Constraint variables of inherited retinal diseases in gnomAD v2.1

Jose S. Pulido, MD,MS,MPH, MBA

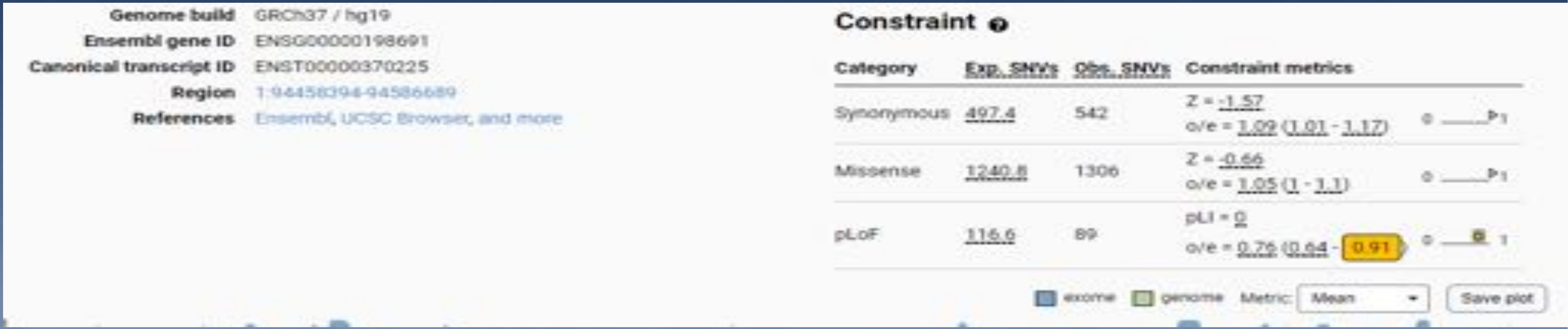
Larry A. Donoso Endowed Chair

Director of the Bower Laboratory for Translational Medicine

Vickie and Jack Farber Vision Research Center at Wills Eye Hospital

- Gavin Arno, PhD
- Hwei Wuen Chan, MD
- Alex Tanner, MD
- Elena Schiff, PhD
- Andrew R. Webster, MD, FRCOphth
- Omar Mahroo, PhD, FRCOphth

gnomAD



Genome build	GRCh37 / hg19
Ensembl gene ID	ENSG00000198691
Canonical transcript ID	ENST00000370225
Region	1:94458394-94586689
References	Ensembl, UCSC Browser, and more

Constraint

Category	Exp. SNVs	Obs. SNVs	Constraint metrics	
Synonymous	497.4	542	Z = -1.57 o/e = 1.09 (1.01 - 1.17)	0 1
Missense	1240.8	1306	Z = -0.66 o/e = 1.05 (1 - 1.1)	0 1
pLoF	116.6	89	pLI = 0 o/e = 0.76 (0.64 - 0.91)	0 1

