TE-Tracker user manual

Arthur Gilly (agilly@genoscope.cns.fr)

Quick Start guide

Read alignment and filtering

Paired reads should be aligned unto a reference sequence using BWA or equivalent, with a high number of alternative mappings allowed. For example :

```
bwa index reference.fa
bwa aln reference.fa read1.fq -l 35 -O 11 -R 10000 > read1.sai
bwa aln reference.fa read2.fq -l 35 -O 11 -R 10000 > read2.sai
bwa sampe reference.fa read1.sai read2.sai read1.fq read2.fq -n 10000 -N
    ↪  10000> data.sam
```

The resulting SAM file should be sorted and cleaned using picard :

```
java -jar SortSam.jar INPUT=data.sam OUTPUT=data.sorted.bam SORT_ORDER=
    ↪ coordinate VALIDATION_STRINGENCY=LENIENT
java -jar CleanSam.jar INPUT=data.sorted.bam OUTPUT=data.sorted.clean.
    ↪ bam
```

Finally optical and PCR duplicates are removed :

```
java -jar MarkDuplicates.jar INPUT=data.sorted.clean.bam OUTPUT=data.
    ↪ sorted.rmdup.bam REMOVE_DUPLICATES=TRUE METRICS_FILE=metrics.txt
    ↪ VALIDATION_STRINGENCY=LENIENT
```

Preprocessing and detecting discordant reads

First, all mappings whose best match contains more than one mismatch are removed using `eris.pl`, and discordant files are generated (please beware that a version of samtools that support the `-c` option has to be provided using the `samtools=` parameter) :

```
eris.pl -bam data.sorted.rmdup.bam -out . -nodiscordant -treat_bam=input
    ↪ :0-1 -chainload=treated
```

If you plan on using donor scoring, you should also generate a discordant BAM file :

```
eris.pl -bam input.0-1.mismatch.sorted.bam -nosort input.0-1.mismatch.
    ↪ ReadOrder.bam out . -samtools=samtools -nodiscordant -discordant=
    ↪ sorted
```

Clustering and calling

Single run with quantile estimation

Provided you generated the previous steps in the `temp/` directory, run :

```
leto.pl -in temp/ -out temp/
```

Single run with Markov-Chain (MC) estimation

Provided you generated the previous steps in the `temp/` directory, run :

```
leto.pl -in temp/ -out temp/ -method=markov,./temp/input.0-1.mismatch.sorted.bam
```

Running on a grid

Provided you generated the previous steps in the `temp/` directory, simply embed `leto.pl` into a loop and change parameters iteratively :

```
for i in {1..30}
  do for j in {1..20}
    do
      X=$[$i*50];
      Y=$[$j*300];
      leto.pl -in temp/ -out auto/$X-$Y -clustering_parameters=$X:$Y rundata=temp/rundata.dat;
  done;
done;
```

In order to compile the insertions found by every sub-run of the grid, simply `cd` to your parent directory and run :

```
traverseResultsF > sv.annotated
```

This program expects to find a sub-directory named `auto` in which it will look for individual sub-run directories and compile the information it finds in them.

Annotating and scoring

At the previous stage, TE-Tracker has already produced a useable output file, `sv.formatted`. However, you can refine your results by performing annotation and scoring. Provided you generated the previous steps in the `temp/` directory, run :

```
metis.pl -discordant=discordant.sorted.bam -in temp/ -out temp/ -reference=reference.fa -score_donors -annotate -acc_annot=acceptor_annotation.bed -donor_annot=donor_annotation.bed
```

where the `acceptor_annotation.bed` and `donor_annotation.bed` are the BED files containing annotation for the acceptor and the donor, respectively. The input directory requires only that a `sv.formatted` file be present, hence it can run on the output of `traverseResultsF`.

