- Getting the data
- Lift over
- Text manipulation
- Filter and sort
- Statistics

The Agenda

The goal of this session is to demonstrate how Galaxy can help you explore and learn options, perform analysis, and then
share, repeat, and reproduce your analyses.

How many SNPs exist in DYRK1A?

Genes & SNPs: A General Plan

• Get all coding SNPs for chromosome 21 of human (HG 19) → UCSC Table Browser.

1	2	3	4
chr21	10906257	10906260	rs146742197
chr21	10906257	10906260	rs146742197
chr21	10906257	10906260	rs146742197
chr21	10906257	10906260	rs146742197
chr21	10906909	10906910	rs144623440
chr21	10906909	10906910	rs144623440
chr21	10906909	10906910	rs144623440
chr21	10906909	10906910	rs144623440
chr21	10906914	10906915	rs169758
chr21	10906914	10906915	rs169758
chr21	10906914	10906915	rs169758
chr21	10906914	10906915	rs169758
chr21	10906921	10906922	rs144163164
chr21	10906921	10906922	rs144163164

- Lift over hg19->hg 38

1	2	3	4
chr21	10606196	10606199	rs146742197
chr21	10606196	10606199	rs146742197
chr21	10606196	10606199	rs146742197
chr21	10606196	10606199	rs146742197
chr21	10605546	10605547	rs144623440
chr21	10605546	10605547	rs144623440
chr21	10605546	10605547	rs144623440
chr21	10605546	10605547	rs144623440
chr21	10605541	10605542	rs169758
chr21	10605541	10605542	rs169758
chr21	10605541	10605542	rs169758
chr21	10605541	10605542	rs169758
chr21	10605534	10605535	rs144163164
chr21	10605534	10605535	rs144163164

- Get functional annotation for all SNPs

1	2	3	4	5	6	7	8	9	10	11	12
#bin	chrom	chromStart	chromEnd	name	transcript	frame	alleleCount	funcCodes	alleles	codons	peptides
664	chr21	10413740	10413741	rs75720584	NM_182482	2	2	8,42,	T,G,	ATG,AGG,	M,R,
664	chr21	10413741	10413742	rs78232757	NM_182482	3	2	8,42,	G,T,	ATG,ATT,	M,I,
664	chr21	10413743	10413744	rs368105863	NM_182482	2	3	8,42,42,	C,G,T,	GCG,GGG,GTG,	A,G,V,
664	chr21	10413744	10413745	rs77011809	NM_182482	3	2	8,3,	G,A,	GCG,GCA,	A,A,
664	chr21	10413747	10413748	rs79633368	NM_182482	3	2	8,3,	T,C,	GCT,GCC,	A,A,
664	chr21	10413748	10413749	rs75103008	NM_182482	1	2	8,41,	G,T,	GGA,TGA,	G,*,
664	chr21	10413752	10413753	rs74415742	NM_182482	2	2	8,42,	T,C,	GTG,GCG,	V,A,
664	chr21	10413753	10413754	rs374894507	NM_182482	3	2	8,3,	G,A,	GTG,GTA,	V,V,
664	chr21	10414834	10414835	rs374827796	NM_182482	1	2	8,42,	C,A,	CTG,ATG,	L,M,

- Annotate your original SNPs using the annotation you obtained (Join)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
rs146742197	chr21	10606196	10606199	#bin	chrom	chromStart	chromEnd	transcript	frame	alleleCount	funcCodes	alleles	codons	peptides
rs1010682	chr21	30396490	30396491	816	chr21	30396490	30396491	NM_181599	3	2	8,41,	T,A,	TGT,TGA,	C,*,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	NM_006031	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	XM_005261124	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	XM_005261125	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	XM_005261126	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	XM_005261127	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	XM_005261128	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	XM_005261129	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	XM_005261130	1	2	8,42,	G,A,	GGC,AGC,	G,S,

