BioSyS

Analysis of Next Generation Sequencing Data

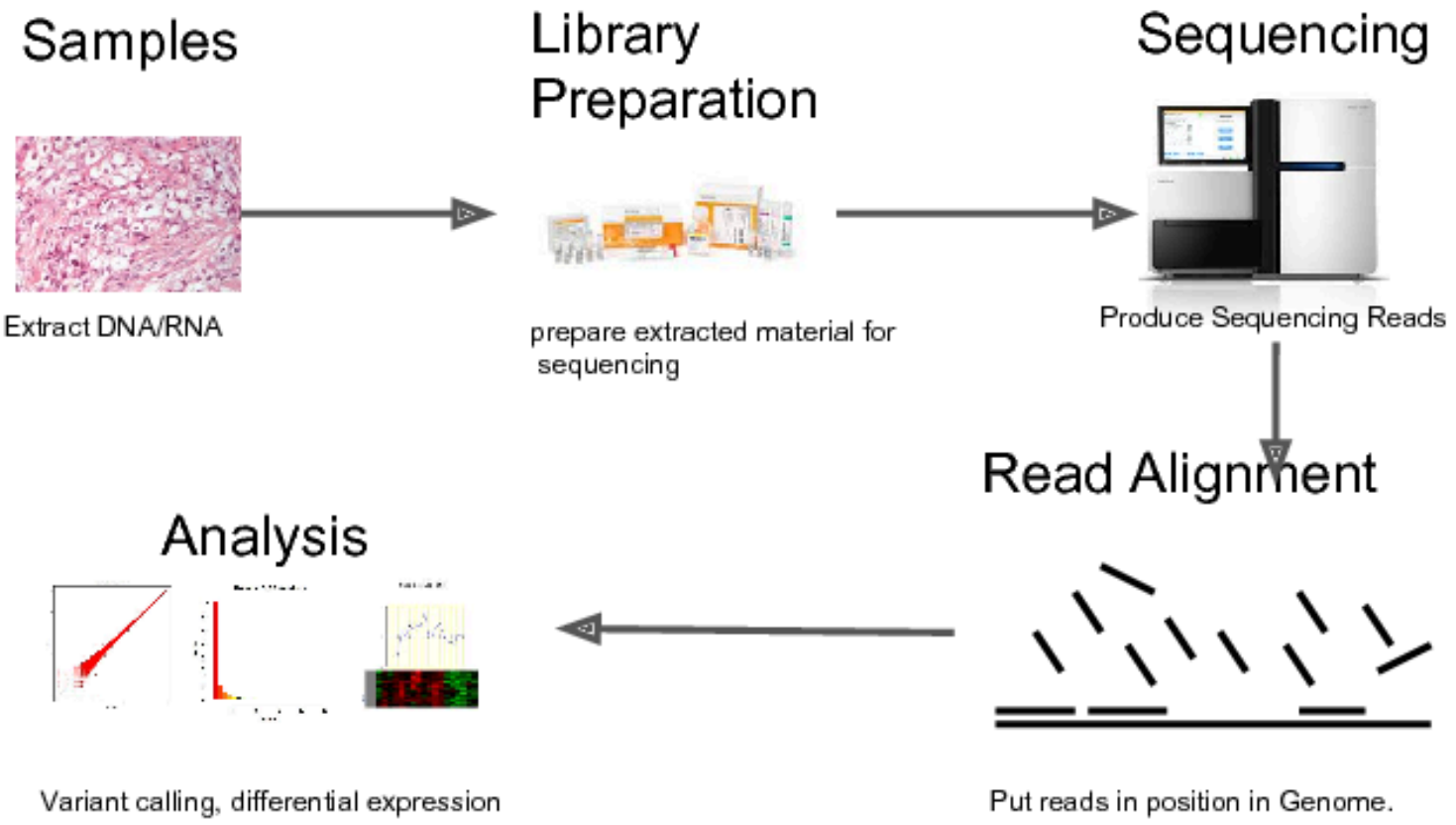
Hugo Martiniano

July 5, 2017

BioISI - Biosystems and Integrative Sciences Institute

Introduction

Ensembl Variant Effect Predictor



Introduction

Data Analysis

- Computationally expensive
- Large data storage demands

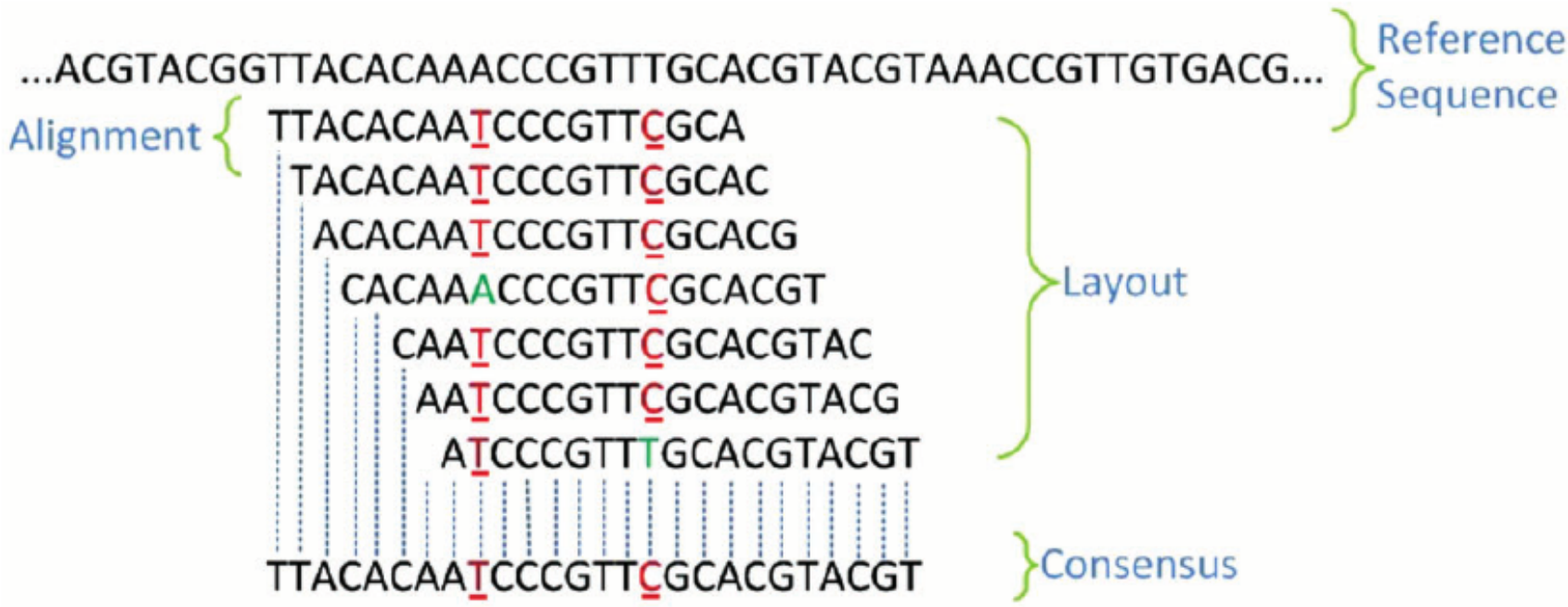
Read Alignment

```
>1:866511 in NA12878/NA12878.bam
  Ref:  TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTCTGCTGTCCGCTGTCTCAGCGTGAGC
                                                                   *

   60>  TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----C
   60<  TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----C
   60>  TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CC
   60<  TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCC
   60<  TCcGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CC
   60>  TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----cCC
   60<  TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----C
   60<   CCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CC
   60>     GAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCt----cCC
   70<     GAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT
   70<     gagaGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT
   60>      AGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCctCCCT----CCCt----ccccc
   70<      AGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----C
   70>       GAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CC
   70<       GAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CC
   60<         GGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCC
   60>             TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCt----CCC
   60>             TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCt----CCC
   60<             TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCC
   60<              ccTGCAGGTAGGaGCCGTGcTGTGCGTGCATAAGAGGGGGCCGTGACTcCCCTCCCTCCCT----CCCT----CCCACCCCT
   60>                TGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCtga
   60>                 GCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----cCC
   60<                 GCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGAC
   60<                 GCAGgtAGGAGCCGTGCTGtGCGTGCATAAGAGGGGGCCGtGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGAC
   60<                  CAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACC
   60<                   AGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCG
   60<                   AGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCG
   60<                         GAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCT
   60<                         gaGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCT
   60<                         gaGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCT
   60>                           GCCGTGCTGTGCGtGCATAAGAGGGGGCCGTGACTCCCCtCCCTCCCT----CCCt----CCCACCCCtGACCgtgccctgc
   60<                           GCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGC
   60<                            CCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCT
   60<                               tgCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTC
   60<                               tgctgtGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTC
   60<                                GCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCaCCCCTGACCGTGCCCTGCTGTCT
   60<                                  TGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTCTGC
   60<                                   GTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTCTGCT
```

