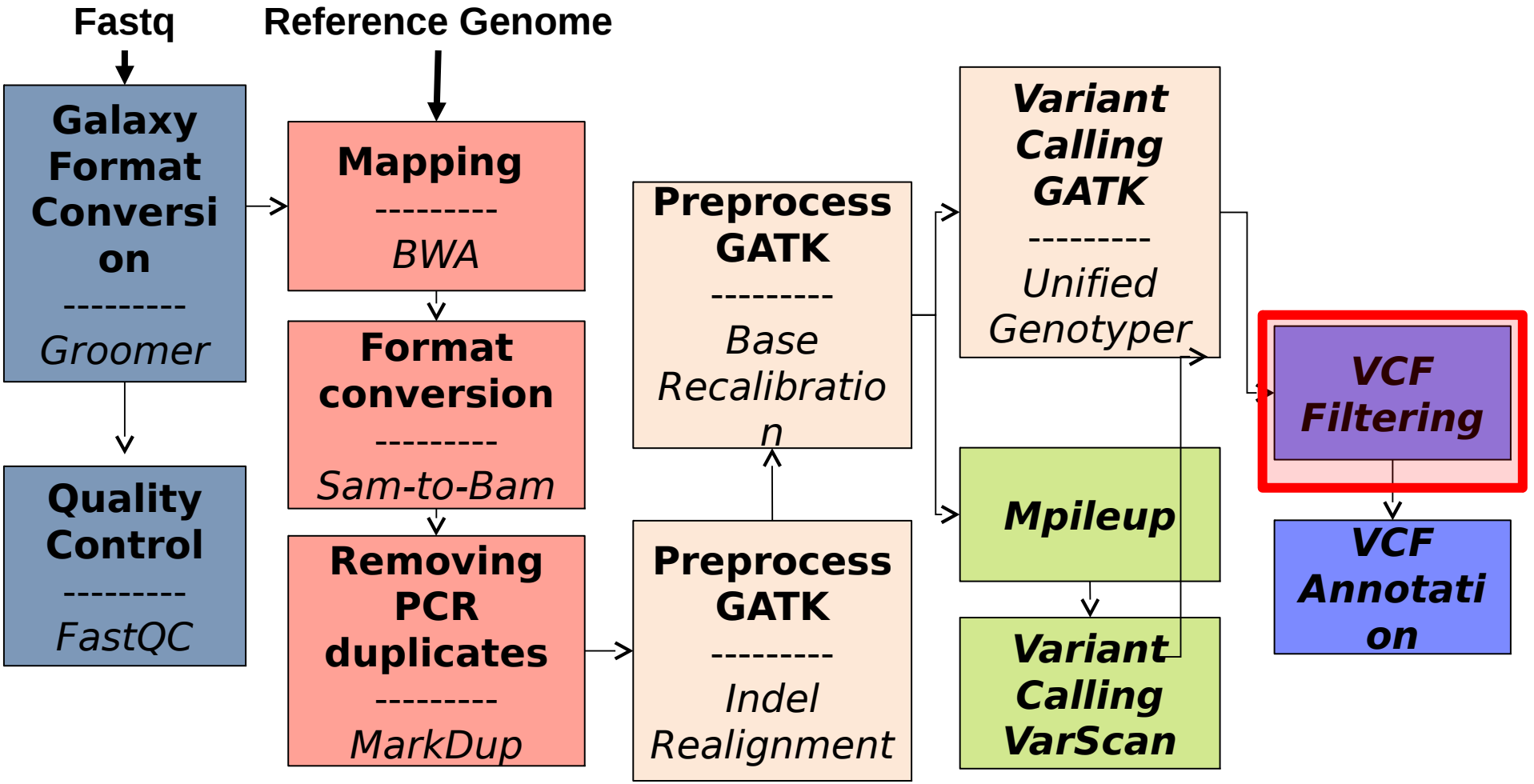
Ecole de bioinformatique - Roscoff - Novembre 2013
Atelier Détection de Variants