- Save only the unique SNPs

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
rs1010682	chr21	30396490	30396491	816	chr21	30396490	30396491	NM_181599	3	2	8,41,	T,A,	TGT,TGA,	C,*,
rs10222114	chr21	46412986	46412987	939	chr21	46412986	46412987	NM_006031	1	2	8,42,	G,A,	GGC,AGC,	G,S,
rs10229	chr21	46235394	46235395	937	chr21	46235394	46235395	NM_003906	2	2	8,42,	C,T,	TCA,TTA,	S,L,
rs1041439	chr21	39199319	39199320	884	chr21	39199319	39199320	NM_018963	2	2	8,42,	T,C,	CTA,CCA,	L,P,
rs1042917	chr21	46125853	46125854	936	chr21	46125853	46125854	NM_001849	2	2	8,42,	G,A,	CGT,CAT,	R,H,
rs1042930	chr21	46132083	46132084	936	chr21	46132083	46132084	NM_001849	3	2	8,3,	G,A,	ACG,ACA,	T,T,
rs10432965	chr21	46137307	46137308	936	chr21	46137307	46137308	NM_006657	3	2	8,3,	C,T,	GGC,GGT,	G,G,
rs1044998	chr21	46416480	46416481	939	chr21	46416480	46416481	NM_006031	2	2	8,42,	T,G,	ATG,AGG,	M,R,
rs1045382	chr21	42769055	42769056	911	chr21	42769055	42769056	NM_001001567	3	2	8,3,	T,C,	CCT,CCC,	P,P,
rs1046210	chr21	41250858	41250859	899	chr21	41250858	41250859	NM_012105	3	2	8,3,	C,T,	GAC,GAT,	D,D,
rs1047036	chr21	29008581	29008582	806	chr21	29008581	29008582	NM_016940	2	2	8,42,	C,T,	GCG,GTG,	A,V,

- Find all the SNPs in the DYRK1A gene. First download all genes.

1	2	3	4	5	6	7	8	9	10
#bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts
623	NM_015259	chr21	+	5022492	5036782	5022679	5034592	7	5022492,5025008,5026279,5027934,5032052,5033407,5034581,
623	NM_001283052	chr21	+	5022492	5036782	5026479	5034592	7	5022492,5025008,5026431,5027934,5032052,5033407,5034581,
623	NM_001283050	chr21	+	5022492	5040668	5022679	5040616	7	5022492,5025008,5026279,5027934,5032052,5033407,5040584,

12	13	14	15	16
score	name2	cdsStartStat	cdsEndStat	exonFrames
0	ICOSLG	cmpl	cmpl	0,2,1,1,1,1,1,
0	ICOSLG	cmpl	cmpl	-1,-1,0,1,1,1,1,
0	ICOSLG	cmpl	cmpl	0,2,1,1,1,1,1,
0	C21orf33	cmpl	cmpl	0,2,2,2,0,0,

- Get the gene of interest (get one of the transcripts)

1	2	3	4	5	6	7	8	9	10
108	NM_101395	chr21	+	37367556	37515376	37420374	37506307	13	37367556,37410569,37420298,37472683,37478180,37480637,37486466,37490174,37493016,37496117,37505282,375060
108	NM_130436	chr21	+	37418904	37515376	37420374	37512531	11	37418904,37472683,37478207,37480637,37486466,37490174,37493016,37496117,37505282,37506098,37511910,
108	NM_130438	chr21	+	37420298	37515376	37420374	37511954	10	37420298,37472683,37478180,37480637,37486466,37490174,37493016,37496117,37505282,37511910,
108	NM_001396	chr21	+	37420299	37515376	37420374	37512531	11	37420299,37472683,37478180,37480637,37486466,37490174,37493016,37496117,37505282,37506098,37511910,

12	13	14	15	16
0	DYRK1A	cmpl	cmpl	-1,-1,0,1,0,0,0,1,0,0,0,1,-1,
0	DYRK1A	cmpl	cmpl	0,1,0,0,0,1,0,0,0,1,0,
0	DYRK1A	cmpl	cmpl	0,1,0,0,0,1,0,0,0,1,
0	DYRK1A	cmpl	cmpl	0,1,0,0,0,1,0,0,0,1,0,