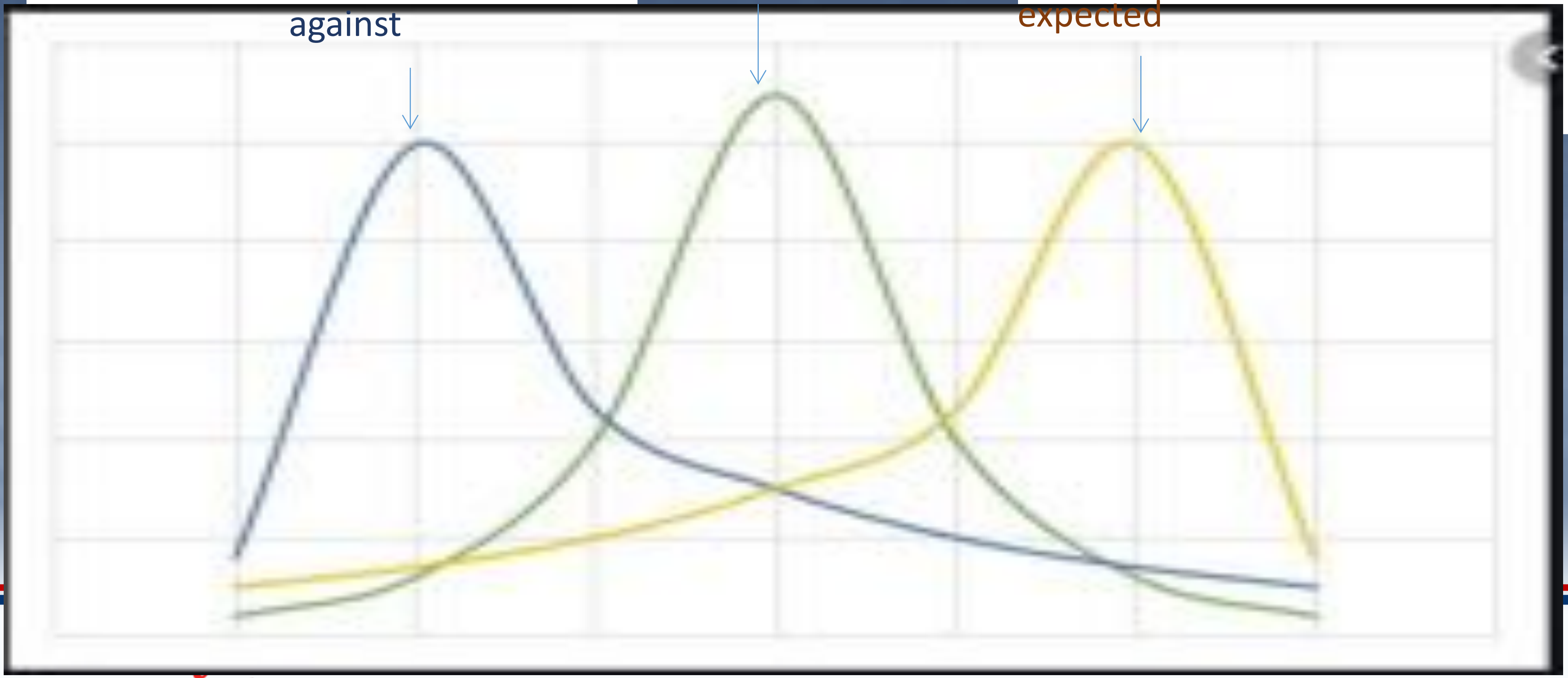
gnomAD

- **141, 456 individuals**
- **125,748 exomes and 15,708 genomes from unrelated individuals aligned against the GRCh37**
- **The average human DNA mutation rate is estimated to be approximately 2.5 x 10(-8) mutations per nucleotide site or 175 mutations per diploid genome per generation**
- **Total 24,754,800 new mutations ct GRCh37 in just this generation**

gnomAD

- If no mutation at any one site, then one peak but the mutations should be dependent upon the number of bases per gene.
- CT GrCh37, there should be a one tailed curve of number of mutations,
- BUT the “reference is actually the human/chimp primate and the dNonsyn/dsyn for every gene is checked as the expected bell shaped curve.
- If the shape shifts towards more mutations then the site is a hot spot for mutations or there is selection for mutation
- if the shape shifts to less mutations then there is selection against mutation or the site is protected against mutation

Normal o/e vs selected against and “selected for or hot spot”

