Swarm Selection Tutorial

Peter Hraber

2015-03-21

Phase I: Select Sites

This phase of analysis requires:

1. A protein alignment, provisionally assumed to be in FASTA format.
2. A way to recognize how the longitudinal samples are labeled.
   By default, sequence names are assumed to be dot-delimited, with the timepoint label in the first (left-most) field.
3. An indication of which sequence is the reference/TF sequence.
   By default, this is taken to be the first sequence in the alignment.

Choose a Cutoff Setting

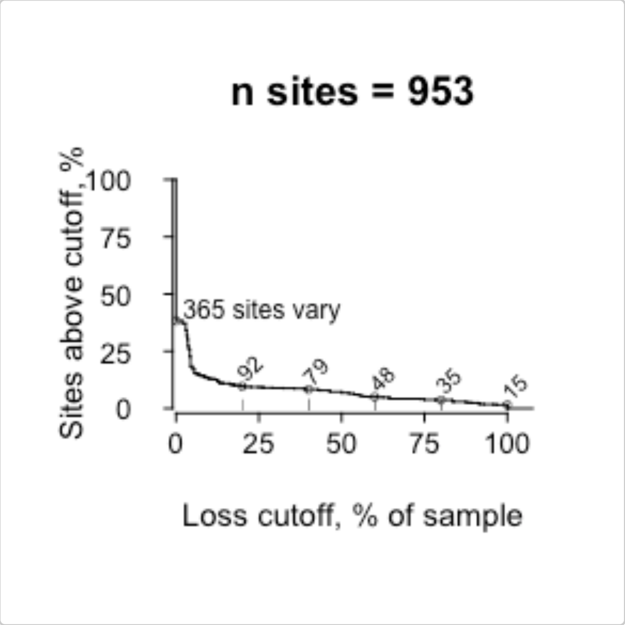
We need to choose a setting for the cutoff parameter value. Let's see how number of sites depends on cutoff threshold. A vector of cutoff values shows how many sites result from multiple settings. Note that the default values work for these data.

In practice, you will not need to use `system.file()` unless you are referring to an example alignment included with the `swarmtools` package.

```
library(swarmtools)
alignment_file <- system.file("extdata", "CH505-gp160.fasta", package="swarmtools")
eg.swarmtools <- swarmtools(aas_file=alignment_file, tf_loss_cutoff=0:5*20)
summary(eg.swarmtools)
```

```
##   %TF Loss at least selected sites, n selected sites, %
## 1                 0                953            100.00
## 2                20                 92              9.65
## 3                40                 79              8.29
## 4                60                 48              5.04
## 5                80                 35              3.67
## 6               100                 15              1.57
```

```
plot(eg.swarmtools)
```



The 80% cutoff value gives 35 sites. Let's go with that.