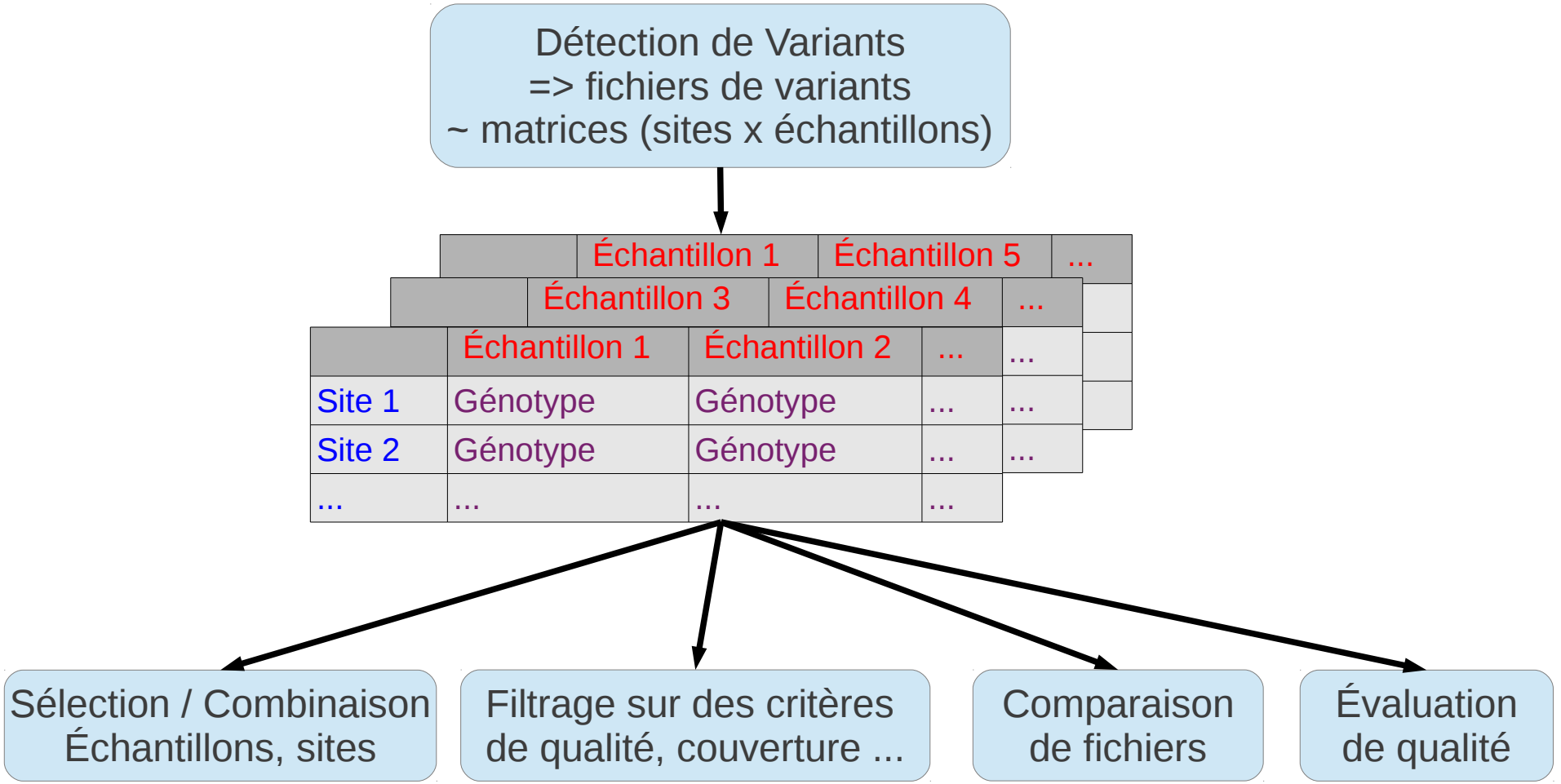
Filtrage de Variants

Sophie Gallina

WorkFlow


Fastq
Reference Genome
Galaxy Format Conversion
Groomer
Quality Control
FastQC
Mapping
BWA
Format conversion
Sam-to-Bam
Removing PCR duplicates
MarkDup
Preprocess GATK
Indel Realignment
Preprocess GATK
Base Recalibration
Variant Calling GATK
Unified Genotyper
Mpileup
Variant Calling VarScan
VCF Filtering
VCF Annotation

