Title: **Webinar 21 – A Primer on Brain Proteomics and protein-QTL Analysis for Substance Use Disorders**

Date: Friday, October 8th at 10am PDT/ 11am MDT/ 12pm CDT/ 1pm EDT, 1-hour presentation followed by 30 minutes of discussion

Authors: Xusheng Wang (xusheng.wang.und.edu) and Rob W Williams (rwilliams@uthsc.edu)

To: Any one interested in brain proteomics, substance use disorder, dopamine, causal

Versions: 08Oct2021 v3; original 14Feb2021 v1

Support: NIDA P30

BACKGROUND and acknowledgements: Xusheng Wang and colleagues Ling Li, Zhiping Wu, Ariana Mancieri, and He Huang at the University of North Dakota have released a cool new rat HRDP brain proteomics data set. Other key collaborators include Junmin Peng (St. Jude Children's Research Hospital) and Michal Pravenec (Czech Academy of Science). The proteomics data set is still growing—so come back in a few months to see the increase in sample size for the **hybrid rat divesity panel** (HRDP) whole brain proteome. And yes, we hope to have proteomes for smaller chunks of the brain and perhaps even data on effects of age and treatments.

The genetic analysis of the proteomics uses new and much improved genotypes that are courtesy of Hao Chen and Hakan Gunturun at UTHSC. We are also using the new GEMMA mapping algorithm courtesy of Pjotr Prins at UTHSC. Data entry was by Arthur Centeno and coding mainly by Zach Sloan.

And finally, a big thank you to NIDA for all of the support over more than 20 years that has made this primer possible—starting with the NIH Human Brain Project that supported GeneNetwork and continuing now with a NIDA P30. Without NIDA's support, rodent neurogenetics would be lost. I hope NIDA staff and teams will be impressed not only by the data that Xusheng Wang has generated, but by the FAIR-ness and ease of analysis of these highly valuable **smart quadratic data***.

> **And what the heck is *smart quadratic data*?** Please see an extended discussion on this topic (verging on a rant in sections) given by RWW at the University of Virginia in the Data Science program, 20Nov2020: ***https://youtu.be/4ZhnXU8gV44***

STATUS of DATA: This is an open, **but not yet final**, quantitative proteomics data for the whole brain of 21 strains of rat (male and female isogenic littermates) from the HXB/BXH family—part of the HRDP. We will have data for more than 33 strains and eventually for four or more subregions of the brain. There is also a companion brain metabolomic data set—a first for rat. And as of August 2021, there is now also **a companion brain metabolomics data set.** Not covered here, but most of the same workflows described below also apply. If you do look at the metabolomics data, why not search for "dopamine" (Rob added a bit on metabolomics at the very end of this document)***.***

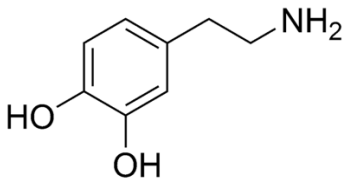
Many of you have heard or even given talks in which you consider a hierarchy of traits—DNA on up to phenotypes. I remember a lovely talk Nora Volkow gave to the Human Brain Project teams in about 2005. Her first or second slide was the hierarchy of biological scales—from gene variants at the bottom to variable drug abuse outcome measures at the top—susceptible vs resistant; fast vs slow metabolizer; will relapse, won't replase. I would say we have made at best only very modest progress at true holistic integration, and few biomedical researcher know much about causal quantitative modeling.

WHY PROTEOMICS? We absolutely need the proteomic tier to model addiction, and we need proteomes from dozen of brain regions and hundreds, if not thousands of individuals to model risk and make reasonable predictions. Otherwise we are just flapping our hands and lips. Transciptomics has dominated almost all omics for 20 years for the simple reason that we can use arrays and high throughput sequencer. But franking proteomics is where most of the cellular action happens.

The work by Xusheng Wang and others shows that we are finally ready to come out of the proteomics "winter". The technolgy is mature; batch effect is well controlled; cost is about the same as Affymetrix arrays were in 2008. Several new proteomic data sets in GeneNetwork prove it, but only Xusheng's data is directly relevant to addiction.

End of context; on to the topic at hand:

One small molecule of great fame—dopamine—and its modulation, variation, and contribution to addiction



QUESTION:

> **What proteins related to dopamine and its many roles in behavior are strongly modulated by DNA variants, and can we determine what gene variants are related both to dopamine function and substance use disorders.**

The Red Hot Chili Peppers ask this question in *This is the Place.*

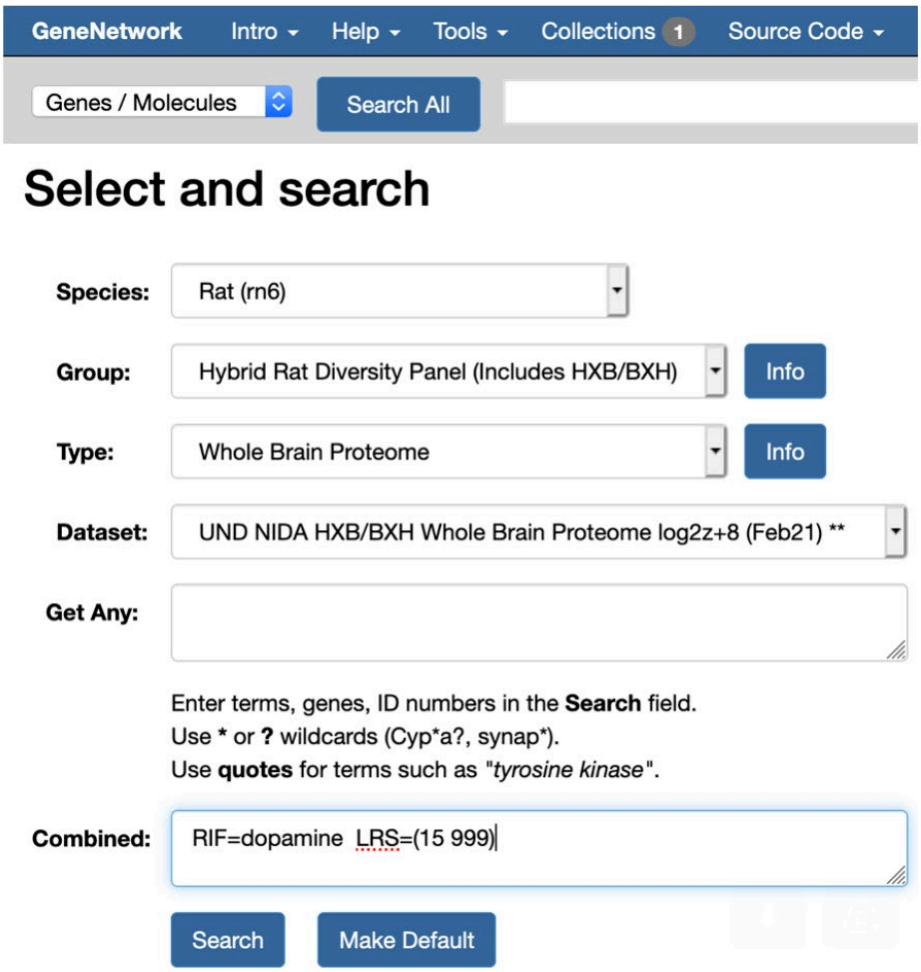
> *"Can I isolate your gene? Can I kiss your dopamine?*
> *...*
> *A master piece of DNA caught in a flashing ray"*