Having Chosen a Cutoff Setting, List the Selected Sites

```
eg.swarmtools <- swarmtools(aas_file=alignment_file, tf_loss_cutoff=80)
print(eg.swarmtools)
```

```
## Loss cutoff = 80%.
## Selected 35 sites:
##       aln hxb2.l hxb2.r hxb2.aa peak_tf_loss when_up   tf_area
## N279  357    279    279       D        95.83       4  3410.644
## H417  499    417    417       P        91.18       7 10103.671
## V281  359    281    281       A       100.00       9  1732.931
## T413  495    413    413       T        88.24       9  6179.279
## N332  412    332    332       N       100.00      20  1734.220
## O334  414    334    334       S       100.00      20  1734.220
## -144g 214    144    145       -       100.00      20  2093.700
## -144h 215    144    145       -       100.00      20  2093.700
## -144f 213    144    145       -       100.00      20  3408.520
## V756  844    756    756       I        92.86      20  4903.194
## A145  216    145    145       G        96.77      20  8750.560
## D325  405    325    325       N        83.33      20 10922.040
## Y330  410    330    330       H       100.00      22  2784.660
## N300  379    300    300       N       100.00      30  2868.218
## T234  312    234    234       N       100.00      30  2917.280
## K302  381    302    302       N       100.00      30  2917.280
## -465  552    465    465       S       100.00      30  3071.900
## -464  551    464    464       E       100.00      30  3145.650
## N462  544    462    462       N        89.29      30 11450.590
## -463e 550    463    464       -       100.00      53  5433.226
## K460  542    460    460       N      100.00      53  6272.314
## O398  480    398    398       S       91.30      53  7020.274
## K347  427    347    347       S       83.33      53  8187.620
## I151  222    151    151       K       83.33      53  8298.458
## H356  437    356    356       N      100.00      78  6617.376
## E275  353    275    275       V       91.67      78  8185.790
## G471  558    471    471       G       87.50      78  8966.026
## O130  193    130    130       K       87.50      78  9129.296
## O147  218    147    147       M       91.67      78  9312.777
## E640  727    640    640       S       83.87      78  9441.390
## T132  195    132    132       T       83.33      78  9724.856
## D185  259    185    185       D       83.33      78 10942.110
## M4      4      4      4       K       87.50     100 10805.110
## G620  707    620    620       E       91.67     136 10612.342
## R412  494    412    412       D       83.33     160 13949.102
```

Phase II: Select Clones

This phase of analysis requires a SwarmTools object created in Phase I.
The SwarmTools object must have a list of selected sites, which happens only when it was created using a single `tf_loss_cutoff` value. Let’s just go with the defaults.

```
eg.swarmset <- swarmset(eg.swarmtools, is_verbose=F)
```

