# VARIANT ANNOTATION

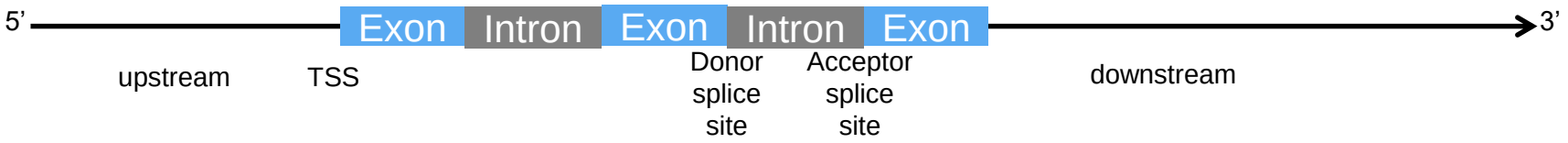
Vivien Deshaies

*vivien.deshaies@icm-institute.org*

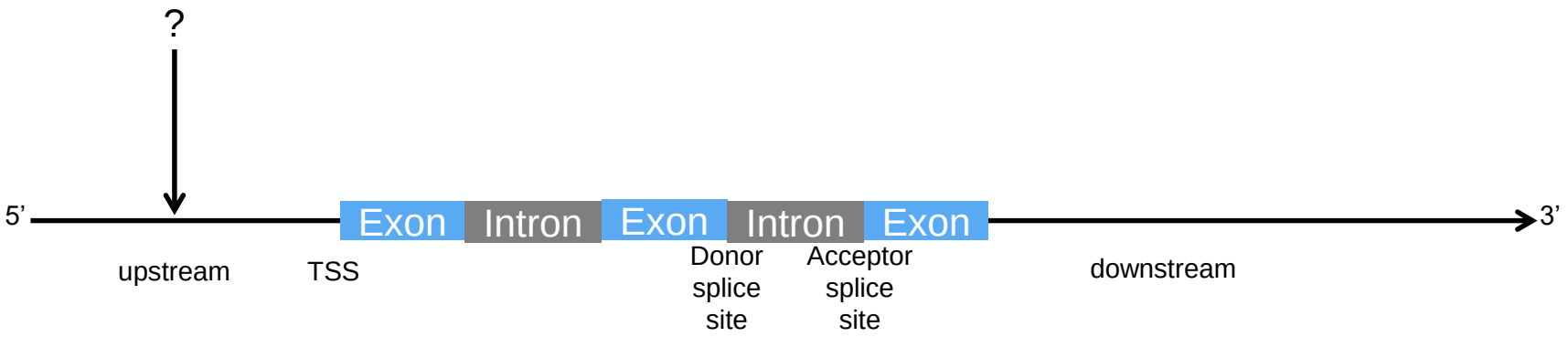
# Goal

Add meta-information on variant to facilitate interpretation

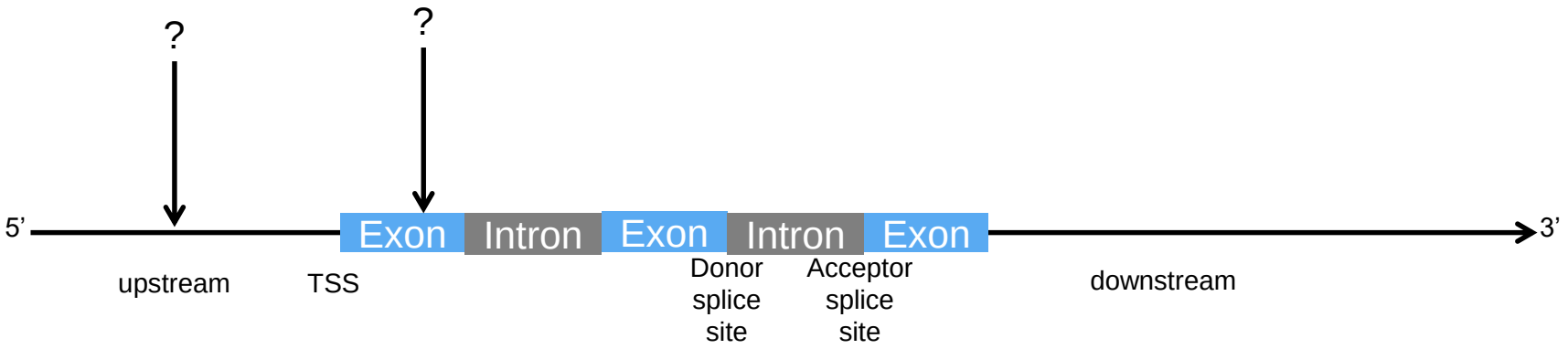
# Location


5’
Exon
Intron
Exon
Intron
Exon
3’
upstream
TSS
Donor splice site
Acceptor splice site
downstream