(The lyrics are on the horror of drug addiction. The lead, Anthony Kiedis, has relapsed several times. The PG version of the song: ***https://www.youtube.com/watch?v=gqgm7ViA2Ag*** and the typical RHCP shirtless version for the cool kids: ***https://www.youtube.com/watch?v=8Dkvwu3aWkY***

Step 1. To answer the BIG Question, we are going to review all genes/proteins in NCBI **Gene Reference into Function**—RIF for short—that are related in some way to *dopamine*.

There are two ways to do this:

1. Link to ***https://www.genenetwork.org*** and set up the **Select and search** screen to look as shown below:



Note that in the **Combined** field above, I have entered the string

RIF=dopamine LRS=(15 999)

This string will retrieve all proteins in the Hybrid Rat Diversity Panel (the HXB/BXD family in this specific case) that are expressed reasonably well (just over 8,000 proteins and over 200,000 peptide fragments) in the whole brain.

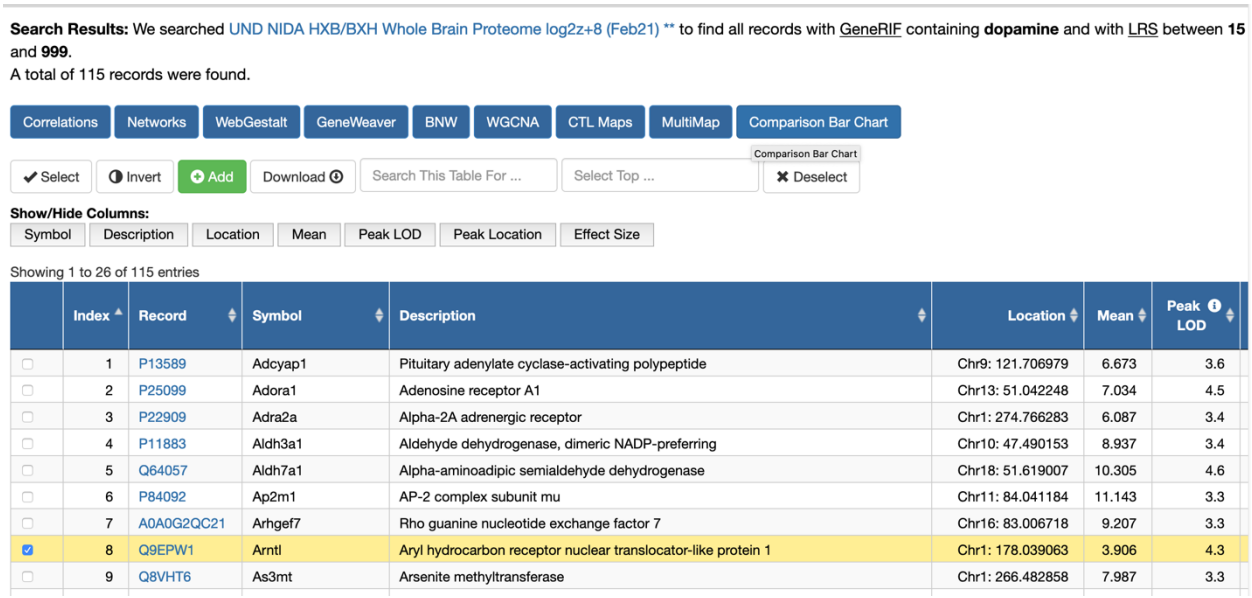
The second part of the search string (LRS...) finds all proteins that have strong linkage—a likelihood ratio statistic score of at least 15. This is equivalent to a LOD score of 3.3, and this is a value that is often close to the genome-wide significance level. The other value, 999, is just a high upper limit.

The second way to find these proteins is a bit easier—just paste this URL into your browser:

https://genenetwork.org/search?species=rat&group=HXBBXH&type=Whole+Brain+Proteome&dataset=UND_NIDA_HXB-BXH_WBPr_log2z8_0221&search_terms_or=&search_terms_and=RIF%3Ddopamine+LRS%3D%2815+999%29&FormID=searchResult

(This link can be shared, and will work *in perpetuity throughout the known universe;* a phrase I steal from the Walt Disney Company legal department with trepidation.)

Step 2. At this point, if you are following along, you should have a list of 115 proteins that are abundantly expressed in brain AND are linked to *dopamine* AND that have reasonable genetic linkage in the HXB family to a particular genome coordinate (usually a SNP). The **Search Results** table should look like the screenshot below.



Search Results: We searched UND NIDA HXB/BXH Whole Brain Proteome log2z+8 (Feb21) ** to find all records with GeneRIF containing **dopamine** and with LRS between **15** and **999**.
A total of 115 records were found.

