An integrated annotation system to support whole-exome sequencing experiments design and data management

Ivan Limongelli, Angelo Nuzzo, Annalisa Vetro, Erika Della Mina, Roberto Ciccone, Orsetta Zuffardi, Riccardo Bellazzi

IRCCS C. Mondino, Pavia
Dipartimento di Genetica Medica
Dipartimento di Informatica e Sistemistica

Exome Sequencing – Analysis Workflow

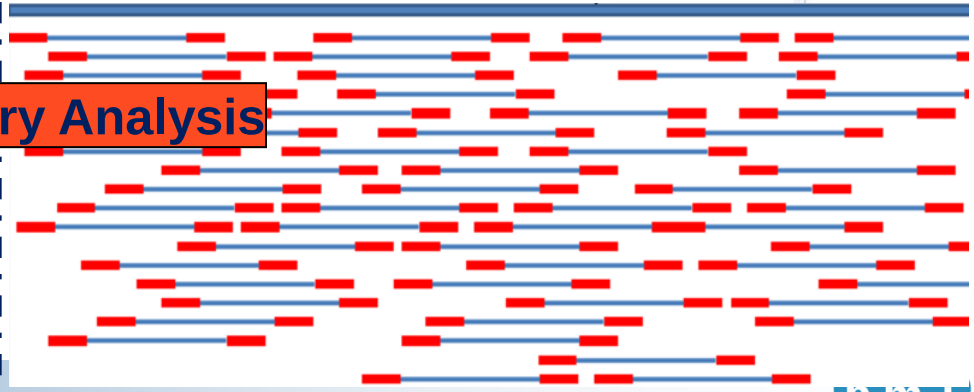
Nucleotide sequences (short reads)

```
@HWUSI-EAS703_0001:1:1:1009:17571#0/2
AAGCAAAAGAGTCCATAGCCAGCAGACCAAATGTTGAAATCTCTGGGCTAATTTGT
+
DDD5DDB:?AADDDBD?DAD:D:?DADBDDDB=DDDDDB=AD=;
0001:1:1:1009:12086#0/2
TCCACAG GCATGGCTCAGCTCTATTTCTTTCACTTCCTAGGGGGCACCGAGTGT
+
:CDACC5C?C-:-AAC5CCCB?DD?D=ADDDBADCD=AA:DDB=CDD?DC:-D-5:
@HWUSI-EAS703_0001:1:1:1010:6169#0/2
GGTCAGTCAGCCTTGTCACCTCTCCAGGATTTACGTAACTGACTACGTACAGGTGA
+
```

Primary Analysis

Reads Mapping

Reference genome sequence

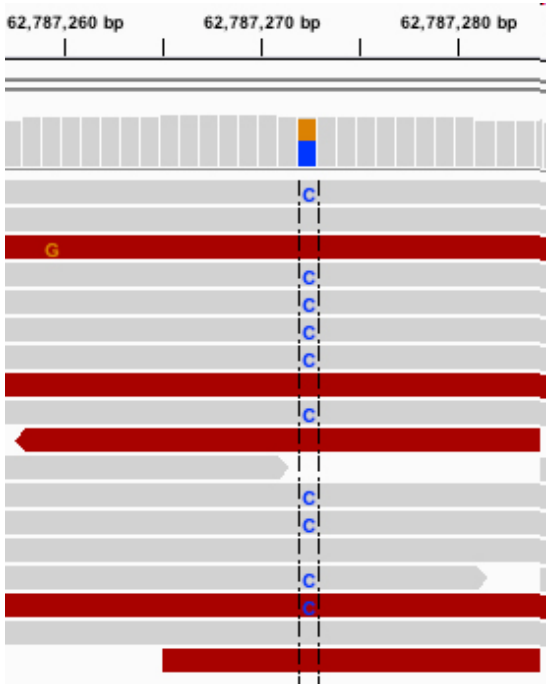


Secondary Analysis

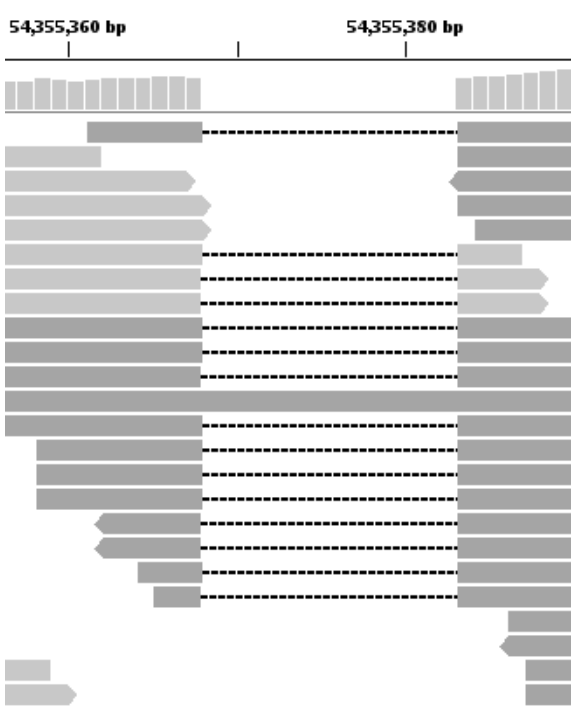
Mapping Analysis

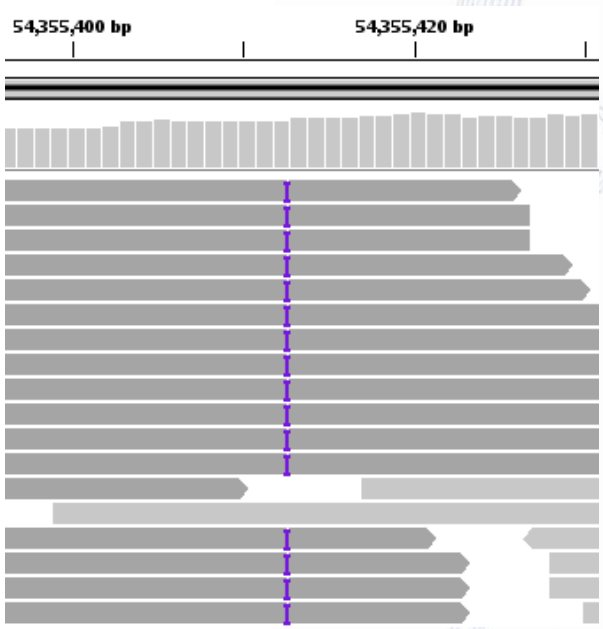
Exome Sequencing – Secondary Analysis

- Variants Detection



SNVs





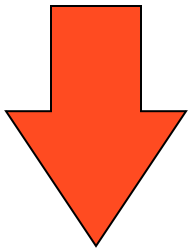
Short In-dels

Secondary Analysis – Variants Annotation

chr1 , 153170600 , A>G ,

Public bio-databases

NM_015383, NBPF14, I >R , ...



Does this variant affect the product (protein structure and function) coded in this region?

Secondary Analysis – Variants Prediction

- Some open tools addressing this aim:
  - *Polyphen2, Mutation Taster , SIFT, Annovar, Sequence variant Analyzer*
- Suitable for loss-of-function mutations
- Score assignment to each mutation corresponding to its probability to damage protein
- Principally based on type of amino acid substitution, conservation across species, polymorphisms databases

Secondary Analysis – Data management

- Some considerations
  - Illumina GAIIx platform: max 8 whole-exome samples sequenced in each experiment
  - 13-18K variants per sample (SNVs/Indels) are detected, but about 95% of them are common variants
- Needs
  - **I would like** to easily perform cross-samples and cross-experiments analysis
  - **I would NOT like** to annotate and predict changes again for previously (already annotated) identified variants

Secondary Analysis – Data management

Output by software device manufacturer

Pipeline 1 | Pipeline 2 | .... | Pipeline N

MiddelWare
(data uploads)

DB

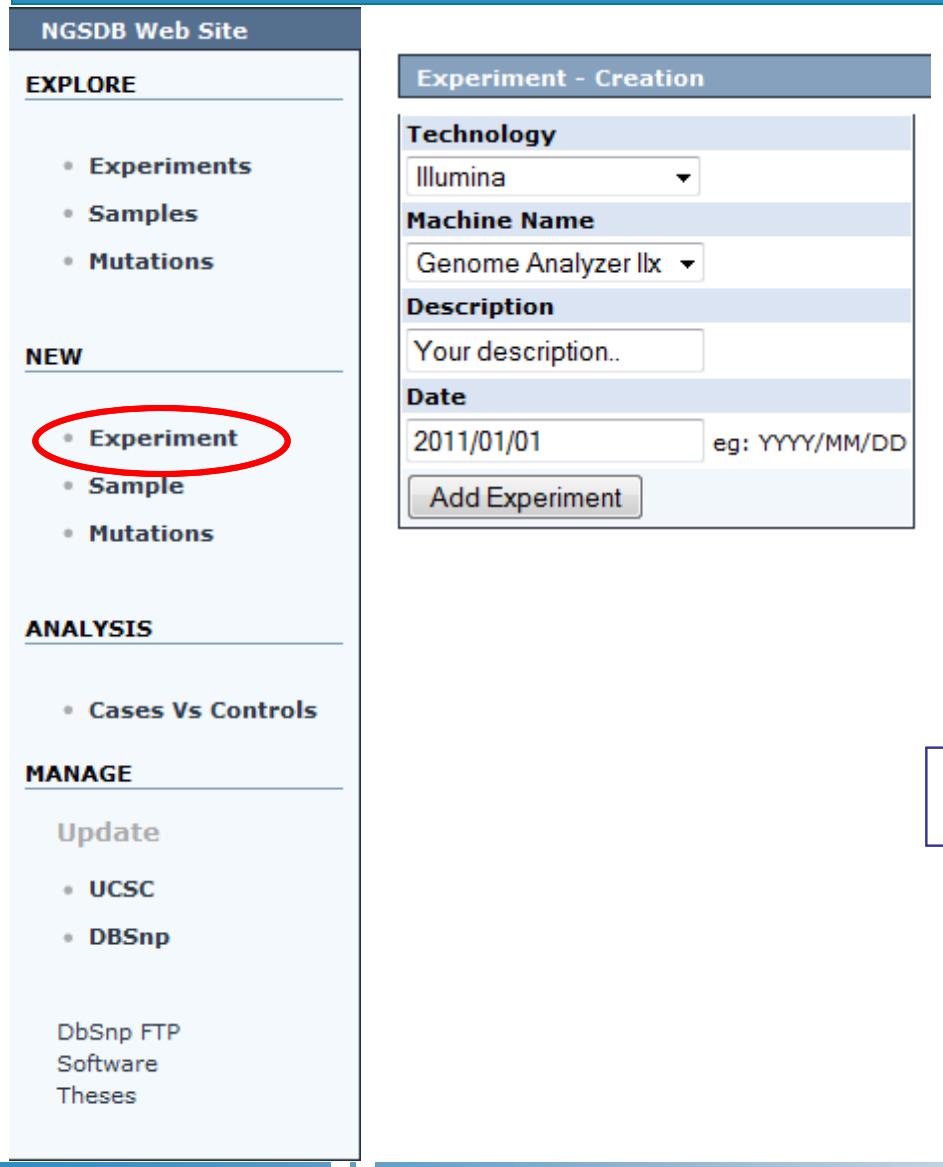
-omics DB

Synchronising

Applications Module

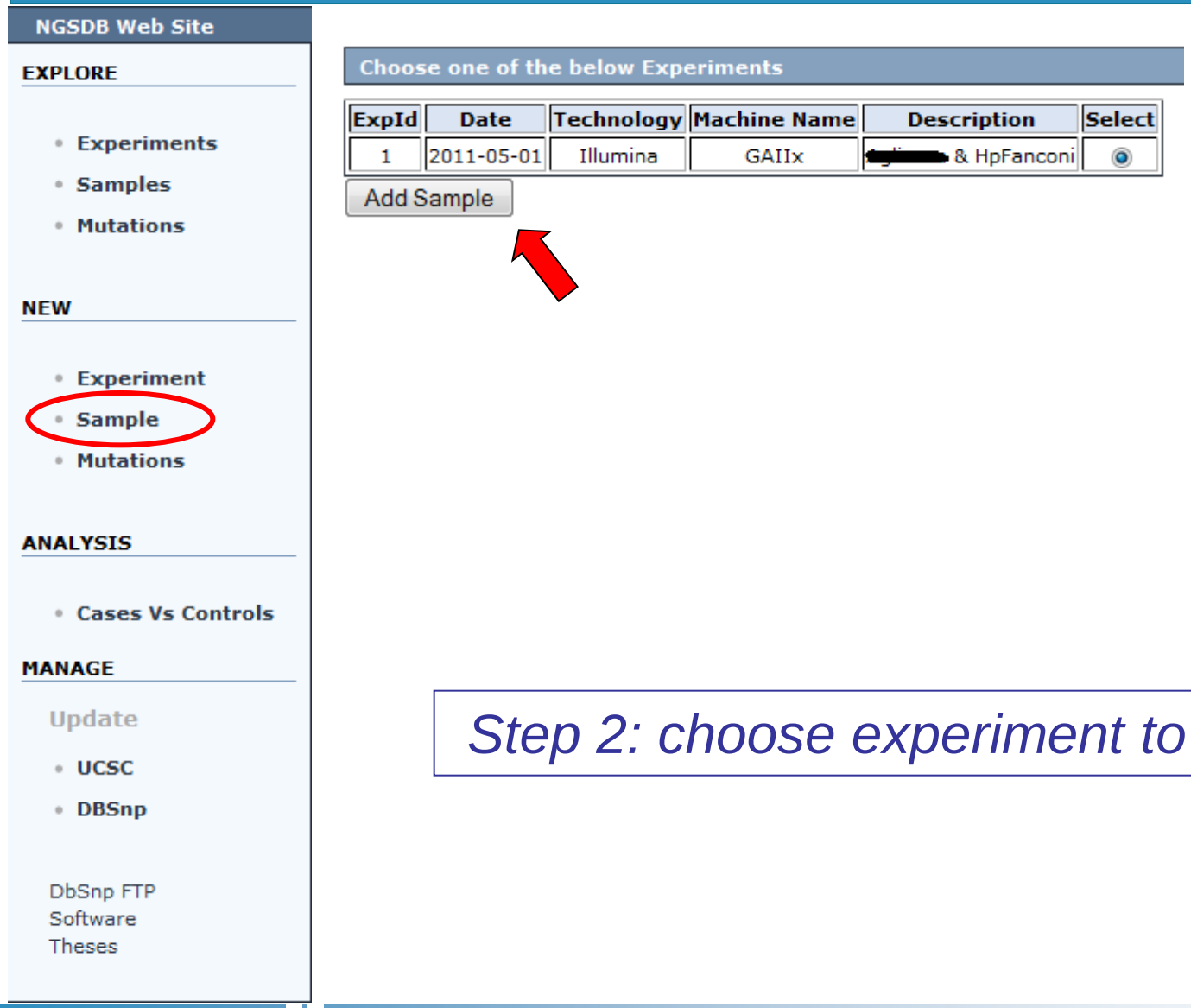
Functional prediction Tools

Secondary Analysis – Data management



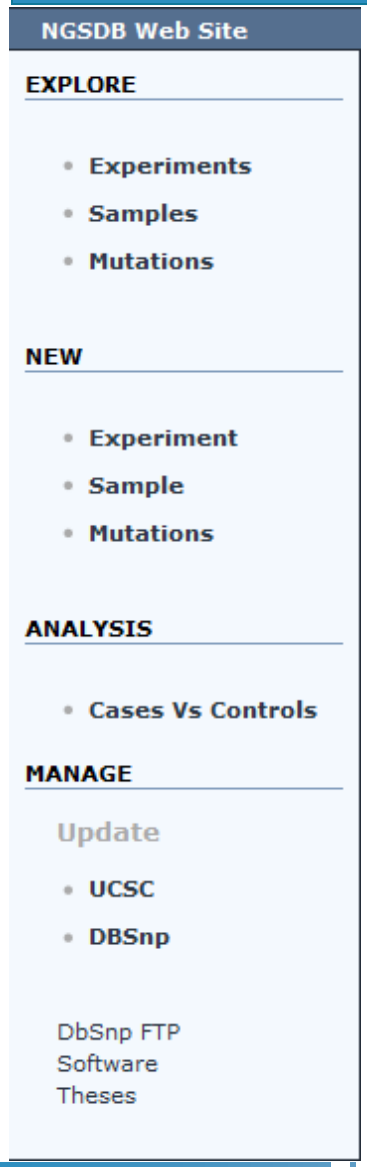
Step 1: create experiment

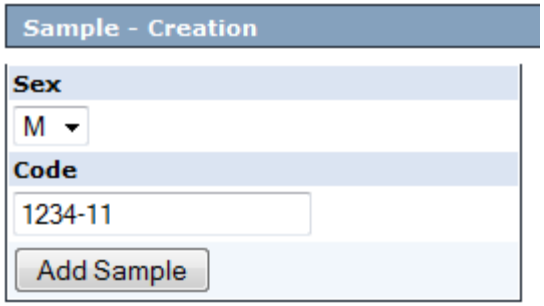
Secondary Analysis – Data management



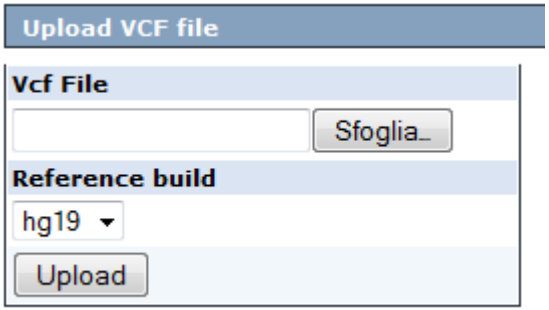
Step 2: choose experiment to add a sample

Secondary Analysis – Data management



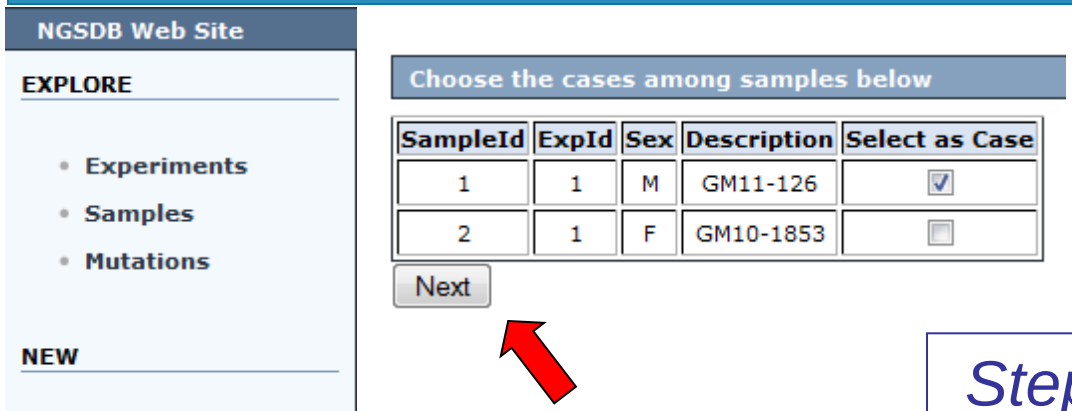


Step 3: create sample



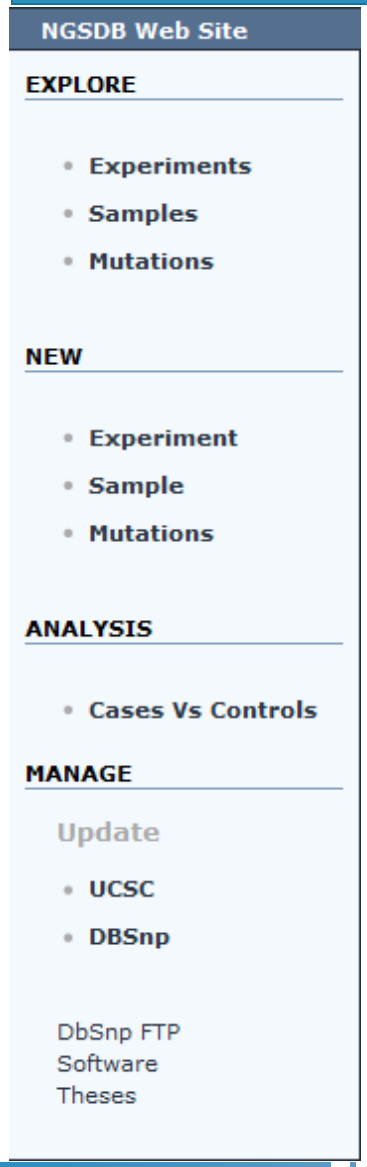
Step 4: upload mutation data for that sample

Secondary Analysis – Data management



Step 5: select cases/controls

Secondary Analysis – Data management



Set Filtering Parameters for Case Samples

SampleId	ExpId	Sex	Description	Minimum Mutation Quality	Minimum Mutation Coverage	Maximum Total Coverage
1	1	M	GM11-126	20	5	5000

Set Filtering Parameters for Control Samples

Minimum Mutation Quality	Minimum Mutation Coverage	Maximum Total Coverage
50	8	5000

File Name	

Filter!

Step 5: filtering parameters setup

Secondary Analysis – Data management

NGSDB Web Site

EXPLORE

- Experiments
- Samples
- Mutations

NEW

- Experiment
- Sample
- Mutations

ANALYSIS

- Cases Vs Controls

MANAGE

Update

- UCSC
- DBSnp

DbSnp FTP
Software
Theses

					Marker Description			
SampleDescription	Reference	Chromosome	Position	refUCSC	Observed	Genotype	Quality	Depth OF Coverage Reference
GM11-126	hg19	1	566048	G	A	hom	42.2	0
GM11-126	hg19	1	567579	C	T	hom	49.3	0
GM11-126	hg19	1	808631	G	A	het	36.65	1
GM11-126	hg19	1	909238	G	C	hom	43.95	0
GM11-126	hg19	1	909238	G	C	hom	43.95	0
GM11-126	hg19	1	909238	G	C	hom	43.95	0
GM11-126	hg19	1	1956399	C	T	het	1089.48	48
GM11-126	hg19	1	1956579	G	A	het	464.87	20
GM11-126	hg19	1	1957037	T	C	het	2970.61	111
GM11-126	hg19	1	154541971	T	G	het	142.39	56
GM11-126	hg19	1	154744807	C	G	het	6490.54	243
GM11-126	hg19	1	154744807	C	G	het	6490.54	243
GM11-126	hg19	1	154744807	C	G	het	6490.54	243
GM11-126	hg19	1	154744852	A	G	het	6271.88	202
GM11-126	hg19	1	154744852	A	G	het	6271.88	202
GM11-126	hg19	1	154744852	A	G	het	6271.88	202
GM11-126	hg19	1	154744937	C	T	het	2371.62	121
GM11-126	hg19	1	154744937	C	T	het	2371.62	121
GM11-126	hg19	1	154744937	C	T	het	2371.62	121
GM11-126	hg19	1	154745031	T	G	het	1249.55	46
GM11-126	hg19	1	154745031	T	G	het	1249.55	46
GM11-126	hg19	1	154745031	T	G	het	1249.55	46

Secondary Analysis – Data management

NGSDB Web Site

EXPLORE

- Experiments
- Samples
- Mutations

NEW

- Experiment
- Sample
- Mutations

ANALYSIS

- Cases Vs Controls

MANAGE

Update

- UCSC
- DBSnp

DbSnp FTP
Software
Theses

	Annotation								
epth OF overage lutation	UCSC transcript	UCSC gene symbol	UniProt Protein	IdDbSnp	Region	change_type	AA_change	Polyphen Prediction	pph_prob
2	unknown	null	null	rs6421780	genomic	na	na	null	null
2	uc001aaz.2	BC018860		rs112232512	utr	na	na	null	null
2	uc001abt.3	FAM41C		rs11240779	intron	na	na	null	null
2	uc001acd.2	PLEKHN1	Q494U1-2	rs3829740	exon	missense	R>P	benign	0.0
2	uc001ace.2	PLEKHN1	Q494U1	rs3829740	exon	missense	R>P	benign	0.0
2	uc001acf.2	PLEKHN1	Q494U1-3	rs3829740	exon	missense	R>P	benign	0.0
43	uc001aip.2	GABRD	O14764	rs79685811	exon	synonymous	G>G	null	null
21	uc001aip.2	GABRD	O14764	null	intron	na	na	null	null
127	uc001aip.2	GABRD	O14764	rs2229110	exon	synonymous	G>G	null	null
68	uc001ffg.2	CHRNB2	Q5SXY3	null	exon	missense	V>G	probably damaging	0.96
197	uc001ffo.2	KCNN3	Q8WXG7	rs1051614	exon	synonymous	L>L	null	null
197	uc001ffp.2	KCNN3	Q6JXY2	rs1051614	exon	synonymous	L>L	null	null
197	uc009wox.1	KCNN3	Q6JXY2	rs1051614	exon	synonymous	L>L	null	null
208	uc001ffo.2	KCNN3	Q8WXG7	rs1131820	exon	synonymous	N>N	null	null
208	uc001ffp.2	KCNN3	Q6JXY2	rs1131820	exon	synonymous	N>N	null	null
208	uc009wox.1	KCNN3	Q6JXY2	rs1131820	exon	synonymous	N>N	null	null
96	uc001ffo.2	KCNN3	Q8WXG7	rs11589471	intron	na	na	null	null
96	uc001ffp.2	KCNN3	Q6JXY2	rs11589471	intron	na	na	null	null
96	uc009wox.1	KCNN3	Q6JXY2	rs11589471	intron	na	na	null	null
44	uc001ffo.2	KCNN3	Q8WXG7	rs11584403	intron	na	na	null	null
44	uc001ffp.2	KCNN3	Q6JXY2	rs11584403	intron	na	na	null	null
44	uc009wox.1	KCNN3	Q6JXY2	rs11584403	intron	na	na	null	null

Secondary Analysis – Data management

- Current prototype implementation allows to:
  - Store experiments and samples data
  - Store identified variants (SNVs/Indels) and their reliability parameters (VCF 4.0 currently supported)
  - Annotate variants
  - Predict their probability to damage protein and store results (Polyphen2, Mutation Taster, SIFT)
  - Control-case studies modelling

The end..

Thank you for your attention!