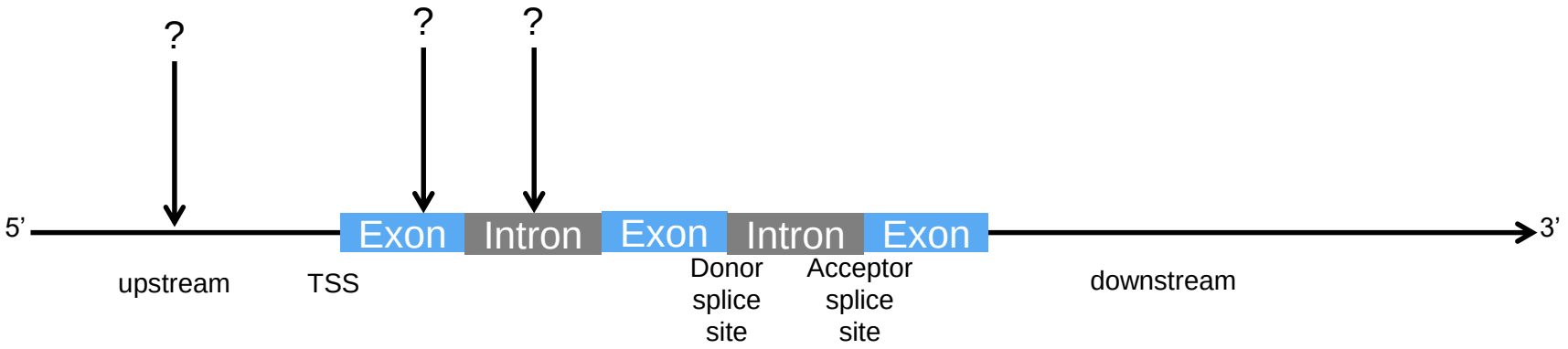
# Location


?
5’
Exon
Intron
Exon
Intron
Exon
3’
upstream
TSS
Donor splice site
Acceptor splice site
downstream

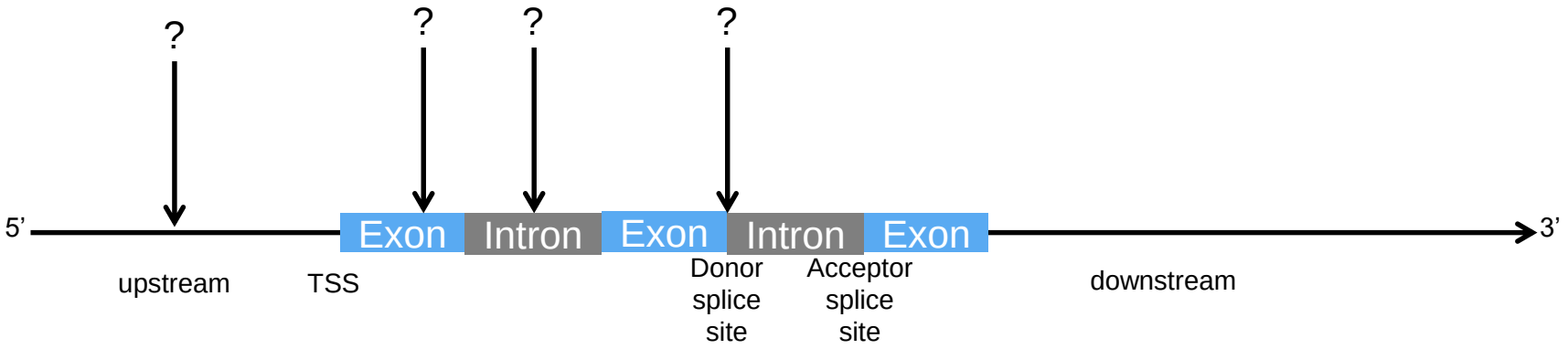
# Location


?
?
5’
Exon
Intron
Exon
Intron
Exon
3’
upstream
TSS
Donor splice site
Acceptor splice site
downstream