All Together Now

Got that? Here’s the whole workflow:

```
library(swarmtools)
alignment_file <- system.file("extdata", "CH505-gp160.fasta", package="swarmtools")
eg.swarmtools <- swarmtools(aas_file=alignment_file, tf_loss_cutoff=0:5*20)
summary(eg.swarmtools)
```

```
##   %TF Loss at least selected sites, n selected sites, %
## 1                 0               953            100.00
## 2                20                92              9.65
## 3                40                79              8.29
## 4                60                48              5.04
## 5                80                35              3.67
## 6               100                15              1.57
```

```
eg.swarmtools <- swarmtools(aas_file=alignment_file, tf_loss_cutoff=80)
print(eg.swarmtools)
```

```
## Loss cutoff = 80%.
## Selected 35 sites:
##       aln hxb2.l hxb2.r hxb2.aa peak_tf_loss when_up   tf_area
## N279  357    279    279       D        95.83       4  3410.644
## H417  499    417    417       P        91.18       7 10103.671
## V281  359    281    281       A       100.00       9  1732.931
## T413  495    413    413       T        88.24       9  6179.279
## N332  412    332    332       N       100.00      20  1734.220
## O334  414    334    334       S       100.00      20  1734.220
## -144g 214    144    145       -       100.00      20  2093.700
## -144h 215    144    145       -       100.00      20  2093.700
## -144f 213    144    145       -       100.00      20  3408.520
## V756  844    756    756       I        92.86      20  4903.194
## A145  216    145    145       G        96.77      20  8750.560
## D325  405    325    325       N        83.33      20 10922.040
## Y330  410    330    330       H       100.00      22  2784.660
## N300  379    300    300       N       100.00      30  2868.218
## T234  312    234    234       N       100.00      30  2917.280
## K302  381    302    302       N       100.00      30  2917.280
## -465  552    465    465       S       100.00      30  3071.900
## -464  551    464    464       E       100.00      30  3145.650
## N462  544    462    462       N        89.29      30 11450.590
## -463e 550    463    464       -       100.00      53  5433.226
## K460  542    460    460       N       100.00      53  6272.314
## O398  480    398    398       S        91.30      53  7020.274
## K347  427    347    347       S        83.33      53  8187.620
## I151  222    151    151       K        83.33      53  8298.458
## H356  437    356    356       N       100.00      78  6617.376
## E275  353    275    275       V        91.67      78  8185.790
## G471  558    471    471       G        87.50      78  8966.026
## O130  193    130    130       K        87.50      78  9129.296
## O147  218    147    147       M        91.67      78  9312.777
## E640  727    640    640       S        83.87      78  9441.390
## T132  195    132    132       T        83.33      78  9724.856
## D185  259    185    185       D        83.33      78 10942.110
## M4      4      4      4       K        87.50     100 10805.110
## G620  707    620    620       E        91.67     136 10612.342
## R412  494    412    412       D        83.33     160 13949.102
```

```
eg.swarmset <- swarmset(eg.swarmtools)
```

```
##     TF w000.TF        NHVTNO---VADYNTK--N-KOKIHEGOOETDMGR
## Number of mutations to be represented is now 92
## t=w004, n=8 viable clones
##   column  1: N357[KD]
##     B1 w004.54        K..................................
## Number of mutations to be represented is now 91
##   column  3: V359[SGDA]
##   column 22: O480[T-]
##     B1 w004.31        .....................-.............
## Number of mutations to be represented is now 90
## t=w007, n=16 viable clones
##   column  1: N357[D]
##   column  2: H499[SRN]
##     B1 w007.8         KR.................................
## Number of mutations to be represented is now 89
##   column  4: T495[OI]
##     A0 w007.34        ...I...............................
