

VARIANT ANNOTATION

Vivien Deshaies

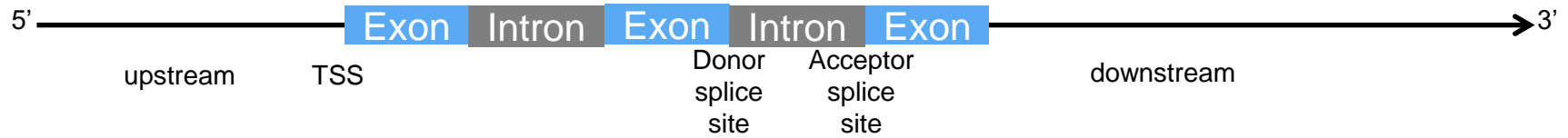
vivien.deshaies@icm-institute.org



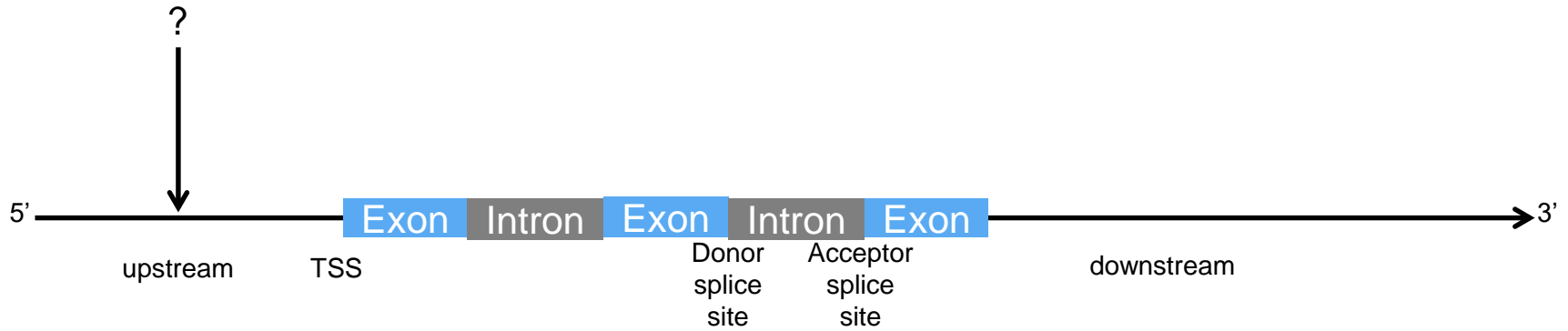
Goal

Add meta-information on variant to
facilitate interpretation

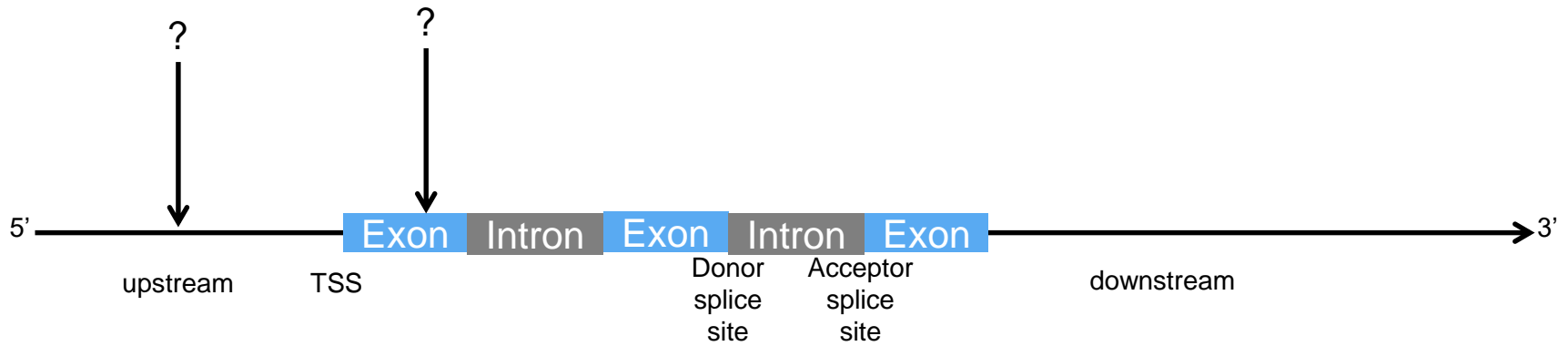
Location



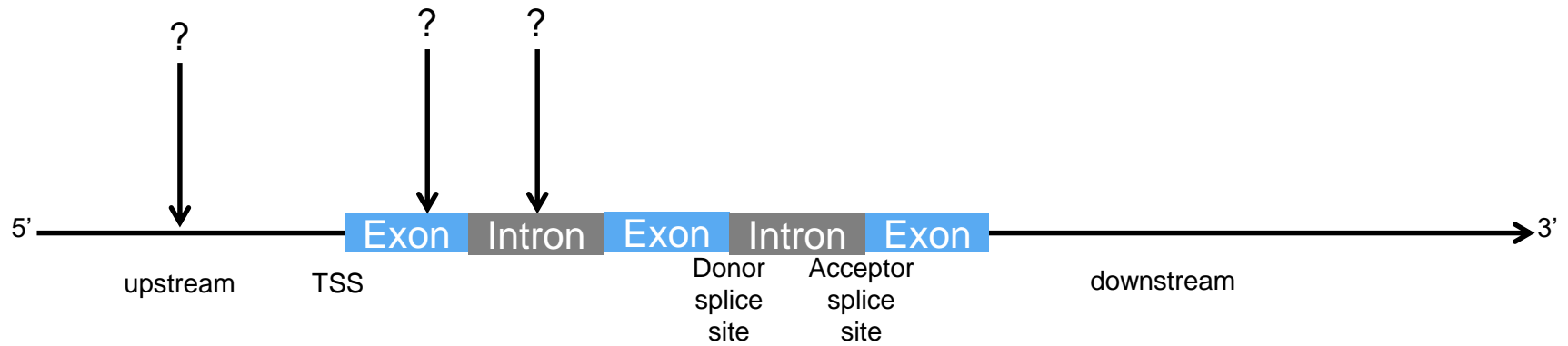
Location



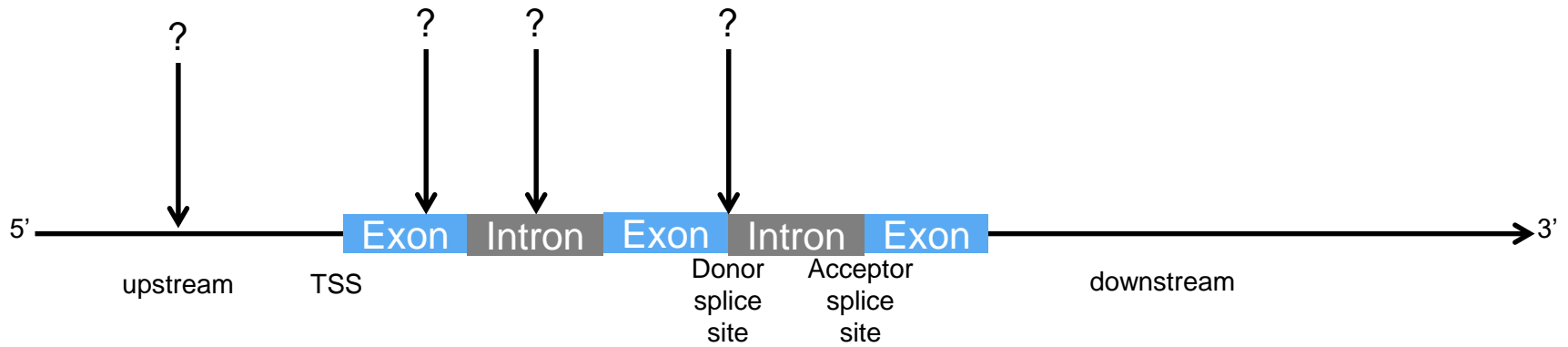
Location



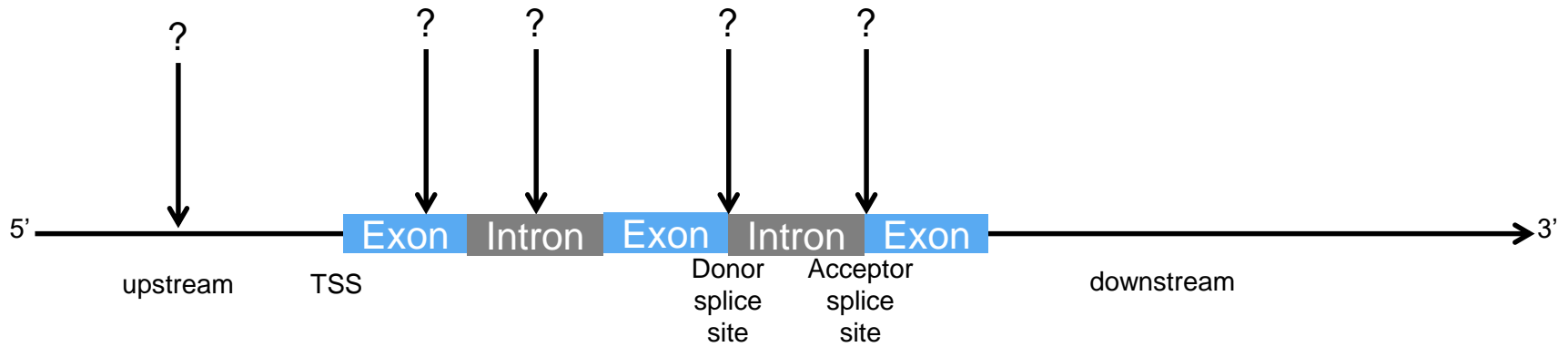
Location



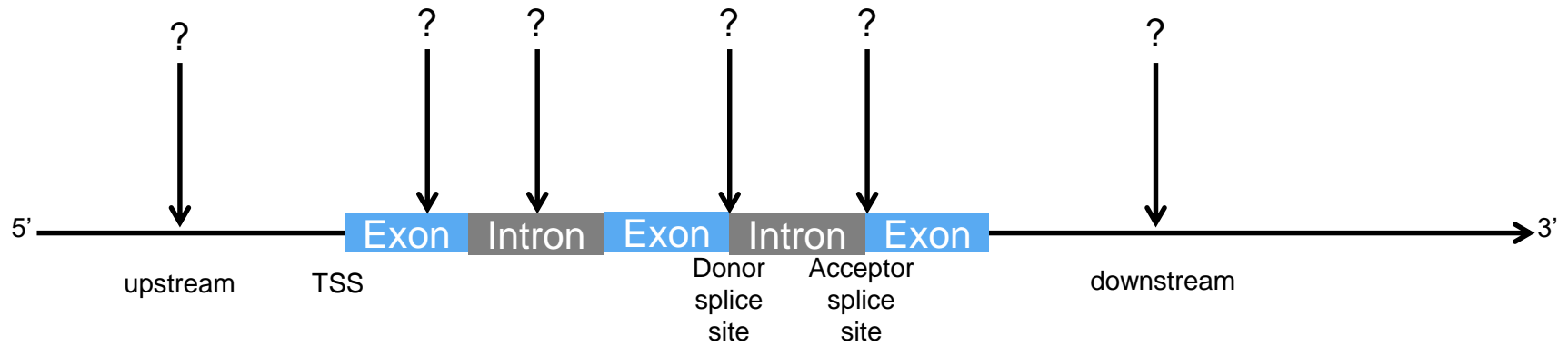
Location



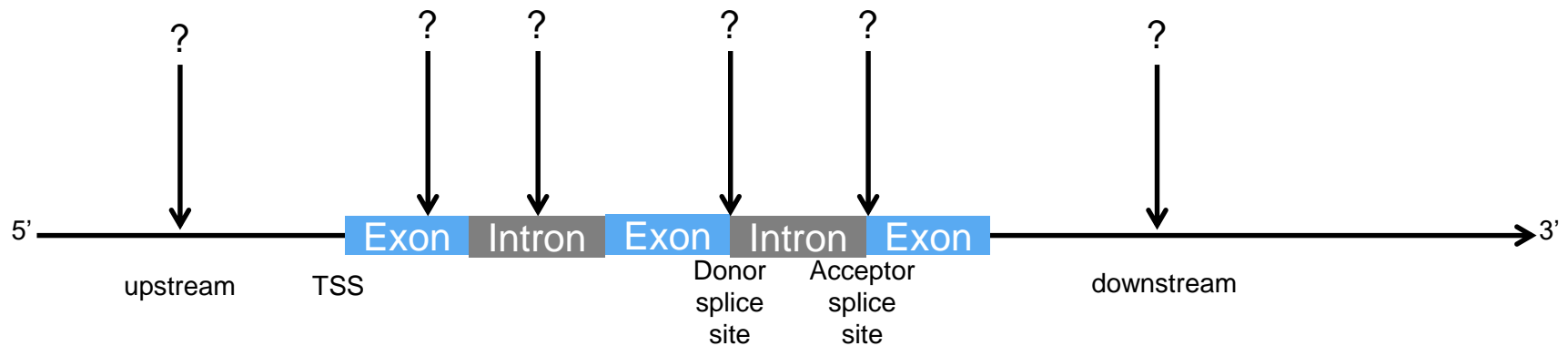
Location



Location

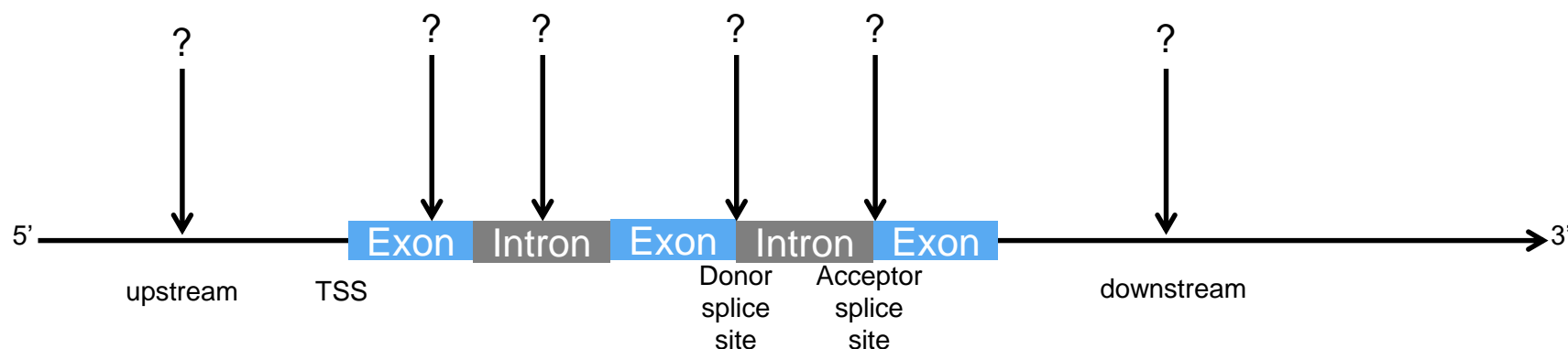


Location



- Regulatory element
 - TFBS
 - miRNA binding
 - Histone mark
 - ...

Location



- Regulatory element

- TFBS
- miRNA binding
- Histone mark
- ...

- Databases : RefSeq, ensembl, encode, motif ...

Known variant ?

- dbSNP :
 - Database of nucleotide variant for 39 species
- Clinvar
 - Relationship between variant and human phenotype
- COSMIC
 - Catalogue of somatic mutation in cancer

Frequency in population

- Rare variant : allele frequency in population $< 5\%$
- 1000 Genome project :
 - 2500 individuals
 - 26 populations
- Exome Sequencing Project (ESP) :
 - ~6500 Exomes
 - 2 populations
- Exome Aggregation Consortium (ExAC) :
 - ~60000 Exomes
 - 7 populations

Functionnal impact

- Impact on protein :
 - synonymous / non synonymous
 - Active domain
 - Protein-protein interaction
 - Structure
- Splicing :
 - Exon skipping
 - Intron retention
 - Alternative splice site
- Transcript expression :
 - Transcription factor binding
 - miRNA binding
- Chromatin conformation
- Pathway

Annotation tools

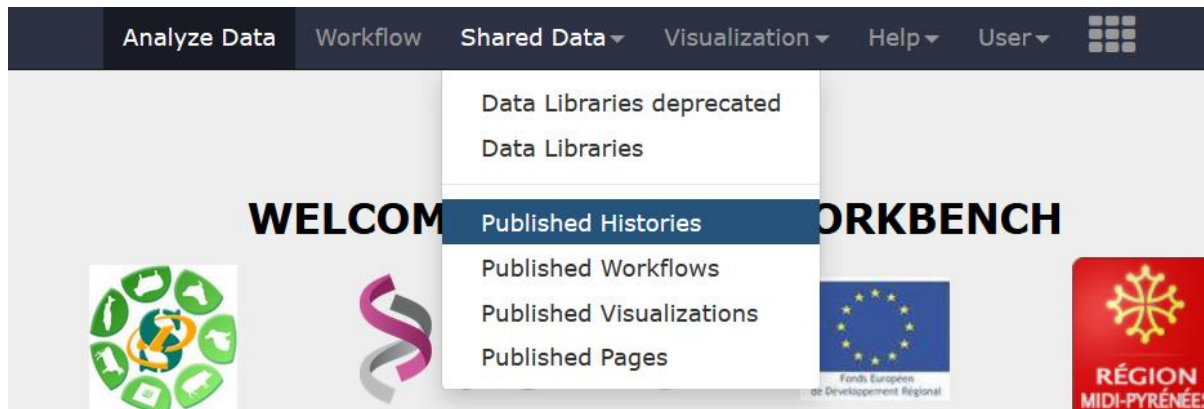
- Annovar
- Ensembl variant_effect_predictor
- snpEff / SnpSift

Snpeff databases

- pre-compiled databases 20 000 species :
 - Downloadable with snpEff download
- snpEff compile :
 - Possibility to compile your own database from a fasta file and a gff

Get data




- In galaxy shared data import : [Variant annotation - Files](#)



Launch snpeff

SnpEff Variant effect and annotation (Galaxy Version 4.1.0) Versions Options

Sequence changes (SNPs, MNPs, InDels)

   1: ex1.vcf

Input format

VCF

Output format

VCF (only if input is VCF)

Genome source

Locally installed reference genome

Genome

Homo sapiens : GRCh37.75

Additional annotations

Select/Unselect all

nextprot
 motif

These are available for only a few genomes

Non-coding and regulatory annotation

Select/Unselect all

CD4
 GM06990
 GM12878
 H1ESC
 HeLa-S3
 HepG2
 HMEC
 HSMM
 HUVEC
 IMR90
 K562
 NH-A
 NHEK

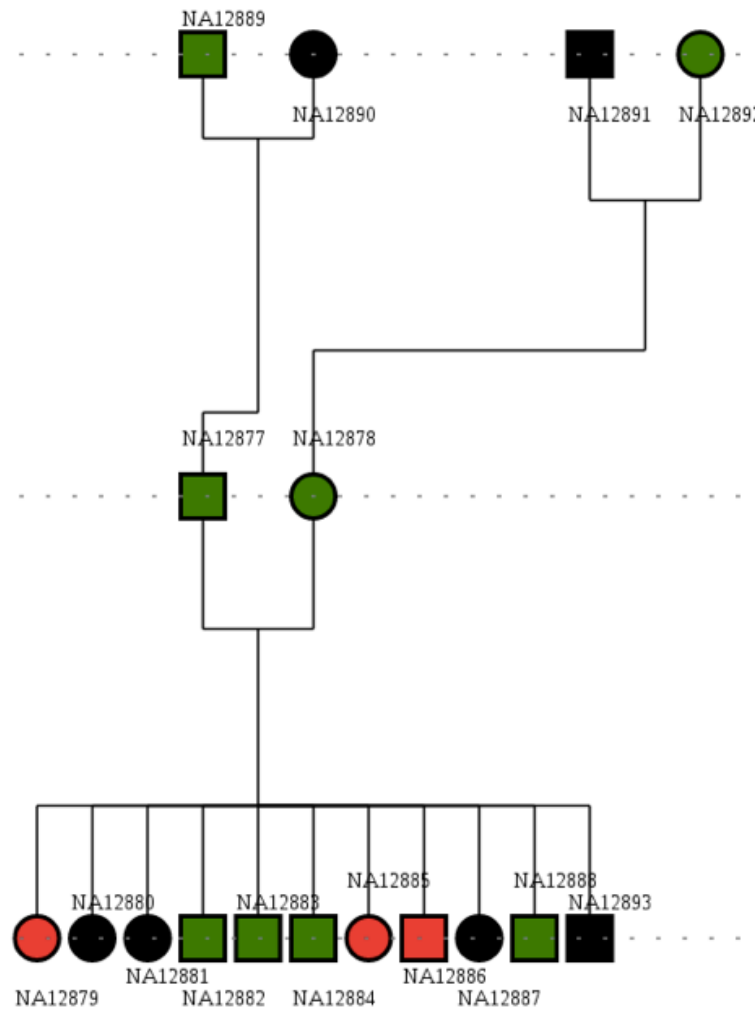
Snpeff output (1/2)

- 'ANN' field in vcf INFO
 - Allele : A, T, C, G, [C-chr1:123456_A>T](#) ...
 - Annotation : missense_variant, synonymous,splice_site ...
 - Annotation_Impact : MODIFIER, LOW, MODERATE, HIGH
 - Gene_Name : HGNC term
 - example : KDM5A
 - Gene_ID : ensembl id (example : [ENSG0000007361](#))
 - Feature_Type : transcript, motif, miRNA ...
 - Feature_ID : [ENST00000399788](#)
 - Transcript_BioType : protein_coding, noncoding ...
 - Rank : Intron or exon rank / total number of introns or exons (eg. 19/28)
 - HGVS.c : c.2594T>C
 - HGVS.p : p.Met865Thr
 - cDNA.pos / cDNA.length : 2957/10763
 - CDS.pos / CDS.length : 2594/5073
 - AA.pos / AA.length : 865/1690
 - ERRORS / WARNINGS / INFO

Snpeff output (2/2)

- Distance :
 - Up/Downstream: Distance to first / last codon
 - Intergenic: Distance to closest gene
 - Distance to closest Intron boundary in exon (+/up/downstream)
 - Distance to closest exon boundary in Intron (+/up/downstream)
 - Distance to first base in MOTIF
 - Distance to first base in miRNA
 - Distance to exon-intron boundary in splice_site or splice _region
 - ChipSeq peak: Distance to summit (or peak center)
 - Histone mark / Histone state: Distance to summit (or peak center)

Cohort pedigree



Launch snpsift CaseControl

Snpsift CaseControl Count samples are in 'case' and 'control' groups. (Galaxy Version 4.1.0) Versions Options

Variant input file in VCF format

3: SnpEff on data 1

Case Control defined in

TFAM file

PLINK TFAM file

2: pedigree.tfam

Read more about TFAM at <http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#tr>

name

name to append to the 'Cases' or 'Controls' tags

- This tool add 2 info :
 - Number of variants in cases: Hom, Het, Count
 - Number of variants in controls: Hom, Het, Count

Snpsift filter

- Allow variant filtering with all info fields using boolean expressions
- All variant that pass previous filters :
(FILTER = 'PASS')
- **And** quality > 30
(FILTER = 'PASS') & (QUAL > 30)
- **Or** quality >= 20 for indels
(FILTER = 'PASS') & ((QUAL > 30) | ((exists INDEL) & (QUAL >= 20)))

SnpSift filter magic functions

- ANN field filter :

(ANN[0].EFFECT has 'frameshift_variant')

(ANN[*].EFFECT has 'frameshift_variant')

- Genotype field :

(isHom(GEN[0]) & isVariant(GEN[0]) & isRef(GEN['sampleX']))


- More info at :

<http://snpeff.sourceforge.net/SnpSift.html#filter>

Filter variants

Snpsift Filter Filter variants using arbitrary expressions (Galaxy Version 4.1.0)

 Versions

 Options

Variant input file in VCF format

   6: Snpsift CaseControl on data 2 and data 3 

Filter criteria

(Cases[0] = 3) & (Controls[0] = 0)

Need help? See below a few examples.


Inverse filter

Yes No

Show lines that do not match filter expression

Filter mode




Retain entries that pass filter, remove other entries 

 Execute

Launch snpsift dbNSFP

SnpSift dbNSFP Add Annotations from dbNSFP or similar annotation DBs (Galaxy Version 4.1.0) Options

Variant input file in VCF format

   7: SnpSift Filter on data 6

dbNSFP

Locally installed dbNSFP database

Genome

GRCh37_dbNSFP2.9

- Annotate with :
 - Uniprot_id
 - Interpro_domain
 - Ensembl_geneid
 - Ensembl_transcriptid
 - MetaSVM_pred
 - MetaLR_pred
 - Reliability_index
 - Cadd_phred
 - 1000Gp1_AF
 - 1000Gp1_EUR_AF
 - ESP6500_EA_AF
 - ExAC_AF
 - ExAC_NFE_AF
 - Clinvar_rs
 - Clivar_clnsig
 - Clinvar_trait

dbNSFP

- Database of non-synonymous variant functional prediction in human
- 83,422,341 pre-annotated SNV
- 17 functional predictions algorithms
- 6 conservation scores
- Allele frequency in populations

SnpSift Extract fields

SnpSift Extract Fields from a VCF file inot a tabular file (Galaxy Version 4.1.0) Options

Variant input file in VCF format

7: SnpSift Filter on data 6

Extract

CHROM POS ID REF ALT ANN[*].GENE ANN[*].GENEID ANN[*].FEATURE ANN[*].FEATUREID ANN[*].EFFECT ANN[*].IMPACT ANN[*].BIOTYPE

Need help? See below a few examples.

One effect per line

Yes No

When variants have more than one effect, lists one effect per line, while all other parameters in the line are repeated across mutiple lines

multiple field separator

;

Separate multiple fields in one column with this character, e.g. a comma, rather than a column for each of the multiple values

empty field text

NA

Represent empty fields with this value, rather than leaving them blank

```
CHROM POS ID REF ALT ANN[*].GENE ANN[*].GENEID ANN[*].FEATURE
ANN[*].FEATUREID ANN[*].EFFECT ANN[*].IMPACT ANN[*].BIOTYPE ANN[*].RANK
ANN[*].HGVS_C ANN[*].HGVS_P ANN[*].DISTANCE ANN[*].ERRORS dbNSFP_Uniprot_id
dbNSFP_clinvar_rs dbNSFP_clinvar_clnsig dbNSFP_clinvar_trait dbNSFP_ExAC_AF
dbNSFP_ExAC_NFE_AF dbNSFP_1000Gp1_AF dbNSFP_1000Gp1_EUR_AF
dbNSFP_ESP6500_EA_AF dbNSFP_Ensembl_geneid dbNSFP_MetaSVM_pred
dbNSFP_MetaLR_pred dbNSFP_Reliability_index dbNSFP_CADD_phred
dbNSFP_Ensembl_transcriptid dbNSFP_Interpro_domain GEN[*].GT
```

Conclusion

- snpEff / snpSift : very versatile tools
- A lot of annotation source for human
 - Choose wisely
- Other organism
 - No population allele frequency
 - Less annotation

Thank you for your attention