École de bioinformatique - Roscoff - Novembre 2013

Atelier Détection de Variants

TP Filtrage de Variants

Sophie Gallina

Objectif :

Filtrer et comparer des listes de variants générés par les TP précédents, sur les données exome et RNAseq de la publication Pickrell 2012, avec 2 méthodes de détection de variants : VARSCAN et GATK.

Fichiers utilisés

- chr12.fa : référence du chromosome 12
- pe12_gatk.vcf : variants détectés par GATK sur les données exome
- pe12_varscan.vcf : variants détectés par VARSCANsur les données exome
- pr12_varscan.vcf : variants détectés par VARSCAN sur les données RNA-seq
- dbsnp_137.vcf : version 137 de la base de données dbSNP

Parcourir les données

- Est-ce que les 2 fichiers pe12_gatk.vcf et pe12_varscan.vcf sont basés sur le même assemblage ? *Regarder database, ou bien les attributs du fichier, icône «* crayon » « edit attributes » / database/Build.
- Combien de variants ont été détectés pour les données exome par chaque méthode (GATK et VARSCAN) ? *Regardez le nombre de lignes dans les datasets Galaxy GATK : 21.071, VARSCAN : 6,171. La stratégie GATK/UnifiedGenotyper est d'augmenter la sensitivité (pour limiter les faux négatifs, donc le nombre de variants manqués), en comptant sur un filtrage ultérieur avec GATK/VariantFiltration pour réduire les faux positifs.*
- Quels sont les champs INFO et FORMAT de ces fichiers vcf (*annexes*) ?

Filtrer les variants exome GATK sur la qualité

Utilisez le champs QUAL avec un seuil de 40 (1 erreur pour 10000, 99,99%).

Pour GATK, utilisez aussi 2 champs de la colonne INFO :

- MQ pour filtrer les sites dont la qualité est inférieure à 30
- DP pour filtrer les sites dont la couverture est inférieure à 20.
- Attention, le champ MQ est du type 'Float', il faudra donc utiliser une valeur de ce type pour l'expression JEXL, c'est à dire 30.0 et non pas 30

```
MQ 1 Float   RMS Mapping Quality
DP 1 Integer Approximate read depth; some reads may have been filtered
```

Tous les outils utilisés dans ce TP sont dans l'onglet Mardi 19 GATK Tools.

Mardi 19 GATK Tools / Variant Filtration on VCF files
Choose the source for the reference list: History
Variant file to annotate: pe12_gatk.vcf
Using reference file: : chr12.fa
Variants Filters : Add new Variant Filters
Filter expression : QUAL<40
FilterName : VariantQuality
Filter expression : MQ<30.0
FilterName : MappingQuality
Filter expression : DP<20
FilterName : Cov20
Provide a Mask reference-ordered data file: Don't set mask
Execute

Renommer le résultat pe12_gatk_filtered.vcf

- Le nombre de variants dans le fichier a-t-il changé ?
- Regardez le contenu du fichier, y a-t-il des sites filtrés ?

Compter le nombre de variants restant après filtrage

Mardi 19 GATK Tools / Select Variant from VCF files
Choose the source for the reference list: History
Variant file to annotate: pe12_gatk_filtered.vcf
Using reference file: : chr12.fa
Don't include filtered loci in the analysis: (-ef,--excludeFiltered) : coché
Execute

Renommer le résultat pe12_gatk_ok.vcf

- Combien de variants reste-t-il ? 4659

Filtrer les variants exome VARSCAN sur la qualité

- Quels critères de filtrage peut-on utiliser sur le fichier VCF généré par VARSCAN ? *Les sites du fichier VARSCAN ne sont pas renseignés pour la colonne QUAL (. = valeur manquante). Ils ne possèdent pas non plus le champs MQ. Mais on peut utiliser le champs ADP pour filtrer les sites ayant une couverture < à 50 bases de qualité >= à 15.*

```
ADP 1 Integer Average per-sample depth of bases with Phred score >= 15
```

ADP est du type 'Integer', donc l'expression sera ADP < 30.

