SeqKat

Fouad Yousif, Xihui Lin, Fan Fan, Christopher Lalansingh

2017-06-08

Background

Kataegis is a localized hypermutation occurring when a region is enriched in somatic SNVs (*Nik-Zainal S. et al 2012*). Kataegis can result from multiple cytosine deaminations catalyzed by the AID/APOBEC family of proteins (*Lada AG et al 2012*). A first step to understand kataegis requires the ability to reproducibly and reliability identify it. Although a formal, quantifiable definition of kataegis has not been reached, we have provided the first operational definition in the form of SeqKat, a R package that predicts kataegis from paired tumour normal human whole genome samples. This package contains functions to detect kataegis from SNVs in BED format.

Approach

SeqKat uses a sliding window (of fixed width) approach to test deviation of observed SNV trinucleotide content and inter-mutational distance from expected by chance alone. Additionally, an exact binomial test is performed to test that the proportion of each of the 32 tri-nucleotides within each window is higher than expected. The resulting p-values are then adjusted for multiple hypothesis testing using FDR. Hypermutation and kataegic scores are calculated for each window as follows

$$hypermutation\ score = -log_{10}(\text{binomial } p_{adj}) * \frac{NobservedMutations}{NexpectedMutations} \text{ [Equation 1]}$$

$$kataegis\ score = hypermutation\ score * \frac{NTCXbases}{NexpectedTCXbases} \text{ [Equation 2]}$$

SeqKat reports both hypermutation score and an APOBEC mediated kataegic score along with the start and end position of each detected event. A reference paper will be added upon publication in an upcoming version of this package.

Input

SeqKat accepts a SNV BED file per patient with the following columns:

- chromosome
- position
- reference base
- alternate base

Running SeqKat

```
seqkat(sigcutoff = 5,
       mutdistance = 3.2,
       segnum = 4,
       ref.dir = NULL,
       bed.file = "./",
       output.dir = "./",
```

```
chromosome = "all",
chromosome.length.file = NULL,
trinucleotide.count.file = NULL
)
```

- **sigcutoff**: The minimum hypermutation score used to classify the windows in the sliding binomial test as significant windows. The score is calculated following [Equation 1]. *Recommended value*: 5
- **mutdistance**: The maximum intermutational distance allowed for SNVs to be grouped in the same kataegic event. *Recommended value*: 3.2
- **segnum**: Minimum mutation count. The minimum number of mutations required within a cluster to be identified as kataegic. *Recommended value*: 4
- **ref.dir**: Path to a directory containing the reference genome. Each chromosome should have its own .fa file and chromosomes X and Y are named as chr23 and chr24. The fasta files should contain no header.
- **bed.file**: Path to the SNV BED file. The BED file should contain the following information: Chromosome, Position, Reference allele, Alternate allele. The file should be named {SAMPLE_NAME}_snvs.bed. Below is an example of a BED file:

```
chr4    17185   G   A
chr4    38640   T   C
chr4    52598   C   T
chr4    53102   C   G
chr4    71989   G   A
chr4    91099   C   G
chr4    91139   G   C
chr4    192852  G   C
chr4    201573  G   C
chr4    212498  C   G
```

- **output.dir**: Path to a directory where output will be created
- **chromosome**: The chromosome to be analysed. This can be (1, 2, ..., 23, 24) or "all" to run sequentially on all chromosomes
- **chromosome.length.file** *(provided)*: A tab separated file containing the lengths of all chromosomes in the reference genome. Below is an example of a chromosome.length.file for hg19:

```
"num" "length"
"1" 249250621
"2" 243199373
"3" 198022430
"4" 191154276
"5" 180915260
"6" 171115067
"7" 159138663
"8" 146364022
"9" 141213431
"10" 135534747
"11" 135006516
"12" 133851895
"13" 115169878
"14" 107349540
```

```
"15" 102531392
"16" 90354753
"17" 81195210
"18" 78077248
"19" 59128983
"20" 63025520
"21" 48129895
"22" 51304566
"23" 155270560
"24" 59373566
"sum.f" 3036303846
"sum.m" 3095677412
```