1	2	3	4	5	6	7	8	9	10
108	NM_101395	chr21	+	37367556	37515376	37420374	37506307	13	37367556,37410569,37420298,37472683,37478180,37480637,37486466,37490174,37493016,37496117,3

- Find the SNPs harbored in that gene (filter to get positions)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
rs1049753	chr21	37472717	37472718	870	chr21	37472717	37472718	NM_001396	3	2	8,3,	T,C,	GTT,GTC,	V,V,
rs1049754	chr21	37472732	37472733	870	chr21	37472732	37472733	NM_001396	3	2	8,3,	A,G,	TCA,TCG,	S,S,
rs1049755	chr21	37472735	37472736	870	chr21	37472735	37472736	NM_001396	3	2	8,3,	T,C,	TTT,TTC,	F,F,
rs1049756	chr21	37472767	37472768	870	chr21	37472767	37472768	NM_001396	2	2	8,42,	G,C,	GGA,GCA,	G,A,
rs1049757	chr21	37472780	37472781	870	chr21	37472780	37472781	NM_001396	3	2	8,3,	T,C,	CAT,CAC,	H,H,
rs1049758	chr21	37472786	37472787	870	chr21	37472786	37472787	NM_001396	3	2	8,3,	T,C,	CAT,CAC,	H,H,
rs1049759	chr21	37472810	37472811	870	chr21	37472810	37472811	NM_001396	3	2	8,3,	A,G,	CCA,CCG,	P,P,
rs1049760	chr21	37472812	37472813	870	chr21	37472812	37472813	NM_001396	2	2	8,42,	A,G,	AAC,AGC,	N,S,
rs1049761	chr21	37472825	37472826	870	chr21	37472825	37472826	NM_001396	3	2	8,3,	A,G,	CAA,CAG,	Q,Q,
rs1049762	chr21	37472831	37472832	870	chr21	37472831	37472832	NM_001396	3	2	8,3,	T,G,	GTT,GTG,	V,V,
rs1049763	chr21	37472841	37472842	870	chr21	37472841	37472842	NM_001396	1	2	8,42,	T,C,	TCA,CCA,	S,P,
rs1049764	chr21	37478233	37478234	870	chr21	37478233	37478234	NM_001396	3	2	8,3,	C,T,	GAC,GAT,	D,D,
rs1049765	chr21	37478248	37478249	870	chr21	37478248	37478249	NM_001396	3	2	8,3,	C,T,	CCC,CCT,	P,P,
rs1049766	chr21	37478260	37478261	870	chr21	37478260	37478261	NM_001396	3	2	8,3,	T,C,	CTT,CTC,	L,L,
rs1049767	chr21	37478266	37478267	870	chr21	37478266	37478267	NM_001396	3	2	8,3,	T,G,	GTT,GTG,	V,V,
rs1049768	chr21	37480677	37480678	870	chr21	37480677	37480678	NM_001396	2	2	8,42,	A,G,	CAG,CGG,	Q,R,
rs1049769	chr21	37480690	37480691	870	chr21	37480690	37480691	NM_001396	3	2	8,3,	T,C,	TCT,TCC,	S,S,
rs1049770	chr21	37480705	37480706	870	chr21	37480705	37480706	NM_001396	3	2	8,3,	A,G,	GAA,GAG,	E,E,