Mardi 19 GATK Tools / Variant Filtration on VCF files
Choose the source for the reference list: History
Variant file to annotate: pe12_varscan.vcf
Using reference file: : chr12.fa
Variants Filters : Add new Variant Filters
Filter expression : ADP < 30
FilterName : SampleDepth
Provide a Mask reference-ordered data file: Don't set mask
Execute

Renommer le résultat pe12_varscan_filtered.vcf

Compter le nombre de variants restant après filtrage

Mardi 19 GATK Tools / Select Variant from VCF files
Choose the source for the reference list: History
Variant file to annotate: pe12_varscan_filtered.vcf
Using reference file: : chr12.fa
Don't include filtered loci in the analysis: (-ef,--excludeFiltered) : coché
Execute

Renommer le résultat pe12_varscan_ok.vcf

- Combien de variants reste-t-il ? 3483

Filtrage sans Sélection

Il n'est pas forcément utile d'utiliser SelectVariant après VariantFiltration.

- La plupart des outils, dont CombineVariant que nous allons utiliser ensuite ont une option permettant d'ignorer les variants filtrés.
- L'information de filtrage est utile, elle est par exemple utilisée par EvalVariants.

Comparer les 2 résultats

Pour savoir combien de variants ont été détectés par les 2 méthodes, ou quels sont les variants détectés uniquement par une des 2 méthodes, utilisez CombineVariant, puis SelectVariants.

Mardi 19 GATK Tools / Combine Variants
Choose the source for the reference list: History

Variants to Merge 1 **:** pe12_gatk_filtered.vcf
Varian name : GATK
Add new Variants to Merge
Variants to Merge 2 **:** pe12_varscan_filtered.vcf
Varian name : VARSCAN
Using reference file: : chr12.fa
Advanced Analysis options :
If true, then filtered VCFs are treated as uncalled, so that filtered set annotation don't appear in the combined VCF: check
Execute

Renommer le résultat pe12_combine.vcf

- Combien de lignes contient pe12_combine.vcf ? *5,056 cela correspond à l'union des 2 listes de variants*
- Dans la colonne INFO, un nouveau champ « set » a été ajouté pour chaque site. Il contient une des 3 valeurs GATK, VARSCAN ou Intersection.

Utilisez GATK/SelectVariants pour extraire les sites communs aux 2 méthodes et les sites spécifiques à chaque méthode. Attention le nom du champ (set) est en minuscules dans le fichier VCF, il doit donc être donné en minuscule dans l'expression JEXL. De même, la valeur Intersection doit être écrite telle quelle dans l'expression.

Mardi 19 GATK Tools / Select Variant from VCF files
Choose the source for the reference list: History
Variant file to annotate: pe12_combined.vcf
Using reference file: : chr12.fa
Add new criteria to use when selecting the datas
JEXL expression : set == "Intersection"
Execute

Renommer le résultat pe12_Intersection.vcf

- Combien de sites sont communs ? 3,086

Même opération en sélectionnant les variants uniquement détectés par GATK. A partir du dataset du précédent résultat, utiliser le bouton « Run this job again », ce qui ouvre à nouveau le questionnaire avec les valeurs utilisées précédemment, modifiez uniquement l'expression JEXL (set == "GATK"), tous les autres paramètres restent identiques, puis bouton « execute ». Ce qui va créer un nouveau dataset pour le résultat avec les nouveaux paramètres.
Renommer le résultat pe12_uniq_GATK.vcf

- Combien de sites sont détectés uniquement par GATK ?

Retrouver les paramètres des résultats déjà générés

- Quel programme a été utilisé pour générer le fichier pe12_combine.vcf ? Quels fichiers ont été utilisés pour produire ce résultat ? R*egarder les informations du fichier (icone « view details »)*

- Quels paramètres ont été utilisés pour créer le fichier combine.vcf *? Utilisez le bouton « Run this job again », ce qui ouvre le questionnaire avec les valeurs utilisées pour l'exécution.*

- Comment Récupérer mon résultat ? Utilisez le bouton « Download ».

Comptage pour évaluer les variants

Tester 2 stratégies :

1. Évaluer 2 fichiers : pe12_gatk_filtered.vcf et pe12_varscan_filtered.vcf
2. Évaluer 1 seul fichier : pe12_combine.vcf, et stratifier sur les 3 groupes (Intersection, GATK et VARSCAN) utilisant des expressions JEXL.

