Simulating pathways, mutual exclusivity, et al.

Raquel Blanco

1/15/2021

Contents

1 Mutal exclusivity and modules

1.1 A probabilistic model of mutually exclusive linearly ordered driver pathways

Mohaghegh Neyshabouri et al. (1) propose a probabilistic model of mutually exclusive linearly ordered driver pathways and analyze two large datasets of colorectal adenocarcinoma (COADREAD) and glioblastoma (GBM) from IntOGen-mutations database. Their model assumes driver genes are over-represented among those mutated across a large tumor collection and, thus, they can be identified in terms of frequency. Also, those participating of the same pathway are mutated in a mutually exclusive manner because more than one mutation in a pathway does not give any selective advantage to the clone.

Like with previous generative models, we map the COADREAD and GMB generative models to actual evolutionary models using different **OncoSimulR** functionalities. This time, we extent what authors model using the mutator and frequency-dependent fitness specifications, to illustrate how differently fitness landscapes evolve even though they are built from exact CPMs when we consider these additional evolutionary phenomena.

1.1.1 Colorectal adenocarcinoma (COADREAD) dataset

1.1.1.1 Modelling mutual exclusivity and order restrictions Figure 7.C from (1) shows the CPM inferred from the COADREAD dataset, consisting of seven modules with between 1 to 4 genes each. The model clearly reconstructs the well-known initiator events in colorectal cancer, including mutations in *APC*, *TP53* and *KRAS* **???**. Using the DAG of restrictions as starting point[1], the evolutionary model is created specifying same genotype fitness for all modules as authors do not state any differences in fitness for when the restrictions in the DAG are satisfied (`s`). However, based on the confidence parameter used by the authors to assess the reliability of modeled restrictions, different fitness are set when the DAG of restrictions is not satisfied (`sh`) (Table 1). Since this method reconstructs linear models (*i.e.* oncogenic trees), there is no need to specify any particular type of dependency between modules (`typeDep`), so we set it to monotonic (MN) as it is a mandatory argument for `allFitnessEffects` function.

[1] Due to memory exhaustion, the following genes from the dataset have been removed: FAT4 (module E), CTNNB1 (module F), TCF7L2 (module E)

Table 1: confidence parameter for each module transition

Module	Confidence parameter (%)
APC	100
TP53	100
KRAS	100
PIK3CA, *NRAS*, *LRP1B*	100
FBXW7, *TCF7L2*, *FAT4*, *ARID1A*	87.7
ATM, *SMAD2*, *ERBB3*, *MTOR*, *CTNNB1*	86.9
SOX9, *SMAD4*	66.7

```r
# Loading library (REMOVE)
library(OncoSimulR)
```

```r
## Restriction table, including DAG of restrictions specifications and associated fitness
COADREAD_rT <- data.frame(parent = c("Root", "A", "B", "C", "D", "E", "F"), # Parent nodes
                          child = c("A", "B", "C", "D", "E", "F", "G"), # Child nodes
                          s = 0.5,
                          sh = c(rep(-1, 4), rep(-.5, 2), -.2),
                          typeDep = "MN")


## Create fitness specifications from DAG of restrictions considering modules
COADREAD_fitness <- allFitnessEffects(COADREAD_rT,
                                      geneToModule = c( "Root" = "Root",
                                                        "A" = "APC",
                                                        "B" = "TP53",
                                                        "C" = "KRAS",
                                                        "D" = "PIK3CA, NRAS, LRP1B",
                                                        "E" = "FBXW7, ARID1A",
                                                        "F" = "ATM, SMAD2, ERBB3, MTOR",
                                                        "G" = "SOX9, SMAD4")) # Modules


## DAG of restrictions representation
plot(COADREAD_fitness, expandModules = TRUE, autofit = TRUE)
```

```r
# Evaluation of all possible genotypes fitness under the previous fitness specifications
COADREAD_FL <- evalAllGenotypes(COADREAD_fitness, max = 131072)

# Fitness landscape representation
plotFitnessLandscape(COADREAD_FL)
```

maybe add here an extended explanation (?)