## Number of mutations to be represented is now 88
##   column 29: O218[SNKD]
##     A0 w007.25        ............................N......
## Number of mutations to be represented is now 87
##   column 35: R494[Q-]
##     A0 w007.21        ..................................Q
## Number of mutations to be represented is now 86
## t=w008, n=12 viable clones
##   column  1: N357[D]
##   column  2: H499[SN]
##   column  3: V359[SGDA]
##     A0 w008.20        ..A................................
## Number of mutations to be represented is now 85
##   column 29: O218[SKD]
##   column 35: R494[-]
## t=w009, n=14 viable clones
##   column  1: N357[D]
##   column  2: H499[SN]
##   column  3: V359[SGD]
##     A1 w009.19        ..G................................
## Number of mutations to be represented is now 84
##   column  4: T495[O]
##   column 35: R494[-]
## t=w010, n=13 viable clones
##   column  1: N357[D]
##   column  2: H499[SN]
##     A1 w010.7         .N.................................
## Number of mutations to be represented is now 83
##   column  3: V359[SD]
##   column  4: T495[O]
## t=w014, n=6 viable clones
##   column  2: H499[S]
##   column  3: V359[SD]
##   column  4: T495[O]
## t=w020, n=23 viable clones
##   column  2: H499[S]
##   column  3: V359[SD]
##   column  4: T495[O]
##   column  5: N412[O]
##     B0 w020.15        ....OS....T........................
## Number of mutations to be represented is now 80
##   column  7: -214[TOIA]
##     B1 w020.25        .R....ATO..........................
## Number of mutations to be represented is now 77
##   column  8: -215[VSKEA]
##   column  9: -213[IA]
##   column 10: V844[A]
##     B0 w020.24        .RA......AT........................
## Number of mutations to be represented is now 76
##   column 11: A216[VSGD-]
##   column 12: D405[N]
##     C1 w020.11        ..........TN.......................
## Number of mutations to be represented is now 75
##   column 29: O218[SKD]
## t=w022, n=15 viable clones
##   column  2: H499[S]
##   column  3: V359[SD]
##   column  4: T495[O]
##     A0 w022.22        ...O...............................
## Number of mutations to be represented is now 74
##   column  7: -214[TOI]
##   column  8: -215[VSKEA]
##   column  9: -213[IA]
##   column 13: Y410[H]
##     B1 w022.6         ..AIOSATO...H......................
## Number of mutations to be represented is now 73
##   column 14: N379[SG]
##     A0 w022.5         .RA.OSATO....S.....................
## Number of mutations to be represented is now 72
##   column 24: I222[VL-]
##     A0 w022.9         ..GIOS.................-...........
## Number of mutations to be represented is now 71
## t=w030, n=24 viable clones
##   column  1: N357[D]
##     A0 w030.17        D.GOOSATO.......TD.................
## Number of mutations to be represented is now 68
##   column  2: H499[S]
##   column  3: V359[SD]
##     A1 w030.13        ..DPOS.............................
## Number of mutations to be represented is now 67
##   column  7: -214[TOI]
##   column  8: -215[VSKEA]
##   column  9: -213[IA]
##   column 11: A216[VSGD-]
##     A0 w030.26        .RG.OS....-....NTD.................
## Number of mutations to be represented is now 65
##   column 14: N379[G]
##   column 15: T312[O]
##     A1 w030.36        ....OSATO.....O.TD.................
## Number of mutations to be represented is now 64
##   column 17: -552[PF]
##   column 18: -551[TNGE]
##     A0 w030.20        ..AIOSATOA......TNTO...............
## Number of mutations to be represented is now 61
##   column 19: N544[SOD-]
##     A1 w030.32        .RG.OSATO.......TD-................
## Number of mutations to be represented is now 60
##   column 20: -550[TKE]
##   column 21: K542[ONE-]
##     B0 w030.21        .RA.OSATO...........E..............
## Number of mutations to be represented is now 59
## t=w053, n=21 viable clones
##   column  2: H499[S]
##   column  3: V359[S]
##   column  7: -214[TOI]
##     B1 w053.22        DRGIOSIEIAG.HSONFT.E.TE............