Utilisez dbsnp_137.vcf comme fichier de comparaison (Comp), ce qui permettra de compter le nombre de variants connus / nouveaux.

Utiliser 2 modules d'évaluation :

- CompOverlap : The overlap between eval and comp sites
- TiTvVariantEvaluator : Ti/Tv Variant Evaluator

Pour une explication sur les propriétés et les attendus sur les variants chez l'homme, cf http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties

Comptage à partir de 2 fichiers VFC

Mardi 19 GATK Tools / Eval Variants
Choose the source for the reference list: History
Input Variant file : pe12_gatk.filtered.vcf
Add new Variant
Input Variant file : pe12_varscan.filtered.vcf
Using reference file: : chr12.fa
Set dbSNP
dbSNP ROD file : dbsnp_137.hg19.chr12.vcf
Advanced analysis options
Eval modules to apply to the eval track(s): cochez CompOverlap et TiTvVariantEvaluator
Do not use the standard eval modules by default: Cochez
Execute

Résultats :

Stratification 4 niveaux : Comp, Eval, Jexl, Novelty.

1. Comp = 1 seul fichier de comparaison, dbsnp
2. Eval = 2 fichiers à évaluer, input_0 = pe12_gatk_filtered.vcf, input_1 = pe12_varscan.vcf
3. JEXL = aucune valeur
4. Novelty : 3 valeurs : all, known (= in dnSNP), novell (= not in dbSNP)

1. 2. 3. 4.

Comp Overlap	Comp Rod	Eval Rod	Jexl Expression	Novelty	N Eval Variants	Novel Sites	N Variants At Comp	Comp Rate	N Concordant	Concordant Rate
Comp Overlap	dbsnp	input_0	none	all	4659	70	4589	98.50	4588	99.98
Comp Overlap	dbsnp	input_0	none	known	4589	0	4589	100.00	4588	99.98
Comp Overlap	dbsnp	input_0	none	novel	70	70	0	0.00	0	0.00
Comp Overlap	dbsnp	input_1	none	all	3483	53	3430	98.48	3417	99.62
Comp Overlap	dbsnp	input_1	none	known	3430	0	3430	100.00	3417	99.62
Comp Overlap	dbsnp	input_1	none	novel	53	53	0	0.00	0	0.00

2.

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

TiTvVariantEvaluator	Comp Rod	Eval Rod	Jexl Expression	Novelty	nTi	nTv	TiTv Ratio	N TiIn Comp	N Tv In Comp	TiTv Ratio Standard
TiTvVariantEvaluator	dbsnp	input_0	none	all	3316	1343	2.47	1398901	692873	2.02
TiTvVariantEvaluator	dbsnp	input_0	none	known	3281	1308	2.51	3241	1290	2.51
TiTvVariantEvaluator	dbsnp	input_0	none	novel	35	35	1.00	1395660	691583	2.02
TiTvVariantEvaluator	dbsnp	input_1	none	all	2295	835	2.75	1398901	692873	2.02
TiTvVariantEvaluator	dbsnp	input_1	none	known	2278	828	2.75	2274	844	2.69
TiTvVariantEvaluator	dbsnp	input_1	none	novel	17	7	2.43	1396627	692029	2.02

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

Comptage à partir du fichier combiné, stratifié sur les 3 groupes

Mardi 19 GATK Tools / Eval Variants
- Choose the source for the reference list: History
- Variant file to annotate: pe12_combine.vcf
- Using reference file: : chr12.fa
- Set dbSNP
- dbSNP ROD file : dbsnp_137.hg19.chr12.vcf
- Advanced analiysis options
- Add new stratification
- Stratification expression : set == "Intersection"
- Name Intersection
- Stratification expression : set == "GATK"

Name GATK
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: CoSchez
Execute

Résultats :

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 dnSNP), novell (= not in dbSNP)