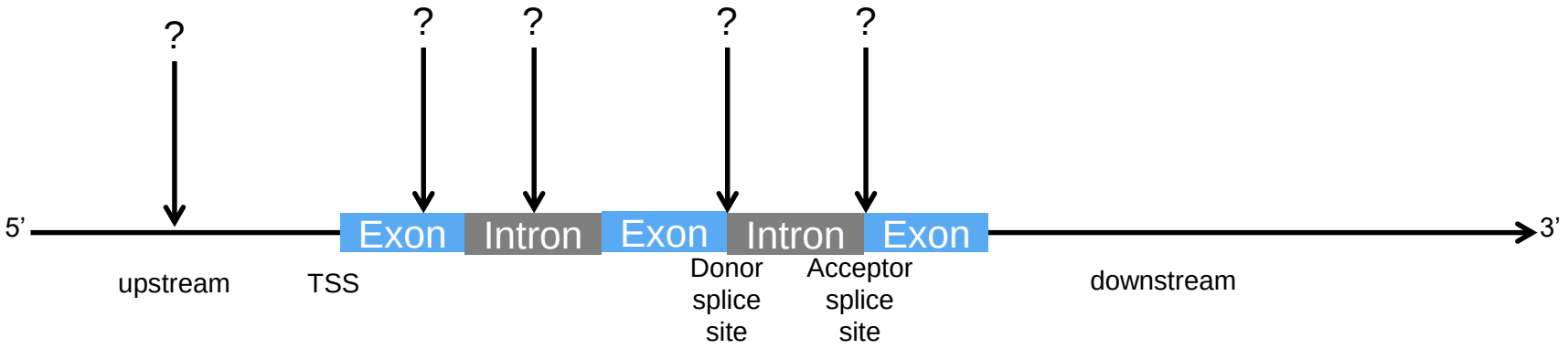
# Location


?
?
?
5’
Exon
Intron
Exon
Intron
Exon
3’
upstream
TSS
Donor splice site
Acceptor splice site
downstream

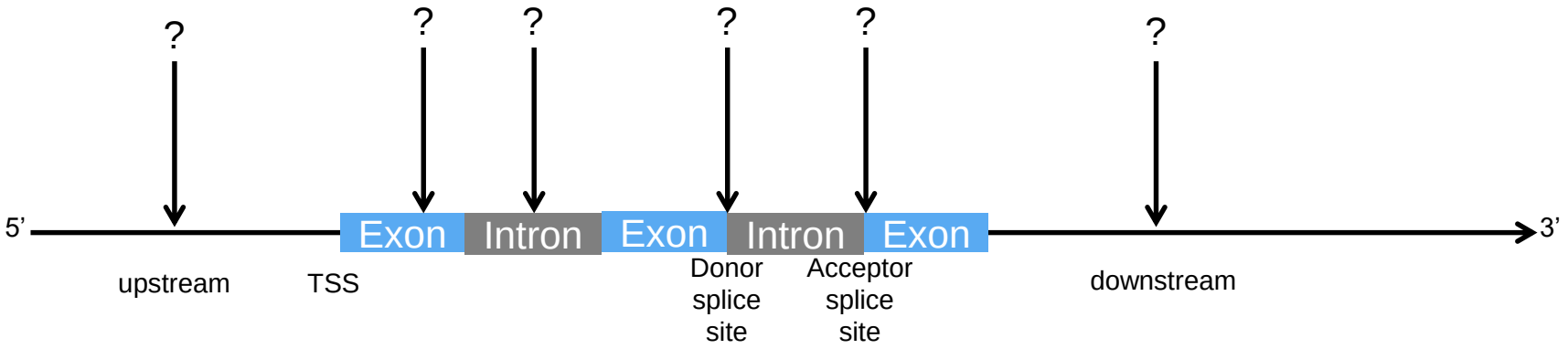
# Location


?
?
?
?
5’
Exon
Intron
Exon
Intron
Exon
3’
upstream
TSS
Donor splice site
Acceptor splice site
downstream

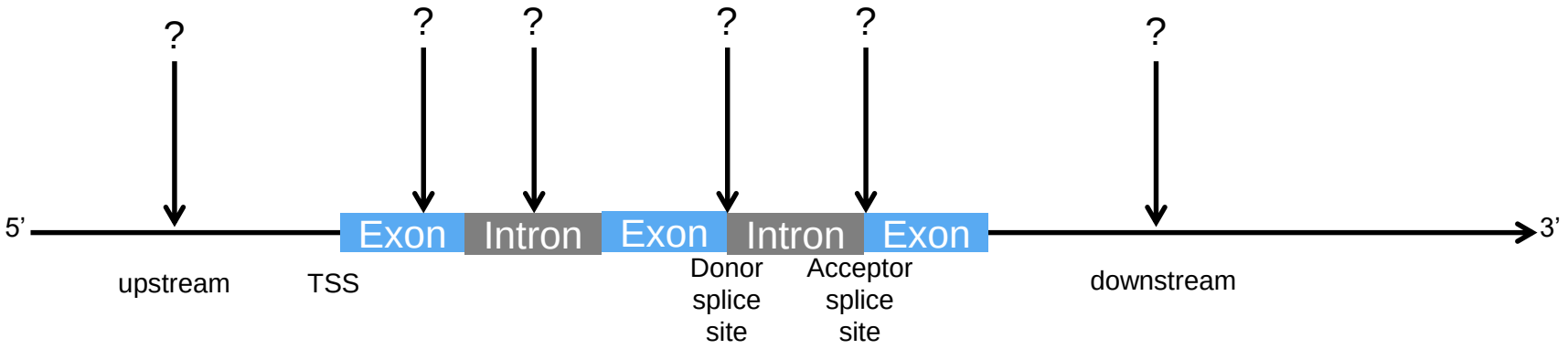
# Location


?
?
?
?
?
5’
Exon
Intron
Exon
Intron
Exon
3’
upstream
TSS
Donor splice site
Acceptor splice site
downstream