Après la détection de variants



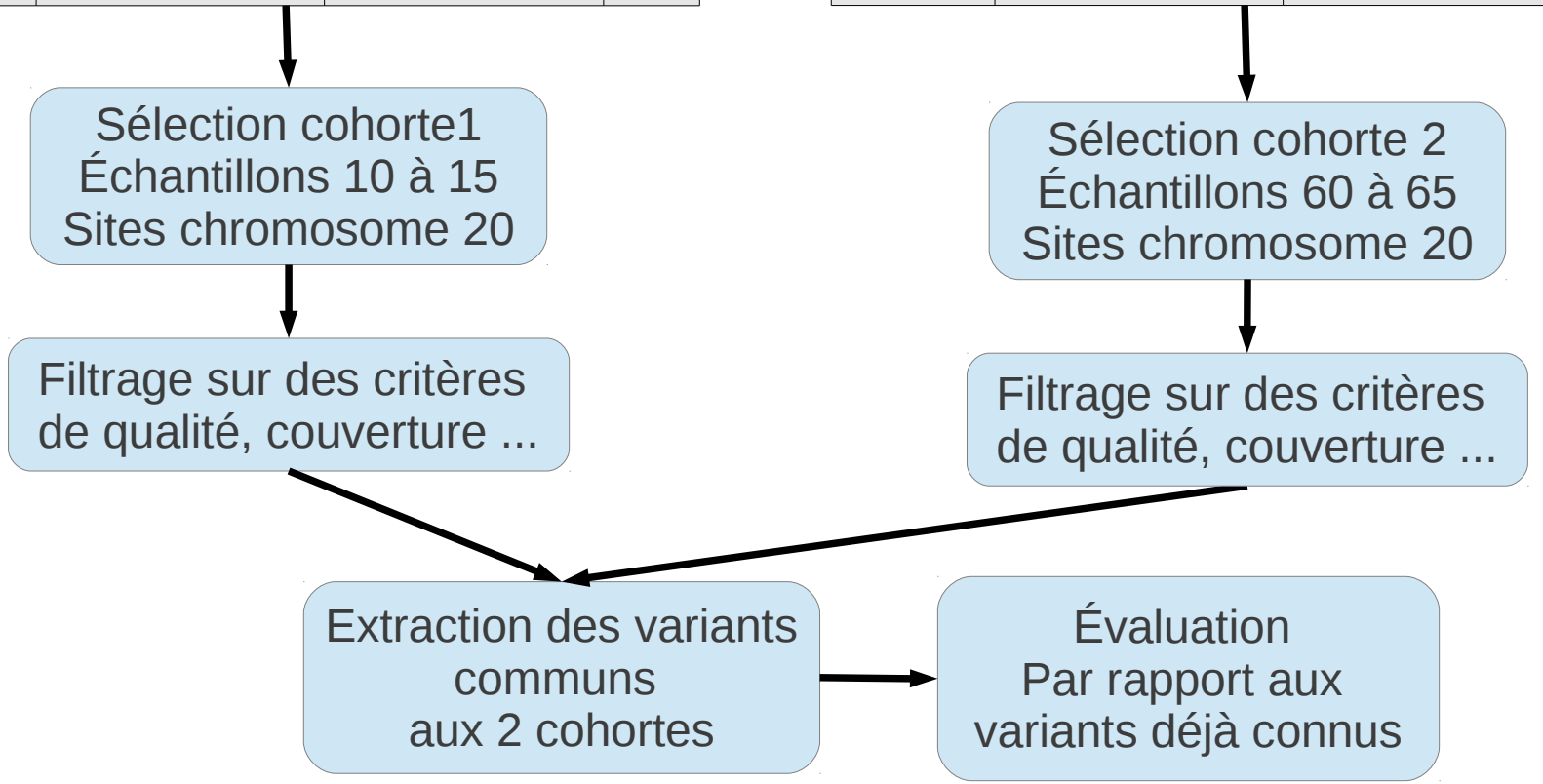
Exemples de tâches

- Extraire les variants de type SNP multi-allelic
- Extraire les variants pour un sous-ensemble d'échantillons
- Combiner les variants de plusieurs groupes d'échantillons
- Comparer les résultats de N groupes d'échantillons ou de plusieurs méthodes sur le même groupe
- Trouver les variants nouveaux par rapport à une liste de référence (ex dbSNP)
- Extraire un sous-ensemble de variants pour
  - une région génomique d'intérêt
  - une liste de sites => Comparaison avec des résultats de puces de génotypage
  - une liste de régions génomiques => comparaison à des résultats d'exome ou de transcriptome

Exemple

	Échantillon 1	Échantillon 2	...
Site 1	Génotype	Génotype	...
Site 2	Génotype	Génotype	...
...	...	...	...

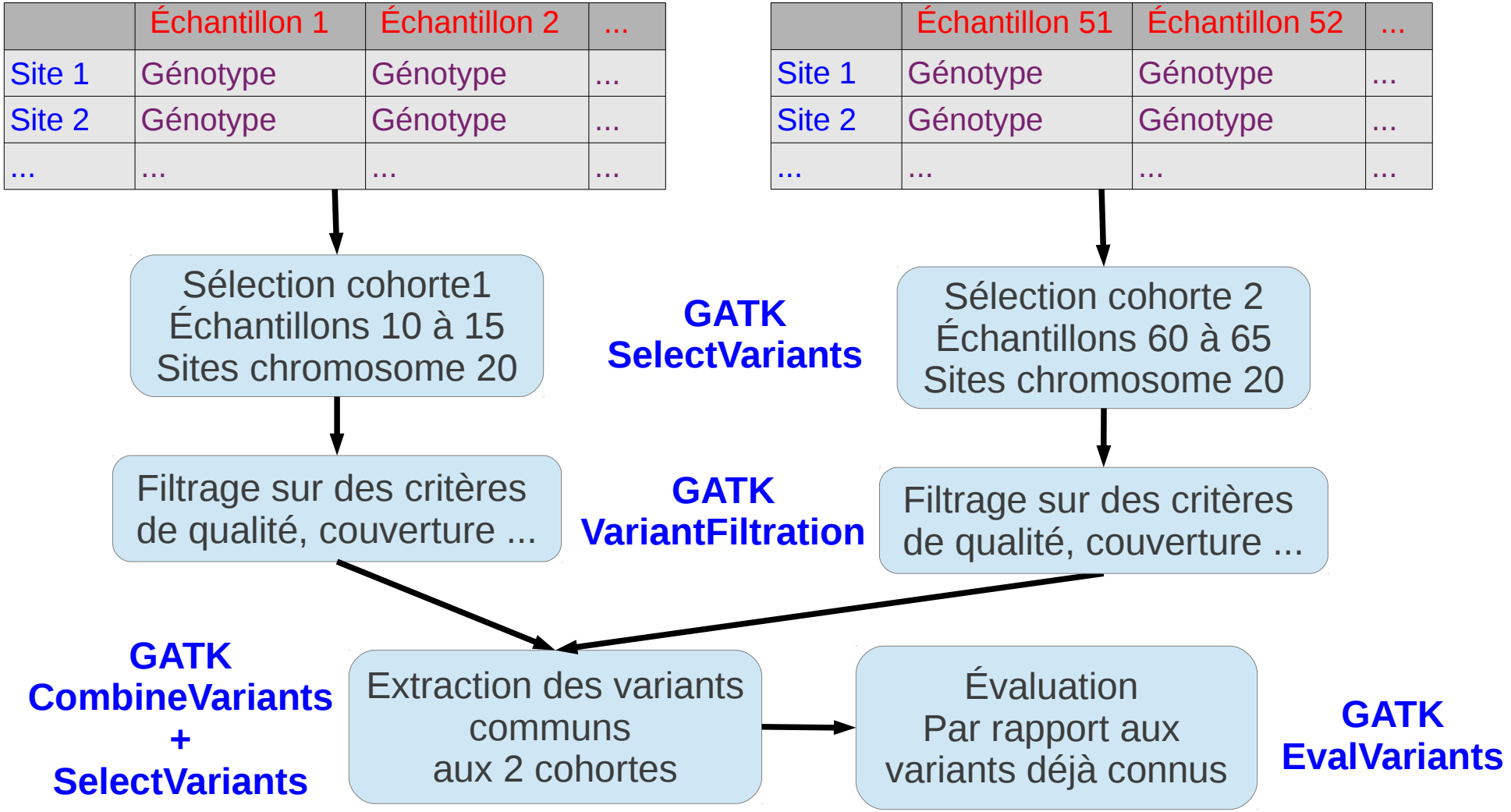
	Échantillon 51	Échantillon 52	...
Site 1	Génotype	Génotype	...
Site 2	Génotype	Génotype	...
...	...	...	...