Comparaison de densité

Une première méthode est d'utiliser les densités estimées. Mettons que l'on veuille comparer notre distribution empirique à celle générée par une variable dont on connaît la distribution : il suffit de tracer les deux densités estimées correspondantes. Si les deux densités ont un tracé très proche, notre variable suit peut-être la loi en question, et il faut confirmer ce résultat par un test.

Q-Q plot

Une manière plus élégante de faire est d'ordonner les vecteurs de données et de représenter leurs histogrammes l'un contre l'autre. La figure qui en résulte est appelée un diagramme quantile-quantile, ou Q-Q plot de son nom anglais. Elle s'obtient à l'aide de la commande `qqplot` de R.

Si deux séries de données sont issues de deux distributions strictement identiques, on s'attend à obtenir des points plus ou moins disposés autour de la droite d'équation $y = x$ [1]. Cela signifie que pour chaque intervalle de Ω, on s'attend à trouver un nombre similaire d'observations pour les deux variables, autrement dit les histogrammes des deux distributions sont identiques.

Remarque 1. *Tout comme pour le tracé simultané de deux estimations de densité, rien n'indique que les deux distributions d'un Q-Q plot doivent être empiriques. Il est donc tout à fait possible d'utiliser cette représentation dans un but inférentiel, pour comparer comme nous l'avons fait une distribution issue des données et une distribution issue de la simulation d'une variable théorique. Si l'on obtient un tracé droit pour le Q-Q plot, nos données suivent vraisemblablement une loi de la même famille que celle que nous avons simulée.*

Remarque 2. *On a vu que les membres d'une même famille de distributions partageaient la même expression fonctionnelle f de leur densité, mais se distinguaient par un jeu de paramètres θ. Imaginons que nous comparons des données issues de deux distributions d'une même famille, de densités respectives $f_{\theta_1}(x)$ et $f_{\theta_2}(x)$. A priori, il existe un nombre réel k tel que $\theta_2 = k.\theta_1$. Cela ne veut pas pour autant dire que les fonctions quantiles $Q_{\theta_2}(x) = k.Q_{\theta_1}(x)$ [2]. Cela signifie que dans un Q-Q plot, la variable issue de $f_{\theta_2}(x)$ (même famille que $f_{\theta_1}(x)$, mais de paramètres différents) pourra apparaître tout aussi différente de $f_{\theta_1}(x)$ qu'une variable issue d'une autre famille. Un Q-Q plot ne permet donc pas en général* [3] *de vérifier l'appartenance à une famille, mais seulement de vérifier que deux distributions sont identiques* [4].

1. Cette droite très particulière est appelée la première diagonale du plan en mathématiques.

2. il faut pour cela que Q possède la propriété de linéarité, ce qui n'est presque jamais le cas. En effet, $Q = F^{-1}$, c'est à dire que c'est la fonction inverse de la c.d.f. qui donne les quantiles. Ces fonctions sont généralement complexes et très fortement non-linéaires.

3. à l'exception notable de la loi normale, pour laquelle $\Phi^{-1} = Q$ est linéaire. Cette fonction a une importance capitale en statistiques, il s'agit de la fonction probit.

4. Il est néanmoins possible de contourner ce problème s'il existe une transformation T_{θ_2} qui ramène toute fonction f_θ à une fonction f_{θ_0} connue. Par exemple, si une variable X suit une loi normale, il suffit de poser $x'_i = (x_i - \mu)/\sigma$ pour que X' suive une loi normale centrée réduite $\mathcal{N}(0,1)$. Il deviendrait donc possible de tester l'appartenance de n'importe quel jeu de données à la loi normale en réalisant un Q-Q plot de X' contre $\mathcal{N}(0,1)$, même si Φ^{-1} n'était pas linéaire.

Méthodes numériques : la corrélation

La corrélation est une mesure de la dépendance existant entre deux séries de réalisations. Il s'agit d'un indicateur statistique calculé sur deux variables.

