

# Package ‘SMASH’

May 7, 2026

**Type** Package

**Title** Subclone Multiplicity Allocation and Somatic Heterogeneity

**Version** 1.0.0

**Date** 2025-02-07

**Description**

Cluster user-supplied somatic read counts with corresponding allele-specific copy number and tumor purity to infer feasible underlying intra-tumor heterogeneity in terms of number of sub-clones, multiplicity, and allocation (Little et al. (2019) <[doi:10.1186/s13073-019-0643-9](https://doi.org/10.1186/s13073-019-0643-9)>).

**License** GPL (>= 3)

**Imports** Rcpp, stats, smarter, reshape2, ggplot2

**LinkingTo** Rcpp, RcppArmadillo

**Encoding** UTF-8

**LazyData** true

**Depends** R (>= 2.10)

**RoxygenNote** 7.2.3

**Suggests** knitr, devtools

**VignetteBuilder** knitr

**NeedsCompilation** yes

**Author** Paul Little [aut, cre]

**Maintainer** Paul Little <[pllittle321@gmail.com](mailto:pllittle321@gmail.com)>

**Repository** CRAN

**Date/Publication** 2025-02-27 16:40:06 UTC

## Contents

eS	2
gen_ITH_RD	2
gen_subj_truth	3
grid_ITH_optim	3
ITH_optim	4
vis_GRID	6

**Index**[7](#)


---

eS	<i>A collection of pre-defined subclone configurations.</i>
----	-------------------------------------------------------------

---

**Description**

A R list containing subclone configurations in matrix form for 1 to 5 subclones. For each matrix, each column corresponds to a subclone and each row corresponds to a variant's allocation across all subclones. For example, the first row of each matrix is a vector of 1's to represent clonal variants, variants present in all subclones.

**Usage**

```
eS
```

**Format**

An object of class `list` of length 5.

---

gen_ITH_RD	<i>gen_ITH_RD</i>
------------	-------------------

---

**Description**

Simulates observed alternate and reference read counts

**Usage**

```
gen_ITH_RD(DATA, RD)
```

**Arguments**

DATA	The output data.frame from <code>gen_subj_truth</code>
RD	A positive integer for the mean read depth generated from the negative binomial distribution

**Value**

A matrix of simulated alternate and reference read counts.

---

gen_subj_truth	<i>gen_subj_truth</i>
----------------	-----------------------

---

**Description**

Simulates copy number states, multiplicities, allocations

**Usage**

```
gen_subj_truth(mat_eS, maxLOCI, nCN = NULL)
```

**Arguments**

mat_eS	A subclone configuration matrix pre-defined in R list eS
maxLOCI	A positive integer number of simulated somatic variant calls
nCN	A positive integer for the number of allelic copy number pairings to sample from. If NULL, it will be randomly sampled between 1 and 5.

**Value**

A list containing the following components:

subj_truth	dataframe of each variant's simulated minor (CN_1) and major (CN_2) copy number states, total copy number (tCN), subclone allocation (true_A), multiplicity (true_M), mutant allele frequency (true_MAF), and cellular prevalence (true_CP)
purity	tumor purity
eta	the product of tumor purity and subclone proportions
q	vector of subclone proportions

---

grid_ITH_optim	<i>grid_ITH_optim</i>
----------------	-----------------------

---

**Description**

This function performs a grid search over enumerated configurations within the pre-defined list eS

**Usage**

```
grid_ITH_optim(
  my_data,
  my_purity,
  list_eS,
  pi_eps0 = NULL,
  trials = 20,
  max_iter = 4000,
  my_epsilon = 1e-06
)
```

**Arguments**

my_data	A R dataframe containing the following columns: tAD tumor alternate read counts tRD tumor reference read counts CN_1 minor allele count CN_2 major allele count, where CN_1 <= CN_2 tCN CN_1 + CN_2
my_purity	A single numeric value of known/estimated purity
list_eS	A nested list of subclone configuration matrices
pi_eps0	A user-specified parameter denoting the proportion of loci not explained by the combinations of purity, copy number, multiplicity, and allocation. If NULL, it is initialized at 1e-3. If set to 0.0, the parameter is not estimated.
trials	Positive integer, number of random initializations of subclone proportions
max_iter	Positive integer, preferably 1000 or more, setting the maximum number of iterations
my_epsilon	Convergence criterion threshold for changes in the log likelihood, preferably 1e-6 or smaller

**Value**

A R list containing two objects. GRID is a dataframe where each row denotes a feasible subclone configuration with corresponding subclone proportion estimates  $q$  and somatic variant allocations  $alloc$ . INFER is a list where INFER[[ $i$ ]] corresponds to the  $i$ -th row or model of GRID.

