Description of expands

Noemi Andor

October 8, 2013

Contents

1 Introduction

Expanding Ploidy and Allele Frequency on Nested Subpopulations (ExPANdS) characterizes genetically diverse subpopulations in a tumor using copy number and allele frequencies derived from exome- or whole genome sequencing input data [1]. Given a set of somatic point mutations, detected in a tumor sample and the copy number of the mutated loci, ExPANdS identifies the number N of clonal expansions within the tumor, the relative size of the resulting subpopulations in the tumor bulk and the mutations habitant in each subpopulation. Sequencing errors, mapping errors and germline mutations have to be filtered first. The remaining set of somatic mutations can be extended to contain loss of heterozygosity (LOH), that is loci with heterozygous germline polymorphisms where the mutated allele is overrepresented in the cancer cell. For tumor types with a low number of somatic point mutations, this approach can provide a sufficient number of somatic events for the subsequent procedure [1]. The model predicts subpopulations based on two assumptions:

- Two independent driver-events of the same type will not target the same genomic position in two different cells. Therefore, no more than two distinct cell types exist with respect to a specific locus.
- Multiple passenger mutations accumulate in a cell before a driver mutation causes a clonal expansion. Thus, each clonal expansion is marked by multiple mutations.

These two assumptions are translated into the ExPANdS model in four main steps: cell frequency estimation, clustering, filtering and assignment of mutations to clusters. The following example demonstrates each of these steps separately. All steps are merged in the main function `runExPANdS` provided with the package *expands*. The robustness of the subpopulation predictions by ExPANdS increases with the number of mutations provided. It is recommended that at least 200 mutations are used as an input to obtain stable results.

2 Data

We illustrate the utility of ExPANdS on data derived from exome sequencing of a Glioblastoma tumor (TCGA-06-0152-01) from TCGA. Somatic mutations and LOH have been obtained by applying MuTutect [2] on the tumor derived BAM file and the patient-matched normal BAM file. Copy number segments have been obtained by a circular binary segmentation algorithm. We load the data into the workspace and assign each mutation the copy number of the segment in which the mutation is embedded:

```
> library(expands)
> ##loading mutations
> data(snv);
> ## use only a subset of all mutations (for performance reasons).
> set.seed(6); idx=sample(1:nrow(snv), 130, replace=FALSE); snv=snv[idx,];
> ##loading copy number segments
> data(cbs);
> ##assign copy number to mutations
> dm=assignQuantityToMutation(snv,cbs,"CN_Estimate");

[1] "Assigning copy number to mutations..."
[1] "Finding overlaps for CBS segment 100 out of  120 ..."
[1] "... Done."
```

Note that we limit the number of mutations used to 130 to accelerate the computation. In practice however, the inclusion of all available mutations is recommended, as the robustness and accuracy of the algorithm depends on the completeness of the input.

3 Parameter Settings

Next we set the parameters for the subsequent prediction. Type `help(runExPANdS)` for more information on these parameters.

```
> ##parameters
> max_PM=6; maxScore=2.5; precision=0.018;
> plotF=1;
> ##the name of the sample
> snvF="TCGA-06-0152-01";
```

4 Predicting coexisting subpopulations with ExPANdS

Now we are ready to predict the number of clonal expansions in TCGA-06-0152-01, the size of the resulting subpopulations in the tumor bulk and which mutations accumulate in a cell prior to its clonal expansion.

4.1 Cell frequency estimation

First we calculates P - the probability density distribution of cellular frequencies for each single mutation separately. For each cellular frequency f, the value of $P(f)$ reflects the probability that the mutation is present in a fraction f of cells. For more information see `help(cellfrequency_pdf)`. This step may take several minutes to complete.

```
> ##compute the cell frequency probability distribution for each mutation
> cfd=computeCellFrequencyDistributions(dm, max_PM, precision)
```

```
[1] "Computing cell-frequency probability distributions..."
[1] "Processed 20 out of  130 SNVs --> success:  20 / 20"
[1] "Processed 40 out of  130 SNVs --> success:  40 / 40"
[1] "Processed 60 out of  130 SNVs --> success:  60 / 60"
[1] "Processed 80 out of  130 SNVs --> success:  80 / 80"
[1] "Processed 100 out of  130 SNVs --> success:  100 / 100"
[1] "Processed 120 out of  130 SNVs --> success:  120 / 120"
[1] "...Done."
```

In the subsequent step - `clusterCellFrequencies` - we will use only those mutations for which the cell frequency estimation was successful:

```
> ##cluster mutations with valid distributions
> toUseIdx=which(apply(is.finite(cfd$densities),1,all) )
```

In this case the cell-frequency probability distributions could be estimated for all mutations.

4.2 Clustering and Filtering

Next we find overrepresented cell frequencies using a two-step clustering procedure. Based on the assumption that passenger mutations occur within a cell prior to the driver event that initiates the expansion, each clonal expansion should be marked by multiple mutations. Thus SNVs and CNVs that took place in a cell prior to a clonal expansion should be present in a similar fraction of cells and leave a similar trace in the subsequent clonal expansion. The aim is to find common peaks in the distribution of $P_l(f)$ for multiple mutated loci l. In the first step, mutations with similar $P_l(f)$ are grouped together by hierarchical cluster analysis of the probability distributions $P_l(f)$ using the Kullback-Leibler divergence as a distance measure. This step may take several minutes to complete, depending on the number of mutations provided. In the second step, each cluster is extended by members with similar distributions in an interval around the cluster-maxima (core-region). Clusters are pruned based on statistics within and outside the core region [1]. All these steps are performed within the function `clusterCellFrequencies`:

```
> SPs=clusterCellFrequencies(cfd$densities[toUseIdx,], precision, label=snvF)
```

```
[1] "Clustering  130 probability distributions..."
[1] "0 SNVs excluded due to non-finite pdfs"
[1] "Done"
[1] "Filtering Clusters..."
[1] "0 % completed"
[1] "10 % completed"
[1] "20 % completed"
[1] "30 % completed"
[1] "40 % completed"
[1] "50 % completed"
[1] "60 % completed"
[1] "70 % completed"
[1] "80 % completed"
[1] "90 % completed"
[1] "Done."
```

At this point we already know that five subpopulations have been predicted to coexist in this tumor:

```
> print(SPs)
```

```
     Mean Weighted     score   x_p nMutations
[1,]         0.154 2.6800081 0.018          2
[2,]         0.280 1.7520438 0.018          2
[3,]         0.388 2.2608224 0.018          7
[4,]         0.712 2.0797624 0.018         16
[5,]         0.964 0.7171765 0.018          6
```

4.3 Assignment of SNVs to clusters

Now, all that remains to be done is to assign each mutated locus to one of the predicted subpopulations. A mutated locus l is assigned to the subpopulation C, whose size is closest to the maximum likelyhood cellular frequency of l: $C := argmin_C|argmax_f P_l(f) - f^C|$, where $P_l(f)$ is the probability distribution of cellular frequencies as computed by `cellfrequency_pdf` and f^C is the size of subpopulation C. The mutated loci assigned to each subpopulation cluster represent the genetic profile of each predicted subpopulation.

```
> ##assign mutations to subpopulations
> aM= assignMutations( dm, SPs,cfd$densities)
```

`aM$dm` contains the input matrix `snv` with two additional columns: `subpopulation` - the size of the subpopulation to which the mutation has been assigned; and `%maxP` - confidence of the assignment.

5 Visualization of predicted subpopulations

Finally we plot the coexistent subpopulations predicted in the previous steps.

```
> plotSPs(aM$dm, snvF,cex=1.9)
```

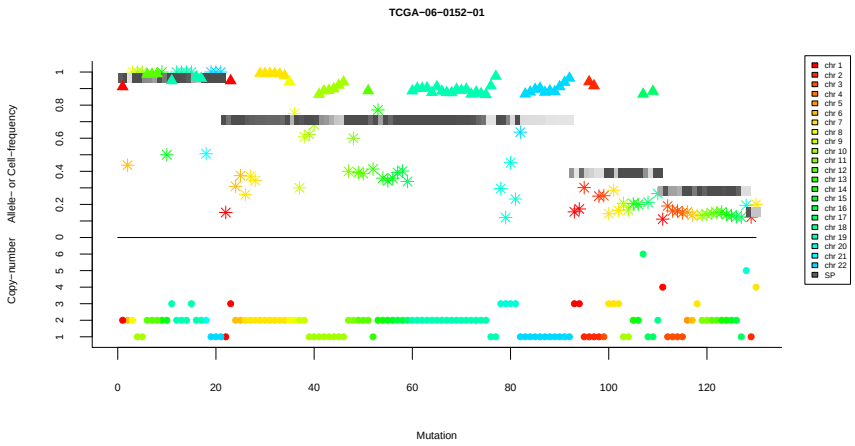

Figure 1: Coexistent subpopulations determined by ExPANdS in an Glioblastoma genome. Five subpopulations were identified based on the allele-frequency and copy number of 130 mutations detected within the cancer-genome. Subpopulations were present in 96%, 71%, 38%, 28% and 15% of the sample (y-axis). For each of the 130 exonic mutations (x-axis) we show: - the subpopulation to which the mutation has been assigned (squares), - the ploidy of the locus in that subpopulation and - the allele frequency of the mutation. Allele frequencies and ploidities are colored based on the chromosome on which the mutation is located (stars - somatic SNVs, triangles - LOH). Subpopulations are colored based on the confidence with which the mutation has been assigned to the subpopulation (black - highest, white - lowest).

6 Acknowledgements

Special thanks to Dr. Ruchira S. Datta for her valuable contributions to the structure and presentation of this manuscript.

References

[1] Noemi Andor, Julie Harness, Hans Werner Mewes and Claudia Petritsch. *ExPANdS: Expanding Ploidy and Allele Frequency on Nested Subpopulations.* Bioinformatics

(2013). In Review.

[2] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotech (2013).