Correlations | Networks | WebGestalt | GeneWeaver | BNW | WGCNA | CTL Maps | MultiMap | Comparison Bar Chart

Select | Invert | Add | Download | Search This Table For ... | Select Top ... | Deselect

Show/Hide Columns: Symbol | Description | Location | Mean | Peak LOD | Peak Location | Effect Size

Showing 1 to 26 of 115 entries

	Index	Record	Symbol	Description	Location	Mean	Peak LOD
☐	1	P13589	Adcyap1	Pituitary adenylate cyclase-activating polypeptide	Chr9: 121.706979	6.673	3.6
☐	2	P25099	Adora1	Adenosine receptor A1	Chr13: 51.042248	7.034	4.5
☐	3	P22909	Adra2a	Alpha-2A adrenergic receptor	Chr1: 274.766283	6.087	3.4
☐	4	P11883	Aldh3a1	Aldehyde dehydrogenase, dimeric NADP-preferring	Chr10: 47.490153	8.937	3.4
☐	5	Q64057	Aldh7a1	Alpha-aminoadipic semialdehyde dehydrogenase	Chr18: 51.619007	10.305	4.6
☐	6	P84092	Ap2m1	AP-2 complex subunit mu	Chr11: 84.041184	11.143	3.3
☐	7	A0A0G2QC21	Arhgef7	Rho guanine nucleotide exchange factor 7	Chr16: 83.006718	9.207	3.3
☑	8	Q9EPW1	Arntl	Aryl hydrocarbon receptor nuclear translocator-like protein 1	Chr1: 178.039063	3.906	4.3
☐	9	Q8VHT6	As3mt	Arsenite methyltransferase	Chr1: 266.482858	7.987	3.3

I have highlighted the row 8—the ARNTL protein—a major transcription factor involved in circadian rhythms that is upregulated by DRD2 signaling (PMID: 16606840 in PNAS, 2006)

Step 3. To begin to answer the second question—is there a major modulator of multiple dopamine-associated proteins—we need to re-sort this table using the column labeled **Peak Location**. In this screenshot below I have scrolled over to the right to display the **Peak Location** column after having performing the sort. All of these proteins map to Chr 1 at about 43.7 megabases (Mb).

Symbol	Description	Location	Mean	Peak LOD	Peak Location	Effect Size
Arntl	Aryl hydrocarbon receptor nuclear translocator-like protein 1	Chr1: 178.039063	3.906	4.3	Chr1: 43.724577	1.351
Caly	Neuron-specific vesicular protein calcyon	Chr1: 212.537848	3.824	4.1	Chr1: 43.724577	1.255
Syt7	Synaptotagmin-7	Chr1: 226.435979	9.454	6.8	Chr1: 43.724577	0.072
Trpc4	Short transient receptor potential channel 4	Chr2: 143.433102	3.795	4.1	Chr1: 43.724577	1.222
Ret	Proto-oncogene tyrosine-protein kinase receptor Ret	Chr4: 150.202058	3.508	3.4	Chr1: 43.724577	0.887
Axin1	Axin-1	Chr10: 15.163684	5.761	6.3	Chr1: 43.724577	-2.111
Ednrb	Endothelin receptor type B	Chr15: 88.006977	5.020	7.1	Chr1: 43.724577	-1.690
Pgpep1	Pyroglutamyl-peptidase 1	Chr16: 20.521956	4.025	4.4	Chr1: 43.724577	1.490
Dtnbp1	Dysbindin	Chr17: 19.685218	5.789	6.1	Chr1: 43.724577	-2.060

We see ARNTL again and eight other proteins that are genetically downstream of one or many DNA variants located on the proximal part of chromosome 1 (Chr 1). The **Peak LOD** scores range between 4.1 and 7.1.

If you scroll down this list (and you should), you will find another region of the rat genome that is highly linked with dopamine-associated proteins—Chr 19 at about 60 Mb. But before we head to

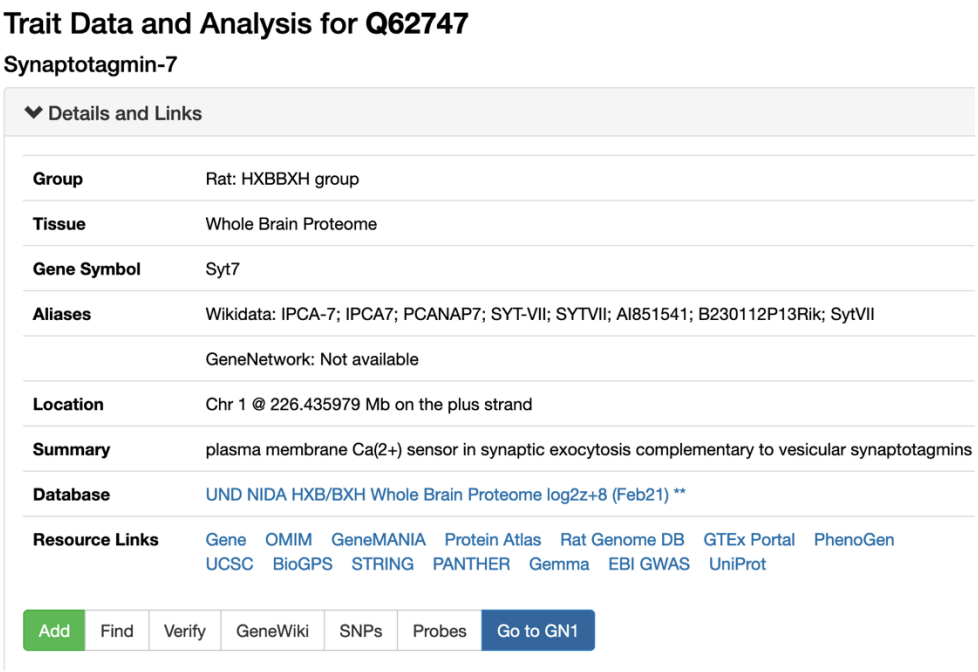
Chr 19, let's continue to work with this proximal part of Chr 1 and try to figure out why the variation in expression of this band of nine proteins maps to this part of the rat genome. Step 3 below is a long step—my apology. Perhaps time for a coffee break.

Step 3 involves mapping one or more of these nine proteins. I will pick SYT7 since it has the highest expression (9 log2 units of expression) and the second highest LOD score (6.8).

