École de bioinformatique - Roscoff – Octobre 2015

Atelier Détection de Variants

TP Filtrage de Variants

Nicolas Lapalu

Objectif

Filtrer les variants générés par les TP précédents, sur les données Exome-seq de la publication Pickrell 2012, avec la suite GATK. Comparer les variants obtenus avec une autre méthode de variant calling (Samtools Mpileup / Varscan) sur les données RNA-seq de Pickrell 2012. (Données fournies non générées par les TP précédents).

Fichiers de départ

Vous allez trouver ci-dessous, la liste des fichiers nécessaires à ce TP. Vous avez soit déjà importé ou généré l'ensemble de ces fichiers dans les TPs précédents, à l'exception du fichier **pr_varscan_chr12_targeted.vcf**. Vous trouverez ci-dessous le lien web pour récupérer ce fichier, ainsi que les autres fichiers nécessaires dans le cas où vous les auriez supprimés.

- pr_varscan_chr12_targeted.vcf (à récupérer - dispo ici)
- pe_gatk_chr12_targeted.vcf (généré au TP précèdent – dispo ici)
- dbsnp137.hg19_chr12.vcf (importé au début du TP – dispo ici)
- chr12.fa (importé au début du TP – dispo ici)

Question : Combien de variants ont été détectés pour les données exome GATK (pe_gatk_chr12_targeted.vcf) et Varscan (pr_varscan_chr12_targeted.vcf)? 688 et 499

Filtrage des variants exome issus de GATK

Nous allons réaliser trois filtres de façon simultanée. Les trois filtres seront appliqués sur les données initiales ce qui permettra de visualiser l'action séparée de chaque filtre (colonne FILTRE remplie avec potentiellement plusieurs filtres). Nous allons filtrer sur la qualité du variant (colonne QUAL), sur la qualité de mapping et la profondeur du variant (colonne INFO, Tags MQ et DP).

Tool : gatk2 – Variant Filtration on VCF files

Choose the source for the reference list: **History**
Variant file to annotate: **pe_gatk_chr12_targeted.vcf**
Using reference file: **chr12.fa**
Variant Filters : Add new Variant Filters
Filter expression : **QUAL < 40**
Filter name : **VariantQuality**
Filter expression : **MQ < 30.0**
Filter name : **MappingQuality**

Filter expression : **DP < 20**
Filter name : **Cov20**
Provide a Mask reference-ordered data file: **Don't set mask**

Renommez le fichier vcf de sortie **: pe_gatk_chr12_filtered.vcf**

Question : Le tag DP est aussi présent dans la colonne FORMAT. Comment faire pour que le filtre s'applique au niveau de l'échantillon (FORMAT) plutôt qu'au niveau du variant (INFO) ?

Tool: gatk2 – Select Variant from VCF files

Choose the source for the reference list: **History**
Variant file to select: **pe_gatk_chr12_filtered.vcf**
Using reference file: **chr12.fa**
Don't include filtered loci in the analysis: **coché**

Renommez le fichier vcf de sortie **: pe_gatk_chr12_filtered_ok.vcf**

Question : Combien de variants reste-il ? 557

Filtrage des variants RNA-Seq issus de VARSCAN

Les données de variants issues de VARSCAN ont été générées indépendamment des TP précédents. Elles sont issues d'un variant calling fait par l'outil samtools Mpileup / Varscan avec un mapping préalable par TopHat.

Question : Quels sont les différences que vous pouvez potentiellement observées entre des données Exome-Seq et RNA-Seq. (cf : Chee-Seng Ku, 2012, Expert Reviews)

Les Tags présents dans le fichier VCF Varscan sont différents de GATK. On se propose ici de filtrer sur le tag ADP. A quoi correspond ce tag?

Il est possible d'utiliser directement l'outil Select Variant pour filtrer les données. Cela peut permettre de limiter la taille du fichier de variants obtenu en le limitant aux données désirées.

Tool: gatk2 – Select Variant from VCF files

Choose the source for the reference list: **History**
Variant file to select: **pr_varscan_chr12_targeted.vcf**
Using reference file: **chr12.fa**
Criteria to use when selecting the datas: Add new criteria to use when selecting the data
JEXL expression: **ADP > 30**
Don't include filtered loci in the analysis: **coché**

Renommez le fichier vcf de sortie **: pr_varscan_chr12_filtered_ok.vcf**

Question : Combien de variants reste-il ? 385

Comparaison des résultats GATK Exome-Seq vs VARSCAN RNA-Seq

Il est possible de pouvoir identifier les variants détectés spécifiquement par une méthode ou une autre. Pour cela utilisez l'outil Combine Variant.