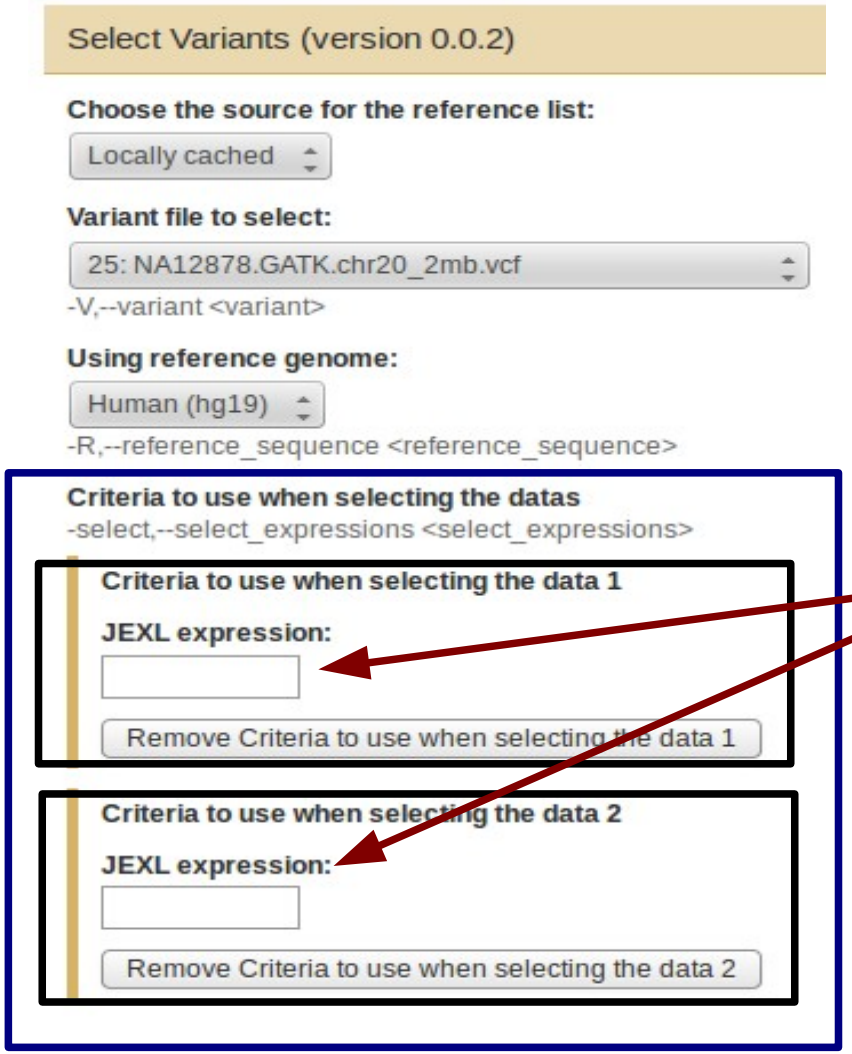
Exemple + Outils GATK

	Échantillon 1	Échantillon 2	...
Site 1	Génotype	Génotype	...
Site 2	Génotype	Génotype	...
...	...	...	...

	Échantillon 51	Échantillon 52	...
Site 1	Génotype	Génotype	...
Site 2	Génotype	Génotype	...
...	...	...	...



GATK Select Variants - Critères de sélection



- Plusieurs jeux de critères de sélection
- Chaque critère = *expression JEXL ?*

GATK Variant Filtration - Critères de filtrage

- Expression JEXL
  - **AB < 0,2 || MQ**

Champ AB < 0,2 — ou — Champ MQ présent

- Liste des champs ?
  - Nom, types de valeurs
- Opérateurs (<, || …) ?

Champs utilisables comme critère

- Fichiers de variants => Format VCF (Variants Call Format)
- Certains champs sont réservés (ex DP = *read depth*)
- D'autres champs sont spécifiques d'une méthode ou d'un outil
  - Par exemple
  - Samtools : PV4 *P-values for strand bias, baseQ bias, mapQ bias and tail distance bias*
  - Varscan : HOM *Number of samples called homozygous-variant*
  - GATK : BaseQRankSum *Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities*
  - FreeBayes : ODDS *The log odds ratio of the best genotype combination to the second-best*
- Ils sont déclarés dans le fichier VCF

Plan

- Format VCF

- Outils GATK
  - VariantFiltration
  - SelectVariants
  - CombineVariants
  - VariantEval

VCF

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis Advance Access publication June 7, 2011

The variant call format and VCFtools

Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

- Format de fichier permettant de stocker des variants
- SNP, indel, variants structuraux … par rapport à 1 seule référence
- Développé pour le projet 1000 génomes
  - Plusieurs millions de sites
  - Plusieurs milliers d'échantillons
- Champs réservés (ex DP *read depth*)
- Extensible : ajout de nouveaux champs, par exemple
  - PV4 (samtools), HOM (Varscan), BaseQRankSum (GATK), ODDS (FreeBayes)

	Echantillon 1	Echantillon 2	...
Site 1	Génotype	Génotype	...
Site 2	Génotype	Génotype	...
...	...	...	...