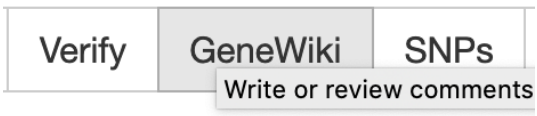
You can either click on the UNIPROT identifier—**Q62747** in the window, or you can just paste this URL command into a browser:

> ***https://genenetwork.org/show_trait?trait_id=Q62747&dataset=UND_NIDA_HXB-BXH_WBPr_log2z8_0221***

If all goes well, your browser will display this content (and much more too):



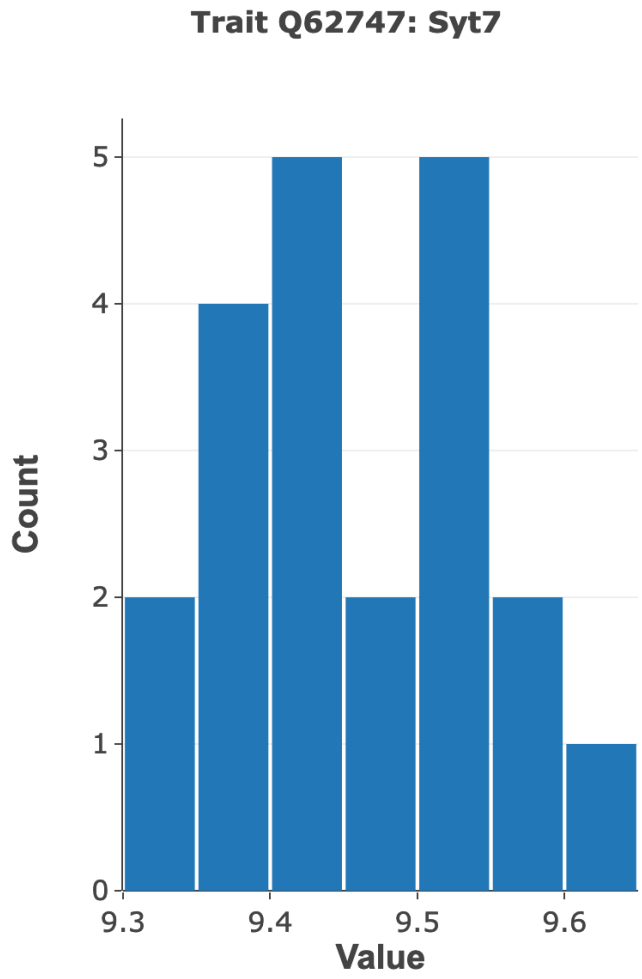
Before we map SYT7 protein expression, you may be curious to know how this protein has been linked to dopamine.
The answer is one click away. Tap on the **GeneWiki** button, highlighted below in grey.



A **GeneWiki** window will open, and RIF number 18 explains the association with *dopamine* and also links to a 2011 paper (PMID 21576241) on somatodendritic dopamine release and the involvement of synaptotagmin 7 (SYT7).

Again we pause briefly for "data due diligence". In the **Statistics histogram** window you will

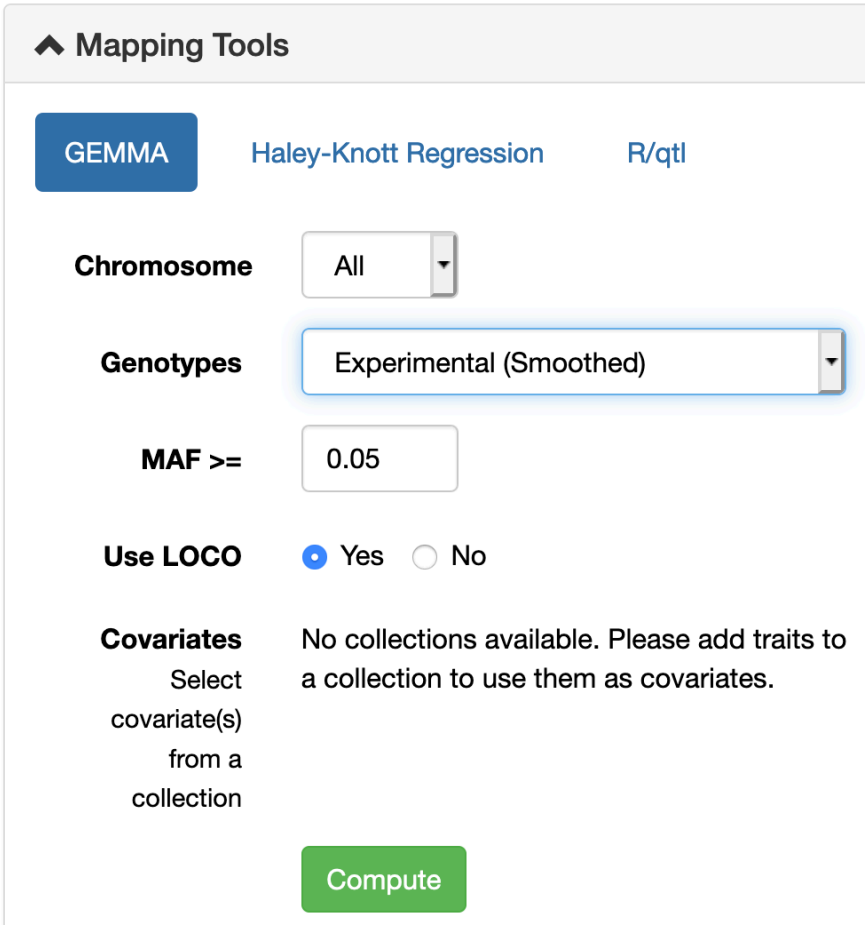
note that the distribution of SYT7 protein levels in 21 strains has a hint of bimodality—that is a good thing.



There are no outliers, so we can map these logged protein expression data "as given" without further normalization.

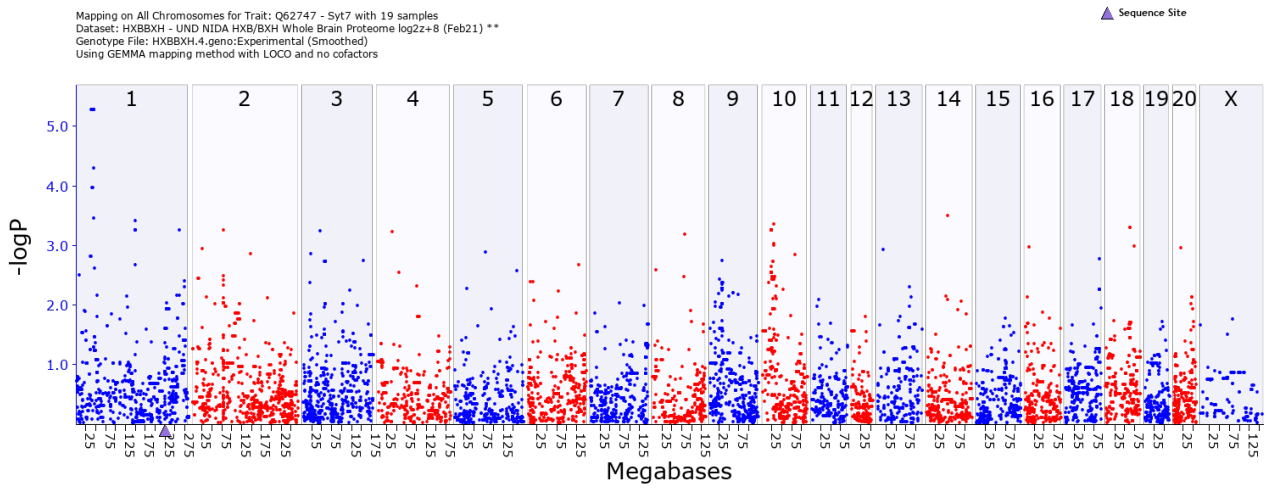
We can now finally proceed to the actual mapping of variation in protein expression—using for the first time "infinite marker maps" for all chromosome of all members of the HXB/BXH family, and using the updated GEMMA linear mixed model mapping function in GeneNetwork.