Tool: gatk2 – Combine Variant

Choose the source for the reference list: **History**
Variants to Merge 1:
Input variant file: **pe_gatk_chr12_filtered_ok.vcf**
Variant Name: **Gatk**
Variants to Merge 2:
Input variant file: **pr_varscan_chr12_filtered_ok.vcf**
Variant Name: **Varscan**
Using reference file: **chr12.fa**

Renommez le fichier vcf de sortie **: combined_chr12_filtered.vcf**

Question : Combien de variants contient le fichier combiné ? 813

La colonne INFO du nouveau fichier contient un nouveau tag « **set** » avec la valeur **Intersection**, **GATK** ou **VARSCAN**. Il est donc possible de filtrer sur ce nouveau tag pour ne garder que les résultats d'une des 2 méthodes.

Tool: gatk2 – Select Variant from VCF files

Choose the source for the reference list: **History**
Variant file to select: **combined_chr12_filtered.vcf**
Using reference file: **chr12.fa**
Criteria to use when selecting the datas : Add new criteria to use when selecting the data
JEXL expression: **set == "Varscan"**
Don't include filtered loci in the analysis: **coché**

Renommez le fichier vcf de sortie **: pr_varscan_chr12_specific.vcf**

Question : Combien de variants sont spécifiques aux données RNA-Seq (Varscan) ? 256

L'obtention des fichiers de variants a été réalisée sans utiliser de banque de référence de SNP connus comme dbSNP. Il est possible d'utiliser l'outil Variant Annotator pour ajouter l'information avant de regarder les résultats.

Tool: gatk2 – Variant Annotator

Choose the source for the reference list: **History**
Variant file to annotate: **combined_chr12_filtered.vcf**
Using reference file: **chr12.fa**

Provide a dbSNP Reference-Ordered Data (ROD) file: Set dbSNP
dbSNP ROD file: dbsnp137.hg19_chr12.vcf
dbsnp ROD name: dbsnp

Renommez le fichier vcf de sortie **: combined_chr12_filtered_dbsnp.vcf**

Vous obtenez maintenant un fichier avec des informations renseignées dans la 3eme colonne du VCF (ID) avec les identifiants connus dans dbSNP.

Questions : combien de variants ne sont pas connus dans dbSNP ? (37) Comment feriez-vous pour ne récupérer que ceux-là dans un nouveau fichier vcf ?

Regardez le variant obtenu avec GATK à la position **chr12-4461695** et connu dans dbSNP avec l'identifiant **rs75781974**. Dans la colonne INFO, ce variant obtient un score de FS (p-value du test de Fisher en Phred-score) supérieur à 60, soit p=0,000001. Ce test statistique permet de mettre en évidence les SNPs avec un biais de brin (SNP trouvé sur un seul brin), dû à une séquence particulière en amont (cf Meacham (2011) *BMC Bioinformatics*). A la vue du résultat obtenu pour ce variant, que concluez-vous ? Vous pouvez consulter la fiche de ce variant sur http://www.ncbi.nlm.nih.gov/snp?term=rs75781974 , cela conforte-t-il votre hypothèse ? Vous pouvez aussi ouvrir IGV à la position du variant et regarder les résultats de mapping des TPs précédents.

Evaluation des variants

Un outil propose d'évaluer les variants obtenus sur différents critères : taux de Transition/Transversion, leur présence dans dbSNP, … . L'analyse sera effectuée sur une stratification sur les 3 groupes identifiés (Intersection, Gatk, Varscan) et sur l'ensemble (none).

Tool: gatk2 – Eval Variants

Choose the source for the reference list: **History**
Variants : Variant 1
Input Variant file : **combined_chr12_filtered_dbsnp.vcf**
Using reference file : **chr12.fa**
Provide a dbSNP Reference-Ordered Data (ROD) file: **Set dbSNP**
dbSNP ROD file : **dbsnp137.hg19_chr12.vcf**
dbsnp ROD name: **dbsnp**
Basic or Advanced Analysis options: **Advanced**
Add new stratification
Stratification 1
Stratification Expression : **set == "Intersection"**
Name **Intersection**
Stratification 2
Stratification Expression : **set == "Gatk"**
Name **Gatk**
Stratification 3
Stratification Expression : **set == "Varscan"**
Name **Varscan**
Eval modules to apply to the eval track(s): cochez **CompOverlap** et **TiTvVariantEvaluator**