# Location

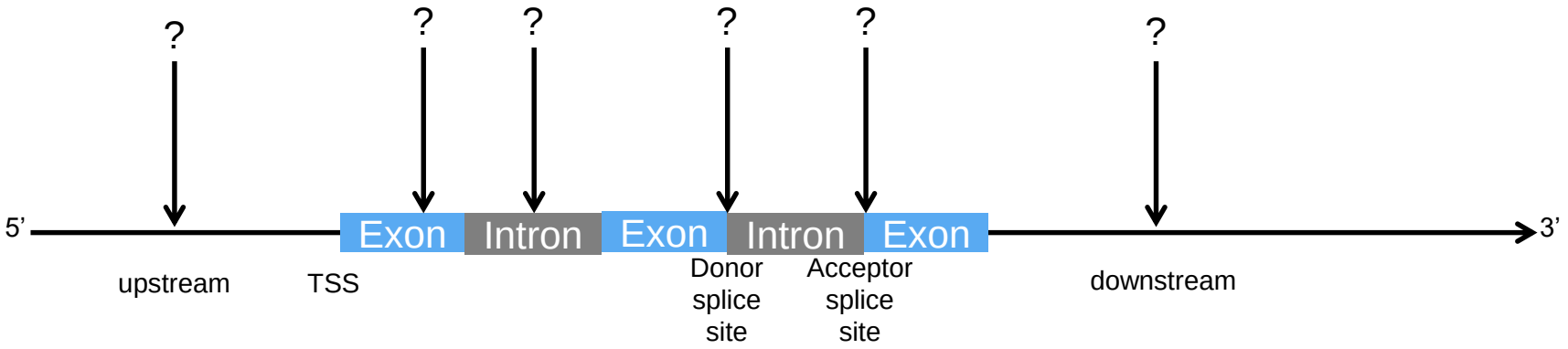

?
?
?
?
?
?
5’
Exon
Intron
Exon
Intron
Exon
3’
upstream
TSS
Donor splice site
Acceptor splice site
downstream

# 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 retension
  - Alternative splice site
- Transcript expression :
  - Transcription factor binding
  - miRNA binding
- Chromatin conformation
- Pathway

# Annotation tools

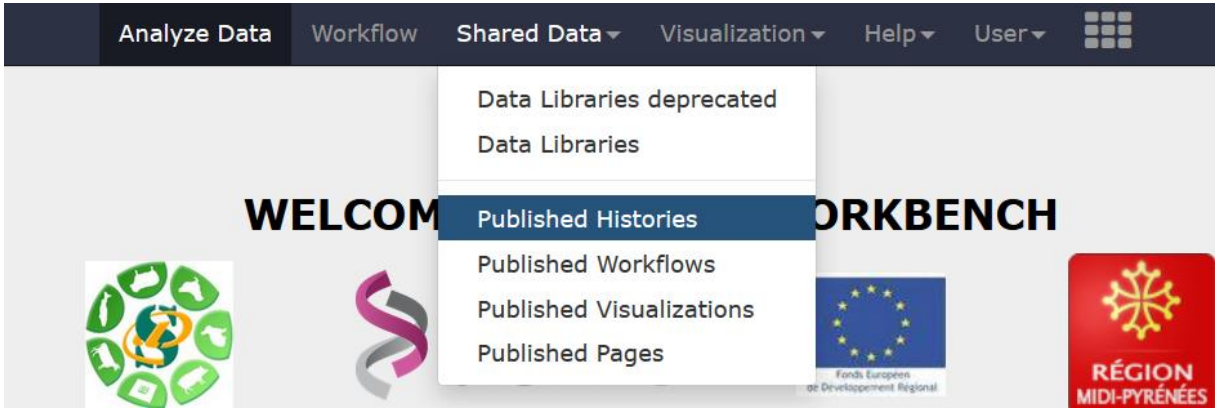
- Annovar
- Ensembl variant_effect_predictor
- snpEff / SnpSift