Comp Overlap	Comp Rod	Eval Rod	Jexl Expression	Novelty	N Eval Variants	Novel Sites	N Variants At Comp	Comp Rate	N Concordant	Concordant Rate
Comp Overlap	dbsnp	input_0	GATK	all	1573	48	1525	96.95	1525	100.00
Comp Overlap	dbsnp	input_0	GATK	known	1525	0	1525	100.00	1525	100.00
Comp Overlap	dbsnp	input_0	GATK	novel	48	48	0	0.00	0	0.00
Comp Overlap	dbsnp	input_0	Intersection	all	3086	22	3064	99.29	3063	99.97
Comp Overlap	dbsnp	input_0	Intersection	known	3064	0	3064	100.00	3063	99.97
Comp Overlap	dbsnp	input_0	Intersection	novel	22	22	0	0.00	0	0.00
Comp Overlap	dbsnp	input_0	VARSCAN	all	397	31	366	92.19	354	96.72
Comp Overlap	dbsnp	input_0	VARSCAN	known	366	0	366	100.00	354	96.72
Comp Overlap	dbsnp	input_0	VARSCAN	novel	31	31	0	0.00	0	0.00
Comp Overlap	dbsnp	input_0	none	all	5056	101	4955	98.00	4942	99.74
Comp Overlap	dbsnp	input_0	none	known	4955	0	4955	100.00	4942	99.74
Comp Overlap	dbsnp	input_0	none	novel	101	101	0	0.00	0	0.00

TiTvVariantEvaluator	Comp Rod	Eval Rod	Jexl Expression	Novelty	nTi	nTv	TiTv Ratio	N tiIn Comp	N TvIn Comp	TiTv Ratio Standard	N Ti Derived	N Tv Derived	TiTv Derived Ratio
TiTvVariantEvaluator	dbsnp	input_0	GATK	all	1044	529	1.97	1015	498	2.04	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	GATK	known	1024	501	2.04	1015	498	2.04	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	GATK	novel	20	28	0.71	0	0	0.00	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	Intersection	all	2271	814	2.79	2226	792	2.81	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	Intersection	known	2256	807	2.80	2226	792	2.81	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	Intersection	novel	15	7	2.14	0	0	0.00	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	VARSCAN	all	24	20	1.20	51	53	0.96	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	VARSCAN	known	22	20	1.10	48	52	0.92	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	VARSCAN	novel	2	0	2.00	3	1	3.00	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	none	all	3339	1363	2.45	1398901	692874	2.02	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	none	known	3302	1328	2.49	3289	1342	2.45	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	none	novel	37	35	1.06	1395612	691532	2.02	0	0	0.00

Ajouter les données de RNA-seq pour l'évaluation

Utilisez le bouton « Run this job again ». Lancer le filtrage à partir du dataset pe12_varscan.vcf, changez uniquement le nom du fichier vcf de départ (pr12_varscan.vcf à la place de pe12_varscan.vcf), conservez les mêmes paramètres pour le reste. Pour le report, utiliser le résultat de EvalVariant avec 2 fichiers et ajoutez le 3ème fichier, sans changer les autres paramètres.

CompOverlap	CompRod	EvalRod	JexlExpression	Novelty	nEvalVariants	novelSites	nVariantsAtComp	compRate	nConcordant	concordantRate
CompOverlap	dbsnp	input_0	none	all	4659	70	4589	98.50	4588	99.98
CompOverlap	dbsnp	input_0	none	known	4589	0	4589	100.00	4588	99.98
CompOverlap	dbsnp	input_0	none	novel	70	70	0	0.00	0	0.00
CompOverlap	dbsnp	input_1	none	all	3483	53	3430	98.48	3417	99.62
CompOverlap	dbsnp	input_1	none	known	3430	0	3430	100.00	3417	99.62
CompOverlap	dbsnp	input_1	none	novel	53	53	0	0.00	0	0.00
CompOverlap	dbsnp	input_2	none	all	1950	251	1699	87.13	1694	99.71
CompOverlap	dbsnp	input_2	none	known	1699	0	1699	100.00	1694	99.71
CompOverlap	dbsnp	input_2	none	novel	251	251	0	0.00	0	0.00