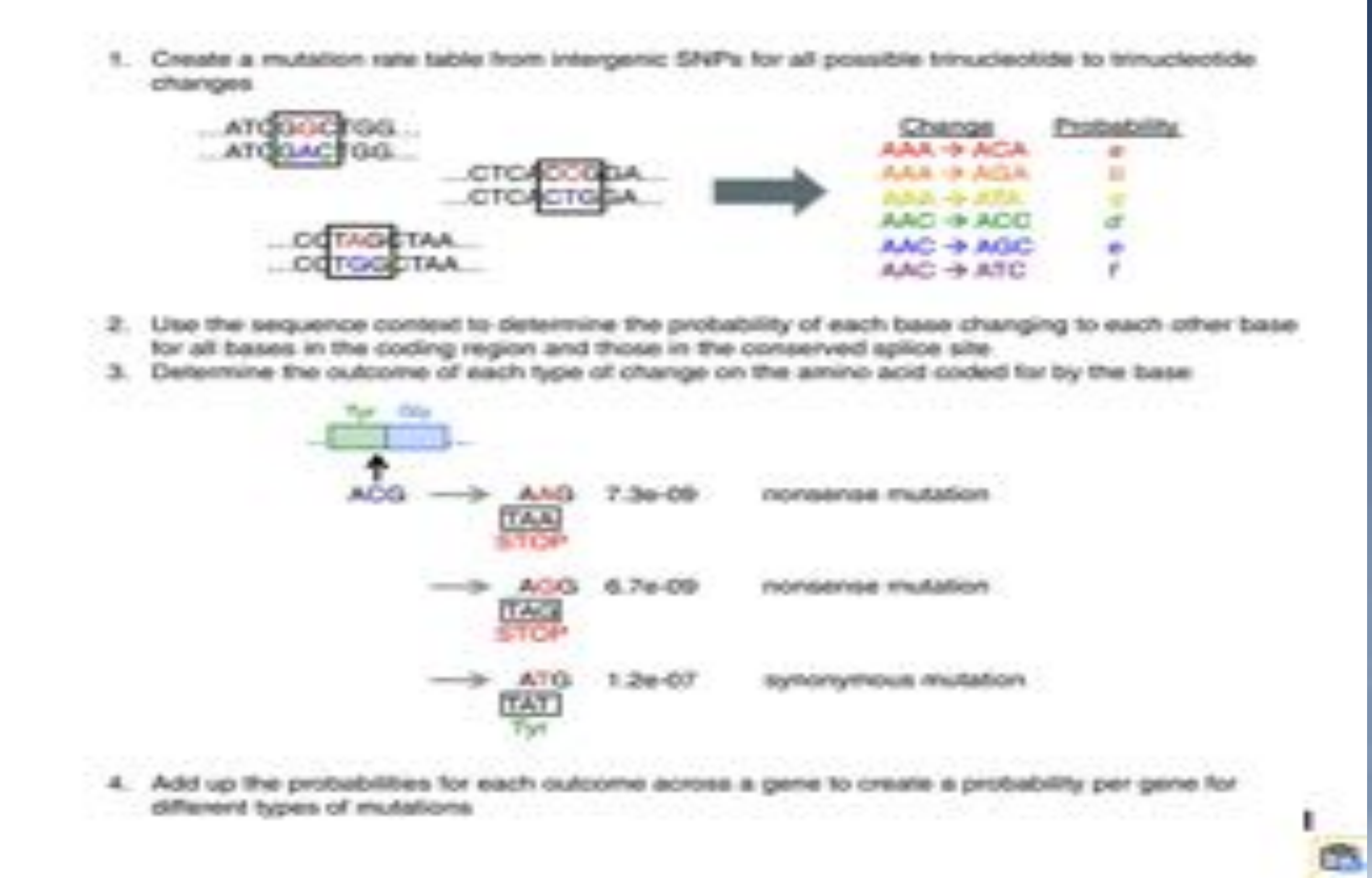


Probability of loss of function intolerance (pli)

- Loss-of-function variants include frameshifting and stop variants and are of particular interest because of their potentially profound impact on the mRNA transcript and translated protein
- >0.9 is considered significant in EXAC and still used but now also
- 0/e upper CI<0.4 (no CI in the past)
- We have used both

- A framework for the interpretation of de novo mutation in human disease Kaitlin E Samocha Nat Gen 2014

PLI continued



1. Create a mutation rate table from intergenic SNPs for all possible trinucleotide to trinucleotide changes

...ATCGGCTGG...
...ATCGACTGG...
...CTCACCGGA...
...CTCACTGGA...
...CCTAGCTAA...
...CCTGGCTAA...

Change	Probability
AAA → ACA	a
AAA → AGA	b
AAA → ATA	c
AAC → ACC	d
AAC → AGC	e
AAC → ATC	f

2. Use the sequence context to determine the probability of each base changing to each other base for all bases in the coding region and those in the conserved splice site
3. Determine the outcome of each type of change on the amino acid coded for by the base

ACG → AAG 7.3e-09 nonsense mutation
TAA
STOP

→ ACG 6.7e-09 nonsense mutation
TAG
STOP

→ ATG 1.2e-07 synonymous mutation
TAT
Tyr

4. Add up the probabilities for each outcome across a gene to create a probability per gene for different types of mutations

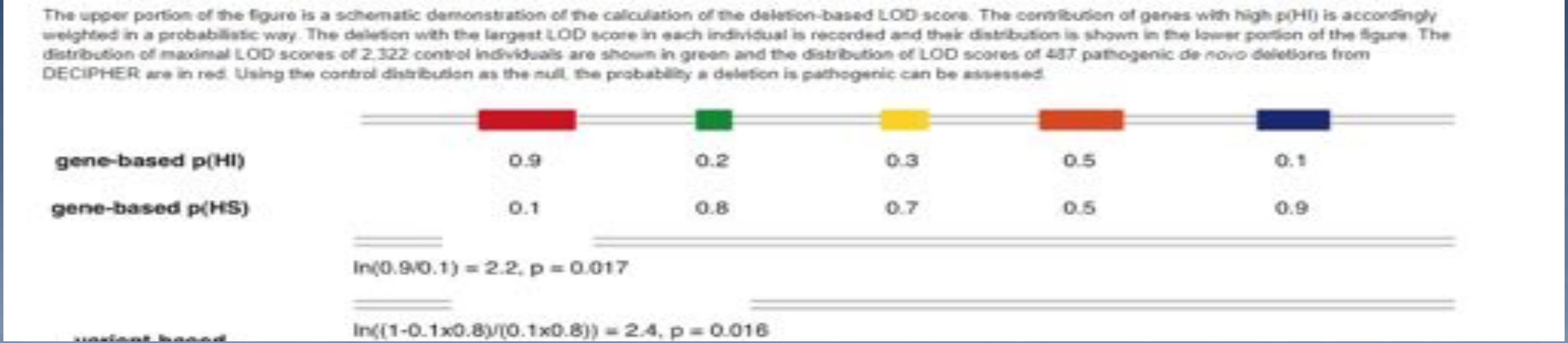
HI

- Huang and colleagues made this metric by using properties of established haploinsufficient and haplosufficient genes to train a predictive model. The properties included in the final model were “dN/dS between human and macaque, promoter sequence, embryonic expression and network proximity to known HI [haploinsufficient] genes

doi:10.1038/nature19057

RESEARCH SUPPLEMENTARY INFORMATION

pHI



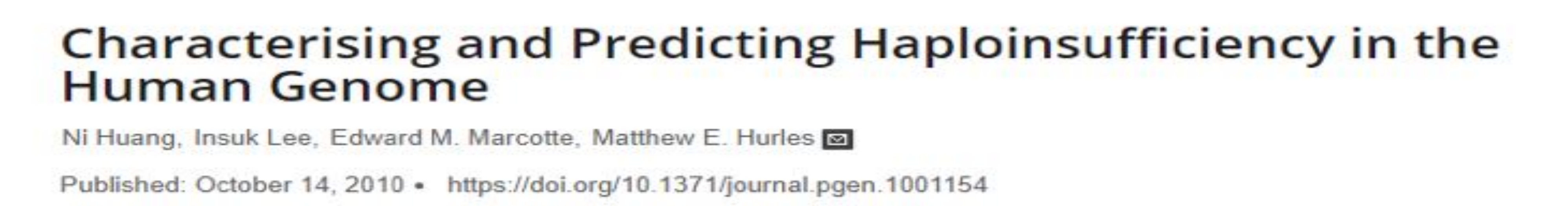
The upper portion of the figure is a schematic demonstration of the calculation of the deletion-based LOD score. The contribution of genes with high p(HI) is accordingly weighted in a probabilistic way. The deletion with the largest LOD score in each individual is recorded and their distribution is shown in the lower portion of the figure. The distribution of maximal LOD scores of 2,322 control individuals are shown in green and the distribution of LOD scores of 487 pathogenic *de novo* deletions from DECIPHER are in red. Using the control distribution as the null, the probability a deletion is pathogenic can be assessed.

	🟥	🟩	🟨	🟧	🟦
gene-based p(HI)	0.9	0.2	0.3	0.5	0.1
gene-based p(HS)	0.1	0.8	0.7	0.5	0.9

ln(0.9/0.1) = 2.2, p = 0.017

ln((1-0.1x0.8)/(0.1x0.8)) = 2.4, p = 0.016

variant-based

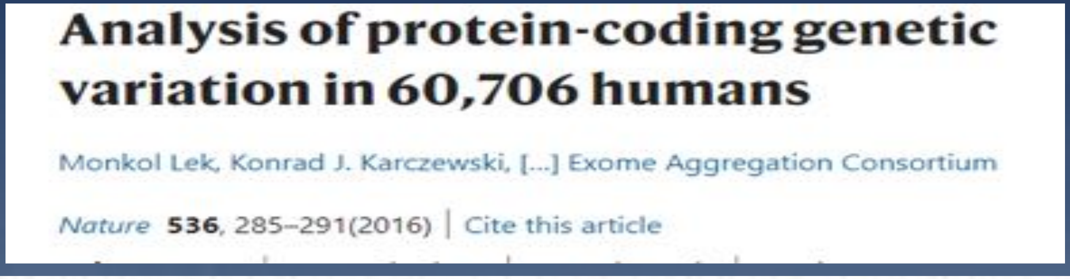


Characterising and Predicting Haploinsufficiency in the Human Genome

Ni Huang, Insuk Lee, Edward M. Marcotte, Matthew E. Hurles

Published: October 14, 2010 • https://doi.org/10.1371/journal.pgen.1001154

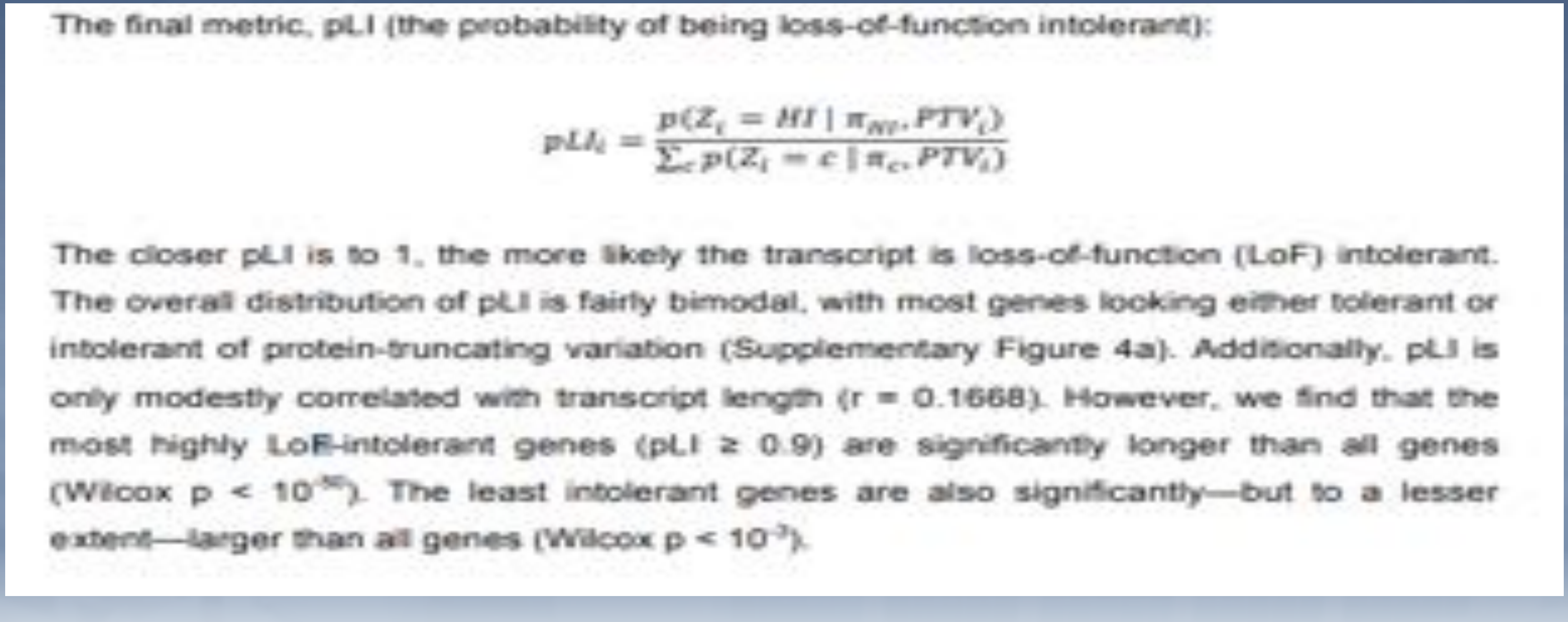
pli further information-related to HI



Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek, Konrad J. Karczewski, [...] Exome Aggregation Consortium

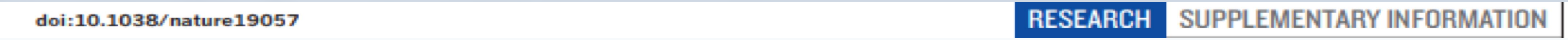
Nature **536**, 285–291(2016) | Cite this article



The final metric, pLI (the probability of being loss-of-function intolerant):

$$pLI_i = \frac{p(Z_i = HI \mid \pi_{HI}, PTV_i)}{\sum_c p(Z_i = c \mid \pi_c, PTV_i)}$$

The closer pLI is to 1, the more likely the transcript is loss-of-function (LoF) intolerant. The overall distribution of pLI is fairly bimodal, with most genes looking either tolerant or intolerant of protein-truncating variation (Supplementary Figure 4a). Additionally, pLI is only modestly correlated with transcript length (r = 0.1668). However, we find that the most highly LoF-intolerant genes ($pLI \geq 0.9$) are significantly longer than all genes (Wilcox $p < 10^{-50}$). The least intolerant genes are also significantly—but to a lesser extent—larger than all genes (Wilcox $p < 10^{-7}$).



doi:10.1038/nature19057

RESEARCH SUPPLEMENTARY INFORMATION

Methods

- 312 genes found to be associated with IRD on RetNet were evaluated for their constraint variables using gnomAD v2.1 and DECIPHER
- For LOF variants constraint was PLI >0.9 and highest CI was 0.35 for o/e
- DECIPHER is based on children who were sequenced-HI<10 and PLI>0.9
- For MS and synonymous variants Z>2.99 or <-2.99 ie less than .0014 in the distribution

DECIPHER

- suffering from Rare Disease
- 33,000 cases from 250 centers
- Uses HI (haploinsufficiency index) 0-10% and pli>0.9 quite haploinsufficient
- HI=known haploinsufficient genes and genes disrupted by unambiguous loss-of-function variants in at least two apparently healthy individuals. Percentages refer to genome-wide percentiles of genes ranked according to their haploinsufficient score.
- Pli= Genes with high pLI scores (pLI ≥ 0.9) are extremely LoF intolerant, whereby genes with low pLI scores (pLI ≤ 0.1) are LoF tolerant.

Rules

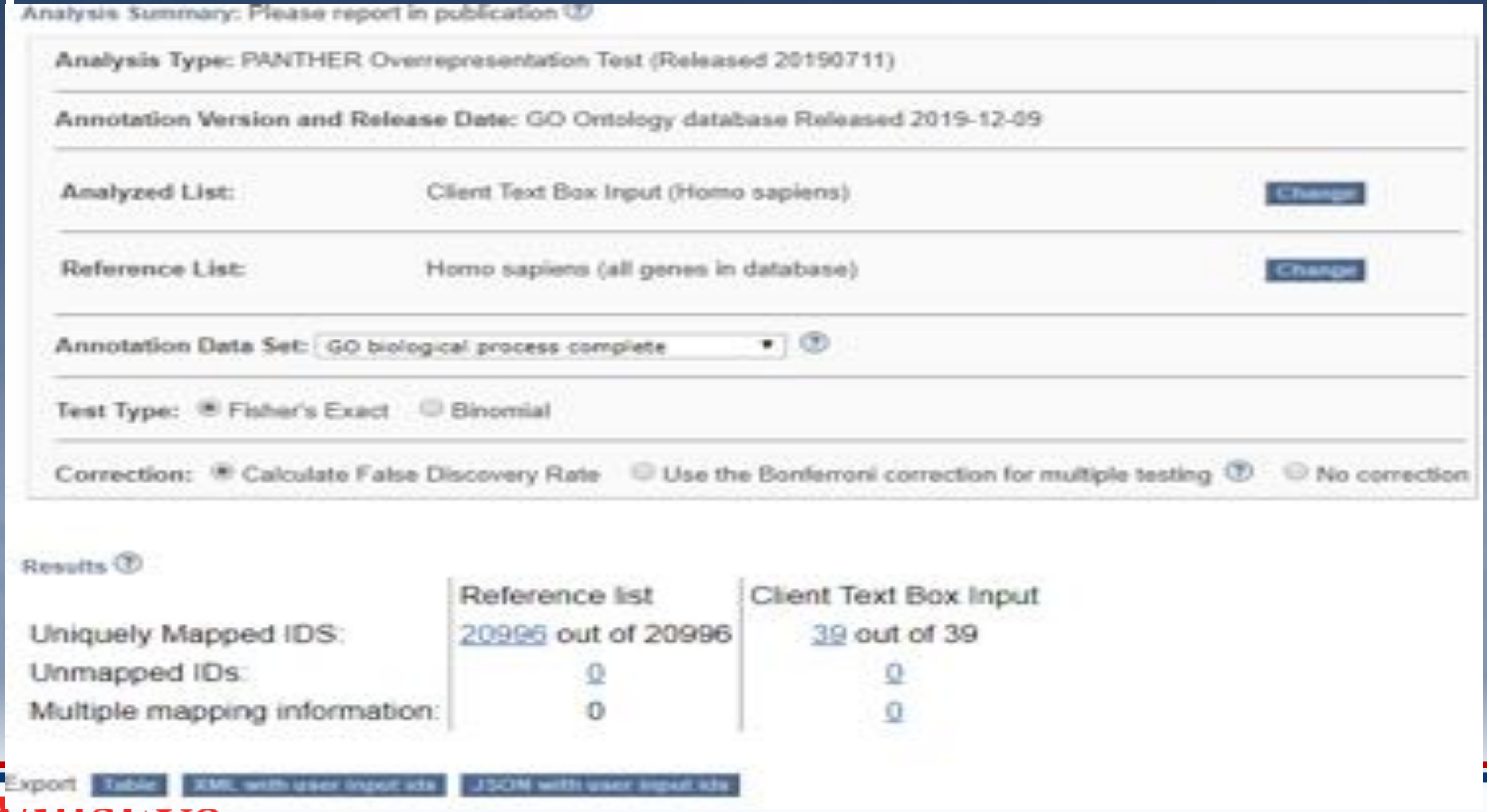
- We also show that longer genes are, in general, more depleted of protein-truncating variation (observed/expected), which can explain the enrichment of long genes in the set of genes with pLI ≥ 0.9. There is a relationship between deciles of gene length (bins of increasing gene length) and the observed depletion of PTVs in that bin: longer genes (deciles closer to 1) have a significantly lower rate of observed/expected (p < 10-50)
- Given that the X chromosome is hemizygous in males, we expect that genes on the X would be more constrained than those on autosomes. As expected, we find the genes on the X chromosomes are significantly more constrained than those genes on the autosomes for missense and loss-of-function (synonymous p = 0.0223; missense p = 4.43x10-8; loss-of-function p = 2.50x10-75). The high correlation between the observed and expected number of synonymous variants on the X chromosome (r = 0.9677 vs 0.9777 for autosomes) indicates that this difference in constraint is not due to a calibration issue

Results-either in gnomAD and/or DECIPHER

Loss of function variants-these are selected against-39 genes

RPGR
RS1
CHM
PRPF31
OPA1
RP2
EFEMP1
PRPF8
KIF11
LRP5*
SNRNP200
PRPF3
FZD4
COL11A1
TOPORS
COL2A1
OPN1LW*
ATXN7*
RIMS1*
JAG1
TEAD1
PITPNM3
OTX2
CCT2
GDF6
CTNNA1
VCAN
MFN2
NR2F1
HK1
PRPF4
AHR
OFD1
PRPS1
ZNF423
RB1
DMD
C3
FBLN5

Gene Ontology Panther (GO Panther) over-representation test



Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (Released 20190711)

Annotation Version and Release Date: GO Ontology database Released 2019-12-09

Analyzed List: Client Text Box Input (Homo sapiens) Change

Reference List: Homo sapiens (all genes in database) Change

Annotation Data Set: GO biological process complete

Test Type: ● Fisher's Exact ○ Binomial

Correction: ● Calculate False Discovery Rate ○ Use the Bonferroni correction for multiple testing ○ No correction

Results

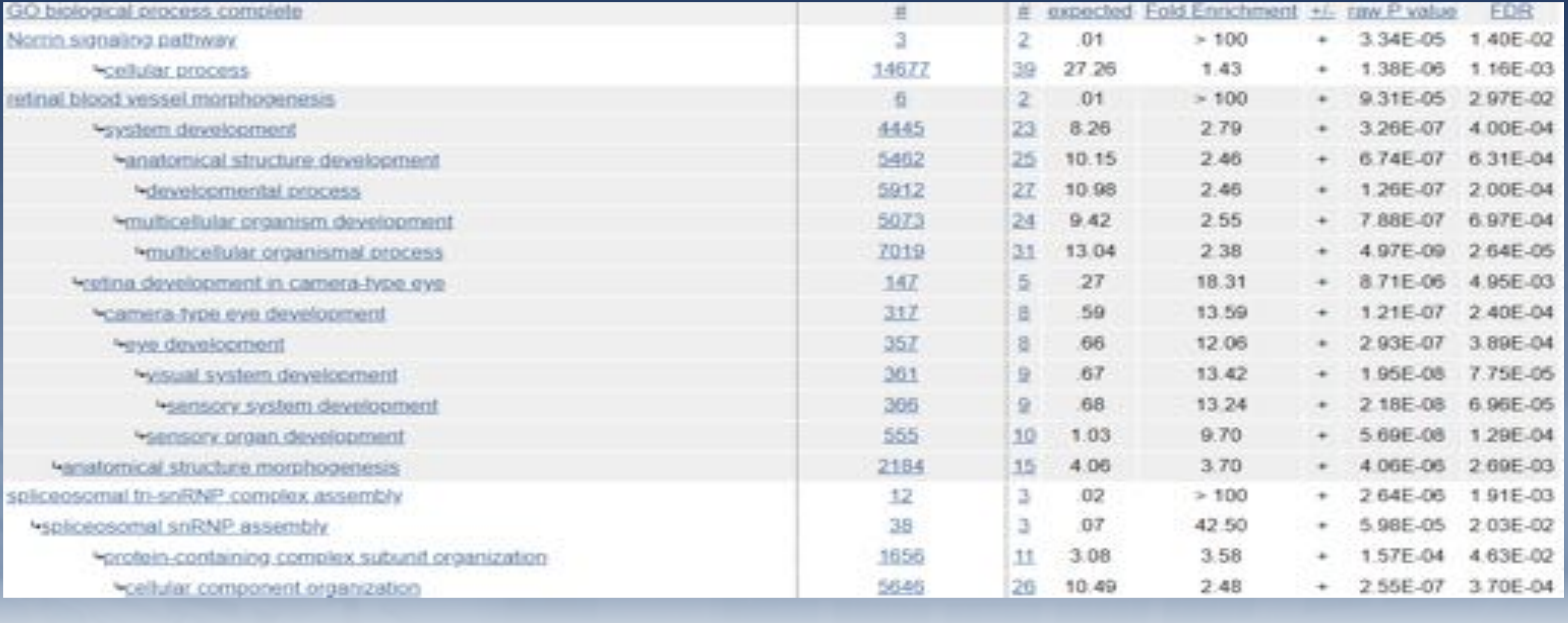
	Reference list	Client Text Box Input
Uniquely Mapped IDS:	20996 out of 20996	39 out of 39
Unmapped IDs:	0	0
Multiple mapping information:	0	0