Introduction

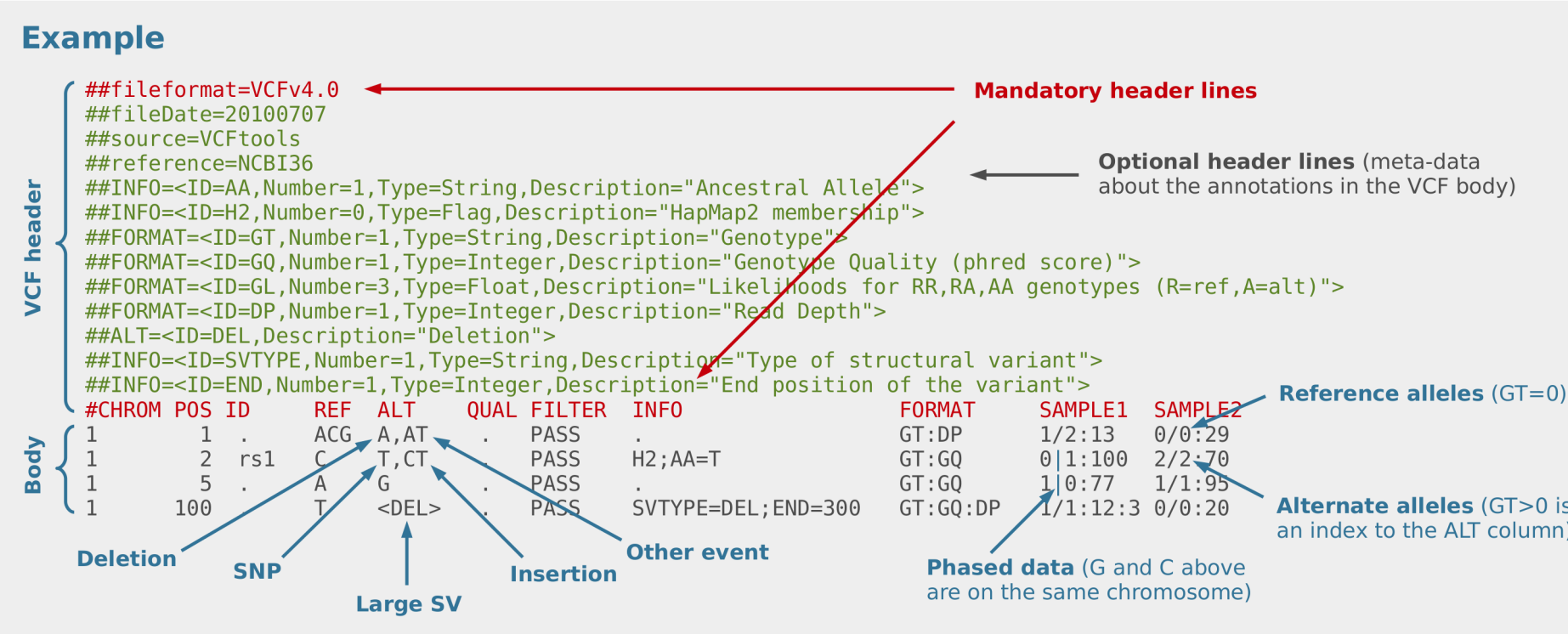
Variant Calling



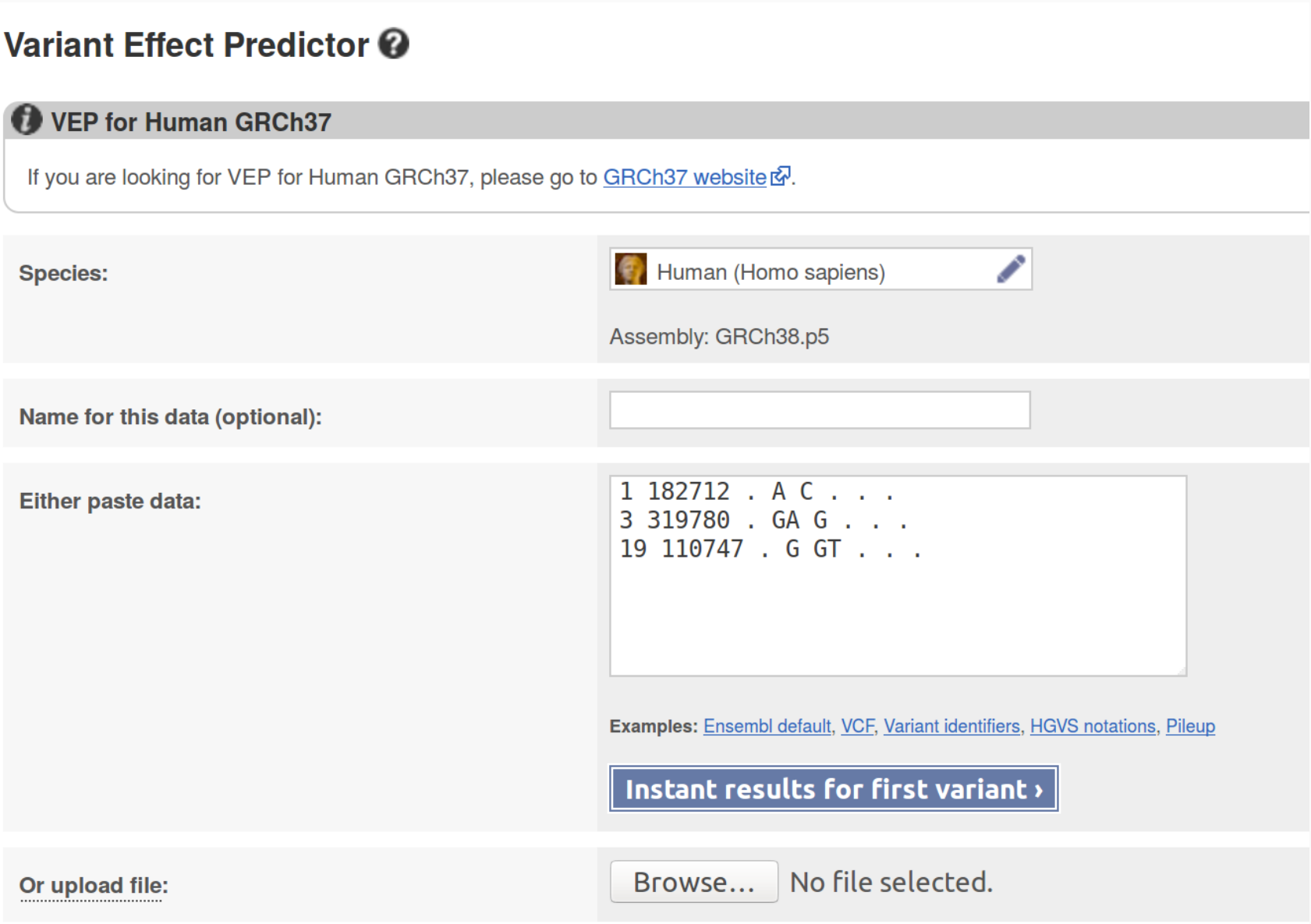
Introduction

Variant Call Format (VCF) file format

Standardized text file format for representing genetic variant data