Format VCF – 3 parties

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sap
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Meta-données = description du contenu des données

Titres des colonnes

En-Tête

Données

Données : Sites, Échantillons, Génotypes

Informations sur chaque site	Génotypes: format + n échantillons

#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5 ,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0,017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3::5:65,3	0/01:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10,AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3,DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2

- Chromosome
- Position
- Nom
- Allèle sur la référence
- Allèle(s) alternatifs

- **QUAL** Qualité
- **FILTER**
- **INFO**

Informations sur le génotype
(site x échantillon)

QUAL = phred-scale quality score on ALT
-10log_10 proba(call in ALT is wrong, ie no variant)

Critères de sélection ou de filtrage :

- Les informations générales du site (lignes)
- Les échantillons (colonnes)
- Les génotypes (sites x échantillons)

Rappel : Phred Quality score

Les score de qualité phred sont reliés de façon logarithmique à la probabilité d'erreur d'identification
http://fr.wikipedia.org/wiki/Score_de_qualit%C3%A9_phred

Score de qualité phred	Probabilité d'une identification incorrecte	Précision de l'identification d'un base
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	99,9 %
40	1 pour 10000	99,99 %
50	1 pour 100000	99,999 %

Sites : colonne FILTER

Informations sur chaque site	Génotypes: format + n échantillons

#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5 ,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0,017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3::5:65,3	0/01:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10,AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3,DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2

FILTER :

- **PASS = OK (le site a passé les filtres)**
- **. = valeur manquante**
  - **=> aucun filtrage n'a été effectué**
- **Autre valeur => Failed**
  - choisie par un programme
  - Exemple ici : q10
  - Signification : cf méta données
  - ##FILTER=<ID=q10,Description="Quality below 10">

VarianFiltration :
Filtrer = écrire une valeur
dans la colonne FILTER

- Ne supprime pas les lignes

SelectVariant, CombineVariants
Filtrer = garder ou utiliser un site
si FILTER == PASS

Eval Variant : utilise FILTER pour stratifier les décomptes

Sites : colonne INFO

Informations sur chaque site	Génotypes: format + n échantillons

#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5 ,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0,017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3::5:65,3	0/01:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10,AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3,DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2

INFO : **Liste de champs + valeurs**
- **champ1=xx:champ2=yy:champ3=zz**
- Exemple ici : NS, DP, AF, AA, DB, H2
  - Signification : cf méta données

```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
```

Les champs ont des formes différentes
- DP (Total depth) contient 1 entier (number=1, type=Integer)
- AF (Allele Frequency) contient une liste de valeurs décimales car il peut y avoir plus d'1 allèle alternatif (number=Array, type=float)
- H2 (Hapmap2 membership) est présent ou absent (number=0, type=flag)

Les sites n'ont pas tous les mêmes champs
- Comment traiter les données manquantes lors du filtrage ? => option des outils

INFO – ID réservés, mais optionnels

AC allele count in genotypes, for each ALT allele, in the same order as listed
AF allele frequency for each ALT allele in the same order as listed: use this when estimated from primary data, not called genotypes
AN total number of alleles in called genotypes
BQ RMS base quality at this position
CIGAR cigar string describing how to align an alternate allele to the reference allele
DB dbSNP membership
DP combined depth across samples
END end position of the variant described in this record (for use with symbolic alleles)
H2 membership in hapmap2
H3 membership in hapmap3
MQ RMS mapping quality, e.g. MQ=52
MQ0 Number of MAPQ == 0 reads covering this record
NS Number of samples with data
SB strand bias at this position
SOMATIC indicates that the record is a somatic mutation, for cancer genomics
VALIDATED validated by follow-up experiment
1000G membership in 1000 Genomes

Génotypes : colonne FORMAT

Informations sur chaque site	Génotypes: format + n échantillons

#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5 ,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0,017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3::5:65,3	0/01:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10,AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3,DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2

Format des informations des colonnes génotype de tous les échantillons

- **champ1:champ2:champ3**
- Exemple ici : GT, GQ, DP, HQ
  - Signification : cf méta données

Valeurs des génotypes ; site x échantillon

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Les informations ont des formes différentes

- DP (Read deepth) contient 1 entier (number=1, type=Integer)
- GT (Génotype) contient un 1 mot (number=1, type=string)
- HQ (Haplotype Quality) contient 2 entiers (number=2, type=Integer)

Les sites n'ont pas tous les mêmes champs

- Comment traiter les données manquantes lors du filtrage ? => option des outils

GT :
Unphased : allèle1 / allèle 2
Phased : allèle1 | allèle 2

Allèle :
0=REF, 1=ALT1, 2=ALT2 ...