# Snpeff databases

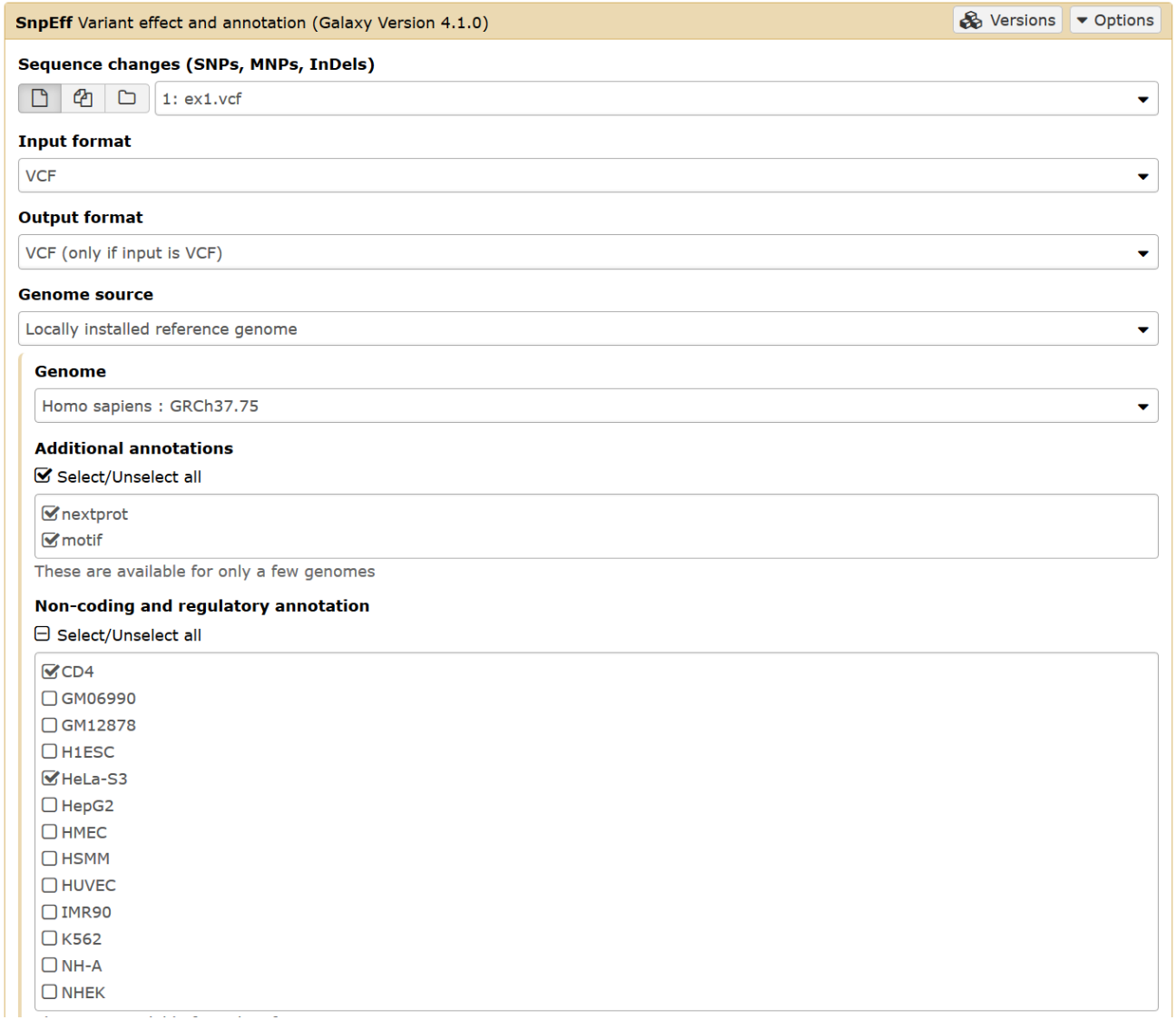
- pre-compiled databases 20 000 species :
  - Downloadable with snpEff download
- snpEff compile :
  - Possibility to compile your own database from a fasta file and a gff

