DOTS-Finder Input-Output Format

INPUT FORMAT

A full explanation of our MARF format is provided.
If the column is set as MANDATORY, no NA is available and all the fields must be filled with proper values.
If the column is set as Optional, a list of accepted NA is provided.

MARF file

Hugo_Symbol	Entrez_Gene_Id	NCBI_Build	Chromosome	Start_position	End_position	Variant_Classification	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2	dbSNP_RS	Tumor_Sample_Barcode	Protein_Change
HK1	3098	37	10	71142350	71142350	Missense_Mutation	C	T	T	rs number / novel	TCGA-A6-3807-01A-01W-0995-10	p.A462V
Unknown	0	MANDATORY	MANDATORY	MANDATORY	MANDATORY	MANDATORY	MANDATORY	MANDATORY	Equal to allele1	empty string or novel	MANDATORY	empty string , NULL

1) Hugo_Symbol (Optional): HGNC symbol according to HUGO Gene Nomenclature Committee (www.genenames.org)
2) Entrez_Gene_Id (Optional): Entrez Gene number according to the NCBI's repository for gene-specific information (www.ncbi.nlm.nih.gov/gene)
3) NCBI_Build (MANDATORY): Human Genome Reference Sequence Assembly. It must be 37, otherwise see lift_over_marf.py script
4) Chromosome (MANDATORY): chromosome where the mutation was found
5) Start_Position (MANDATORY): position starting point of the mutation according to maf specification (0 based)
6) End_Position (MANDATORY): position ending point of the mutation according to maf specification (0 based)
7) Variant_Classification (MANDATORY): type of mutation according to the TCGA maf file specifications. The IGR (Intragenic Region), Intron and RNA mutation will be excluded during the analysis
3'Flank
3'UTR
5'Flank
5'UTR
De_novo_Start_InFrame
De_novo_Start_OutOfFrame
Frame_Shift_Del
Frame_Shift_Ins
IGR
In_Frame_Del
In_Frame_Ins
Intron
Missense_Mutation
Nonsense_Mutation
Nonstop_Mutation
RNA
Silent
Splice_Region
Splice_Site
Translation_Start_Site
8) Reference_Allele(MANDATORY): reference base on the reference genome corresponding to the start position
9) Tumor_Seq_Allele1(MANDATORY): first strand base(s) called in tumor sample
10) Tumor_Seq_Allele2(Optional): second strand base(s) called in tumor sample
11) dbSNP_RS(Optional): rs number if the mutation is also found in the dbSNP database (novel or empty cell otherwise)
12) Tumor_Sample(MANDATORY): unique barcode for the sample/patient in which the mutation was found
13) Protein_Change(Optional): amino acid change in HGVS nomenclature (www.hgvs.org/mutnomen/recs-prot.html)

File Conversion

Any MAF file can be easily converted to MARF in case is malformed or corrupted as long as all the 13 columns specified above. The program accept MAF standard 2.3 and 2.4

A VCF file can be converted to a MARF, but must be annotated using program like Annovar or Oncotator.

A CSV from Annovar can be converted to a MARF following these guidelines:

For column choice, follow this simple header conversion:

MARF - header	Annovar - header	MATCH
Hugo_Symbol	Gene	CORRESPONDENCE
Entrez_Gene_Id	NONE	SET TO '0' OR PROVIDE Hugo to Entrez conversion
NCBI_Build	NONE	SET TO 36/37
Chromosome	Chr	CORRESPONDENCE
Start_Position	Start	CORRESPONDENCE
End_Position	End	CORRESPONDENCE
Variant_Classification	Func + ExonicFunc	FOLLOW CONVERSION MATRIX
Reference_Allele	Ref	CORRESPONDENCE
Tumor_Seq_Allele1	Obs	CORRESPONDENCE
Tumor_Seq_Allele2	NONE	SET EQUAL TO Obs/Tumor_Seq_Allele1
dbSNP_RS	dbSNP135 , dbSNP137 , dbSNP#	CORRESPONDENCE
Tumor_Sample_Barcode	NONE(taken by file name when merged)	SET WHEN SAMPLES ARE AGGREGATE
Protein_Change	AAChange	CORRESPONDENCE

For most of the columns there is a perfect one to one match.
The only column where you have to pay attention is the Variant_Classification. The matrix below provide a simple way to obtain acceptable value for MARF format:

Func - Annovar	ExonicFunc - Annovar	Variant_Classification - MARF
exonic	frameshift insertion	Frame_Shift_Ins
exonic	frameshift deletion	Frame_Shift_Del
exonic	frameshift block substitution	Frame_Shift_Del OR Frame_Shift_Ins
exonic	stopgain	Nonsense_Mutation
exonic	stoploss	Nonstop_Mutation
exonic	nonframeshift insertion	In_Frame_Ins
exonic	nonframeshift deletion	In_Frame_Del
exonic	nonframeshift block substitution	In_Frame_Del OR In_Frame_Ins
exonic	nonsynonymous SNV	Missense_Mutation
exonic	synonymous SNV	Silent
exonic	unknown	NO CORRESPONDENCE
splicing	-	Splice_Site
ncRNA	-	RNA
UTR5	-	5'UTR
UTR3	-	3'UTR
intronic	-	Intron
upstream	-	IGR
downstream	-	IGR
intergenic	-	IGR
exonicsplicing	-	Splice_Site
ncRNA_UTR3	-	3'Flank
ncRNA_UTR5	-	5'Flank
ncRNA_exonic	-	RNA
ncRNA_intronic	-	Intron
upstream;downstream	-	IGR
UTR5;UTR3	-	5'UTR OR 5'UTR
ncRNA_UTR5;ncRNA_UTR3	-	NO CORRESPONDENCE
ncRNA_splicing	-	Splice_Region

For the NO CORRESPONDENCE value, there is no possible translation and therefore those rows must be removed.

OUTPUT FORMAT

Pfx_table.txt

An aggregated table formed by 49 columns. One row per gene.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Tumor	Patients	Gene	Entrez	Chrom	Size	Tot_Freq	NS_freq	Tot_Mutation	NS_Mutation	S_Mutation	Unique_NS	Unique_S	SNP	SNP_freq_unique	Mean_mut_btw_patient	Mean_mutNS_btw_patient	Mean_mutS_btw_patient
LUSC	177	PIK3CA	5290	3	3709	0.15819209	0.152542373	30	29	1	16	1	16	0.235294118	455.1785714	339	169

1) Tumor: tumor/filename
2) Patients: number of samples analyzed and present in the maf file proposed
3) Gene: unique identifier of the line by its hugo name
4) Entrez: unique identifier of the line by its entrez name
5) Chrom: chromosome number in which the gene is
6) Size: minimum length of all the exome component of the gene among all its transcript
7) Tot_Freq: total_sample_mutated / total_sample
8) NS_freq: total_number_of_sample_mutated_non_silently / total_number_of_sample
9) Tot_Mutation: total number of mutations found on the gene
10) NS_Mutation: total number of non silent mutation found on the gene
11) S_Mutation: total number of silent mutation found on the gene
12) Unique_NS: total number of single spots for non silent mutations
13) Unique_S: total number of single spots for silent mutations
14) SNP: total number of single spots mutations with rs reference on dbSNP
15) SNP_freq_unique: number of rs referenced mutation spots / total mutation spots
16) Mean_mut_btw_patient: average number of total mutation among samples with that gene mutated
17) Mean_mutNS_btw_patient: average number of NS mutation among samples with that gene mutated
18) Mean_mutS_btw_patient: average number of S mutation among samples with that gene mutated

19	20	21	22	23
Mean_NS_Tot_rate	Mean_trunc_btw_patient	Mean_miss_btw_patient	Mean_trunc_RATE_btw_patient	Mean_miss_RATE_btw_patient
0.755936916	7.171428571	53.58490566	0.09336381	0.679580505

19) Mean_NS_Tot_rate: average NS/Tot ratio among samples with that gene mutated
20) Mean_trunc_btw_patients: average number of truncating type mutations per samples with that gene mutated (De_novo_Start_OutOfFrame , Frame_Shift_Del , Frame_Shift_In , Nonsense_Mutation , Nonstop_Mutation , Splice_Region , Splice_Site , Translation_Start_Site)
21) Mean_miss_btw_patients: average number of missense type mutations per samples with that gene mutated (De_novo_Start_InFrame , In_Frame_Del , In_Frame_Ins , Missense_Mutation)
22) Mean_trunc_RATE_btw_patients: average number of truncating type mutations over total mutations per samples with that gene mutated (De_novo_Start_OutOfFrame , Frame_Shift_Del , Frame_Shift_In , Nonsense_Mutation , Nonstop_Mutation , Splice_Region , Splice_Site , Translation_Start_Site)
23) Mean_miss_RATE_btw_patients: average number of missense type mutations over total mutations per samples with that gene mutated (De_novo_Start_InFrame , In_Frame_Del , In_Frame_Ins , Missense_Mutation)

24	25	26	27	28	29	30	31	32	33	34	35	36	37
SNV	InDel_Double	AC	AG	AT	CA	CG	CT	GA	GC	GT	TA	TC	TG
30	0	0	2	2	0	0	0	20	1	4	0	1	0

24) SNV: number of single nucleotide variation type of mutation
25) InDel_Double: number of Insertion , Deletion or double nucleotide variations
26-37) AC-TG: number of SNV divided by the actual change of base (A to C , A to G .... T to G)

38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
3Flank	3UTR	5Flank	5UTR	De_novo_Start_InFrame	De_novo_Start_OutOfFrame	Frame_Shift_Del	Frame_Shift_Ins	In_Frame_Del	In_Frame_Ins	Missense_Mutation	Nonsense_Mutation	Nonstop_Mutation	Silent	Splice_Region	Splice_Site	Translation_Start_Site
0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0

38-54) De_novo_Start_InFrame - Translation_Start_Site: number of mutations divided by their Variant_Classification (see INPUT FORMAT -> MARF file)

55	56	57	58	59	60
OncoGene_Entropy_Score	TSG_Score	MissenseType	TruncatingType	Domains_Gini_Index	Most_Affected_Domain
3.900522913	37.08389346	54	70	1	(64, 203)

55) OncoGene_Entropy_Score: see reference paper
56) TSG_Score: see reference paper
57) MissenseType: number of mutations of type missense (De_novo_Start_InFrame , In_Frame_Del , In_Frame_Ins , Missense_Mutation)
58) TruncatingType: number of mutations of type truncating (De_novo_Start_OutOfFrame , Frame_Shift_Del , Frame_Shift_In , Nonsense_Mutation , Nonstop_Mutation , Splice_Region , Splice_Site , Translation_Start_Site)
59) Domains_Gini_Index homogeneity Gini Index where 0 represents no aggregation in a particular superfamily domain while 1 is complete aggregation
60) Most_Affected_Domain amino acid span where the majority of mutations was found

Pfx_patdistr.txt

A summary table for samples/patients mutations distribution.

1	2	3	4	5	6	7
Tumor	Patient	All_mutation	Nonsilent_mutation	Silent_mutation	Missense	Truncating
LUAD	TCGA-BP-4963-01A-01D-1462-08	90	80	10	70	20

1) Tumor: Tumor/filename
2) Patient: sample/patient unique identifier
3) All_mutation: number of total mutation found in the sample/patient
4) Nonsilent_mutation: number of non silent mutation found in the sample/patient
5) Silent_mutation: number of silent mutation found in the sample/patients (Silent)

6) Missense: number of mutation of type missense (De_novo_Start_InFrame , In_Frame_Del , In_Frame_Ins , Missense_Mutation)
7) Truncating number of mutations of type truncating (De_novo_Start_OutOfFrame , Frame_Shift_Del , Frame_Shift_In , Nonsense_Mutation , Nonstop_Mutation , Splice_Region , Splice_Site , Translation_Start_Site)

Pfx_metadata.txt

Some general information about mean, median and variance of the amount of mutation per sample/patient and per gene. Column names self explained.

Pfx_TSG_Driver.txt
Pfx_OncoGene_Driver.txt

The original Pfx_table.txt file + some new columns . Not all the genes are listed, just the ones that pass through the functional score threshold.

61	62	63	64	65	66	67	68	69	70	71	72
Mean_NS_freq_TCGA	Mean_NS_freq_Cosmic	Expected_NS	Expected_S	Annotated_SNP	Cancer_Gene_Census	p_higherfreq	p_ACGT	p_highertumorfreq	p_FI_Total	p_FI_Onco	Global_P_Value
0.017524339	0.02852349	19	5	48	1	2.01E-16	0.009454287	4.77E-09	1.01E-05	0.010498047	0

61) Mean_NS_freq_TCGA: average non silent mutation frequency across different kind of tumors (calculated from TCGA data)
62) Mean_NS_freq_Cosmic: average non silent mutation frequency across different kind of tumors (calculated from COSMIC database)
63) Expected_NS: number of non silent mutation expected by the 79 rule
64) Expected_S: number of silent mutation expected by the 79 algorithm
65) Annotated_SNP: total number of SNP listed for that gene in dbSNP
66) Cancer_Gene_Census: 0/1 if the gene is listed among the 478 somatically mutated known driver genes of CGC
67) p_higherfreq: p-value for the enrichment in observed mutations compared to background rate
68) p_ACGT: p-value for the higher NS/S rate compared to expected NS/S from the 79 rule or to the Expected_NS/Expected_S ratio
69) p_highertumorfreq: p-value for the higher tumor specific frequency of mutation
70) p_FI_Total: p-value for higher impact score for all mutations (used in TSG_Driver)
71) p_FI_Onco: p-value for higher impact score for missense type mutations (used in OncoGene_Driver)
72) Global_P_Value: Stouffer combined p-value with FDR correction (the order of the file is set to this value. We accept a driver if it's below or equal to 0.1)

REFERENCE

Melloni GEM, Ogier AGE, de Pretis S, Mazzarella L, Pelizzola M, Pelicci PG, Riva L. DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes. Genome Medicine 2014, 6:44, DOI: 10.1186/gm563