Données : Sites # Génotypes

Informations sur chaque site	Génotypes: format + n échantillons

#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5 ,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0,017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3::5:65,3	0/01:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10,AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3,**DP=13**;AA=T	GT:GQ:**DP**:HQ	0|0:54:**7**:56,60	0|0:48:**4**:51,51	0/0:61:**2**

INFOrmations sur le site

FORMAT des informations des colonnes génotype

Valeurs des génotypes (site x échantillon)

2 informations différentes peuvent avoir le même nom dans les colonnes INFO et FORMAT

Exemple ici : DP

INFO / DP = Couverture totale du site (avec tous les échantillons)

FORMAT / DP = Couverture du site pour l'échantillon

Cf méta-données :

```
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```

Génotypes – ID réservés, mais optionnels

GT genotype
DP read depth at this position for this sample (Integer)
FT sample genotype filter
GL genotype likelihoods
GLE genotype likelihoods of heterogeneous ploidy
PL phred-scaled genotype likelihoods
GP the phred-scaled genotype posterior probabilities
GQ conditional genotype quality
HQ haplotype qualities, two comma separated phred qualities (Integers)
PS phase set.
PQ phasing quality
EC expected alternate allele counts
MQ RMS mapping quality, similar to the version in the INFO field. (Integer)

Outils GATK

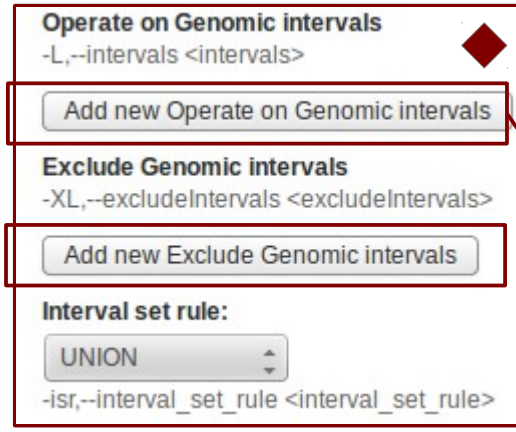
- **SelectVariants**
  - Extraire un sous-ensemble de d'échantillons et de sites
- **CombineVariants**
  - Combiner plusieurs fichiers VCF
- **VariantFiltration**
  - Modification de la colonne FILTER
  - Critère sur les sites (colonne INFO)
  - Critère sur les génotypes (colonnes FORMAT + samples)
- **EvalVariant**
  - Différents comptages sur 1 ou plusieurs vcf, comparés entre eux et avec d'autres sources

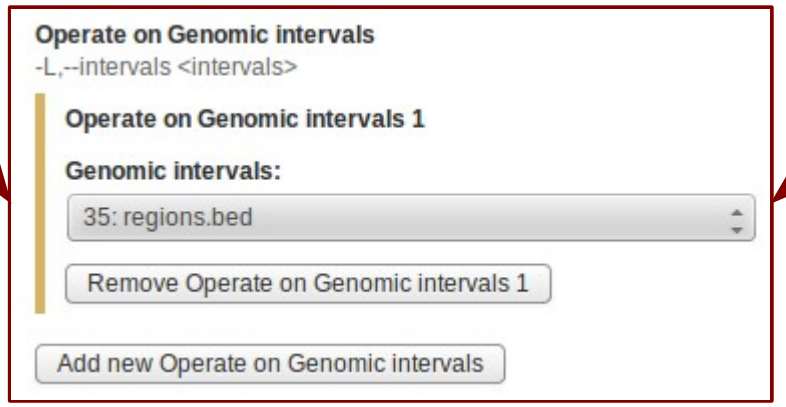
SelectVariants

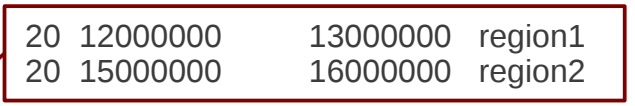
- Extraire un sous-ensemble d'échantillons et de sites
- Critères
  - Régions génomiques
  - Échantillons
  - Sites : champs FILTER et INFO
    - Expression générale sur INFO
    - Critère prédéfini
    - Jonctions avec d'autres fichiers
  - Pas de critères sur FORMAT/GENOTYPE (cf VariantFiltration)

Régions génomiques & samples

Régions génomiques





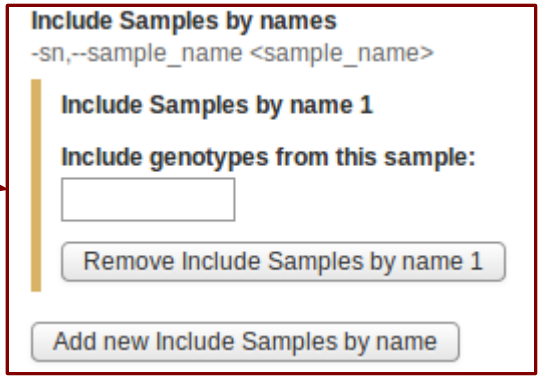


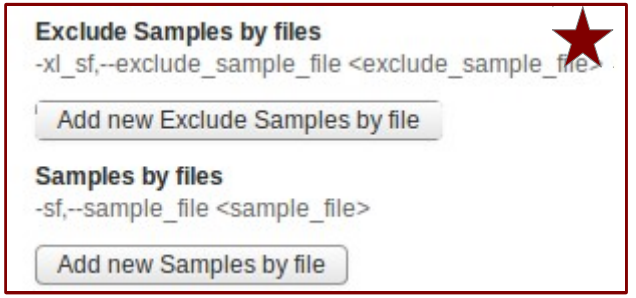
ROD file : Reference Ordered Data

- VCF : variant loci and genotype calls
- dbSNP : UCSC formated dbSNP
- BED : general format for genomic interval

!! bed-format 0-based other 1-based

Échantillons

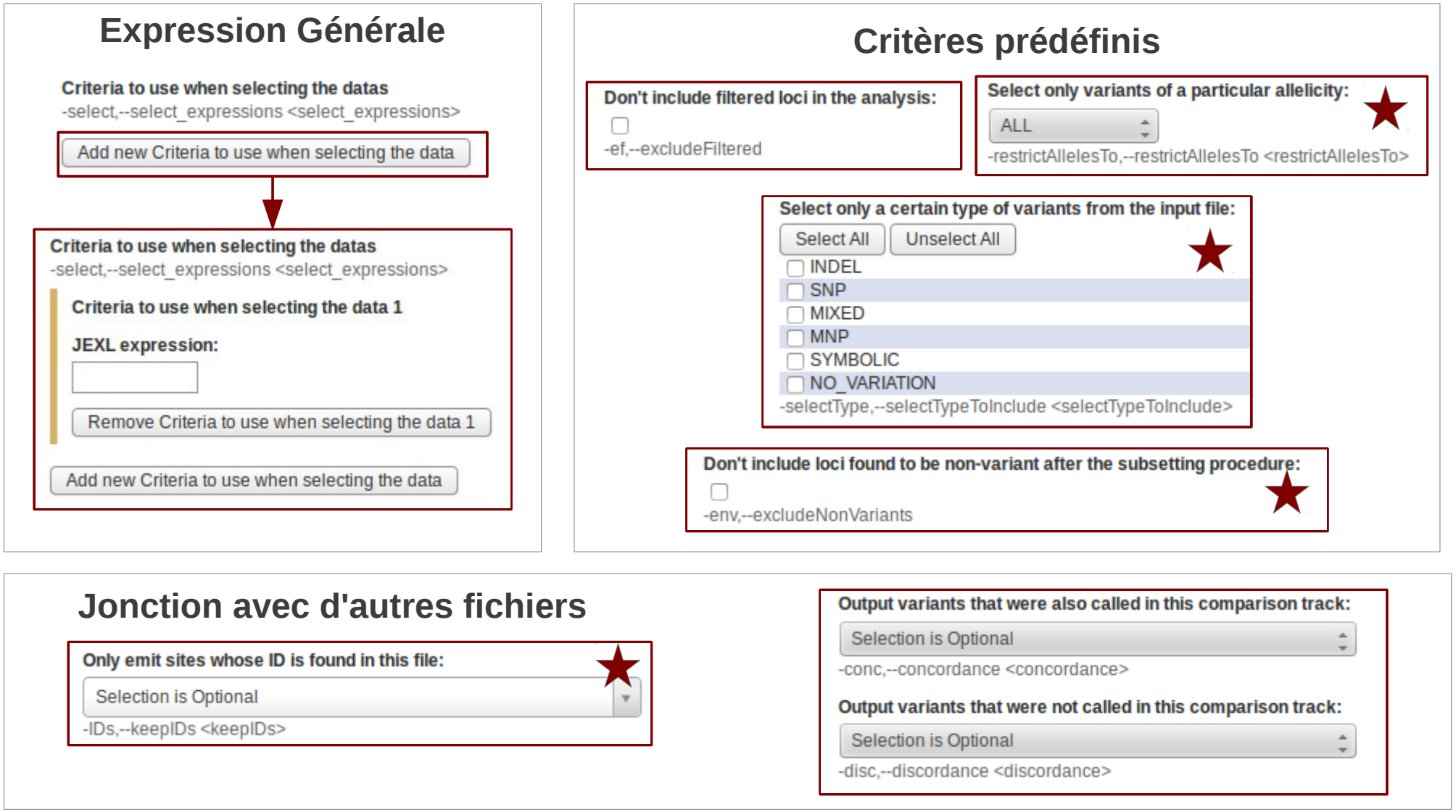




◆ Advanced GATK options

