

# BioSyS

## Analysis of Next Generation Sequencing Data

---

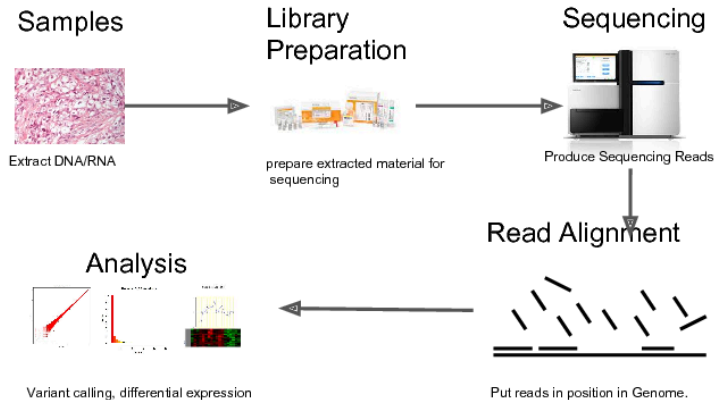
Hugo Martiniano

July 5, 2017

BioISI - Biosystems and Integrative Sciences Institute

# Introduction

## Ensembl Variant Effect Predictor



# Introduction

## Data Analysis

- Computationally expensive
- Large data storage demands



# Introduction

## Read Alignment

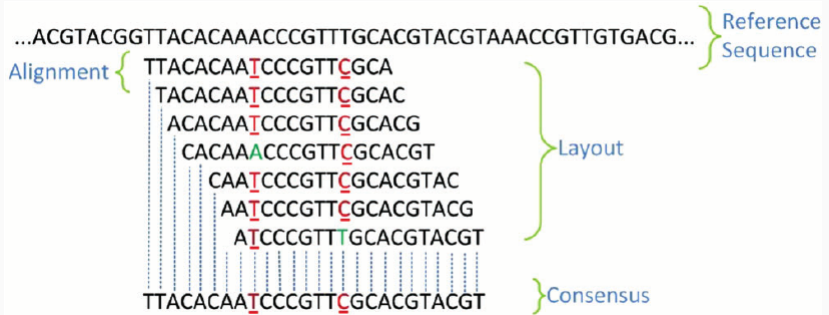
>1:866511 in NA12878/NA12878.bam

Ref: TCCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTGCTGCCGTGCTCAGCGTGAGC

```
60> TCCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----C
60> TCCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----C
60> TCCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CC
60> TCCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCC
60> TcGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CC
60> TCCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----cCC
60> TCCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----C
60> CCGAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CC
60> GAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----cCC
70> GAGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT
70> gagaGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT
70> AGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCctCCCT----CCct----cccc
70> AGAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT
70> GAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CC
70> GAGGCGTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CC
60> GCGTCTCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCC
60> TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCct----CCC
60> TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCct----CCC
60> TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCC
60> cCTGCAGGTAGGAGCCGTGcTGTGCGTGCATAAGAGGGGGCCGTGACTcCCCTCCCTCCCT----CCCT----CCCACCCCT
60> TGAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCTCCCTCCCTCCCT----CCCT----CCCACCCCTga
60> GCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----cCC
60> GCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCTCCCTCCCTCCCT----CCCT----CCCACCCCTGAC
60> GCAGgtAGGAGCCGTGCTgtGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGAC
60> CAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGAC
60> AGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCG
60> AGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCG
60> GAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGCCCT
60> gaGCGCTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGCCCT
60> gaGCGCTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGCCCT
60> GCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGtgccctgc
60> GCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGccctgc
60> CCGTGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGccctgc
60> CcTGTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGccctgcTGTGTC
60> tgcTgtGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGccctgcTGTGTC
60> GCTGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGccctgcTGTGTC
60> TGTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGccctgcTGTGTC
60> GTGCGTGCATAAGAGGGGGCCGTGACTC-----CCCT----CCCACCCCTGACCGTGccctgcTGTGTC
```

# Introduction

## Variant Calling



# Introduction

## Variant Call Format (VCF) file format

Standardized text file format for representing genetic variant data

### Example

**VCF header**

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

**Mandatory header lines**

**Optional header lines (meta-data about the annotations in the VCF body)**

**Body**

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Deletion**

**SNP**

**Large SV**

**Insertion**

**Other event**

**Phased data (G and C above are on the same chromosome)**

## Ensembl Variant Effect Predictor

### Variant Effect Predictor ?

#### VEP for Human GRCh37

If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Species:



Human (Homo sapiens)



Assembly: GRCh38.p5

Name for this data (optional):

Either paste data:

```
1 182712 . A C . . .  
3 319780 . GA G . . .  
19 110747 . G GT . . .
```

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#), [Pileup](#)

**Instant results for first variant ›**

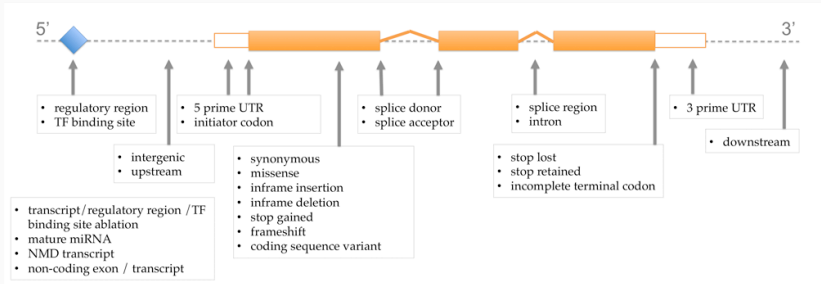
Or upload file:

Browse...

No file selected.

# Introduction

## Variant Types





## Ensembl Variant Effect Predictor

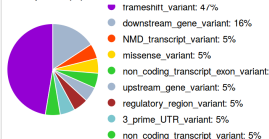
### Variant Effect Predictor results

[Job details](#) 

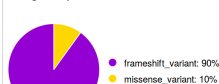
[Summary statistics](#) 

Category	Count
Variants processed	3
Variants remaining after filtering	3
Novel / existing variants	2 (66.7%) / 1 (33.3%)
Overlapped genes	4
Overlapped transcripts	16
Overlapped regulatory features	1

Consequences (all)



Coding consequences



Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype
.	<a href="#">1:182712-182712</a>	C	<a href="#">missense_variant</a>	MODERATE	FO538757.3	<a href="#">ENSG00000279928</a>	Transcript	<a href="#">ENST00000624431</a>	protein_coding
.	<a href="#">1:182712-182712</a>	C	<a href="#">downstream_gene_variant</a>	MODIFIER	FO538757.2	<a href="#">ENSG00000279457</a>	Transcript	<a href="#">ENST00000624735</a>	protein_coding
.	<a href="#">1:182712-182712</a>	C	<a href="#">downstream_gene_variant</a>	MODIFIER	FO538757.2	<a href="#">ENSG00000279457</a>	Transcript	<a href="#">ENST00000623083</a>	protein_coding
.	<a href="#">1:182712-182712</a>	C	<a href="#">downstream_gene_variant</a>	MODIFIER	FO538757.2	<a href="#">ENSG00000279457</a>	Transcript	<a href="#">ENST00000623834</a>	protein_coding
.	<a href="#">3:319780-319781</a>	-	<a href="#">frameshift_variant</a>	HIGH	CHL1	<a href="#">ENSG00000134121</a>	Transcript	<a href="#">ENST00000449294</a>	protein_coding
.	<a href="#">3:319780-319781</a>	-	<a href="#">frameshift_variant</a>	HIGH	CHL1	<a href="#">ENSG00000134121</a>	Transcript	<a href="#">ENST00000397491</a>	protein_coding
.	<a href="#">3:319780-319781</a>	-	<a href="#">frameshift_variant</a>	HIGH	CHL1	<a href="#">ENSG00000134121</a>	Transcript	<a href="#">ENST00000620033</a>	protein_coding
.	<a href="#">3:319780-319781</a>	-	<a href="#">frameshift_variant</a>	HIGH	CHL1	<a href="#">ENSG00000134121</a>	Transcript	<a href="#">ENST00000421198</a>	protein_coding
.	<a href="#">3:319780-319781</a>	-	<a href="#">frameshift_variant</a>	HIGH	CHL1	<a href="#">ENSG00000134121</a>	Transcript	<a href="#">ENST00000427688</a>	protein_coding

Thank you for your attention. Questions?