---

 ITH\_optim

*ITH\_optim*


---

**Description**

Performs EM algorithm for a given configuration matrix

**Usage**

```
ITH_optim(
  my_data,
  my_purity,
  init_eS,
  pi_eps0 = NULL,
  my_unc_q = NULL,
  max_iter = 4000,
  my_epsilon = 1e-06
)
```

**Arguments**

<code>my_data</code>	A R dataframe containing the following columns: <code>tAD</code> tumor alternate read counts <code>tRD</code> tumor reference read counts <code>CN_1</code> minor allele count <code>CN_2</code> major allele count, where $CN_1 \leq CN_2$ <code>tCN</code> $CN_1 + CN_2$
<code>my_purity</code>	A single numeric value of known/estimated purity
<code>init_eS</code>	A subclone configuration matrix pre-defined in R list <code>eS</code>
<code>pi_eps0</code>	A user-specified parameter denoting the proportion of loci not explained by the combinations of purity, copy number, multiplicity, and allocation. If NULL, it is initialized at $1e-3$ . If set to 0.0, the parameter is not estimated.
<code>my_unc_q</code>	An optimal initial vector for the unconstrained <code>q</code> vector, useful after running <code>grid_ITH_optim</code> . If this variable is NULL, then the subclone proportions, <code>q</code> , are randomly initialized. For instance, if <code>my_unc_q = (x1, x2)</code> , then $q = (\exp(x1) / (1 + \exp(x1) + \exp(x2)), \exp(x2) / (1 + \exp(x1) + \exp(x2)), 1 / (1 + \exp(x1) + \exp(x2)))$ .
<code>max_iter</code>	Positive integer, preferably 1000 or more, setting the maximum number of iterations
<code>my_epsilon</code>	Convergence criterion threshold for changes in the log likelihood, preferably $1e-6$ or smaller

**Value**

If the EM algorithm converges, the output will be a list containing

<code>iter</code>	number of iterations
<code>converge</code>	convergence status
<code>unc_q0</code>	initial unconstrained subclone proportions parameter
<code>unc_q</code>	unconstrained estimate of <code>q</code>
<code>q</code>	estimated subclone proportions among cancer cells
<code>CN_MA_pi</code>	estimated mixture probabilities of multiplicities and allocations given copy number states
<code>eta</code>	estimated subclone proportion among tumor cells
<code>purity</code>	user-inputted tumor purity
<code>entropy</code>	estimated entropy
<code>infer</code>	A R dataframe containing inferred variant allocations ( <code>infer_A</code> ), multiplicities ( <code>infer_M</code> ), cellular prevalences ( <code>infer_CP</code> ).
<code>ms</code>	model size, number of parameters within parameter space
<code>LL</code>	The observed log likelihood evaluated at maximum likelihood estimates.
<code>AIC = 2 * LL - 2 * ms</code>	Negative AIC, used for model selection
<code>BIC = 2 * LL - ms * log(LOCI)</code>	Negative BIC, used for model selection
<code>LOCI</code>	The number of inputted somatic variants.

---

`vis_GRID`*vis\_GRID*

---

**Description**

A simple visualization of SMASH's grid of solutions

**Usage**

```
vis_GRID(GRID)
```

**Arguments**

GRID            The GRID object output from `grid_ITH_optim`.

**Value**

A ggplot object for data visualization

# Index

\* **datasets**

eS, [2](#)

eS, [2](#)

gen\_ITH\_RD, [2](#)

gen\_subj\_truth, [3](#)

grid\_ITH\_optim, [3](#)

ITH\_optim, [4](#)

vis\_GRID, [6](#)