## Number of mutations to be represented is now 50
##   column  8: -215[VSKA]
##   column  9: -213[A]
##   column 11: A216[VSD]
##   column 14: N379[G]
##   column 17: -552[P]
##   column 18: -551[GE]
##   column 20: -550[TK]
##   column 21: K542[ON-]
##     A1 w053.31        DRGIOSAT....HSON....N-.............
## Number of mutations to be represented is now 49
##   column 24: I222[VL]
##     B1 w053.15        D.G.OSATOA..HSONFT.E.-.L...........
## Number of mutations to be represented is now 48
##   column 25: H437[QONGD]
##     A0 w053.8         DRGIOSATOA..HSONFT.E.T..Q..........
## Number of mutations to be represented is now 47
##   column 27: G558[VE]
##     A0 w053.29        .RAIOSATOA..HSONFT.E.-....E........
## Number of mutations to be represented is now 46
##   column 28: O193[YSN]
##     A0 w053.9         DRGIOSIEIAG.HSONFT.E.TE....N..I....
## Number of mutations to be represented is now 44
##   column 31: T195[A]
## t=w078, n=33 viable clones
##   column  2: H499[S]
##   column  3: V359[S]
##   column  7: -214[TO]
##     A1 w078.9         DRG.OSTAA.S.HSONFT.E....QK..S......
## Number of mutations to be represented is now 38
##     B0 w078.6         DRGIOSOS.AS.HSONTN.OE-.............
## Number of mutations to be represented is now 36
##   column  8: -215[VK]
##   column 11: A216[VD]
##   column 14: N379[G]
##   column 17: -552[P]
##   column 18: -551[GE]
##   column 19: N544[SOD]
##     B0 w078.26        DRGIOSTAAAS.HSON..S.O.E.NK..S......
## Number of mutations to be represented is now 33
##     A0 w078.33        ..A.OSATOA..HSONT.O.N-.....N..I.T..
## Number of mutations to be represented is now 31
##     A1 w078.15        DRGIOSATOA..HSON..D..TEL.KES.......
## Number of mutations to be represented is now 29
##   column 20: -550[TK]
##   column 21: K542[-]
##     A1 w078.29        DRGIOSTAAAS.HSONTN..-..-L..NS.A....
## Number of mutations to be represented is now 27
##   column 24: I222[V]
##   column 25: H437[OGD]
##     B0 w078.17        DRG.OSATOA..HSONFT.EE..LDK.D..IG...
## Number of mutations to be represented is now 25
##   column 27: G558[V]
##   column 28: O193[Y]
##   column 29: O218[KD]
##     A0 w078.30        ..A.OSATOA..HSON....N-......D......
## Number of mutations to be represented is now 24
##   column 30: E727[ND]
##     A1 w078.36        DRGIOSTAAAS.HSON..S.O-.-....SD.....
## Number of mutations to be represented is now 23
##   column 33:   M4[R]
##     A0 w078.27        DRGIOSATOA..HSONTDD..TEL.KES....R..
## Number of mutations to be represented is now 22
## t=w100, n=26 viable clones
##   column  2: H499[S]
##     A1 w100.B10       DSG.OSATOA..HSONTDD....LDKEN..I....
## Number of mutations to be represented is now 21
##   column  3: V359[S]
##     B0 w100.A11       D.S.OSATOA..HSONTNTOE-..D.E.KD.....
## Number of mutations to be represented is now 19
##   column  8: -215[VK]
##     A1 w100.B2        .RAIOSIK.AG.HSON....N...D.V........
## Number of mutations to be represented is now 17
##   column 11: A216[VD]
##     B0 w100.A13       ..A.OSATOAV.HSONTNTOE-.......D.....
## Number of mutations to be represented is now 16
##   column 14: N379[G]
##   column 17: -552[P]
##   column 18: -551[GE]
##   column 20: -550[TK]
##   column 24: I222[V]
##   column 25: H437[OG]
##   column 28: O193[Y]
##   column 30: E727[N]
##   column 34: G707[SD]
##     A0 w100.T3        DRGIOSATO..NHSONTDD.ETEL.KEN..I.RS.
## Number of mutations to be represented is now 15
##     A0 w100.B4        DRGIOSATO...HSON..S.O...DKE.K....D.
## Number of mutations to be represented is now 14
## t=w136, n=28 viable clones
##   column  8: -215[V]
##     A1 w136.B2        D.G.OSTVAA-.HGONIDOT--E.O.......RD.
## Number of mutations to be represented is now 10
##   column 11: A216[D]
##     A1 w136.B10       D.GIOSATOADNHSONTD.E-TELDKES.DIY.S.
## Number of mutations to be represented is now 9
##   column 17: -552[P]
##   column 18: -551[GE]
##     B0 w136.B5        ..A.OSATOAV.HSONTESK-.E.O..Y.DI....