rs1049771	chr21	37480729	37480730	870	chr21	37480729	37480730	NM_001396	3	2	8,3,	T,C,	TAT,TAC,	Y,Y,
rs1049773	chr21	37480786	37480787	870	chr21	37480786	37480787	NM_001396	3	2	8,3,	C,T,	TAC,TAT,	Y,Y,
rs1049774	chr21	37480792	37480793	870	chr21	37480792	37480793	NM_001396	3	2	8,3,	T,C,	ATT,ATC,	I,I,
rs1049775	chr21	37480801	37480802	870	chr21	37480801	37480802	NM_001396	3	2	8,3,	G,A,	TTG,TTA,	L,L,
rs1049776	chr21	37486504	37486505	870	chr21	37486504	37486505	NM_001396	3	2	8,3,	T,C,	GTT,GTC,	V,V,
rs1049777	chr21	37486516	37486517	870	chr21	37486516	37486517	NM_001396	3	2	8,3,	A,C,	ATA,ATC,	I,I,
rs1049778	chr21	37486519	37486520	870	chr21	37486519	37486520	NM_001396	3	2	8,3,	A,C,	ATA,ATC,	I,I,
rs1049779	chr21	37486534	37486535	870	chr21	37486534	37486535	NM_001396	3	2	8,3,	T,G,	GCT,GCG,	A,A,
rs1049780	chr21	37486564	37486565	870	chr21	37486564	37486565	NM_001396	3	2	8,3,	A,G,	CGA,CGG,	R,R,
rs1049781	chr21	37486567	37486568	870	chr21	37486567	37486568	NM_001396	3	2	8,3,	T,G,	CTT,CTG,	L,L,
rs1049782	chr21	37490220	37490221	871	chr21	37490220	37490221	NM_001396	3	2	8,3,	T,G,	GTT,GTG,	V,V,
rs1049783	chr21	37490238	37490239	871	chr21	37490238	37490239	NM_001396	3	2	8,3,	C,T,	TAC,TAT,	Y,Y,
rs1049784	chr21	37490241	37490242	871	chr21	37490241	37490242	NM_001396	3	2	8,3,	C,T,	AAC,AAT,	N,N,
rs1049785	chr21	37490250	37490251	871	chr21	37490250	37490251	NM_001396	3	2	8,3,	C,T,	GAC,GAT,	D,D,
rs1049786	chr21	37490254	37490255	871	chr21	37490254	37490255	NM_001396	1	2	8,3,	C,T,	CTG,TTG,	L,L,
rs1049787	chr21	37490268	37490269	871	chr21	37490268	37490269	NM_001396	3	2	8,3,	T,C,	AAT,AAC,	N,N,
rs1049788	chr21	37490322	37490323	871	chr21	37490322	37490323	NM_001396	3	2	8,3,	T,A,	ACT,ACA,	T,T,
rs1049789	chr21	37490326	37490327	871	chr21	37490326	37490327	NM_001396	1	2	8,3,	C,T,	CTG,TTG,	L,L,
rs1049791	chr21	37490382	37490383	871	chr21	37490382	37490383	NM_001396	3	2	8,3,	A,G,	GAA,GAG,	E,E,
rs1049792	chr21	37490385	37490386	871	chr21	37490385	37490386	NM_001396	3	2	8,3,	T,C,	AAT,AAC,	N,N,
rs113004433	chr21	37493046	37493047	871	chr21	37493046	37493047	NM_001396	1	2	8,42,	C,T,	CGG,TGG,	R,W,