TiTvVariantEvaluator	CompRod	EvalRod	JexlExpression	Novelty	nTi	nTv	tiTvRatio	nTiInComp	nTvInComp	TiTvRatioStandard
TiTvVariantEvaluator	dbsnp	input_0	none	all	3316	1343	2.47	1398901	692873	2.02
TiTvVariantEvaluator	dbsnp	input_0	none	known	3281	1308	2.51	3241	1290	2.51
TiTvVariantEvaluator	dbsnp	input_0	none	novel	35	35	1.00	1395660	691583	2.02
TiTvVariantEvaluator	dbsnp	input_1	none	all	2295	835	2.75	1398901	692873	2.02
TiTvVariantEvaluator	dbsnp	input_1	none	known	2278	828	2.75	2274	844	2.69
TiTvVariantEvaluator	dbsnp	input_1	none	novel	17	7	2.43	1396627	692029	2.02
TiTvVariantEvaluator	dbsnp	input_2	none	all	1363	489	2.79	1398901	692873	2.02
TiTvVariantEvaluator	dbsnp	input_2	none	known	1169	442	2.64	1139	439	2.59
TiTvVariantEvaluator	dbsnp	input_2	none	novel	194	47	4.13	1397762	692434	2.02

Annexes

Champs INFO et FORMAT des fichiers produits par VARSCAN

info		ADP	1	Integer	Average per-sample depth of bases with Phred score >= 15
info		WT	1	Integer	Number of samples called reference (wild-type)
info		HET	1	Integer	Number of samples called heterozygous-variant
info		HOM	1	Integer	Number of samples called homozygous-variant
info		NC	1	Integer	Number of samples not called
Format	STD	GT	1	String	Genotype
Format	STD	GQ	1	Integer	Genotype Quality
Format		SDP	1	Integer	Raw Read Depth as reported by SAMtools
Format	STD	DP	1	Integer	Quality Read Depth of bases with Phred score >= 15
Format		RD	1	Integer	Depth of reference-supporting bases (reads1)
Format		AD	1	Integer	Depth of variant-supporting bases (reads2)
Format		FREQ	1	String	Variant allele frequency
Format		PVAL	1	String	P-value from Fisher's Exact Test
Format		RBQ	1	Integer	Average quality of reference-supporting bases (qual1)
Format		ABQ	1	Integer	Average quality of variant-supporting bases (qual2)
Format		RDF	1	Integer	Depth of reference-supporting bases on forward strand (reads1plus)
Format		RDR	1	Integer	Depth of reference-supporting bases on reverse strand (reads1minus)
Format		ADF	1	Integer	Depth of variant-supporting bases on forward strand (reads2plus)
Format		ADR	1	Integer	Depth of variant-supporting bases on reverse strand (reads2minus)

Champs INFO et FORMAT des fichiers produits par GATK

info	STD	AC	A	Integer	Allele count in genotypes, for each ALT allele, in the same order as listed
info	STD	AF	A	Float	Allele Frequency, for each ALT allele, in the same order as listed
info	STD	AN	1	Integer	Total number of alleles in called genotypes
info		BaseQRankSum	1	Float	Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities
info	STD	DB	0	Flag	dbSNP Membership
info	STD	DP	1	Integer	Approximate read depth; some reads may have been filtered
info		DS	0	Flag	Were any of the samples downsampled?
info		Dels	1	Float	Fraction of Reads Containing Spanning Deletions
info		FS	1	Float	Phred-scaled p-value using Fisher's exact test to detect strand bias
info		HRun	1	Integer	Largest Contiguous Homopolymer Run of Variant Allele In Either Direction
info		HaplotypeScore	1	Float	Consistency of the site with at most two segregating haplotypes
info		InbreedingCoeff	1	Float	Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation
info	STD	MQ	1	Float	RMS Mapping Quality
info	STD	MQ0	1	Integer	Total Mapping Quality Zero Reads
info		MQRankSum	1	Float	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
info		QD	1	Float	Variant Confidence/Quality by Depth
info		ReadPosRankSum	1	Float	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
Format		AD	.	Integer	Allelic depths for the ref and alt alleles in the order listed
Format	STD	DP	1	Integer	Approximate read depth (reads with MQ=255 or with bad mates are filtered)
Format	STD	GQ	1	Float	Genotype Quality
Format	STD	GT	1	String	Genotype
Format	STD	PL	G	Integer	Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification