes_dysfunction.txt file

CNA/Output/order_genes_dysfunction.txt has information about the order of gene dysfunction during evolution in the next format (only first 10 lines are presented here):

Table 1: **Order of gene dysfunction.**

N_cells	ID	ParentID	Birth_time	type	mut_den	driver_genes	passenger_genes
8995	2	0	0	primary	0.25	0 1 0 0	0 0 0 0
6460	12	2	0	metastatic	0.50	0 1 0 1	0 0 0 0
13048	16	2	1	metastatic	0.50	0 1 1 0	0 0 0 0
37	18	2	1	primary	0.50	1 1 0 0	0 0 0 0
3148	21	1	2	metastatic	0.25	0 0 0 1	0 0 0 0
45	47	2	5	primary	0.25	0 1 0 0	0 0 0 1
15958	49	2	5	metastatic	0.50	0 1 0 1	0 0 0 0
69	54	2	5	primary	0.50	1 1 0 0	0 0 0 0
21	60	2	6	primary	0.50	1 1 0 0	0 0 0 0
7	61	2	6	primary	0.25	0 1 0 0	1 0 0 0

Table 2: **Order of gene dysfunction (continuous).**

N_cells	ID	PointMut_ID	CNA_ID	order
8995	2	1	0	KRAS
6460	12	1, 17	0	KRAS -> PIK3CA
13048	16	1, 21	0	KRAS -> TP53
37	18	1	5	KRAS -> APC
3148	21	27	0	PIK3CA
45	47	1, 73	0	KRAS
15958	49	1, 77	0	KRAS -> PIK3CA
69	54	1, 87	0	KRAS -> APC

N_cells	ID	PointMut_ID	CNA_ID	_____order_____
21	60	1, 97	0	KRAS -> APC
7	61	1, 99	0	KRAS

1. **N_cells** - the number of cells in this clone, e.g. 1000, 2.
2. **ID** - the unique ID of clone, e.g., 1, 50.
3. **Parent_ID** - the parent index, e.g., 0, 45.
4. **Birth_time** - the time step of the clone's birth, e.g., 0, 5.
5. **type** - the type of the cell: 'normal' or 'primary' or 'metastatic'.
6. **mut_den** - the density of mutations for the cell, it equals to ratio a number of mutated driver genes to a number of all the genes, e.g., 0, 0.32.

7. **driver_genes** - the binary numbers indicate the driver mutation at the gene related to order of genes in onco as well as order of the next columns with genes' names, e.g., '1 0 0 0' means that the first gene has a driver mutation and other genes have no.
8. **passenger_genes** - the binary numbers indicate the passenger mutation at the gene related to order of genes in onco as well as order of the next columns with genes' names, e.g., '0 0 1 0' means that the third gene has a passanger mutation and other genes have no.
9. **PointMut_ID** - the index of data row for point mutation data frame saved at the end of simulation in the file **Point_mutations.txt**, e.g., 23, 32.
10. **CNA_ID** - the index of data row for CNA data frame saved at the end of simulation in the file **CNA.txt**, e.g., 44, 21.
11. **order** - Order of gene dysfunction from the first mutated gene to the last one, e.g., 'PIK3CA -> APC -> PIK3CA -> APC' related to driver mutations only.

Variant allele frequencies information

CNA/Output/VAF.txt file has information about the VAFs for each gene and each site in the genes (first 10 lines):

Table 3: **Variant allele frequencies.**

site	Chr	gene	rho	VAF_primary	VAF_primary_numerator	VAF_primary_denominator
25227277	12	KRAS	0	0.5	0.9997295	1.9994589
179229383	3	PIK3CA	0	0.0	0.0000000	0.0000000
7673750	17	TP53	0	0.0	0.0000000	0.0000000
179203635	3	PIK3CA	0	0.0	0.0000000	0.0000000
179199801	3	PIK3CA	0	0.5	0.0044188	0.0088376
179226008	3	PIK3CA	0	0.0	0.0000000	0.0000000
112801321	5	APC	0	0.5	0.0062224	0.0124448
112838706	5	APC	0	0.5	0.0018938	0.0037875
112844068	5	APC	0	0.5	0.0024348	0.0048697
112838623	5	APC	0	0.5	0.0045992	0.0091983