Open the **Mapping Tools** window



In the screenshot above I have mapped variation in SYT7 protein level using the new **Genotypes file: Experimental (smoothed)**
These are genotypes based on whole genome sequencing of the HXB/BXH family using linked-read 10X Chromium DNA libraries at a mean sequence coverage of just over 45X. Libraries were prepared at HudsonAlpha and sequenced on an Illumina Novaseq across the street from NIH at *The American Genome Center* (TAGC, thanks Michal, Melinda, Hao, Clifton, Jonathan, David, Hakan, Tristan, Victor, Jun, many others....).

The Manhattan plot of variation in SYT7 protein expression should look like this:



Beneath the Manhattan plot there is a **Mapping Statistics** table that provides estimates a SNP coordinates (Rnor6 assembly) calculated by GEMMA with –logP values and additive effects (log2 scale).

Mapping Statistics

✔ Select All ✖ Deselect All ↕ Invert ✪ Add

Showing from 1 to 2,000 of 12536 records

	Row	Marker	–logP	Position (Mb)	Add Eff
☑	1	chr1_35661556_A_C	5.27	Chr1: 35.661556	-0.140
☐	2	chr1_37895806_C_T	5.27	Chr1: 37.895806	-0.140
☑	3	chr1_39328636_G_A	5.27	Chr1: 39.328636	-0.140
☐	4	chr1_39517348_A_G	5.27	Chr1: 39.517348	-0.140
☑	5	chr1_39525726_G_C	5.27	Chr1: 39.525726	-0.140
☐	6	chr1_41344047_A_G	5.27	Chr1: 41.344047	-0.140
☑	7	chr1_41984535_A_G	5.27	Chr1: 41.984535	-0.140
☐	8	chr1_41985407_A_G	5.27	Chr1: 41.985407	-0.140
☑	9	chr1_42708786_A_G	5.27	Chr1: 42.708786	-0.140
☐	10	chr1_42720371_A_G	5.27	Chr1: 42.720371	-0.140
☑	11	chr1_42928603_G_T	5.27	Chr1: 42.928603	-0.140
☐	12	chr1_43022071_C_T	5.27	Chr1: 43.022071	-0.140
☑	13	chr1_42711576_C_T	4.27	Chr1: 42.711576	-0.124
☐	14	chr1_39397605_G_C	3.95	Chr1: 39.397605	-0.126

A –logP value of 5.27 is good—normally at or above genome-wide threshold of significance. (This assertion does need more support, and we are testing thresholds using using other mapping methods, including R/qtl's and WebQTL's standard interval mapping methods, and using permutation tests.)

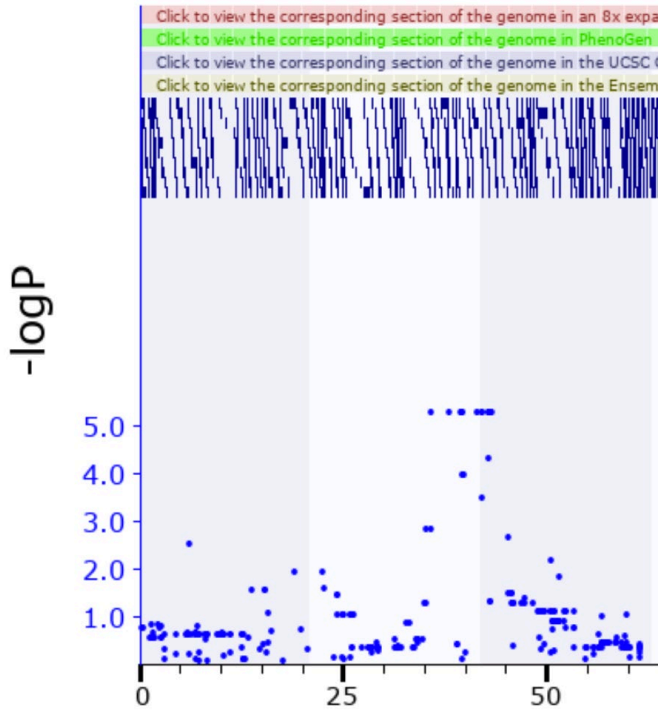
Step 4. What is the approximate confidence interval of the SYT7 protein expression quantitative trait locus (QTL) on Chr 1? To answer this question we need to sort the **Mapping Statistics** by

the **Position** column. Once sorted, we have to decide how wide a confidence interval is appropriate given the density of DNA variants, gene density, and –logP values. Karl Broman and others recommend a drop in the –logP linkage statistic of about 1.5 on either side of the peak, or plateau in this case. For the QTL map of SYT7 the confidence interval encompasses an stretch of DNA from about 35 megabases (Mb) to 43 Mb.

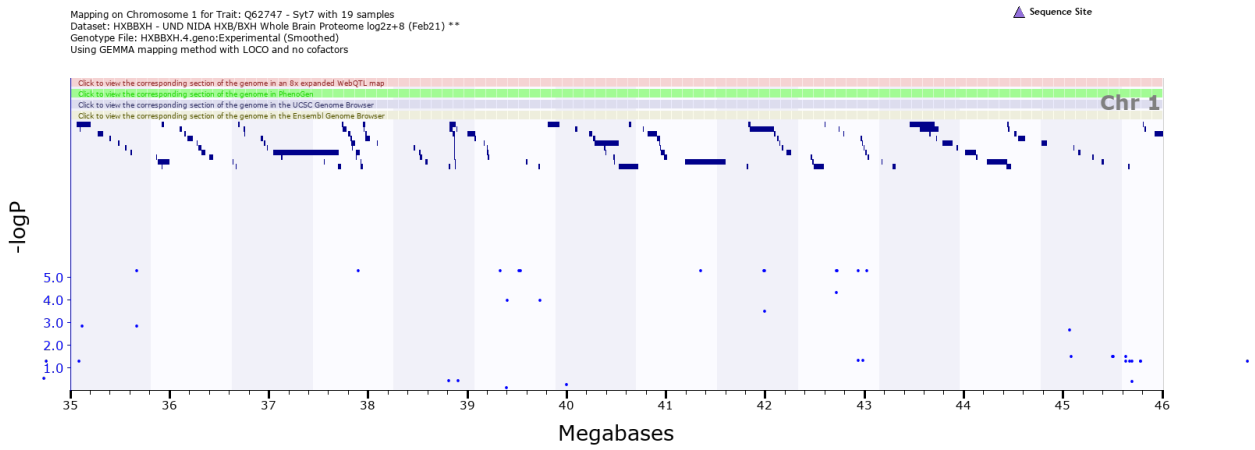
Normally, in an interval this large, we would just hit the pause button and spend more time increasing the sample size (in progress already by Xusheng Wang and colleagues). But for the sake of this GeneNetwork workflow, I am going to forge ahead and get to the box of chocolates—that essential dopamine kiss in nucleus accumbens.

Step 5. What genes are located along this part of Chr 1?
To answer this question, click on the chromosome number, **1** in this case.

This will generate a chromosome-specific view; part shown below.



The QTL peak is a "non-recombinant" plateau that extends from 35 to 45 Mb—confirming visually what we had already determined from the –logP values. The blue blocks along the top are gene "models" and the lighter blue dots are the linkage values at different SNPs. You can zoom to a map with specific start- and end-coordinates.
You can keep zooming in on a specific region of a chromosome by clicking on the pink horizontal bar alonge the top. Here is the plateau region of the SYT7 protein expression QTL, or pQTL.



As you can tell from the screenshot, there are lots of genes—real and putative—that call this part of Chr 1 home.

Underneath each map there is an **Interval Analyst** table of all genes and pseudogenes in a specific interval. In this case, there are about 130 gene, of which 36 are protein-coding.

Let me list them out: from 35.2 to 44.6 Mb.

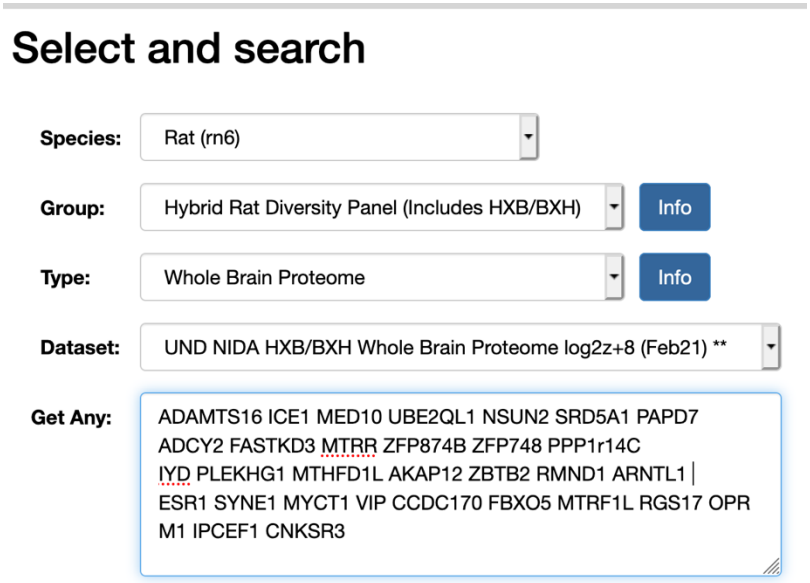
ADAMTS16
ICE1
MED10
UBE2QL1
NSUN2
SRD5A1
PAPD7
ADCY2
FASTKD3
MTRR
ZFP874B
ZFP748
PPP1R14C
IYD
PLEKHG1
MTHFD1L
AKAP12
ZBTB2
RMND1
ARNTL1
ESR1
SYNE1
MYCT1
VIP
CCDC170
FBXO5
MTRF1L
RGS17

OPRM1
IPCEF1
CNKSR3
SCAF8
TIAM2
TFB1M
CLDN20
NOX3

Anything catch your eye? Actually, lots of these genes may catch your eye—perhaps too many.

The gene/protein that most of you will highlight is ***OPRM1***—the mu opioid receptor. Variants in this gene and locus are definitely controllers of morphine response—particularly so in the BXD mouse family (Paige Lemen, Hao Chen, Guy Mittleman, and Price Dickson have a paper in progress on this topic). And this is also true in *Homo sapiens* based on initial GWAS analysis.

Step 6. How do we evaluate the strength of these candidates as controller of some subset of the nine proteins with variable expression that map to this region?
Simple—clip out all of those positional candidate genes and paste them into the search **Get Any** window of GeneNetwork. It should look like this:



About 12 of these proteins have reasonably high expression in the rat brain, and three of these also are associated with reasonably strong cis-acting modulation—FASTKD3, PPP1R14C, and MTRR. That means that DNA variant in or around these genes modulate both mRNA expression but much more importantly, also the protein level.

You can review these three candidates at your leisure.
PPP1R14C (aka KEPI)—see PMID: 11812771
MTRR: not much related to CNS function—mainly cancer and development
FASKD3: not much CNS but key in mitochondrial function

Ok, time to go out and swim.

Any one that made it this far—bravo—you have persistence.

Any questions about the proteomics to Xusheng Wang.

Any questions about the genotypes and HXB sequence to Hao Chen.

Any questions about mapping to Pjotr Prins and me.

Any questions about GeneNetwork user interface to me.