The workflow

Filter - data on any column using simple expressions

Filter : 7: Unique on data 6
With following condition : c3>=37420374 and c4<=37506307
Number of header lines to skip : 0

17: Filter on data 7

Select - lines that match an expression

Select lines from : 9: UCSC Main on Human: refGene (chr21:1-46709983)
that : Matching
the pattern : (NM_101395)

13: Select on data 9

UCSC Main - table browser

GALAXY_URL : *not used (parameter was added after this job was run)*
tool_id : *not used (parameter was added after this job was run)*
sendToGalaxy : *not used (parameter was added after this job was run)*
hgta_compressType : *not used (parameter was added after this job was run)*
hgta_outputType : primaryTable

9: UCSC Main on Human: refGene (chr21:1-46709983)

Unique - occurrences of each record

File to scan for unique values : 6: Join on data 5 and data 3
Ignore differences in case when comparing : False
Column only contains numeric values : False
Advanced Options : advanced
Column start : 1
Column end : 1

7: Unique on data 6

Join - two files

1st file : 3: Convert genome coordinates on data 2 [MAPPED COORDINATES]
Column to use from 1st file : 4
2nd File : 5: UCSC Main on Human: snp142CodingDbSnp (chr21:1-46709983)
Column to use from 2nd file : 5
Output lines appearing in : Both 1st & 2nd file.
First line is a header line : True
Ignore case : False
Value to put in unpaired (empty) fields : 0

6: Join on data 5 and data 3

UCSC Main - table browser

GALAXY_URL : *not used (parameter was added after this job was run)*
tool_id : *not used (parameter was added after this job was run)*
sendToGalaxy : *not used (parameter was added after this job was run)*
hgta_compressType : *not used (parameter was added after this job was run)*
hgta_outputType : primaryTable

5: UCSC Main on Human: snp142CodingDbSnp (chr21:1-46709983)

Convert genome coordinates - between assemblies and genomes

Convert coordinates of : 2: UCSC Main on Human: snp142CodingDbSnp (chr21:1-48129895)
To : /galaxy/data/hg19/liftOver/hg19ToHg38.over.chain
Minimum ratio of bases that must remap : 0.95
Allow multiple output regions? : 0
minSizeQ : 0
minChainQ : 0
minChainT : 0

3: Convert genome coordinates on data 2 [MAPPED COORDINATES]

UCSC Main - table browser

GALAXY_URL : *not used (parameter was added after this job was run)*
tool_id : *not used (parameter was added after this job was run)*
sendToGalaxy : *not used (parameter was added after this job was run)*
hgta_compressType : *not used (parameter was added after this job was run)*
hgta_outputType : bed

2: UCSC Main on Human: snp142CodingDbSnp (chr21:1-48129895)

- Count the occurrence of functional SNPs in this gene

1	2
76	8,3,
1	8,41,
33	8,42,

- Find the table that interprets these codes

http://www.ncbi.nlm.nih.gov/projects/SNP/

dbSNP Column Description for table: SnpFunctionCode

Table name and description

Table Description
SnpFunctionCode table has the lookup table for the snp funtion class code used in table SNPContigLocusId's fxn_class column. Snp function code is assigned after the snp is mapped onto contig and we know for example whether the snp causes amino acid change.

Table column and description

Column	Description	Type	Byte	Order
code	Snp function code. This is referred by SNPContigLocusId.fxn_class.	tinyint	1	1
abbrev	Abbreviation of the snp function.	varchar	20	2
descrip	Description of the meaning of the snp function code.	varchar	255	3
create_time	Time the row is inserted into the table. This is for internal tracking.	smalldatetime	4	4
top_level_class	Group all the snp function into two top level class: cSNP and other. cSNP is a subset of the group "is_coding", the next column.	char	5	5
is_coding	This column is used to group the snp function code into two groups: UTR, intron, splice-site will have is_coding = 0; locus, coding, cds-synon and cds-nonsynon and cds-reference will have is_coding=1.	tinyint	1	6