★ Advanced Analysis options

Sites

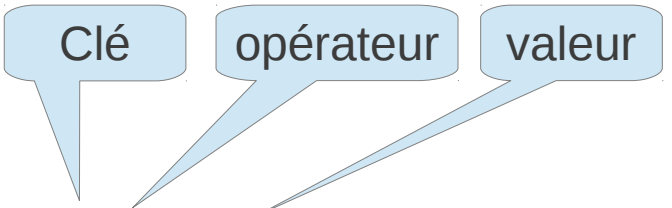


◆ Advanced GATK options
★ Advanced Analysis options

JEXL expression (Java EXpression Language)

- Clé = nom d'un champ
- Valeur = numérique ou "Mot"
- Case-sensitive
  - set / Set / SET => différents
- Type-sensitive
  - QUAL < 50 / QUAL < 50.0
    - Selon le VCF
- Opérateurs
  - Arithmétiques : + - * / %
  - Relationnels : == != < > <= >=
  - Logiques : && (ET) || (OU)

Exemples



- DP < 100
- VT == "INDEL"
- set == "Intersection"
- QUAL / DP < 10
- QUAL < 30 && DP < 10

- ! Ne pas utiliser d'apostrophe
- VT == 'INDEL'
- Pas de message d'erreur mais interprété comme XINDELX

SelectVariants - Exemples

1) Sélection : extraire les variants de type SNP multi-allelic
2) Sélection : extraire
   - Les échantillons de la cohorte 1 (fichier cohorte1.samples)
   - Avec uniquement les sites polymorphes dans cette cohorte
3) Comparaison : extraire les sites concordants ou discordants avec un autre résultat (ex fichier cohorte1_methode2.vcf)
4) Filtrage : extraire les sites qui sont discordants par rapport à une liste de référence => Nouveaux variants
5) Sélection : extraire un sous-ensemble de variants
   - À partir d'une liste de variants (fichier vcf : [chrom pos])
     - => Comparaison avec des résultats de puces de génotypage
   - À partir d'une liste de régions génomiques (fichier bed : [chrom start end])
     - liste de capture de régions => pour comparer à un résultat d'exome
     - liste de transcrits => pour comparer à un résultat de transcriptome

CombineVariants

- Combiner plusieurs fichiers VCF
  - variants de plusieurs groupes d'échantillons
  - variants détectés par des méthodes différentes sur les mêmes échantillons
- Chaque fichier est identifié par un ID choisi par l'utilisateur
- Pour chaque site, CombineVariants ajoute un champ **set** dans INFO
  - set=ID1 : site unique à ID1
  - set=ID1-ID2 : site présent dans ID1 et ID2
  - set=ID1-filteredInID2 : site présent dans ID1 et ID2 mais filtré dans ID2
  - set=Intersection : site commun à tous les fichiers
  - set=filteredInAll : site filtré dans tous les fichiers
- Paramètres
  - minN : Conserve uniquement les variants présent dans N sources
  - Genotypemergeoption : Comment combiner des génotypes différents pour le même échantillon
  - filteredrecordsmergetype : Comment combiner des sites avec des statuts différents
  - FilteredAreUncalled: Traite les sites filtrés comme des sites absents

VariantFiltration

- **VariantFiltration**
- *hard filter* : modification de la colonne FILTER
  - Critère sur les sites (colonne INFO)
  - Critère sur les génotypes (colonnes FORMAT + samples)
  - Plusieurs filtres nommés
  - Traitement des valeurs manquantes
  - --missingValuesInExpressionsShouldEvaluateAsFailing

VariantsEval

- Effectue des comptages
  - sur 1 ou plusieurs vcf
    - avec possibilité de stratifier sur des paramètres ou grâce à des modules de stratification (exemple Novelty, Samples)
  - comparés entre eux et avec d'autres sources
  - en calculant différentes métriques (= modules de comparaison)
    - CompOverlap : The overlap between eval and comp sites
    - CountVariants : Counts different classes of variants in the sample
    - TiTvVariantEvaluator : Ti/Tv Variant Evaluator
    - ValidationReport : Assess site accuracy and sensitivity of callset against follow-up validation assay
    - VariantSummary : 1000 Genomes Phase I summary of variants table

Evaluation des variants

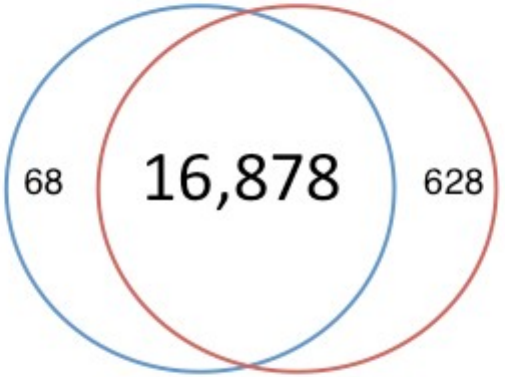
http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties

- Propriétés connues pour l'homme
  - Nombre de variants : 1 pour 1000bp // génome
  - Variants déjà connus (ie présents dans dbSNP)
    - 90% de variants pour 1 individu
    - + si variants communs à de nombreux individus
  - Taux de transition/transversion (Ti/Tv)
    - Ti (A-G C-T) 2 fois plus fréquentes que Tv (A-C A-T G-C G-T)
    - => TiTv > 2 pour les variants de bonne qualité
    - TiTv > 3 variants dans des régions codantes (exome)

VariantEval exemple

http://gatkforums.broadinstitute.org/discussion/48/using-varianteval

Whole-exome NA12878 SNP overlaps: MAQ vs. BWA



MAQ	
Unique calls	68
No. in dbSNP	46 / 68 (68%)
Known Ti/Tv	2.54
No. filtered in other	58 / 68 (85%)
No. in dbSNP of filtered	91 / 249 (37%)

BWA	
Unique calls	628
No. in dbSNP	284 / 628 (45%)
Known Ti/Tv	2.26
No. filtered in other	411 / 628 (65%)
No. in dbSNP of filtered	229 / 1560 (15%)

Intersection	
Common calls	16878/ 17574 (96.7%)
No. in dbSNP	16617 / 16878 (98.5%)
Known Ti/Tv	3.29
Novel Ti/Tv	1.84

Slide design stolen from MAD

VariantEval
Exemples de report obtenu dans le TP

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

TiTvVariantEvaluator	Comp Rod	Eval Rod	Jexl Expression	Novelty	nTi	nTv	TiTv Ratio	N TiIn Comp	N Tv In Comp	TiTv Ratio Standard	N Ti Derived	N Tv Derived	TiTv Derived Ratio
TiTvVariantEvaluator	dbsnp	input_0	none	all	3316	1343	2.47	1398901	692873	2.02	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	none	known	3281	1308	2.51	3241	1290	2.51	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_0	none	novel	35	35	1.00	1395660	691583	2.02	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_1	none	all	2295	835	2.75	1398901	692873	2.02	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_1	none	known	2278	828	2.75	2274	844	2.69	0	0	0.00
TiTvVariantEvaluator	dbsnp	input_1	none	novel	17	7	2.43	1396627	692029	2.02	0	0	0.00