1.1.1.2 Simplified cancer progression model As we did we previous models, for illustrating purposes we designed a simplified version of the CPM by (1) to explore the relationships between modules. Considering

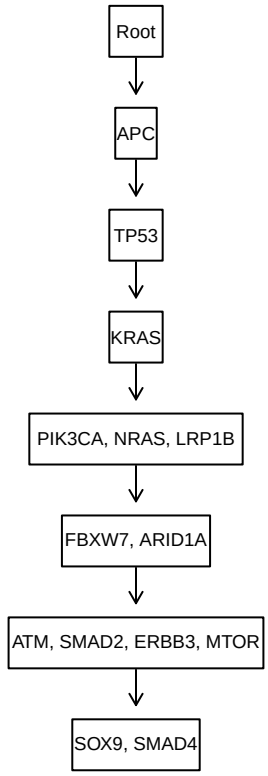

Figure 1: DAG of restrictions for the COADREAD dataset

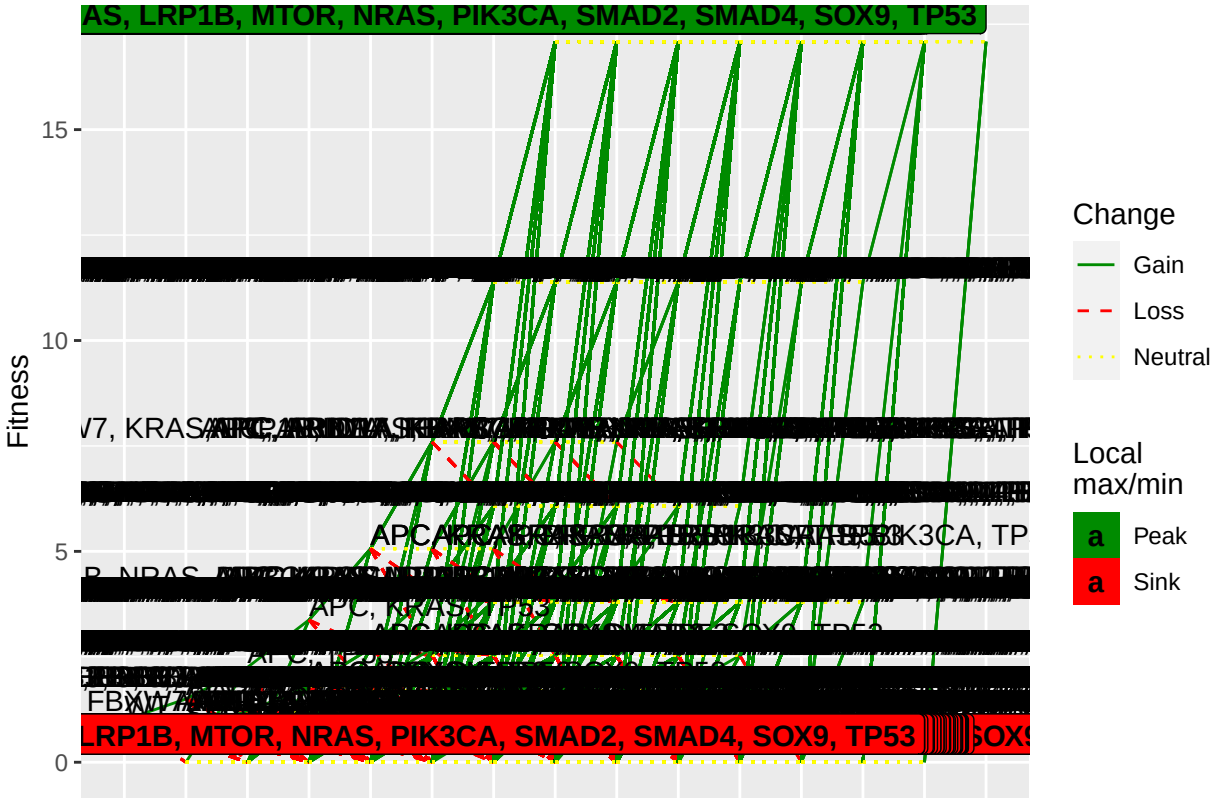

Figure 2: Fitness landscape corresponding with the DAG of restrictions for the COADREAD dataset

each module represents a set of mutually exclusive genes, this simplified version of the model assumes a scenario in which theses genes never mutate at the same time, and thus those genotypes never exist. This way, modules can incorporate a single gene and the relationships between them can be more clearly visualized both in the fitness landscape and in the simulations. Also, we can analyze the different evolutionary scenarios arising from the same set of restrictions but happening for different mutated genes in each module.

In the simplified version of the CPM, we use the dataframe `COADREAD_rT` without the `geneToModule` specification in the `allFitnessEffects` function. Thus, we visualize "module genotypes" instead of genes. In this first example we assume that the effect each gene in a module has on fitness is the same.

```
## Create fitness specifications from somplified DAG of restrictions
COADREADsim_fitness <- allFitnessEffects(COADREAD_rT)

## Simplified DAG of restrictions representation
plot(COADREADsim_fitness)
```

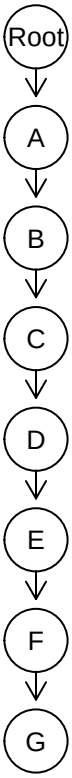

Figure 3: Simplified DAG of restrictions for the COADREAD dataset

```
# Evaluation of all possible genotypes fitness under the previous fitness specifications
COADREADsim_FL <- evalAllGenotypes(COADREADsim_fitness)

# Fitness landscape representation
plotFitnessLandscape(COADREADsim_FL)
```

The fitness landscape in Figure 4 more clearly shows how fitness increases with the accumulation of mutations in the order specified in the DAG of restrictions (maybe discuss this a bit more).

Next, we use the simplified fitness landscape to simulate tumor progression for one individual with the `oncoSimulIndiv` functionality.

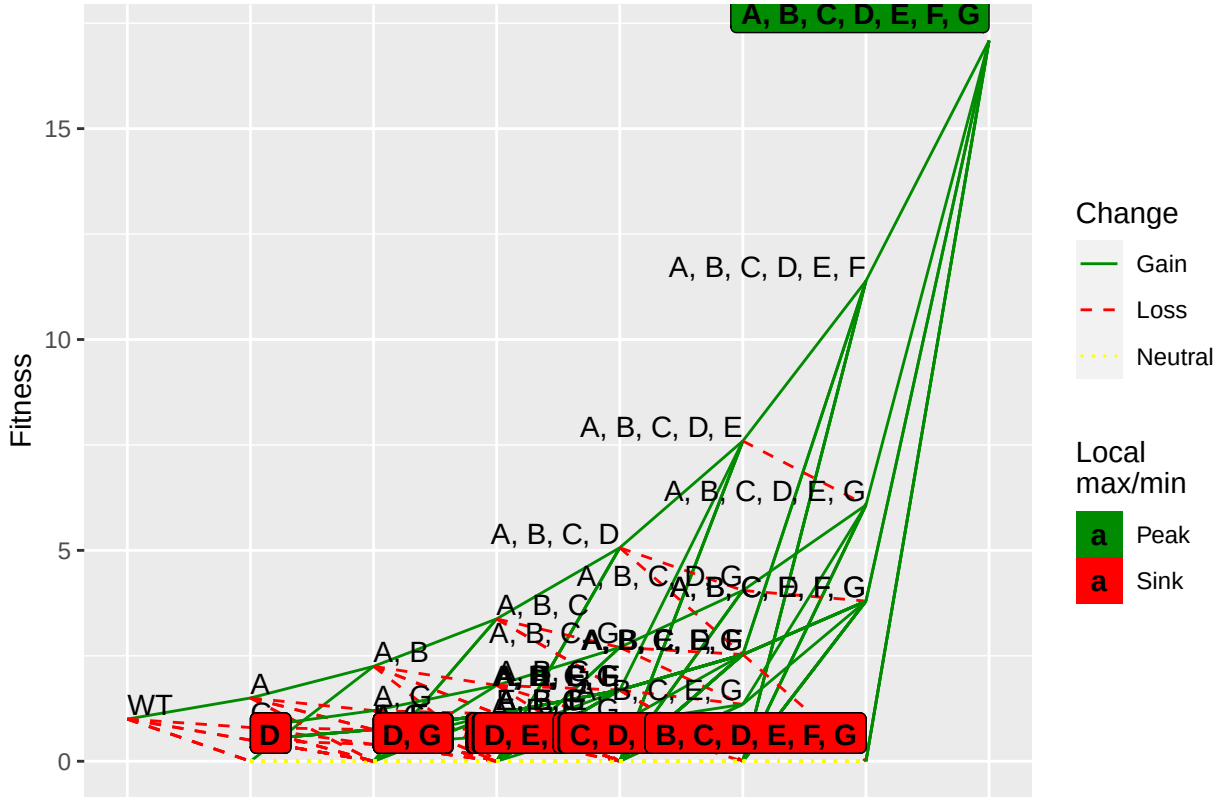

Figure 4: Fitness landscape corresponding with the simplified DAG of restrictions for the COADREAD dataset

```
COADREADsim_Simul <- oncoSimulIndiv(COADREADsim_fitness,
                            model = "McFL", ## Model used
                            mu = 1e-4, ## Mutation rate
                            sampleEvery = 0.02, ## How often the whole population is sampled
                            keepEvery = 1,
                            initSize = 2000, ## Initial population size
                            finalTime = 200,
                            keepPhylog = TRUE, ## Allow to see parent-child relationships
                            onlyCancer = FALSE)

## Plot of simulation for genotypes
plot(COADREADsim_Simul,
show = "genotypes",
type = "stacked")
```

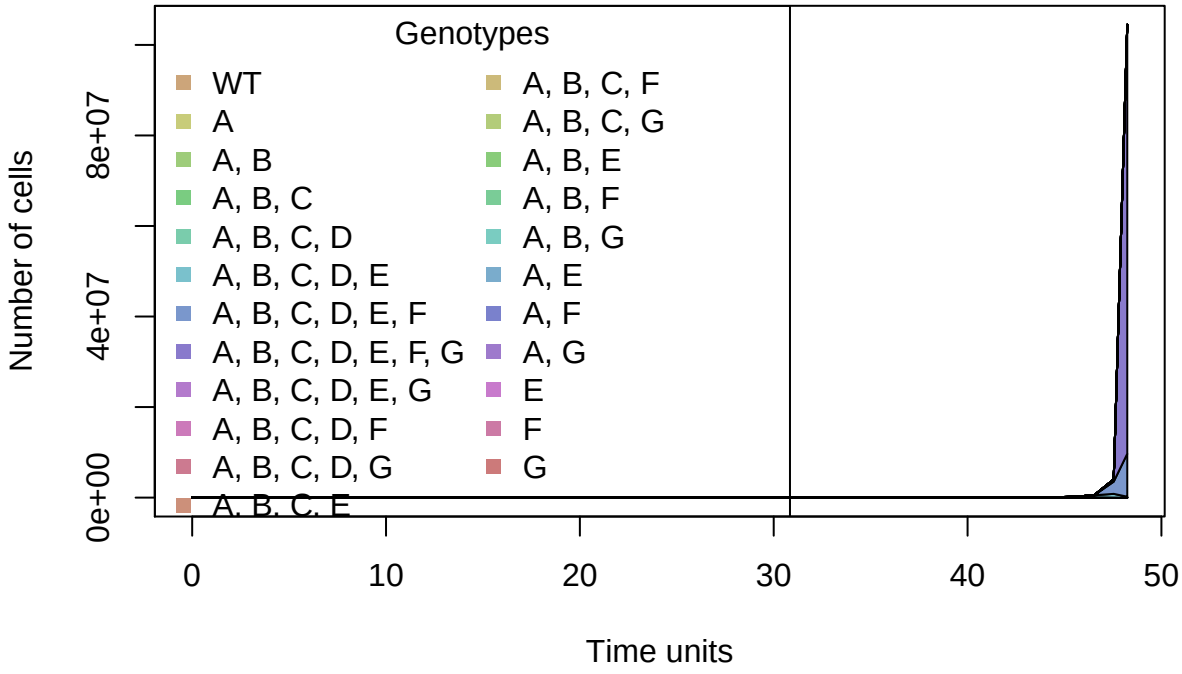

Figure 5: Simulation of cancer progression using the fitness landscape of the simplied model for the COAD-READ dataset (stacked plot)

```
plot(COADREADsim_Simul,
     show = "genotypes",
     type = "line"
)
```

```
## Parent-child relationship derived from simulation
plotClonePhylog(COADREADsim_Simul,
                N = 0, ## Specify clones that exist
                keepEvents = TRUE ## Arrows showing how many times each clones appeared
)
```

1.1.1.3 Frequency-dependent fitness Clones that coexist in a tumor can influence the fitness of each other in a frequency-dependent manner when a mutation produces a phenotype able to modulate the tumor microenvironment. OncoSimulR incorporates the `frecuencyDependentFitness` specification to allow for

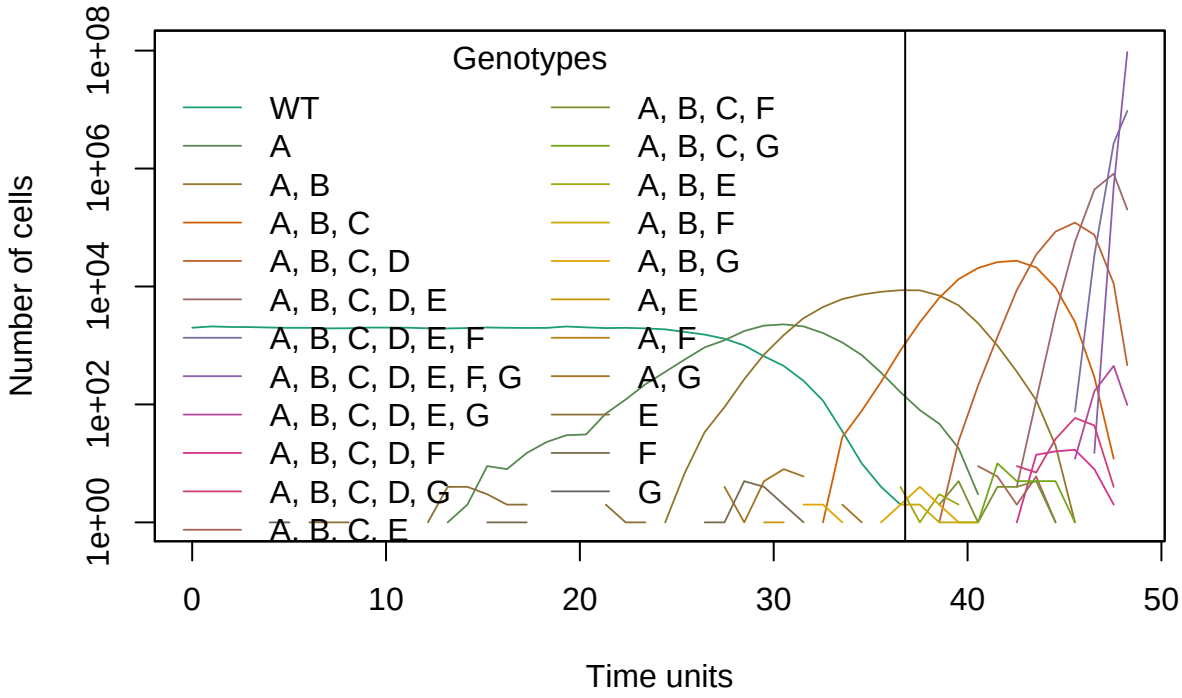

Figure 6: Simulation of cancer progression using the fitness landscape of the simplified model for the COAD-READ dataset (line plot)

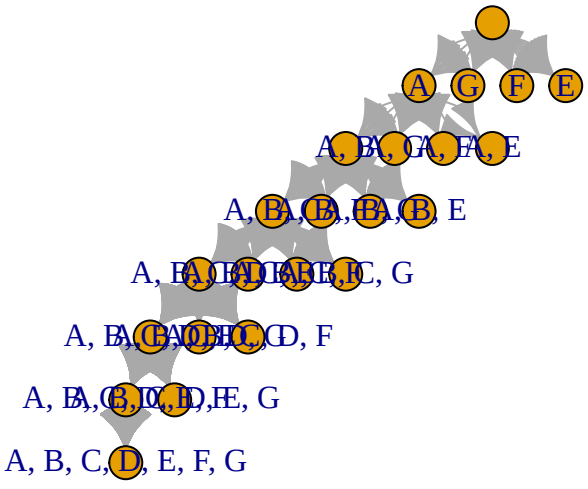

Figure 7: Parent-child relationship derived from simulation

modelling interaction among clones during tumor progression. To further explore this option, we bring back the complete COADREAD evolutionary model and zoom into its five first nodes to model a scenario in which all the possible node-five genotypes coexist at a time and influence each other. For simplicity, we are not using modules this time.

```
# Mapping of genotypes to frequency-dependent fitness
COADREAD5_gen <- data.frame(Genotype = c("APC, TP53, KRAS",
                                         "APC, TP53, KRAS, PIK3CA",
                                         "APC, TP53, KRAS, NRAS",
                                         "APC, TP53, KRAS, LRP1B"),
                            Fitness = c("1 - (f_APC_TP53_KRAS_PIK3CA + f_APC_TP53_KRAS_NRAS +
                                             f_APC_TP53_KRAS_LRP1B)",
                                        "2 + 1.5 * f_APC_TP53_KRAS_LRP1B",
                                        "2 + 1.1 * f_APC_TP53_KRAS_LRP1B",
                                        "2 - f_APC_TP53_KRAS_NRAS"))


## Evaluate all genotypes considering population sizes of the clones
COADREAD5_FL <- evalAllGenotypes(allFitnessEffects(genotFitness = COADREAD5_gen,
                                                   frequencyDependentFitness = TRUE,
                                                   frequencyType = "rel"),
                                 spPopSizes = c("APC, TP53, KRAS" = 100,
                                                "APC, TP53, KRAS, PIK3CA" = 20,
                                                "APC, TP53, KRAS, NRAS" = 20,
                                                "APC, TP53, KRAS, LRP1B" = 30))


# Fitness landscape representation
plotFitnessLandscape(COADREAD5_FL)
```