## Number of mutations to be represented is now 6
##   column 24: I222[V]
##     B1 w136.B23       D.A.OSIK..G.HSONTEST-..VD....N...D.
## Number of mutations to be represented is now 4
##   column 25: H437[G]
## t=w160, n=20 viable clones
##   column 17: -552[P]
##     A1 w160.T4        D.A.OSTVA.S.HSONPD..-...G...DN.....
## Number of mutations to be represented is now 2
##   column 18: -551[G]
##     A1 w160.C1        ..A.OSVTOAV.HSONTGST-...D..Y.D..TV.
## Number of mutations to be represented is now 1
##   column 35: R494[-]
##     C0 w160.T3        D.SIOSATOA.NHSONTD.E-TELDKVND.IGRD-
## Number of mutations to be represented is now 0
```

```
summary(eg.swarmset)
```

```
## Selected n=54 sequences:
## w000.TF
## w004.31, w004.54
## w007.8, w007.21, w007.25, w007.34
## w008.20
## w009.19
## w010.7
## w020.15, w020.11, w020.24, w020.25
## w022.6, w022.5, w022.9, w022.22
## w030.20, w030.17, w030.21, w030.36, w030.26, w030.13, w030.32
## w053.15, w053.29, w053.22, w053.8, w053.31, w053.9
## w078.6, w078.36, w078.9, w078.26, w078.29, w078.30, w078.33, w078.17, w078.15, w078.27
## w100.T3, w100.B10, w100.B2, w100.B4, w100.A11, w100.A13
## w136.B10, w136.B5, w136.B2, w136.B23
## w160.C1, w160.T3, w160.T4
```

```
    print(eg.swarmset)
```

```
## w000.TF  NHVTNO---VADYNTK--N-KOKIHEGOOETDMGR
## w004.31  .....................-.............
## w004.54  K..................................
## w007.8   KR.................................
## w007.21  ..................................Q
## w007.25  ............................N......
## w007.34  ...I...............................
## w008.20  ..A................................
## w009.19  ..G................................
## w010.7   .N.................................
## w020.15  ....OS....T........................
## w020.11  ..........TN.......................
## w020.24  .RA......AT........................
## w020.25  .R....ATO..........................
## w022.6   ..AIOSATO...H......................
## w022.5   .RA.OSATO....S.....................
## w022.9   ..GIOS.................-...........
## w022.22  ...O...............................
## w030.20  ..AIOSATOA......TNTO...............
## w030.17  D.GOOSATO.......TD.................
## w030.21  .RA.OSATO...........E..............
## w030.36  ....OSATO.....O.TD.................
## w030.26  .RG.OS....-....NTD.................
## w030.13  ..DPOS.............................
## w030.32  .RG.OSATO.......TD-................
## w053.15  D.G.OSATOA..HSONFT.E.-.L...........
## w053.29  .RAIOSATOA..HSONFT.E.-....E........
## w053.22  DRGIOSIEIAG.HSONFT.E.TE............
## w053.8   DRGIOSATOA..HSONFT.E.T..Q..........
## w053.31  DRGIOSAT....HSON....N-.............
## w053.9   DRGIOSIEIAG.HSONFT.E.TE....N..I....
## w078.6   DRGIOSOS.AS.HSONTN.OE-.............
## w078.36  DRGIOSTAAAS.HSON..S.O-.-....SD.....
## w078.9   DRG.OSTAA.S.HSONFT.E....QK..S......
## w078.26  DRGIOSTAAAS.HSON..S.O.E.NK..S......
## w078.29  DRGIOSTAAAS.HSONTN..-..-L..NS.A....
## w078.30  ..A.OSATOA..HSON....N-......D......
## w078.33  ..A.OSATOA..HSONT.O.N-.....N..I.T..
## w078.17  DRG.OSATOA..HSONFT.EE..LDK.D..IG...
## w078.15  DRGIOSATOA..HSON..D..TEL.KES.......
## w078.27  DRGIOSATOA..HSONTDD..TEL.KES....R..
## w100.T3  DRGIOSATO..NHSONTDD.ETEL.KEN..I.RS.
## w100.B10 DSG.OSATOA..HSONTDD....LDKEN..I....
## w100.B2  .RAIOSIK.AG.HSON....N...D.V........
## w100.B4  DRGIOSATO...HSON..S.O...DKE.K....D.
## w100.A11 D.S.OSATOA..HSONTNTOE-..D.E.KD.....
## w100.A13 ..A.OSATOAV.HSONTNTOE-.......D.....
## w136.B10 D.GIOSATOADNHSONTD.E-TELDKES.DIY.S.
## w136.B5  ..A.OSATOAV.HSONTESK-.E.O..Y.DI....
## w136.B2  D.G.OSTVAA-.HGONIDOT--E.O.......RD.
## w136.B23 D.A.OSIK..G.HSONTEST-..VD....N...D.
## w160.C1  ..A.OSVTOAV.HSONTGST-...D..Y.D..TV.
## w160.T3  D.SIOSATOA.NHSONTD.E-TELDKVND.IGRD-
## w160.T4  D.A.OSTVA.S.HSONPD..-...G...DN.....
```

```
    plot(eg.swarmset)
```