Do not use the standard eval modules by default: **Cochez**

Stratification 4 niveaux : Comp, Eval, Jexl, Novelty.
Comp = 1 seul fichier de comparaison, dbsnp
Eval = 1 seul fichier à évaluer, input_0 = pe12_combine.vcf
JEXL = 4 valeurs de set (Gatk, Intersection, Varscan, none).
Novelty : 3 valeurs : all, known (= in dbSNP), novell (= not in dbSNP)

```
#:GATKReport.v1.1:2
#:GATKTable:11:12:%s:%s:%s:%s:%s:%d:%d:%d:%.2f:%d:%.2f:;
#:GATKTable:CompOverlap:The overlap between eval and comp sites
```

CompOverlap	CompRod	EvalRod	Jexl Exp	Novelty	nEval Variants	novel Sites	nVariants AtComp	comp Rate	nConcordant	concordant Rate
CompOverlap	dbsnp	input_0	Gatk	all	429	3	426	99.30	426	100.00
CompOverlap	dbsnp	input_0	Gatk	known	426	0	426	100.00	426	100.00
CompOverlap	dbsnp	input_0	Gatk	novel	3	3	0	0.00	0	0.00
CompOverlap	dbsnp	input_0	Intersection	all	128	0	128	100.00	128	100.00
CompOverlap	dbsnp	input_0	Intersection	known	128	0	128	100.00	128	100.00
CompOverlap	dbsnp	input_0	Intersection	novel	0	0	0	0.00	0	0.00
CompOverlap	dbsnp	input_0	Varscan	all	256	34	222	86.72	221	99.55
CompOverlap	dbsnp	input_0	Varscan	known	222	0	222	100.00	221	99.55
CompOverlap	dbsnp	input_0	Varscan	novel	34	34	0	0.00	0	0.00
CompOverlap	dbsnp	input_0	none	all	813	37	776	95.45	775	99.87
CompOverlap	dbsnp	input_0	none	known	776	0	776	100.00	775	99.87
CompOverlap	dbsnp	input_0	none	novel	37	37	0	0.00	0	0.00

```
#:GATKTable:14:12:%s:%s:%s:%s:%s:%d:%d:%.2f:%d:%d:%.2f:%d:%d:%.2f:;
#:GATKTable:TiTvVariantEvaluator:Ti/Tv Variant Evaluator
```

TiTv Variant Evaluator	Comp Rod	Eval Rod	Jexl Exp	Novelty	nTi	nTv	tiTv Ratio	nTiI nComp	nTvIn Comp	TiTvRatio Standard
TiTvVariantEvaluator	dbsnp	input_0	Gatk	all	308	113	2.73	295	110	2.68
TiTvVariantEvaluator	dbsnp	input_0	Gatk	known	305	113	2.70	295	110	2.68
TiTvVariantEvaluator	dbsnp	input_0	Gatk	novel	3	0	3.00	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	Intersection	all	97	30	3.23	90	29	3.10
TiTvVariantEvaluator	dbsnp	input_0	Intersection	known	97	30	3.23	90	29	3.10
TiTvVariantEvaluator	dbsnp	input_0	Intersection	novel	0	0	0.00	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	Varscan	all	190	53	3.58	160	45	3.56
TiTvVariantEvaluator	dbsnp	input_0	Varscan	known	170	44	3.86	160	44	3.64
TiTvVariantEvaluator	dbsnp	input_0	Varscan	novel	20	9	2.22	0	1	0.00
TiTvVariantEvaluator	dbsnp	input_0	none	all	595	196	3.04	1398900	692871	2.02
TiTvVariantEvaluator	dbsnp	input_0	none	known	572	187	3.06	545	183	2.98
TiTvVariantEvaluator	dbsnp	input_0	none	novel	23	9	2.56	1398355	692688	2.02

nEvalVariants : nombre de sites dans le fichier eval (input_N)
nVariantsAtComp : nombre de sites dans le fichier comp (= dbsnp dans ce cas)
NovelSites : nombre de sites dans du fichier eval non retrouvé dans le fichier comp
CompRate : pourcentage de sites du fichier eval retrouvés dans le fichier comp
nConcordant : nombre de sites concordant (sites avec le même ALT allèle que dans le fichier comp)
ConcordantRate : taux de concordance

nTi : nombre de sites avec transition dans le fichier eval
nTv : nombre de sites avec transversion dans le fichier eval
TiTvRatio : rapport transition / transversion dans le fichier eval
nTiInComp : nombre de sites avec transition dans le fichier comp
nTvInComp : nombre de sites avec transversion dans le fichier comp
TiTvRatioStandard : rapport transition / transversion dans le fichier comp

Question : Combien de sites connus (dans dbSNP) sont retrouvés pour GATK et Varscan, avec un allele ALT différent ? 0 et 1