- **trinucleotide.count.file** *(provided)*: A tab seprarated file containing a count of all trinucleotides present in the reference genome. This can be generated with the get.trinucleotide.counts() function in this package. Below is an example of trinuclotide.count.file for hg19:

```
trinucleotide   count
ACA 118307548
ACC 67377361
ACG 15031779
ACT 94371148
ATA 120046083
ATC 78231773
ATG 107040211
ATT 145343907
CCA 107257513
CCC 75979793
CCG 16160851
CCT 103230644
CTA 75285140
CTC 98529730
CTG 118268845
CTT 117139678
GCA 84247974
GCC 68786498
GCG 13995012
GCT 81425256
GTA 65911575
GTC 54961008
GTG 88082699
GTT 86132552
TCA 114943318
TCC 90485129
TCG 13117247
TCT 130196437
TTA 120231763
TTC 117019750
TTG 111364601
TTT 225211336
```

note: sigcutoff, multidistance and segnum default parameters are optimized using *Alexandrov et al*'s "Signatures of mutational processes in human cancer" dataset.

note: trinucleotide.count.file and chromosome.length.file have been provided for GRCh38 reference as well

Output

If Kataegic events are detected, SeqKat generates a tab delimited file that includes details about the detected events. Each line represents one detected hypermutation or kataegic event. The file includes the following columns:

- **sample**: sample name
- **chr**: chromosome
- **start**: coordinate indicating where the event starts
- **end**: coordinate indicating where the event ends
- **variants**: number of SNVs within the event
- **score.hm**: hypermutation score.
- **score.kat**: kataegic score

note: if no event is detected then no file is generated

Example

A subset BED file from the publically available breast cancer sample PD4120a is provided in the package. This BED contains 2804 SNVs in the first 2,000,000 bases of chromosome 4. A subset FASTA file and chromosome length file have also been provided for **testing purposes only**.

```
example.bed.file <- paste0(
    path.package("SeqKat"),
    "/extdata/test/PD4120a-chr4-1-2000000_test_snvs.bed"
    );
example.ref.dir <- paste0(
    path.package("SeqKat"),
    "/extdata/test/ref/"
    );
example.chromosome.length.file <- paste0(
    path.package("SeqKat"),
    "/extdata/test/length_hg19_chr_test.txt"
    );
seqkat(
    5,
    3.2,
    2,
    bed.file = example.bed.file,
    output.dir = ".",
    chromosome = "4",
    ref.dir = example.ref.dir,
    chromosome.length.file = example.chromosome.length.file
    );
```

To view the detected events, you can check the file *PD4120a-chr4-1-2000000_chr4_cutoff5_mutdist3.2_segnum2.txt*

```
sample  chr start   end variants    score.hm    score.kat
PD4120a-chr4-1-2000000  4   1009070 1009541 4   119920.029973896    0
```

In this example, SeqKat detected one hypermutation window on chromosome 4 between 1009070 and 1009541, containing 4 SNVs with a hypermutation score of 119920.03. The kataegic score is 0, indicating that it is not an APOBEC mediated event.

References

- Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jönsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerød A, Tutt A, Martens JW, Aparicio SA, Borg Å, Salomon AV, Thomas G, Børresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR; Breast Cancer Working Group of the International Cancer Genome Consortium.. **Mutational processes molding the genomes of 21 breast cancers**. *Cell.* 2012 May 25;149(5):979-93. doi: 10.1016/j.cell.2012.04.024. Epub 2012 May 17. PubMed PMID: 22608084; PubMed Central PMCID: PMC3414841.

- Lada AG, Dhar A, Boissy RJ, Hirano M, Rubel AA, Rogozin IB, Pavlov YI. **AID/APOBEC cytosine deaminase induces genome-wide kataegis**. *Biol Direct.* 2012 Dec 18;7:47; discussion 47. doi: 10.1186/1745-6150-7-47. PubMed PMID: 23249472; PubMed Central PMCID: PMC3542020

- Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR; Australian Pancreatic Cancer Genome Initiative.; ICGC Breast Cancer Consortium.; ICGC MMML-Seq Consortium.; ICGC PedBrain., Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. **Signatures of mutational processes in human cancer**. *Nature.* 2013 Aug 22;500(7463):415-21. doi: 10.1038/nature12477. Epub 2013 Aug 14. Erratum in: Nature. 2013 Oct 10;502(7470):258. Imielinsk, Marcin [corrected to Imielinski, Marcin]. PubMed PMID: 23945592; PubMed Central PMCID: PMC3776390.