NIG Bioinformatics Training Program

Eli Kaminuma (National Institute of Genetics)

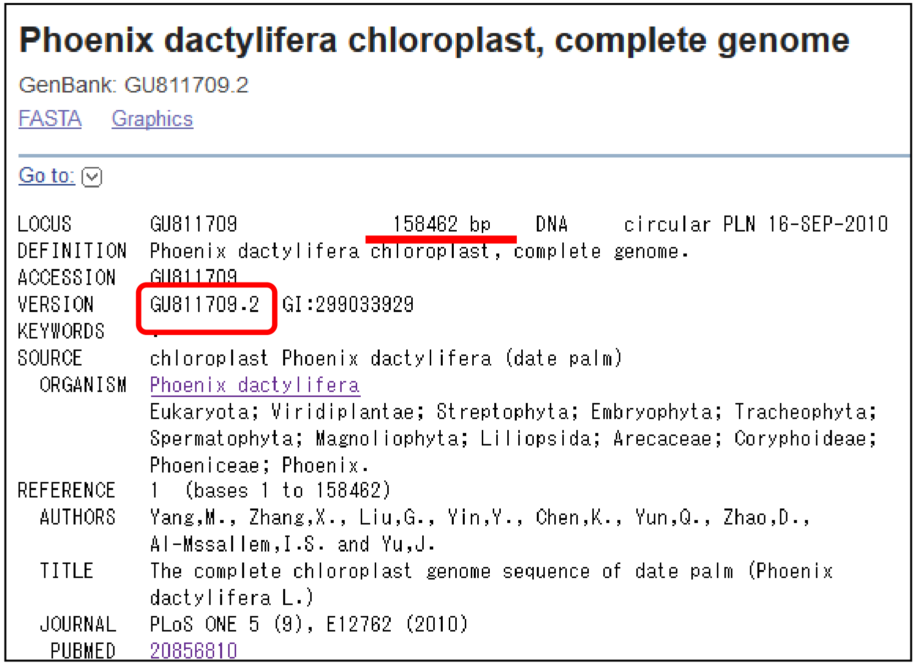
2015-11-27

5. SNP Annotation

- VCF format
- SnpEff tool

Difference between Ref. versions

http://www.ncbi.nlm.nih.gov/nuccore/299033929/



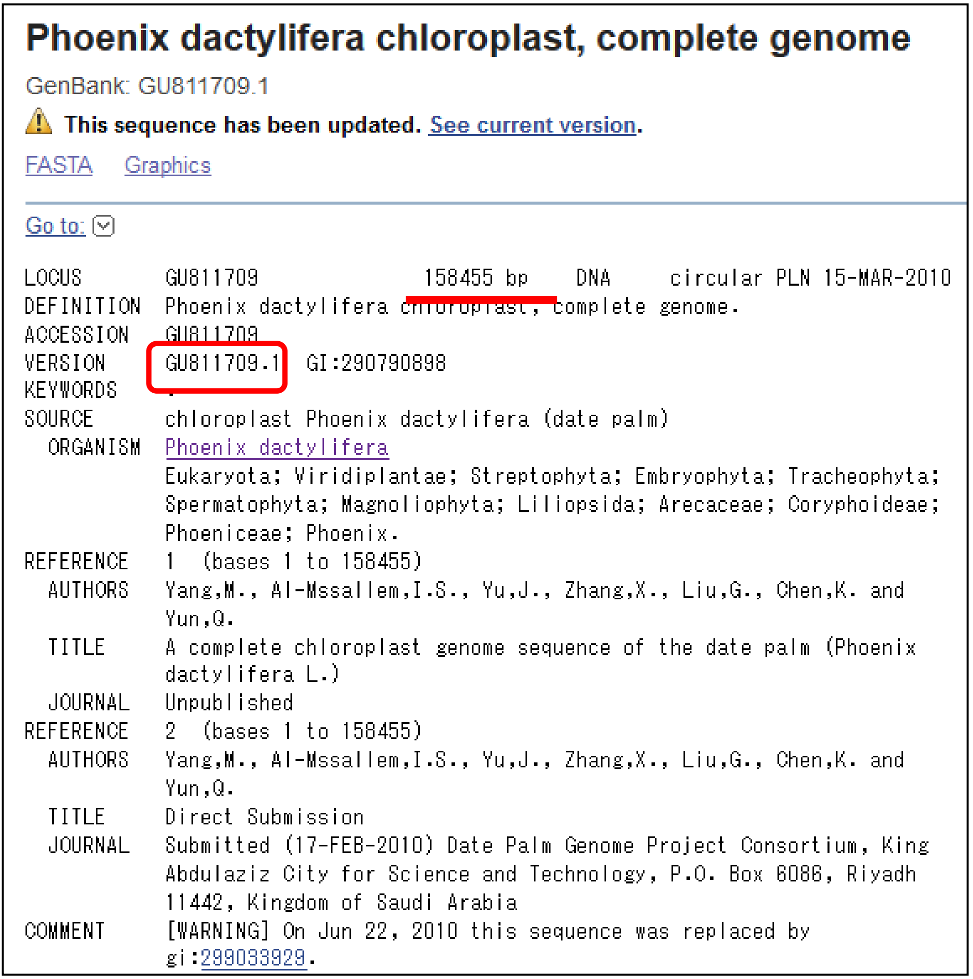
Phoenix dactylifera chloroplast, complete genome

GenBank: GU811709.2

FASTA Graphics

Go to:

```
LOCUS       GU811709              158462 bp    DNA     circular PLN 16-SEP-2010
DEFINITION  Phoenix dactylifera chloroplast, complete genome.
ACCESSION   GU811709
VERSION     GU811709.2  GI:299033929
KEYWORDS    .
SOURCE      chloroplast Phoenix dactylifera (date palm)
  ORGANISM  Phoenix dactylifera
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Arecaceae; Coryphoideae;
            Phoeniceae; Phoenix.
REFERENCE   1  (bases 1 to 158462)
  AUTHORS   Yang,M., Zhang,X., Liu,G., Yin,Y., Chen,K., Yun,Q., Zhao,D.,
            Al-Mssallem,I.S. and Yu,J.
  TITLE     The complete chloroplast genome sequence of date palm (Phoenix
            dactylifera L.)
  JOURNAL   PLoS ONE 5 (9), E12762 (2010)
   PUBMED   20856810
```



Phoenix dactylifera chloroplast, complete genome

GenBank: GU811709.1

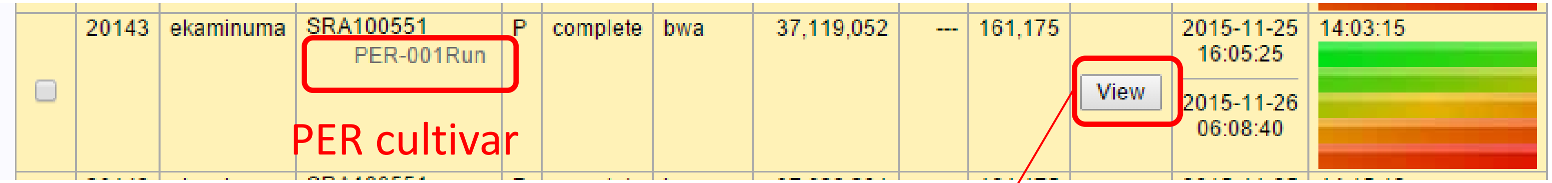
⚠ **This sequence has been updated. See current version.**

FASTA Graphics

Go to:

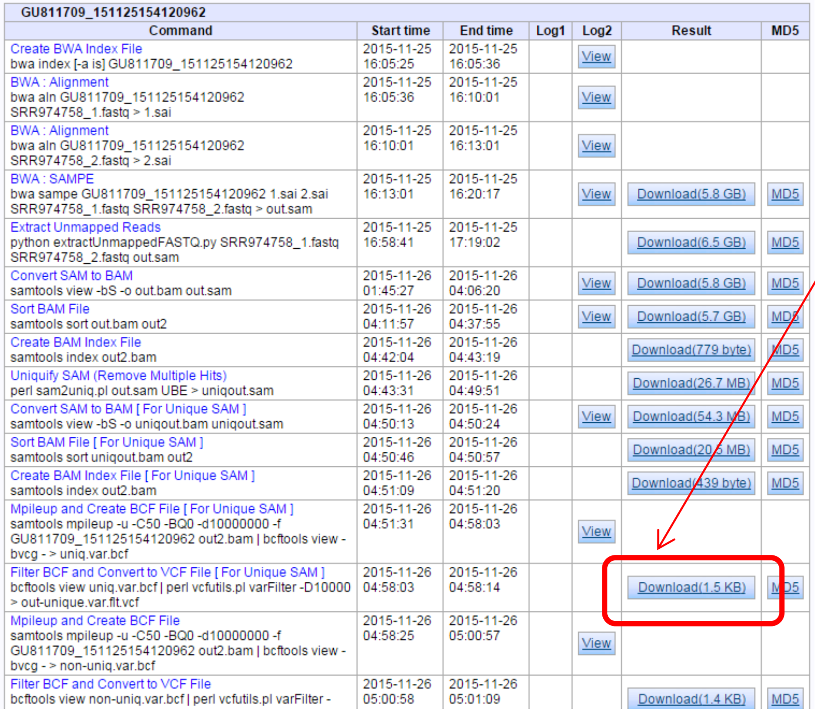
```
LOCUS       GU811709              158455 bp    DNA     circular PLN 15-MAR-2010
DEFINITION  Phoenix dactylifera chloroplast, complete genome.
ACCESSION   GU811709
VERSION     GU811709.1  GI:290790898
KEYWORDS    .
SOURCE      chloroplast Phoenix dactylifera (date palm)
  ORGANISM  Phoenix dactylifera
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Arecaceae; Coryphoideae;
            Phoeniceae; Phoenix.
REFERENCE   1  (bases 1 to 158455)
  AUTHORS   Yang,M., Al-Mssallem,I.S., Yu,J., Zhang,X., Liu,G., Chen,K. and
            Yun,Q.
  TITLE     A complete chloroplast genome sequence of the date palm (Phoenix
            dactylifera L.)
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 158455)
  AUTHORS   Yang,M., Al-Mssallem,I.S., Yu,J., Zhang,X., Liu,G., Chen,K. and
            Yun,Q.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-FEB-2010) Date Palm Genome Project Consortium, King
            Abdulaziz City for Science and Technology, P.O. Box 6086, Riyadh
            11442, Kingdom of Saudi Arabia
COMMENT     [WARNING] On Jun 22, 2010 this sequence was replaced by
            gi:299033929.
```

Variant Call Format(VCF)

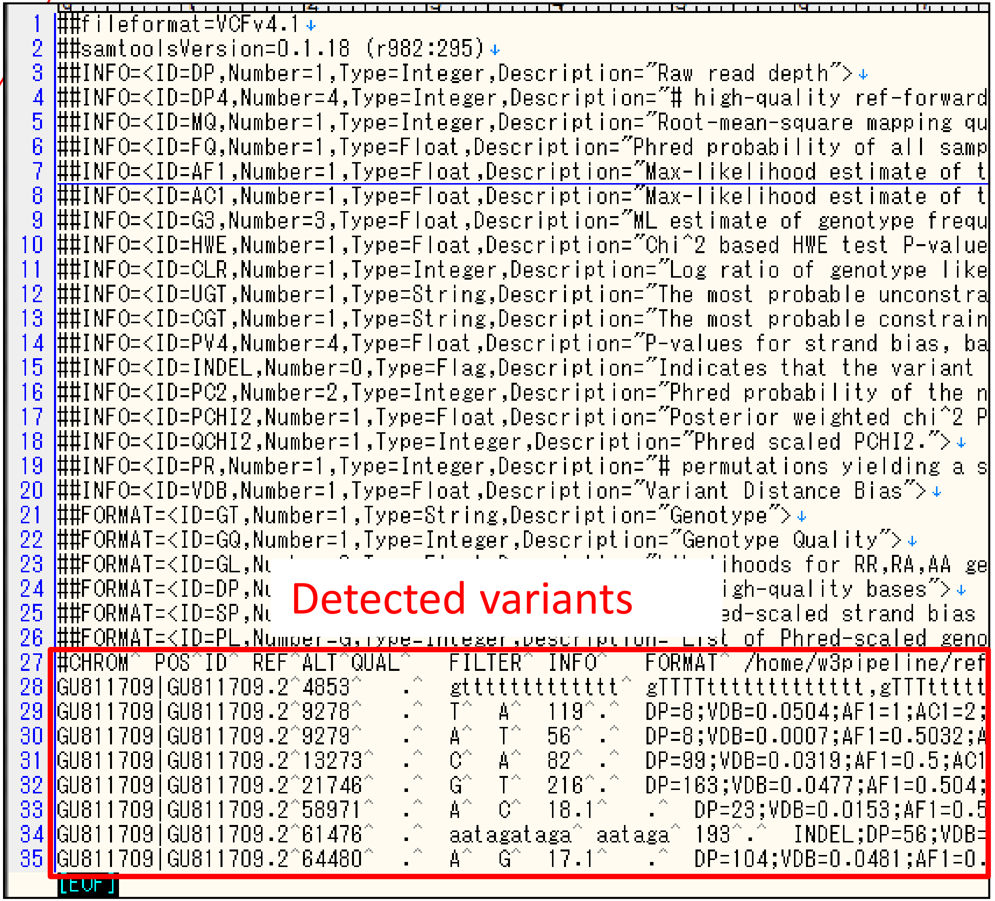


□	20143	ekaminuma	SRA100551 PER-001Run	P	complete	bwa	37,119,052	---	161,175	View	2015-11-25 16:05:25 2015-11-26 06:08:40	14:03:15

PER cultivar



GU811709_151125154120962						
Command	Start time	End time	Log1	Log2	Result	MD5
Create BWA Index File bwa index [-a is] GU811709_151125154120962	2015-11-25 16:05:25	2015-11-25 16:05:36		View		
BWA : Alignment bwa aln GU811709_151125154120962 SRR974758_1.fastq > 1.sai	2015-11-25 16:05:36	2015-11-25 16:10:01		View		
BWA : Alignment bwa aln GU811709_151125154120962 SRR974758_2.fastq > 2.sai	2015-11-25 16:10:01	2015-11-25 16:13:01		View		
BWA : SAMPE bwa sampe GU811709_151125154120962 1.sai 2.sai SRR974758_1.fastq SRR974758_2.fastq > out.sam	2015-11-25 16:13:01	2015-11-25 16:20:17		View	Download(5.8 GB)	MD5
Extract Unmapped Reads python extractUnmappedFASTQ.py SRR974758_1.fastq SRR974758_2.fastq out.sam	2015-11-25 16:58:41	2015-11-25 17:19:02			Download(6.5 GB)	MD5
Convert SAM to BAM samtools view -bS -o out.bam out.sam	2015-11-26 01:45:27	2015-11-26 04:06:20		View	Download(5.8 GB)	MD5
Sort BAM File samtools sort out.bam out2	2015-11-26 04:11:57	2015-11-26 04:37:55		View	Download(5.7 GB)	MD5
Create BAM Index File samtools index out2.bam	2015-11-26 04:42:04	2015-11-26 04:43:19			Download(779 byte)	MD5
Uniquify SAM (Remove Multiple Hits) perl sam2uniq.pl out.sam UBE > uniqout.sam	2015-11-26 04:43:31	2015-11-26 04:49:51			Download(26.7 MB)	MD5
Convert SAM to BAM [For Unique SAM] samtools view -bS -o uniqout.bam uniqout.sam	2015-11-26 04:50:13	2015-11-26 04:50:24		View	Download(54.3 MB)	MD5
Sort BAM File [For Unique SAM] samtools sort uniqout.bam out2	2015-11-26 04:50:46	2015-11-26 04:50:57			Download(20.5 MB)	MD5
Create BAM Index File [For Unique SAM] samtools index out2.bam	2015-11-26 04:51:09	2015-11-26 04:51:20			Download(439 byte)	MD5
Mpileup and Create BCF File [For Unique SAM] samtools mpileup -u -C50 -BQ0 -d10000000 -f GU811709_151125154120962 out2.bam | bcftools view -bvcg - > uniq.var.bcf	2015-11-26 04:51:31	2015-11-26 04:58:03		View		
Filter BCF and Convert to VCF File [For Unique SAM] bcftools view uniq.var.bcf | perl vcfutils.pl varFilter -D10000 > out-unique.var.flt.vcf	2015-11-26 04:58:03	2015-11-26 04:58:14			Download(1.5 KB)	MD5
Mpileup and Create BCF File samtools mpileup -u -C50 -BQ0 -d10000000 -f GU811709_151125154120962 out2.bam | bcftools view -bvcg - > non-uniq.var.bcf	2015-11-26 04:58:25	2015-11-26 05:00:57		View		
Filter BCF and Convert to VCF File bcftools view non-uniq.var.bcf | perl vcfutils.pl varFilter -	2015-11-26 05:00:58	2015-11-26 05:01:09			Download(1.4 KB)	MD5



```
##fileformat=VCFv4.1↓
##samtoolsVersion=0.1.18 (r982:295)↓
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">↓
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping qu
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samp
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of t
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of t
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequ
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value
##INFO=<ID=CLR,Number=1,Type=Integer,Description="Log ratio of genotype like
##INFO=<ID=UGT,Number=1,Type=String,Description="The most probable unconstra
##INFO=<ID=CGT,Number=1,Type=String,Description="The most probable constrain
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, ba
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant
##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the n
##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P
##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">↓
##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a s
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">↓
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">↓
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">↓
##FORMAT=<ID=GL,Nu                                    ihoods for RR,RA,AA ge
##FORMAT=<ID=DP,Nu                                    igh-quality bases">↓
##FORMAT=<ID=SP,Nu                                    ed-scaled strand bias
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled geno
#CHROM^  POS^ID^  REF^ALT^QUAL^     FILTER^  INFO^     FORMAT^  /home/w3pipeline/ref
GU811709|GU811709.2^4853^      .^   gttttttttttttt^  gTTTTtttttttttttt,gTTTtttttt
GU811709|GU811709.2^9278^      .^   T^   A^   119^.^   DP=8;VDB=0.0504;AF1=1;AC1=2;
GU811709|GU811709.2^9279^      .^   A^   T^   56^  .^   DP=8;VDB=0.0007;AF1=0.5032;A
GU811709|GU811709.2^13273^     .^   C^   A^   82^  .^   DP=99;VDB=0.0319;AF1=0.5;AC1
GU811709|GU811709.2^21746^     .^   G^   T^   216^.^   DP=163;VDB=0.0477;AF1=0.504;
GU811709|GU811709.2^58971^     .^   A^   C^   18.1^     .^   DP=23;VDB=0.0153;AF1=0.5
GU811709|GU811709.2^61476^     .^   aatagataga^  aataga^  193^.^   INDEL;DP=56;VDB=
GU811709|GU811709.2^64480^     .^   A^   G^   17.1^     .^   DP=104;VDB=0.0481;AF1=0.
[EOF]
```

Detected variants

Reference:

1) https://en.wikipedia.org/wiki/Variant_Call_Format
2) https://samtools.github.io/hts-specs/VCFv4.2.pdf

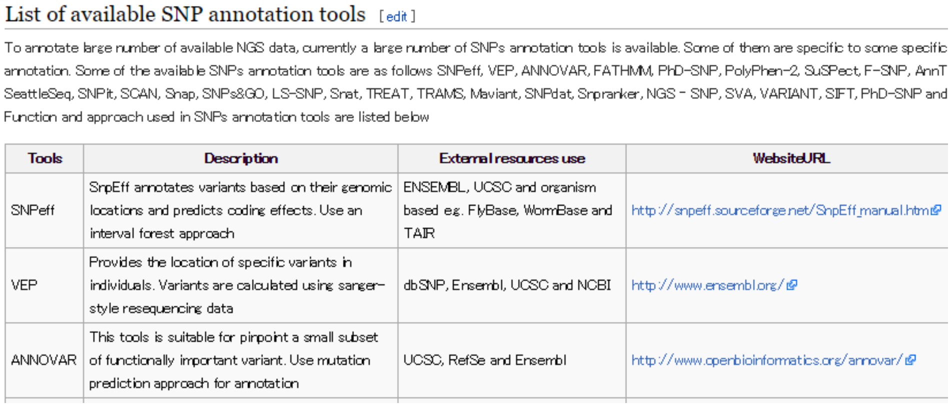
(current version)

Variant Effect Annotation

Predicts coding effects of genomic variants

■List of variant annotation tools

(https://en.wikipedia.org/wiki/SNP_annotation)



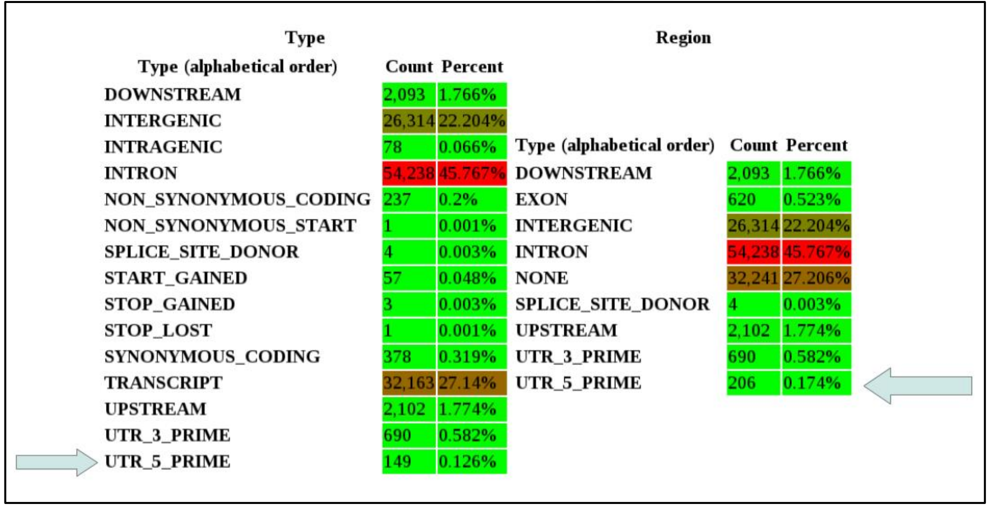
List of available SNP annotation tools [edit]

To annotate large number of available NGS data, currently a large number of SNPs annotation tools is available. Some of them are specific to some specific annotation. Some of the available SNPs annotation tools are as follows SNPeff, VEP, ANNOVAR, FATHMM, PhD-SNP, PolyPhen-2, SuSPect, F-SNP, AnnT SeattleSeq, SNPit, SCAN, Snap, SNPs&GO, LS-SNP, Snat, TREAT, TRAMS, Maviant, SNPdat, Snpranker, NGS - SNP, SVA, VARIANT, SIFT, PhD-SNP and Function and approach used in SNPs annotation tools are listed below

Tools	Description	External resources use	WebsiteURL
SNPeff	SnpEff annotates variants based on their genomic locations and predicts coding effects. Use an interval forest approach	ENSEMBL, UCSC and organism based e.g. FlyBase, WormBase and TAIR	http://snpeff.sourceforge.net/SnpEff_manual.html
VEP	Provides the location of specific variants in individuals. Variants are calculated using sanger-style resequencing data	dbSNP, Ensembl, UCSC and NCBI	http://www.ensembl.org/
ANNOVAR	This tools is suitable for pinpoint a small subset of functionally important variant. Use mutation prediction approach for annotation	UCSC, RefSe and Ensembl	http://www.openbioinformatics.org/annovar/

■SnpEff tool (example outputs※)

※http://snpeff.sourceforge.net/SnpEff_manual.html



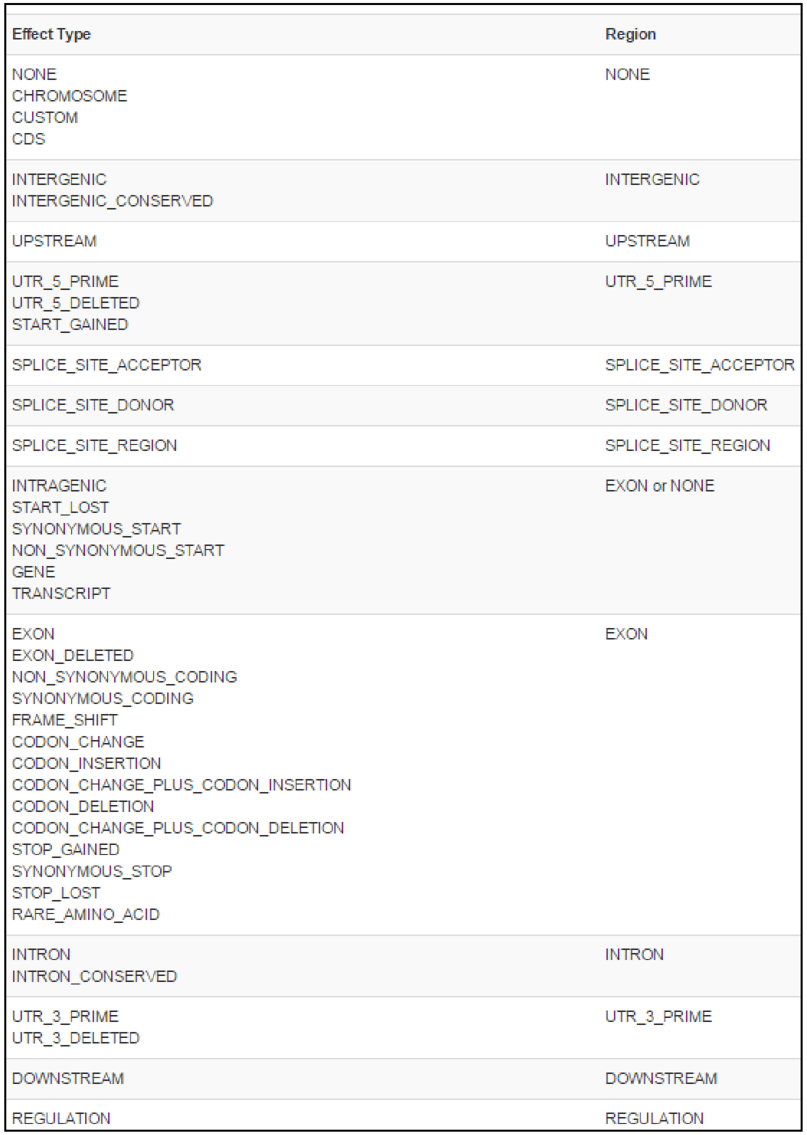
Type

Type (alphabetical order)	Count	Percent
DOWNSTREAM	2,093	1.766%
INTERGENIC	26,314	22.204%
INTRAGENIC	78	0.066%
INTRON	54,238	45.767%
NON_SYNONYMOUS_CODING	237	0.2%
NON_SYNONYMOUS_START	1	0.001%
SPLICE_SITE_DONOR	4	0.003%
START_GAINED	57	0.048%
STOP_GAINED	3	0.003%
STOP_LOST	1	0.001%
SYNONYMOUS_CODING	378	0.319%
TRANSCRIPT	32,163	27.14%
UPSTREAM	2,102	1.774%
UTR_3_PRIME	690	0.582%
UTR_5_PRIME	149	0.126%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	2,093	1.766%
EXON	620	0.523%
INTERGENIC	26,314	22.204%
INTRON	54,238	45.767%
NONE	32,241	27.206%
SPLICE_SITE_DONOR	4	0.003%
UPSTREAM	2,102	1.774%
UTR_3_PRIME	690	0.582%
UTR_5_PRIME	206	0.174%

■Effect type of the SnoEff tool※



Effect Type	Region
NONE CHROMOSOME CUSTOM CDS	NONE
INTERGENIC INTERGENIC_CONSERVED	INTERGENIC
UPSTREAM	UPSTREAM
UTR_5_PRIME UTR_5_DELETED START_GAINED	UTR_5_PRIME
SPLICE_SITE_ACCEPTOR	SPLICE_SITE_ACCEPTOR
SPLICE_SITE_DONOR	SPLICE_SITE_DONOR
SPLICE_SITE_REGION	SPLICE_SITE_REGION
INTRAGENIC START_LOST SYNONYMOUS_START NON_SYNONYMOUS_START GENE TRANSCRIPT	EXON or NONE
EXON EXON_DELETED NON_SYNONYMOUS_CODING SYNONYMOUS_CODING FRAME_SHIFT CODON_CHANGE CODON_INSERTION CODON_CHANGE_PLUS_CODON_INSERTION CODON_DELETION CODON_CHANGE_PLUS_CODON_DELETION STOP_GAINED SYNONYMOUS_STOP STOP_LOST RARE_AMINO_ACID	EXON
INTRON INTRON_CONSERVED	INTRON
UTR_3_PRIME UTR_3_DELETED	UTR_3_PRIME
DOWNSTREAM	DOWNSTREAM
REGULATION	REGULATION

ASSIGNMENT[9]

Variant Effect Annotation(DatePalm vcf)

ASSIGNMENT[9]
Apply SnpEff tool to vcf files at chloroplast genome of date palm dataset (SRA100551) and investigate variant effect types.

■ SNPeff manual = http://snpeff.sourceforge.net/SnpEff_manual.html

■ Install at NIG supercomputer

```
wget http://sourceforge.net/projects/snpeff/files/snpEff_latest_core.zip  (url  in manual)
unzip snpEff_latest_core.zip
cd snpeff
```

■ Generate target database

```
cd data
 mkdir  GU811709
cp SRA100551/snpeff/db/*   ./data/GU811709/
 (add snpeff.config) →
java -Xmx400M  -jar snpEff.jar build  -genbank -v GU811709
```

```
#Phoenix_dactylifera
GU811709.genome : Phoenix_dactylifera
```

■ Apply query vcf files to target database

```
cp SRA100551/snpeff/query/*.vcf  ./
```

edit 1st column from GU811709|GU811709.2 to GU811709 for all *.vcf

remove all comments lines (start #) for all *.vcf

```
java -Xmx400M -jar snpEff.jar  -ud 200 GU811709 query/AJW.vcf > AJW_ann.vcf
```

ASSIGNMENT[9]

Variant Effect Annotation(DatePalm vcf)

■Check variant annotations

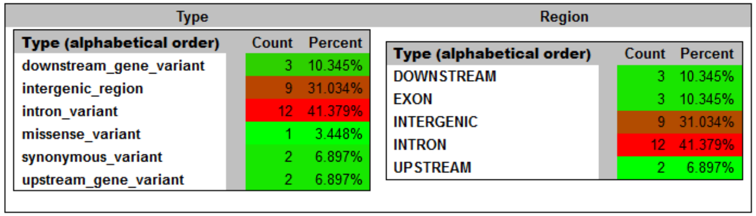
```
GU811709        4853        .        GTTTTTTTTTTTTT        "GTTTTTTTTTTTTTTTTTT,GTTTTTTTTTTTTTTTT,GTTTTTTTTTTT"
                90.5        .        "INDEL;DP=33;VDB=0.0298;AF1=1;AC1=2;DP4=0,1,10,22;MQ=44;FQ=-
59.5;PV4=1,1,0.27,1";ANN="GTTTTTTTTTTTTTTTTTT|downstream_gene_variant|MODIFIER|rps16|Gene_4869_5988|transcript|ADD63156.2|Coding||c.*
4_*17delAAAAAAAAAAAAACinsAAAAAAAAAAAAAAAAAC
```

```
GU811709        21746        .        G        T        225.0        .
                "DP=229;VDB=0.0487;AF1=0.5;AC1=1;DP4=29,27,87,86;MQ=42;FQ=141;PV4=0.88,1,1,0.074";ANN=T|synonymous_variant|LOW|rpoC1|
Gene_21631_24418|transcript|ADD63235.1|Coding|2…
```

```
GU811709        21750        .        T        C        70.0        .
                "DP=231;VDB=0.0403;AF1=0.5;AC1=1;DP4=90,95,25,21;MQ=42;FQ=73;PV4=0.51,0.37,1,1";ANN=C|missense_variant|MODERATE|
rpoC1|…
```

■Check SNP supEff_summary.html

Number of effects by type and region

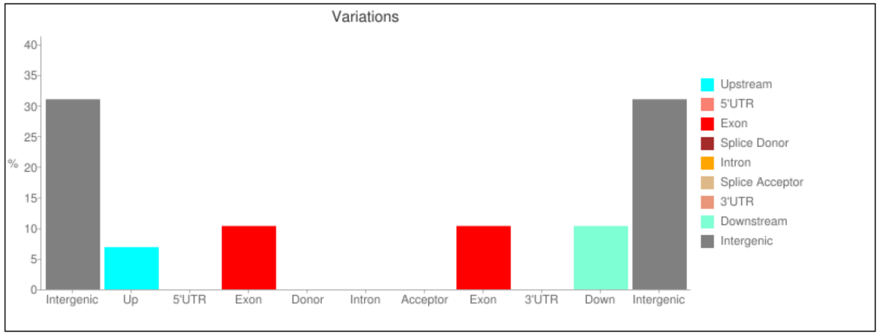


Type

Type (alphabetical order)	Count	Percent
downstream_gene_variant	3	10.345%
intergenic_region	9	31.034%
intron_variant	12	41.379%
missense_variant	1	3.448%
synonymous_variant	2	6.897%
upstream_gene_variant	2	6.897%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	3	10.345%
EXON	3	10.345%
INTERGENIC	9	31.034%
INTRON	12	41.379%
UPSTREAM	2	6.897%



■Count unique annotations for AJW_ann.vcf

AnnType	Count	Percent
downstream gene		
intergenic region		
intron variant		
missense variant		
synonymous variant		
upstream gene		

THANK YOU!