# Get data

- In galaxy shared data import : Variant_annotation - Files



# Launch snpeff



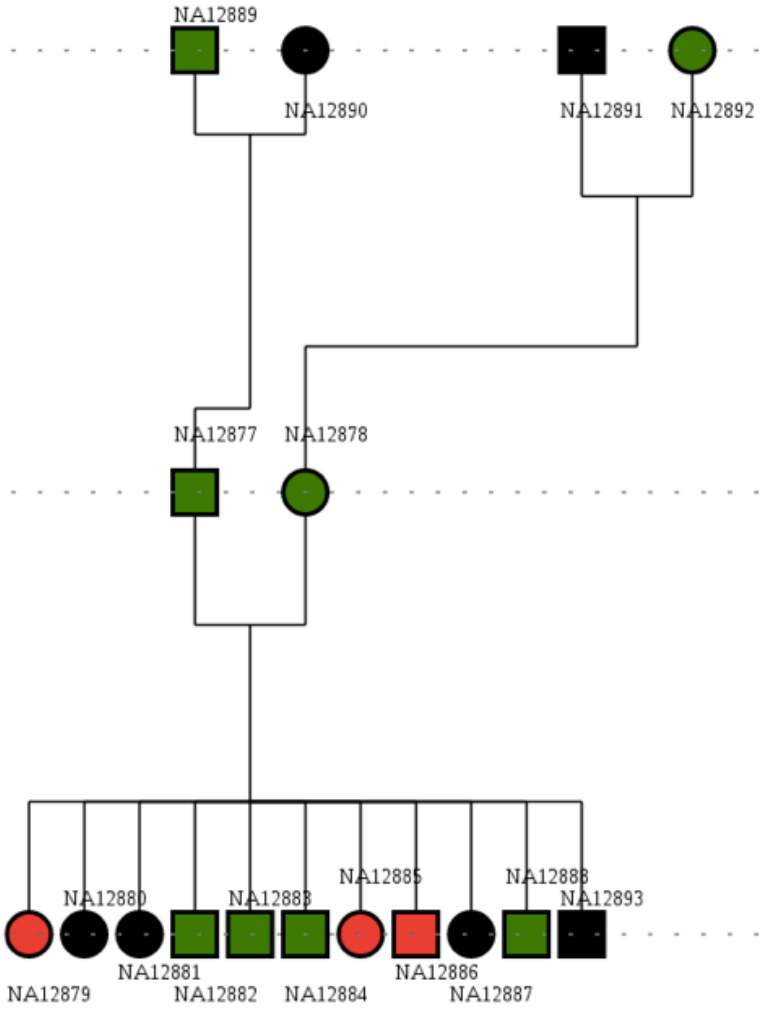
# Snpeff output (1/2)

- ‘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

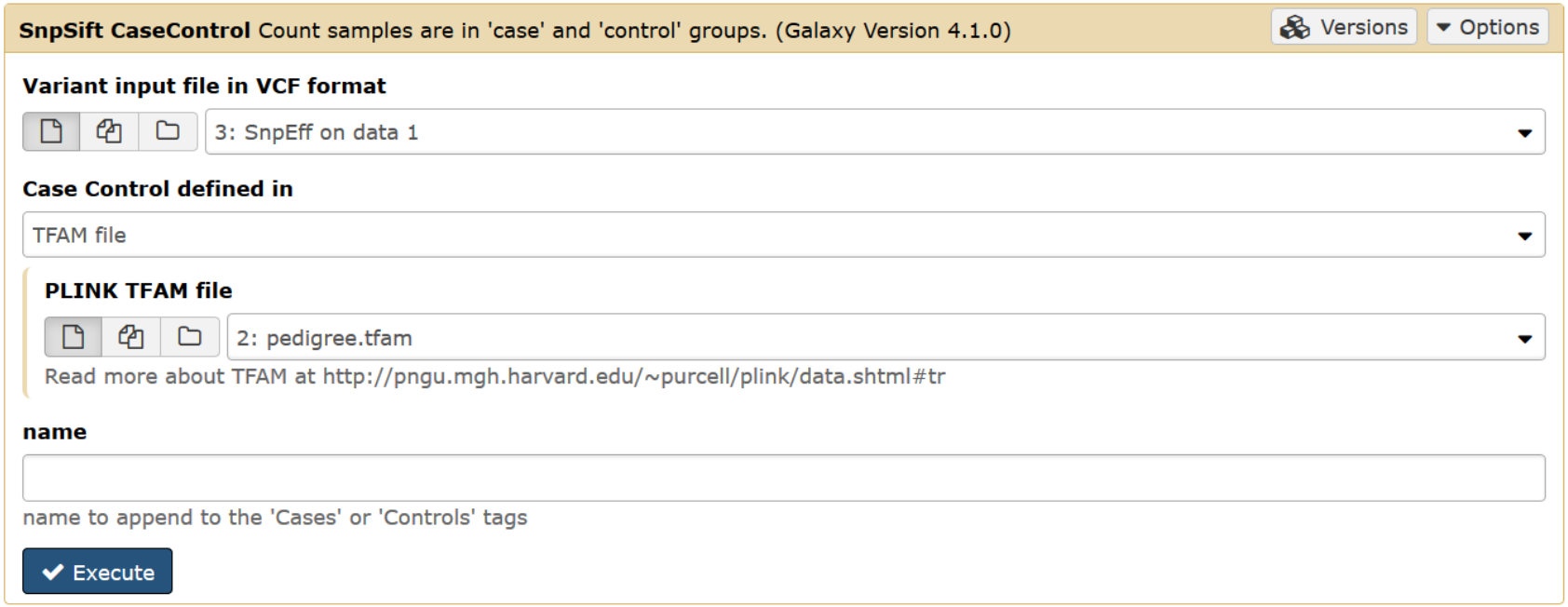
# Snpeff ouput (2/2)