Pour bien comprendre le problème de la corrélation, il est nécessaire de préciser ce qu'on entend par "dépendance". Deux variables aléatoires sont dépendantes si la réalisation de l'une influe sur la réalisation de l'autre : autrement dit, la distribution associée à la deuxième valeur sera différente pour chaque valeur réalisée de la première. Malheureusement, cette définition ne nous dit rien sur la façon dont nos deux séries sont liées. Par exemple, la consommation électrique d'un foyer est clairement dépendante de l'heure : une famille consomme plus d'électricité le matin et le soir que pendant la journée ou la nuit ; la relation est cyclique et suit une courbe sinusoïdale. Par contre, si la force exercée par une pince est bien dépendante de la force exercée par l'opérateur sur le manche, la relation est très différente : la physique nous dit qu'elle est linéaire. Toute la difficulté d'une étude statistique de corrélation consiste donc à déterminer le type de corrélation que l'on souhaite mettre en évidence, et partant, à choisir l'indicateur approprié.

Covariance La covariance est une mesure de la similarité d'évolution entre deux variables. Elle est définie par :

$$\sigma_{XY} = cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Cette formule rappelle beaucoup celle de la variance, il est d'ailleurs très simple de montrer que $\sigma(XX) = Var(X)$. De même que la variance, la covariance possède un estimateur biaisé et un estimateur non-biaisé s_{xy}.

Remarque 3. *si* $[X]$ *est l'unité de* X *et* $[Y]$ *est l'unité de* Y*, alors* $[cov(X,Y)] = [Y].[X]$*. Cette unité n'est pas du tout évidente à utiliser, en particulier il est impossible de comparer deux covariances arbitraires entre elles. La covariance est donc très peu utilisée, à l'image de la variance. Il est donc préférable de ne pas la calculer et de la considérer comme un résultat intermédiaire.*

Dépendance linéaire : le coefficient de corrélation linéaire de Pearson Pour remédier au problème d'unité de la covariance, on définit la corrélation de Pearson (de son nom officiel *Pearson product-moment correlation coefficient*) :

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

C'est simplement le quotient de la covariance non-biaisée par les variances non-biaisées des deux variables. Comme son nom l'indique, la corrélation de Pearson quantifie la linéarité de deux variables : elle n'est valable que si l'on soupçonne (ou souhaite mettre en évidence) une relation du type $Y = a.X + b$ entre X et Y.

Si l'on trace le nuage de points (X,Y), la valeur absolue de r_{xy} sera liée à l'étalement des valeurs de X et Y autour de la droite qui approxime le mieux le nuage, le signe de r_{xy} sera celui du coefficient directeur a dans la modélisation $Y = a.X + b$. Le signe et la valeur du coefficient de Pearson constituent donc deux informations bien distinctes.

Remarque 4. *La corrélation de Pearson est sans unité, elle est toujours comprise entre* -1 *et* 1*. Une corrélation de* -1 *indique une corrélation négative parfaite : plus la valeur de* X *est grande, plus celle de* Y *est petite. Symétriquement pour* $r_{xy} = 1$*, on a une corrélation positive parfaite. Une corrélation de* 0 *signifie qu'il n'existe pas de relation linéaire entre les deux variables, mais cela ne signifie pas que celles-ci sont indépendantes pour autant. Cependant, si deux variables sont indépendantes, leur corrélation de Pearson sera bien nulle.*

Remarque 5. *Il est possible (dans certaines limites) de tester des dépendances non-linéaires avec le coefficient de corrélation de Pearson, et ce à travers un changement de variable. Par exemple si l'on soupçonne une relation logarithmique entre* X *et* Y *telle que* $Y = ln(aX) + b$*, on pourra vérifier la corrélation de* e^Y *avec* X*. En effet, la relation précédente est équivalente à* $e^Y = (ae^b).X$*, qui est une relation linéaire entre* e^Y *et* X*.*