Table 4: **Variant allele frequencies (continuous).**

site	Chr	gene	rho	VAF_metastatic	VAF_metastatic_numerator	VAF_metastatic_denominator
25227277	12	KRAS	0	0.4999422	0.9613210	1.9228641
179229383	3	PIK3CA	0	0.4997532	0.0715425	0.1431557
7673750	17	TP53	0	0.4999317	0.1477899	0.2956202
179203635	3	PIK3CA	0	0.5000000	0.0385982	0.0771965
179199801	3	PIK3CA	0	0.0000000	0.0000000	0.0000000
179226008	3	PIK3CA	0	0.4998254	0.2167175	0.4335864
112801321	5	APC	0	0.0000000	0.0000000	0.0000000
112838706	5	APC	0	0.0000000	0.0000000	0.0000000
112844068	5	APC	0	0.5000000	0.0000101	0.0000202
112838623	5	APC	0	0.0000000	0.0000000	0.0000000

1. **site** - position at mutated site in the gene, e.g., 123, 1028.
2. **Chr** - chromosome where a gene is located, e.g., '12', '3'.
3. **Gene** - name of gene, e.g. TP53, KRAS.

4. **rho** - rho parameter in the formula of VAF calculation in the range [0,1], e.g., 0.0, 0.3.
5. **VAF_Primary** - VAF for cells in the primary tumor = VAF_primary_numerator / VAF_primary_denominator, e.g. 0.2345.
6. **VAF_primary_numerator** - numerator in the formula of VAF calculation for tumor primary cells and speckled normal cells, e.g., 0.9997295.
7. **VAF_primary_denominator** - denominator in the formula of VAF calculation for tumor primary cells and speckled normal cells, e.g., 1.9994589.
8. **VAF_Metastatic** VAF for metastatic cells = VAF_metastatic_numerator / VAF_metastatic_denominator, e.g. 0.35.
9. **VAF_metastatic_numerator** - numerator in the formula of VAF calculation for metastatic cells, e.g., 0.9997295.
10. **VAF_metastatic_denominator** - denominator in the formula of VAF calculation for metastatic cells, e.g., 1.9994589.

CNA/Output/VAF_data.txt file has information about the point mutations and its locations at chromosome, gene etc. (first 10 lines):

Table 5: **Information on point mutations in the clones at final time step.**

PointMut_ID	Parental_1or2	Chr	Ref_pos	Phys_pos	Delta	Copy_number	Gene_name
1	1	12	25227277	[25227277]	[0]	1	KRAS
17	2	3	179229383	[179229383]	[0]	1	PIK3CA
21	2	17	7673750	[7673750]	[0]	1	TP53
27	2	3	179203635	[179203635]	[0]	1	PIK3CA
73	2	3	179199801	[179199801]	[0]	1	PIK3CA
77	1	3	179226008	[179226008]	[0]	1	PIK3CA
87	1	5	112801321	[112801321]	[0]	1	APC
97	1	5	112838706	[112838706]	[0]	1	APC
99	1	5	112844068	[112844068]	[0]	1	APC
117	1	5	112838623	[112838623]	[0]	1	APC

Table 6: **Information on point mutations in the clones at final time step (continuous).**

PointMut_ID	MalfunctionedByPointMut	mut_order	N_speckled_normal	N_primary	N_metastatic	Copy_number_A
1	TRUE	1	0	11086	95201	1
17	TRUE	11	0	0	7075	1
21	TRUE	15	0	0	14634	1
27	TRUE	20	0	0	3823	1
73	FALSE	46	0	49	0	1
77	TRUE	48	0	0	17547	1
87	TRUE	53	0	69	0	1
97	TRUE	59	0	21	0	1
99	FALSE	60	0	27	0	1
117	TRUE	70	0	51	0	1

Table 7: **Information on point mutations in the clones at final time step (continuous).**

PointMut_ID	N_speckled_normal_total	N_primary_total	N_metastatic_total
1	0	11089	99046
17	0	11089	99046
21	0	11089	99046
27	0	11089	99046
73	0	11089	99046
77	0	11089	99046
87	0	11089	99046
97	0	11089	99046
99	0	11089	99046
117	0	11089	99046

1. **PointMut_ID** - ID of point mutation.
2. **Parental_1or2** - indicates either of the two parental chromosomes.
3. **Chr** - name of a chromosome.
4. **Ref_pos** - the reference position of an allele. The reference position is on the coordinate system of the human reference genome.
5. **Phys_pos** - the physical position of an allele. The physical length of a (parental) chromosome is extended or shrunk by CNA duplications or deletions, respectively. When a duplication happens, the reference position is divided into two or more physical positions, which are represented by multiple elements in a vector. When a deletion happens and the allele is lost, the lost is represented by "-" on the coordinate system of physical positions.
6. **Delta** - difference between the reference and physical positions.
7. **Copy_number** - the copy number of an allele at the chromosome where mutation is happened (allele B).
8. **Gene_name** - the name of a gene.
9. **MalfunctionedByPointMut** - logical indicator of whether or not the gene is malfunctioned by the point mutation.
10. **mut_order** - indicator of mutation order in the simulation, it's used to detect order of mutations in the clone at each chromosome.
11. **N_speckled_normal** - number of speckled normal cells at final time step which have that PointMut_ID.
12. **N_primary** - number of primary tumor cells at final time step which have that PointMut_ID.
13. **N_metastatic** - number of metastatic cells at final time step which have that PointMut_ID.
14. **Copy_number_A** - the copy number of an allele at the chromosome where mutation is NOT happened (allele A).
15. **N_speckled_normal_total** - the total amount of speckled normal cells in the pool of simulation at the last time step.
16. **N_primary_total** - the total amount of primary tumor cells in the pool of simulation at the last time step.
17. **N_metastatic_total** - the total amount of metastatic cells in the pool of simulation at the last time step.

4. Figures

The directory **Figures/** contains many output figures, generated during the analysis process of **cloneout.txt** file, including the evolution of the number of primary tumors and metastasis cells (Fig.1 left), hallmarks (Fig.1 right), and probabilities (Fig.2 left).

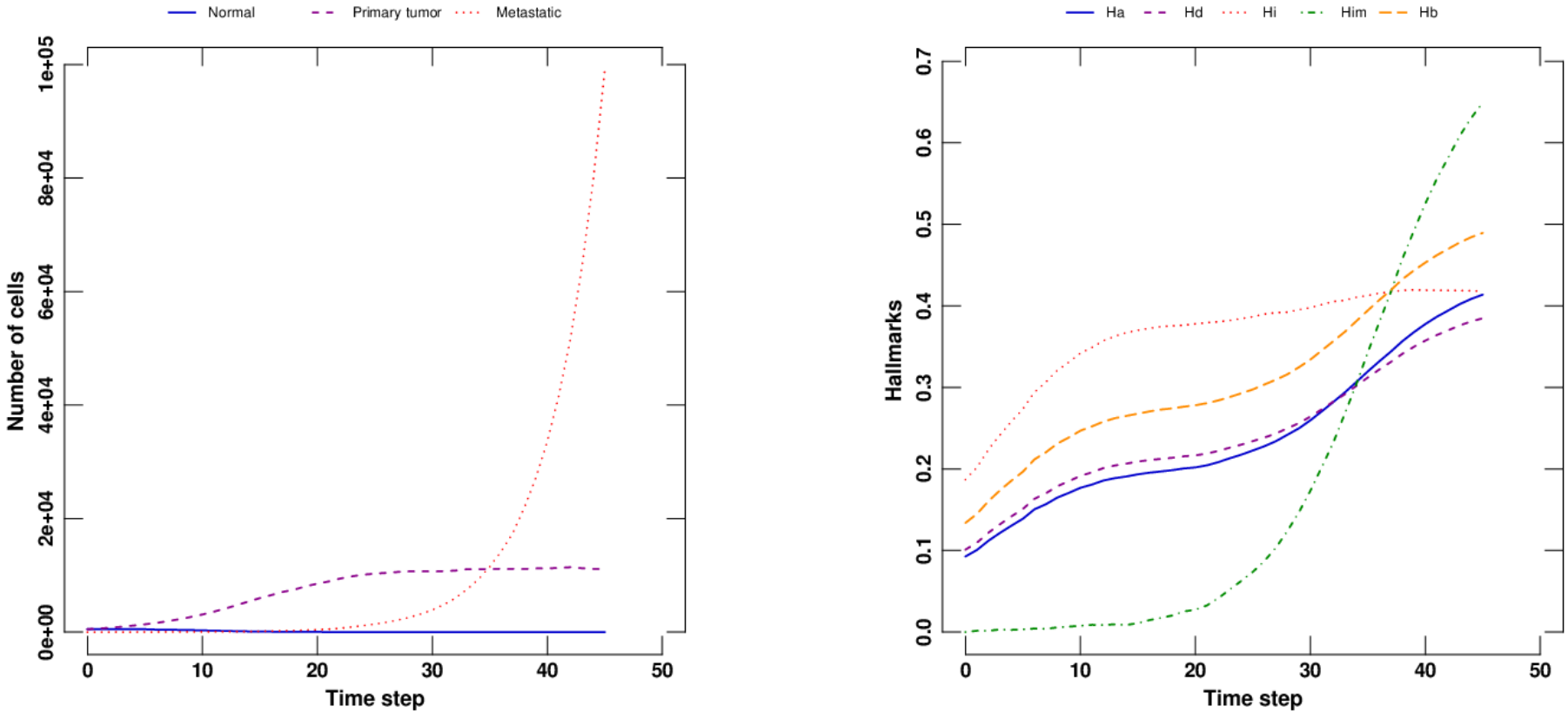
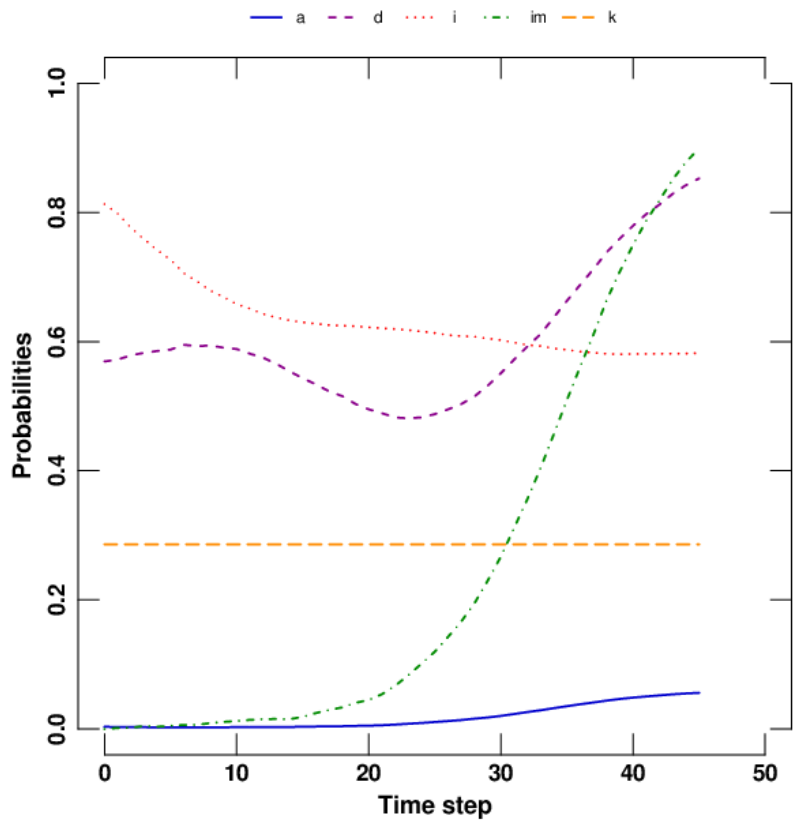


Figure 1: Results of the simulation: left - evolution of number of cells, right - evolution of hallmarks. Files are "Cells evolution.pdf" and "Hallmarks.pdf"

The right side Fig.2 shows the evolution of mutation rate. Fig.3 shows the evolution of clones with different separation of them: for ‘large’ and ‘small’ clones with large and small number of cells respectively.

Final figure shows the list of ‘order of genes dysfunction’ sorted with corresponding number of cells.



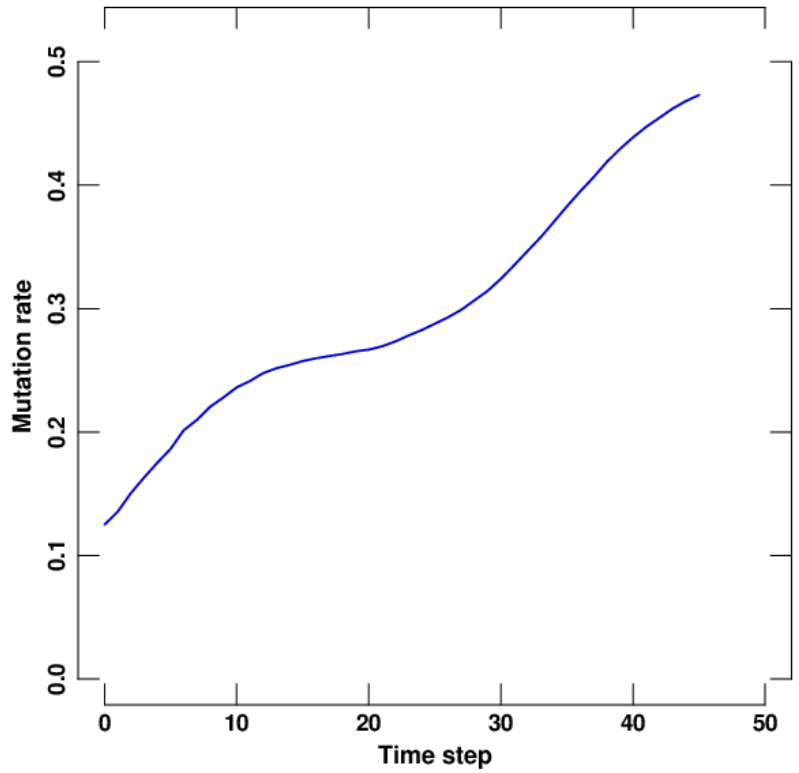


Figure 2: Results of the simulation: left - evolution of probabilities, right - evolution of average mutation rate. Files are "Probabilities.pdf" and "Mutation rate.pdf".