- 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




- 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‘)

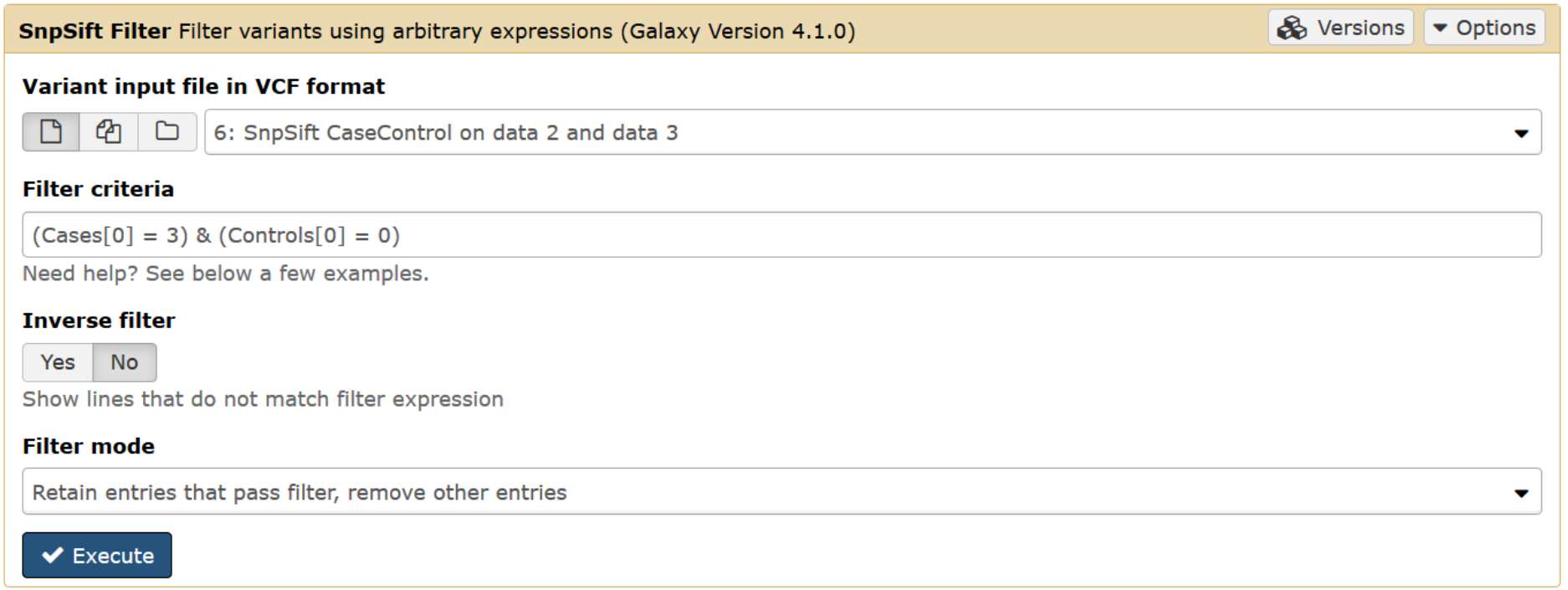
- Genotype field :

  (isHom( GEN[0] ) & isVariant( GEN[0] ) & isRef( GEN[‘sampleX’]))

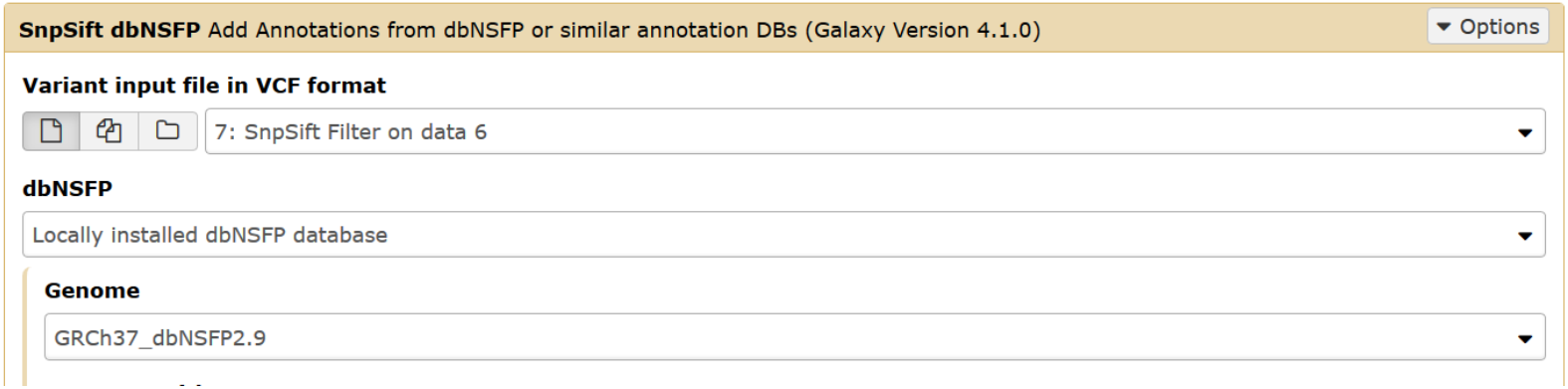
- More info at :

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

# Filter variants



# Launch snpsift dbNSFP

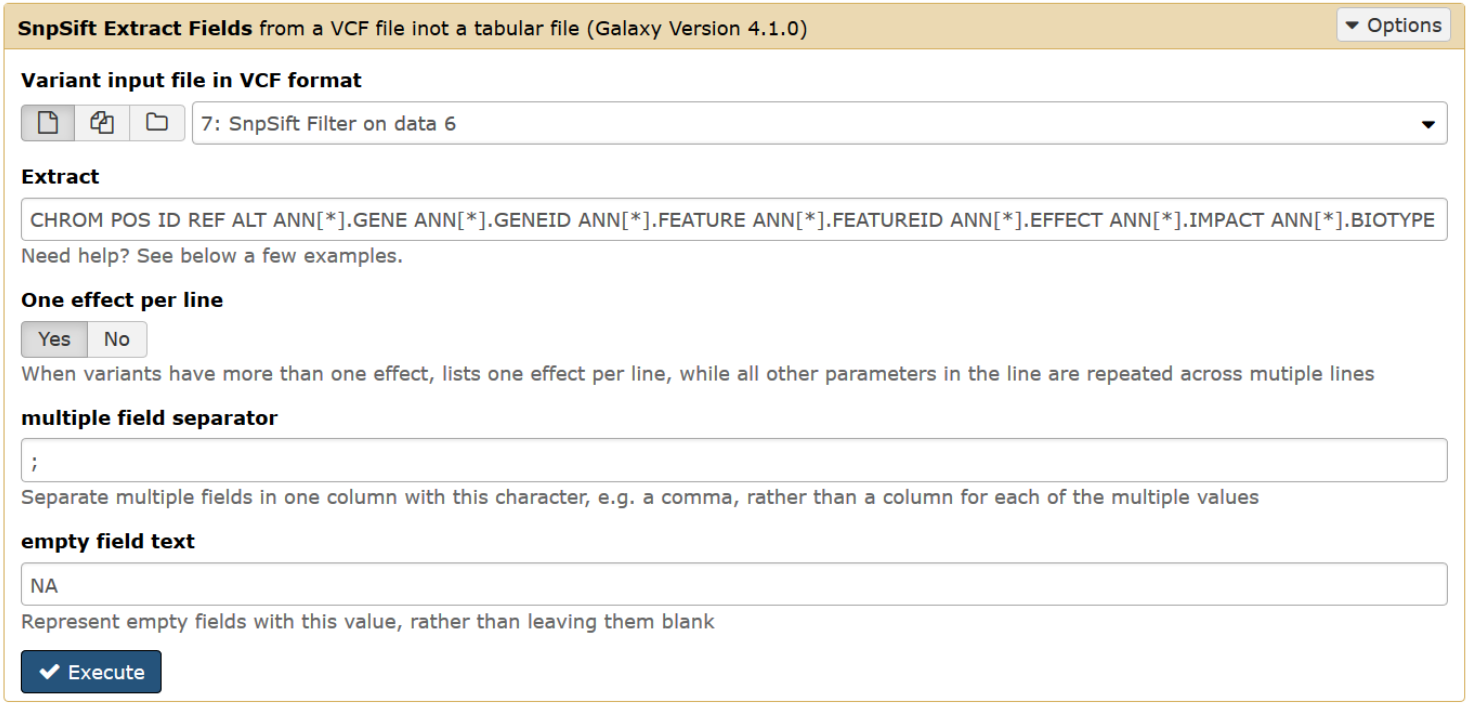



- Annotate with :
  - Uniprot_id
  - Interpro_domain
  - Ensembl_geneid
  - Ensembl_transcriptid
  - MetaSVM_pred
  - MetaLR_pred
  - Reliability_index
  - Cadd_phred
  - 1000Gp1_AF
  - 1000Gp1_EUR_AF
  - ESP6500_EA_AF
  - ExAC_AF
  - ExAC_NFE_AF
  - Clinvar_rs
  - Clivar_clnsig
  - Clinvar_trait

# 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




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