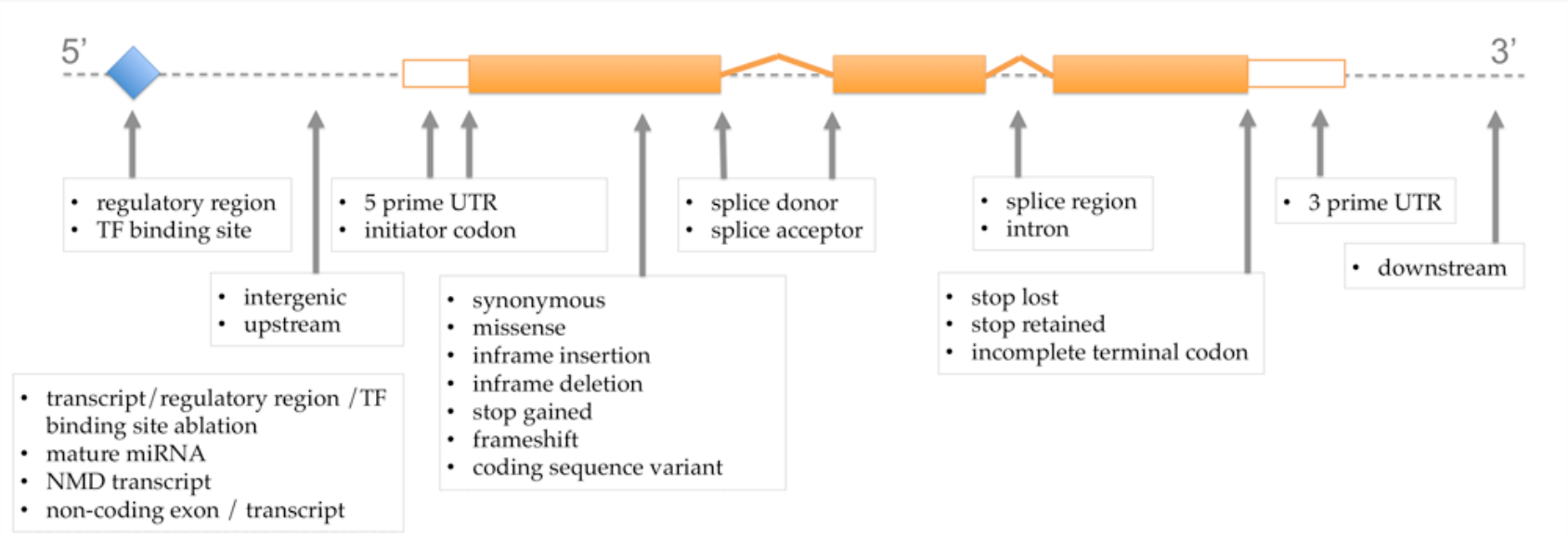


Ensembl Variant Effect Predictor



Introduction

Variant Types



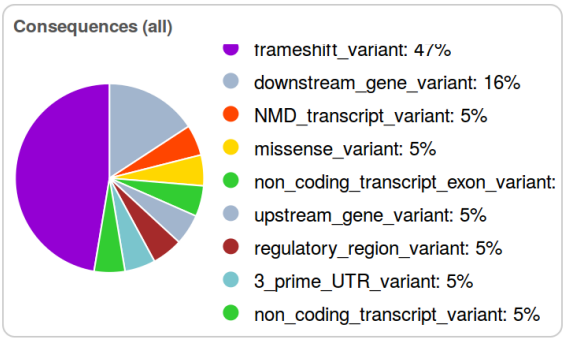
Ensembl Variant Effect Predictor

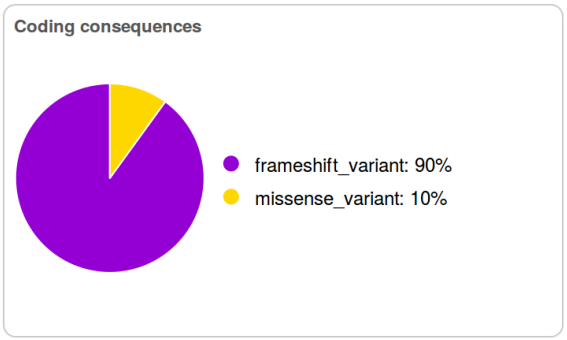
Variant Effect Predictor results

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





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

Thank you for your attention. Questions?