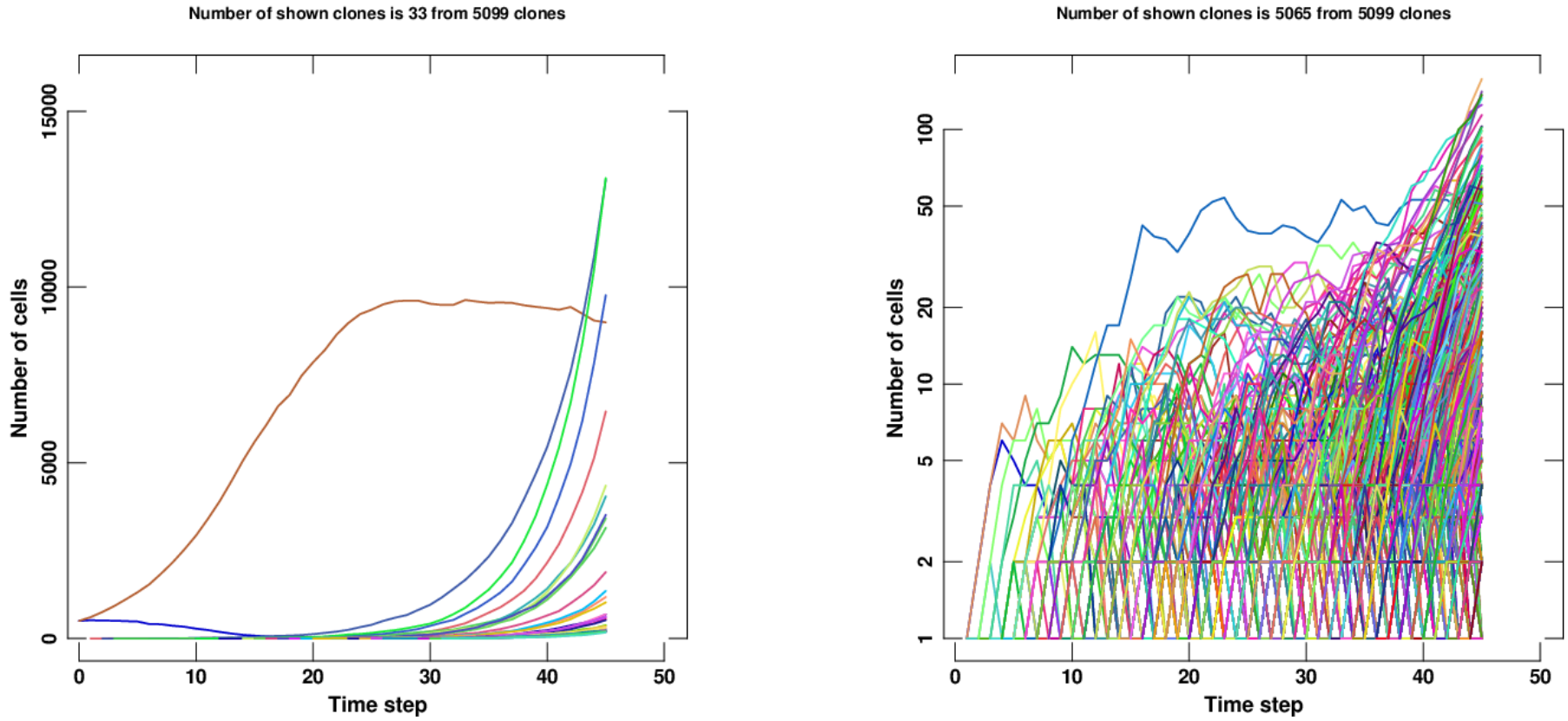


Figure 3: Results of the simulation: left - evolution of number of cells in clones for "large" clones, right - evolution of number of cells in clones for "small" clones (log scale). Files are "Large clones.pdf" and "Small clones.pdf".