Glioblastoma (GBM) dataset

1. Neyshabouri MM, Jun SH, Lagergren J. Inferring tumor progression in large datasets. PLoS Computational Biology. 2020;16(10):1–16. Available from: http://dx.doi.org/10.1371/journal.pcbi.1008183

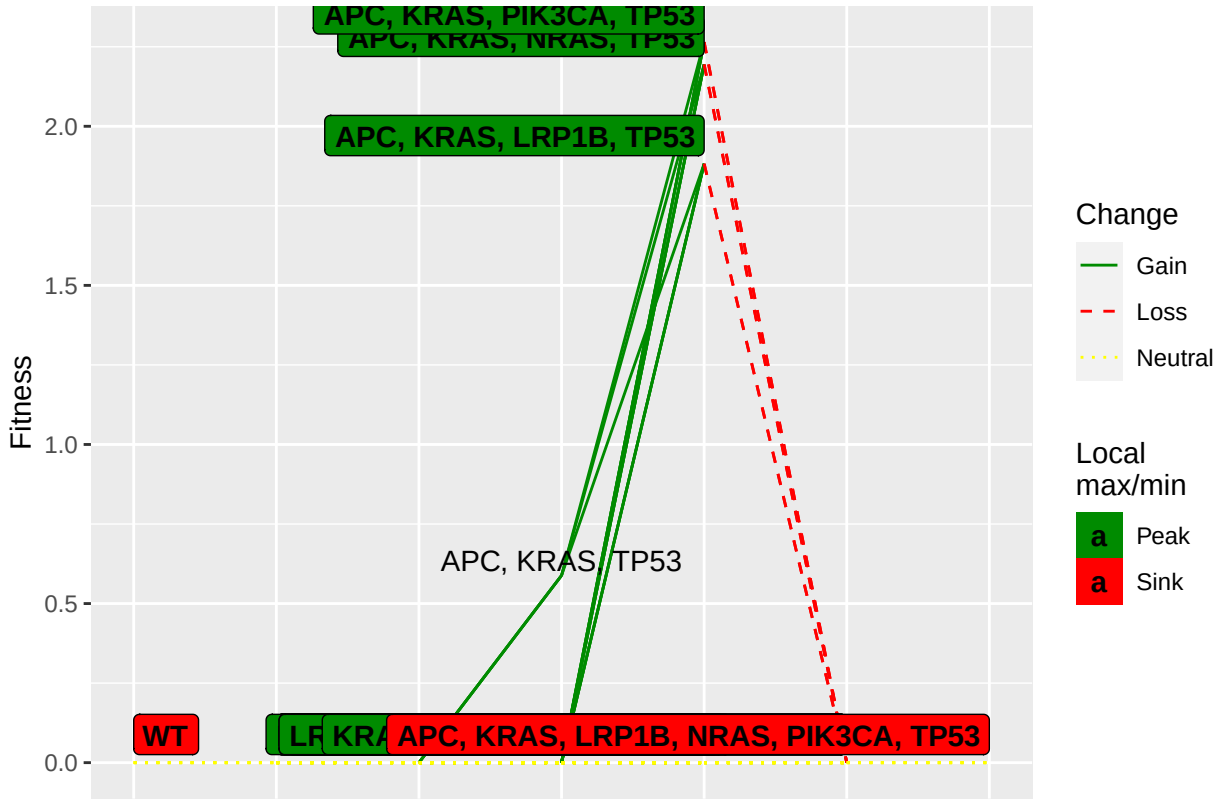

Figure 8: Fitness landscape corresponding with the first-five-nodes possible genotyoes for the COADREAD dataset accounting for frequency-dependent fitness