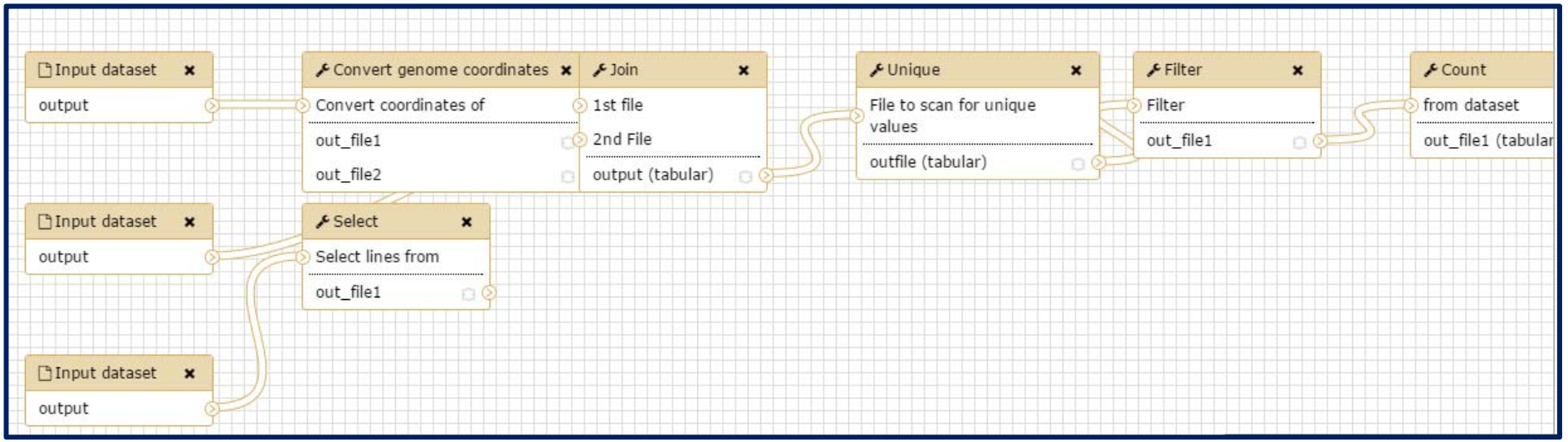
View the data in table: SnpFunctionCode.

SnpFunctionCode.bcp × Document1 *

```
1   3       cds-synon       synonymous change. ex. rs248, GAG->GAA, both produce amino acid: Glu    2003-02-03 16:01:00.0   cSNP    1
2   6       intron  intron. ex. rs249.      2003-02-03 16:01:00.0   other   0                       8       SO:0001627
3   8       cds-reference   contig reference        2003-02-03 16:01:00.0   other   1
4   9       synonymy unknown        coding: synonymy unknown. Not used since 2003.  2003-02-03 16:01:00.0   other   1
5   13      nearGene-3      within 3' 0.5kb to a gene. ex.  rs3916027  is at NT_030737.9 pos7669796, within 500 bp of UTR starts 7669
6   15      nearGene-5      within 5' 2kb to a gene. ex. rs7641128 is at NT_030737.9 pos7641128, with 2K bp of UTR starts 7641510 for
7   20      intergenic      variant between two genes, outside of 2Kb upstream and 500bp downstream of a gene       2012-01-06 11:54:(
8   30      ncRNA   variant on non-coding RNA(NCBI Refseq prefix NR,XR).    2012-01-06 13:04:00.0   other   0                       2(
9   41      STOP-GAIN       changes to STOP codon. ex.  rs328, TCA->TGA, Ser to terminator. 2005-08-01 00:00:00.0   cSNP    1       1
10  42      missense        alters codon to make an altered amino acid in protein product. ex.  rs300, ACT->GCT, Thr->Ala.  2005-08-0
11  43      STOP-LOSS       changes STOP codon to other non stop codon      2010-09-13 16:00:00.0   cSNP    1               256
12  44      frameshift      indel snp causing frameshift.   2005-08-01 00:00:00.0   cSNP    1       1       8               SO:000158
13  45      cds-indel       indel snp with length of multiple of 3bp, not causing frameshift.       2010-01-07 10:12:00.0   cSNP    1
14  53      UTR-3   3 prime untranslated region. ex.  rs3289.       2005-08-01 00:00:00.0   other   0       1               128     SC
15  55      UTR-5   5 prime untranslated region. ex.  rs1800590.    2005-08-01 00:00:00.0   other   0       1               64      SC
16  73      splice-3        3 prime acceptor dinucleotide. The last two bases in the 3 prime end of an intron. Most intron ends with
17  75      splice-5        5 prime donor dinucleotide. 1st two bases in the 5 prime end of the intron. Most intron starts is GU. ex.
```

http://www.ncbi.nlm.nih.gov/SNP/snp_db_table_description.cgi?t=SnpFunctionCode

- Save your workflow



https://test.galaxyproject.org/u/eel/h/unnamed-history

- Publish your workflow and history