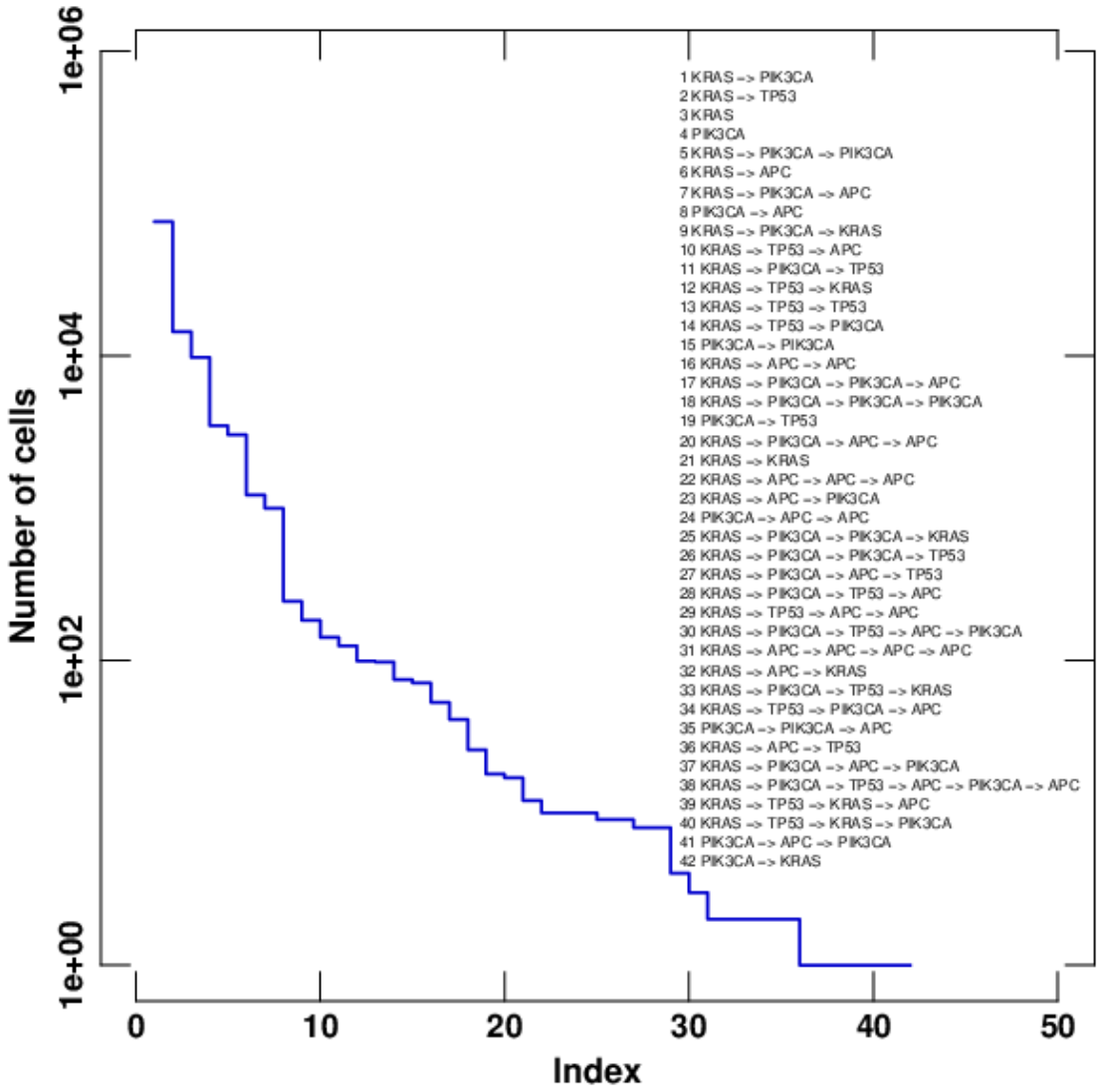


Figure 4: Results of the simulation: order of genes dysfunction as sorted histogram of number of cells for each unique value.

5. Relation to experimental measurements

We here list variables processed from the tugHall outputs that are related to experimental measurements.

Variables processed from the simulator outputs	Relation to experimental measurements
Number of cells	Observed tumor size. 10^9 cells correspond to the tumor tissue diameter of 1 cm. 10^{12} cells correspond to that of 10 cm. $10^{12} - 10^{13}$ cells correspond to lethal burden. See Friberg and Mattson, Journal of Surgical Oncology, 1997.
VAF	VAF calculated from sequence reads in the next-generation sequencer (NGS) under the assumption of 100% tumor purity.
Mutation number per base-pairs	Tumor mutation burden calculated from NGS data.
Number of clones	Number of clones estimated from NGS data by computational tools such as SciClone (Miller et al, PLOS Computational Biology, 2014) and SubClonalSelection (Williams et al, Nature Genetics, 2018).