Remarque 6. *Lorsque nous parlons de "la droite qui approxime au mieux le nuage de points* (X, Y)*", nous réalisons en réalité une régression linéaire, qui sera détaillée au cours du chapitre sur les modèles. En particulier, on verra que le carré de* r_{xy}*, est égal à une grandeur très utilisée dans tous les types de régression, à savoir* R^2*, communément appelé "fraction expliquée de la variance".*

Remarque 7. *Comme pour toute variable aléatoire, on peut calculer des intervalles de confiance pour les* r *de Pearson ; et plus généralement, la distribution de* r *est connue pour un échantillonnage donné. Cela permet en particulier de réaliser des tests de significativité, ou de déterminer à partir de quelle valeur* r *est considéré comme élevé à un niveau de confiance* α*.*

Dépendance non-linéaire Il existe plusieurs indicateurs capables de mesurer une dépendance non-linéaire entre deux variables. Leur inconvénient majeur, comme pour tout indicateur exotique, est d'être moins utilisés, donc moins documentés. Les plus connus sont :

la corrélation *rho* (ρ) de Spearman , qui est une corrélation de Pearson calculée non pas sur les valeurs des variables mais sur les rangs de ces variables au sein de l'échantillon. Son interprétation est similaire à celle de r_{xy}, sauf qu'on n'évalue pas la linéarité de la relation mais sa monotonie (Y s'exprime comme une fonction très régulière de X). Cet indicateur est très robuste aux valeurs extrêmes mais se comporte mal en présence d'ex-aequo (couples de mesures identiques).

la corrélation *tau* (τ) de Kendall , qui est un indicateur de rang beaucoup plus robuste aux ex-aequo. Contrairement au ρ de Spearman, qui comme r peut se rapporter à une proportion de variance expliquée, τ représente une probabilité d'association. Il est réputé posséder des intervalles de confiance plus robustes que le ρ de Spearman. De plus, certains statisticiens considèrent que les tests de rang tels que ρ et τ sont plus robustes à la non-normalité des données que le coefficient de Pearson. Enfin, contrairement à r et ρ, il n'est pas basée sur une régression, elle ne fait donc pas intervenir les moindres carrés, qui dans certains cas peuvent introduire un biais difficile à quantifier.

le critère d'information mutuel (*Mutual information criterion*, MIC) , qui permet de détecter à peu près n'importe quelle dépendance continue entre deux variables. Tel quel, il peut être utilisé pour des tests d'indépendance : si le MIC vaut 0, les données sont indépendantes. Il peut également être associé à des procédures exploratoires complexes comme la méthode MINE [5].

Remarque 8. *Les mesures de dépendances non-linéaires sont des outils puissants, mais parfois ambigus. Que faire en présence d'un τ élevé ? La seule chose qu'on sait, c'est que nos deux variables sont liées par une relation monotone : est-elle polynômiale, exponentielle, logarithmique ? Chacune de ces hypothèses nécessite une inférence complexe, ou bien un changement de variable qui permette de se ramener au r de Pearson. Cet exemple illustre parfaitement le fait que r est déjà à moitié un modèle, alors que ρ et τ servent surtout à rejeter un modèle, celui de l'indépendance.*

Autres corrélations

Comme pour presque toute quantité d'intérêt en statistiques, il existe un grand nombre d'indicateurs spécifiques mesurant la corrélation entre deux variables dans des cas plus particuliers. Citons le D de Somers, le Γ de Goodman et Kruskal, la corrélation polychorique, le λ symétrique ou asymétrique, les coefficients d'incertitude... Bien que ces chiffres puissent dans certains cas donner une mesure plus pertinente de la relation entre deux variables, il est souvent nécessaire de s'immerger dans leur définition mathématique pour les interpréter correctement : comme ils sont moins utilisés, il n'y a souvent pas de consensus au sujet de leur interprétation.

Corrélation et causalité

5. Voir http ://www.exploredata.net/ .