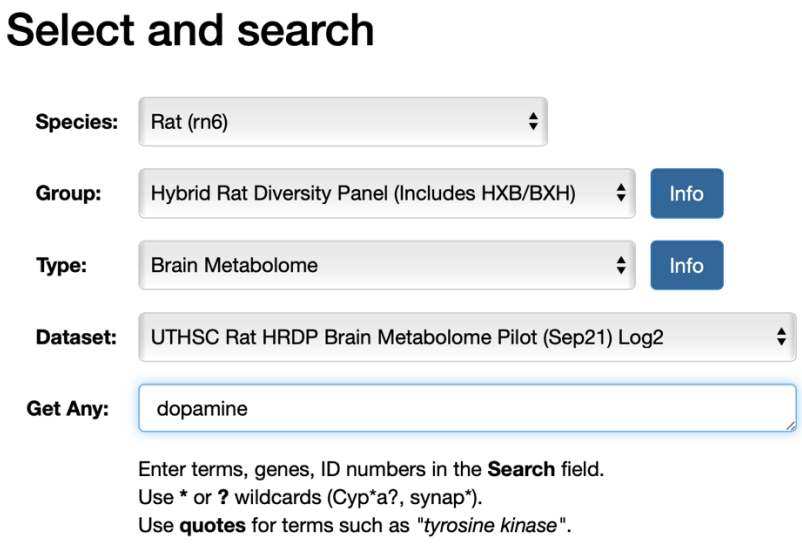
> Can I isolate your gene? Can I kiss your dopamine?....
> A perfect piece of DNA caught in a flashing ray
> A master piece of DNA caught in a flashing ray

Thanks RHCP for thinking of us NIDA- and NIAAA-funded genetics researchers.

ps. You may want to know about OPRM1 as a great position and biological candidate gene—is it causal? Unfortunately expression is not consistently high in this proteomics analysis and we will have to look at bit harder to find peptide fragments for this protein. Coming soon to a webservice near you. But Hao Chen does know that there are high impact variants in OPRM1 in the HRDP, so one could test the hypothesis that the variant is causal by a CRISPR-Cas9 allele swap.

A bit on the new Metabolomics Data

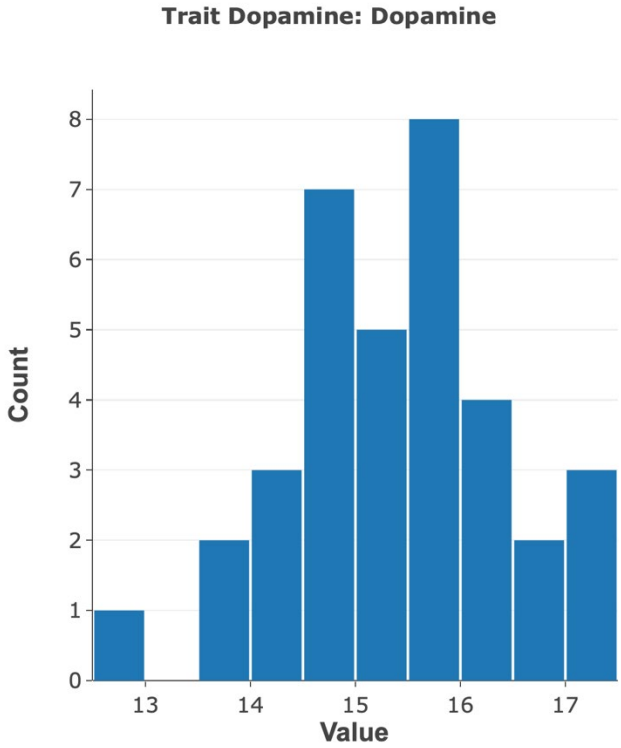
1. Link to



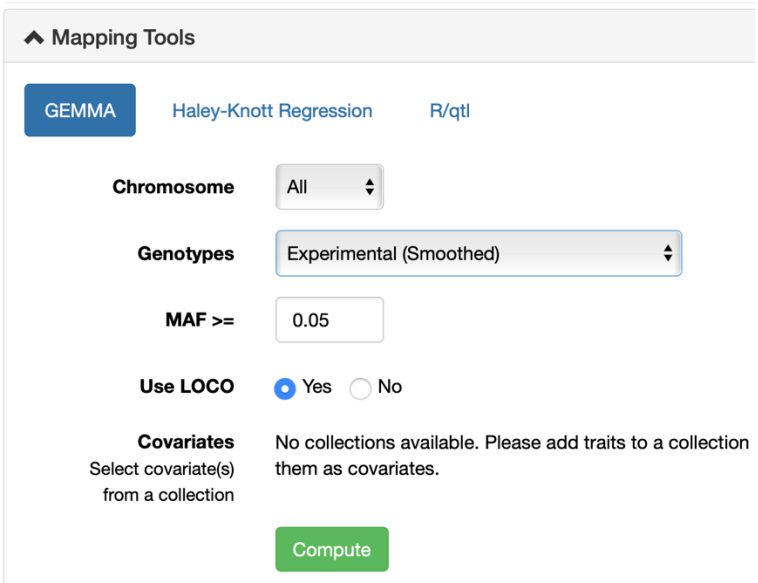
or just directly to

https://genenetwork.org/show_trait?trait_id=Dopamine&dataset=UTHSC_Rat_HRDP_Meta_log2_0921

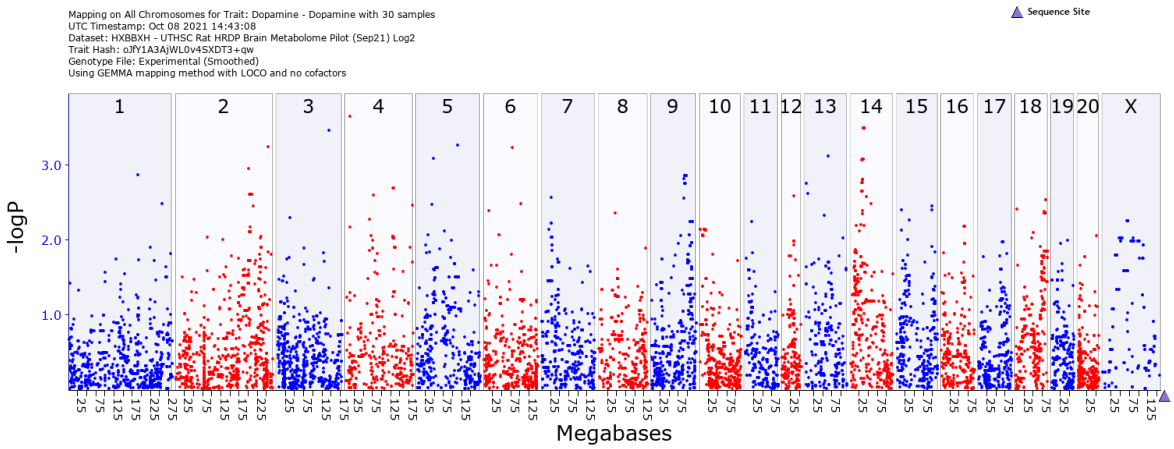
Here are expression levels of dopamine across the whole brain in 35 HRDP strains—all replicated male-female pairs. Scale in log2 level.



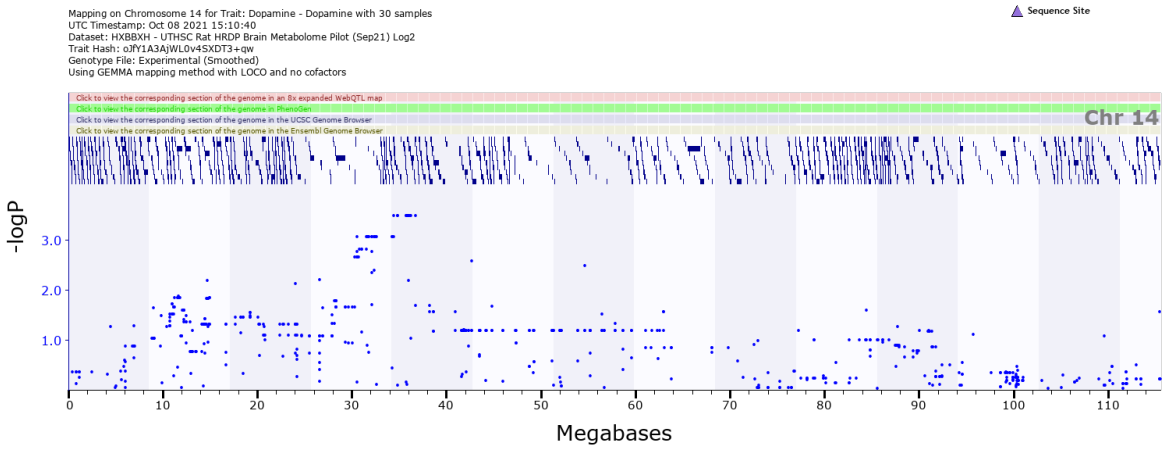
Map dopamine expression level using new experimental genotypes.



Below is the whole genetic map of modulation of DA levels.



One of the prominent peaks is on Chr 14 at about 34 Mb with a –logP scale is 3.5.



Let's look at genes in this interval. Do any of the genes in the table below have a role in DA ADME in brain?

☐	54	LOC102556008	34.273286	64.965						uncharacterized LOC102556008
☑	55	Nmu	34.353683	28.285		5	76.333494	4	56.302325	neuromedin U
☑	56	Pdcl2	34.384338	20.582		5	76.312115	4	56.263620	phosducin-like 2
☐	57	LOC102556353	34.414320	38.449						uncharacterized LOC102556353
☑	58	Clock	34.418226	83.992		5	76.209867	4	56.139587	clock circadian regulator
☐	59	LOC108352740	34.481340	4.073						uncharacterized LOC108352740
☐	60	Tmem165	34.503037	25.225		5	76.183879			transmembrane protein 165
☑	61	Srd5a3	34.554769	15.654		5	76.140272			steroid 5 alpha-reductase 3
☑	62	Kdr	34.727677	59.450		5	75.933269	4	55.785576	kinase insert domain receptor
☐	63	LOC102553643	34.960158	7.880						uncharacterized LOC102553643
☑	64	Kit	35.072131	77.507		5	75.574986	4	55.365033	KIT proto-oncogene receptor tyrosine kinase
☐	65	LOC679908	35.161235	0.377						similar to CG1249-PA
☐	66	LOC102553856	35.181886	22.876						uncharacterized LOC102553856
☐	67	LOC103692794	35.472712	0.619						NADH dehydrogenase [ubiquinone] 1 alpha subcompl
☑	68	Pdgfra	35.527926	53.204		5	75.156165	4	54.936374	platelet derived growth factor receptor alpha
☐	69	LOC102554024	35.556016	1.375						platelet-derived growth factor receptor alpha-like
☐	70	Gsx2	35.650985	1.724		5	75.075600			GS homeobox 2
☐	71	LOC102556465	35.679517	3.091						histone-binding protein RBBP4-like
☑	72	Chic2	35.683657	34.348		5	75.005003	4	54.716885	cysteine-rich hydrophobic domain 2
☐	73	LOC102554098	35.721135	1.708						uncharacterized LOC102554098
☐	74	LOC102554165	35.848301	7.643						uncharacterized LOC102554165
☐	75	LOC102554235	35.861836	13.602						uncharacterized LOC102554235
☐	76	LOC108352741	35.879034	10.743						uncharacterized LOC108352741
☐	77	LOC102554353	36.012187	1.952						uncharacterized LOC102554353
☑	78	Lnx1	36.047136	102.604		5	74.597103			ligand of numb-protein X 1
☑	79	Fip1l1	36.150381	58.014		5	74.535481	4	46.379466	factor interacting with PAPOLA and CPSF1
☑	80	Scfd2	36.216002	327.659		5	74.204815	4	53.580079	sec1 family domain containing 2
☐	81	LOC103692809	36.377567	5.465						uncharacterized LOC103692809
☐	82	LOC103692806	36.423277	2.801						uncharacterized LOC103692806
☑	83	Rasl11b	36.550353	4.221		5	74.195325	4	53.569422	RAS-like family 11 member B
☐	84	LOC102554412	36.560700	31.487						uncharacterized LOC102554412
☑	85	Dancr	36.664036	0.980		5	74.093082			differentiation antagonizing non-protein coding RNA
☑	86	Usp46	36.687134	68.226		5	74.000037	4	53.302137	ubiquitin specific peptidase 46
☐	87	LOC102554514	36.700564	7.337						uncharacterized LOC102554514
☐	88	LOC364143	36.850517	1.990						similar to WW domain binding protein 11
☐	89	LOC100363494	36.865771	0.287						ubiquitin B-like
☐	90	LOC108352851	36.886081	4.431						uncharacterized LOC108352851
☐	91	LOC108352742	36.969456	3.846						uncharacterized LOC108352742
☐	92	LOC103690248	36.971622	24.629						uncharacterized LOC103690248
☐	93	LOC102554979	36.983738	4.115						uncharacterized LOC102554979
☑	94	Spata18	37.081529	26.716		5	73.651379	4	52.758520	spermatogenesis associated 18
☐	95	LOC103692822	37.110015	2.934						uncharacterized LOC103692822
☑	96	Sgcb	37.113194	15.429		5	73.632748	4	52.727799	sarcoglycan, beta
☑	97	Lrrc66	37.128715	25.175		5	73.606641			leucine rich repeat containing 66