Export Table XML with user input ids JSON with user input ids

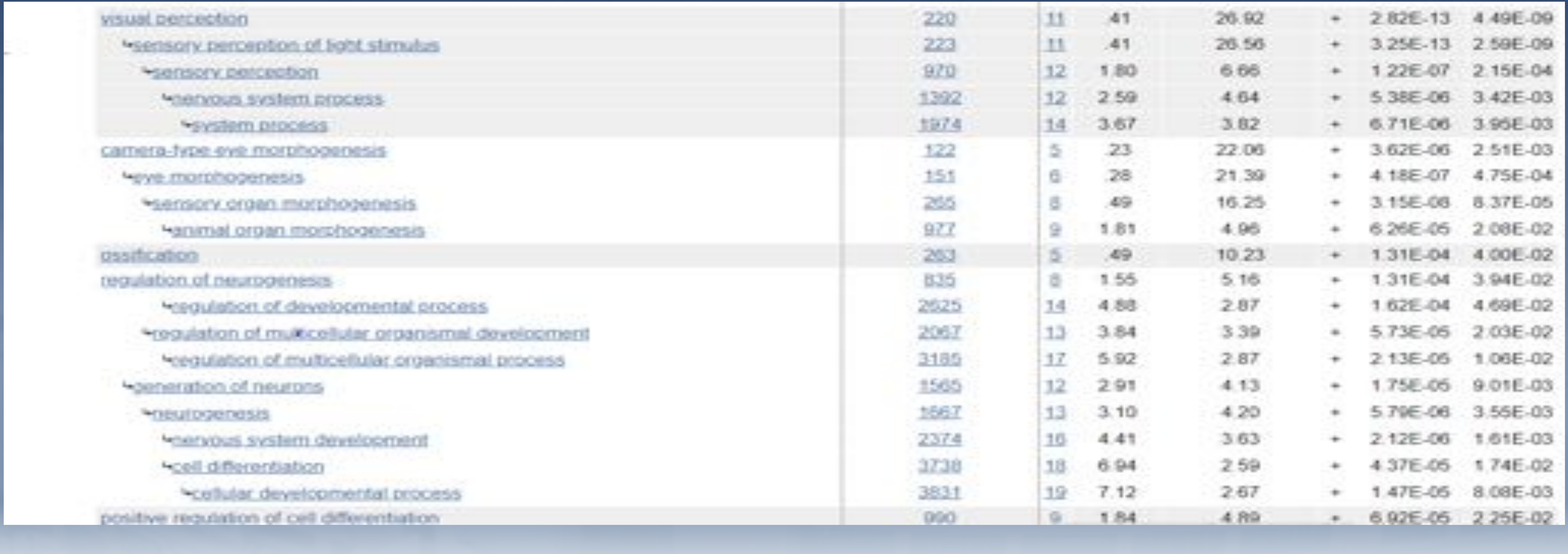
WillsEye Hospital

- *P-value* is the probability or chance of seeing at least x number of genes out of the total n genes in the list annotated to a particular GO term, given the proportion of genes in the whole genome that are annotated to that GO Term. That is, the GO terms shared by the genes in the user's list are compared to the background distribution of annotation. The closer the p-value is to zero, the more significant the particular GO term associated with the group of genes is (i.e. the less likely the observed annotation of the particular GO term to a group of genes occurs by chance).

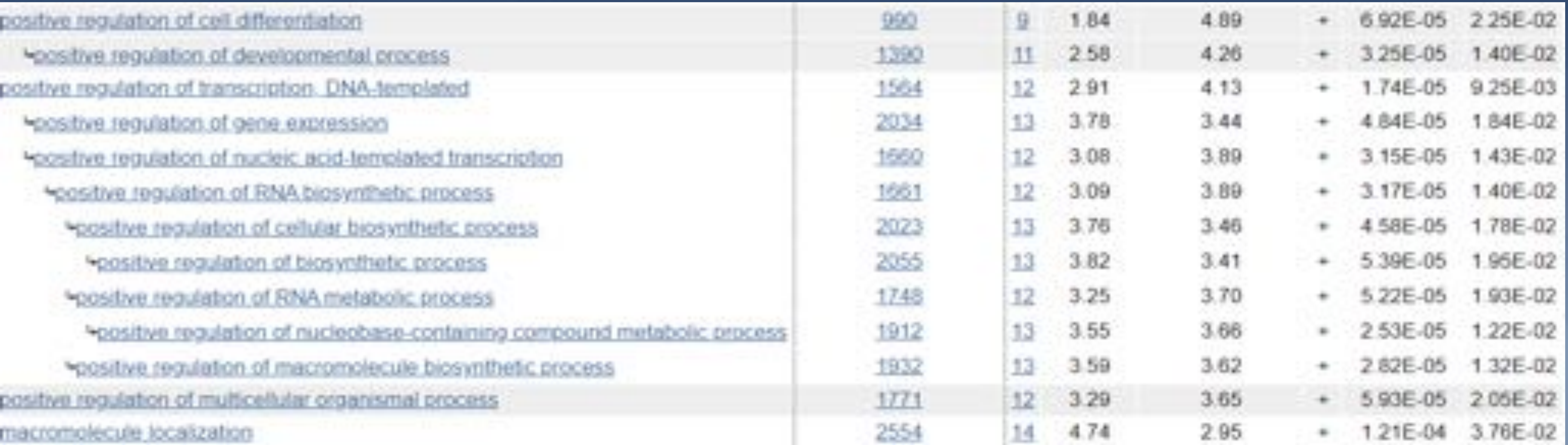
Norrin signalling-2, Spliceosomal 3



GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
Norrin signaling pathway	3	2	.01	> 100	+	3.34E-05	1.40E-02
↳cellular process	14677	39	27.26	1.43	+	1.38E-06	1.16E-03
retinal blood vessel morphogenesis	6	2	.01	> 100	+	9.31E-05	2.97E-02
↳system development	4445	23	8.26	2.79	+	3.26E-07	4.00E-04
↳anatomical structure development	5462	25	10.15	2.46	+	6.74E-07	6.31E-04
↳developmental process	5912	27	10.98	2.46	+	1.26E-07	2.00E-04
↳multicellular organism development	5073	24	9.42	2.55	+	7.88E-07	6.97E-04
↳multicellular organismal process	7019	31	13.04	2.38	+	4.97E-09	2.64E-05
↳retina development in camera-type eye	147	5	.27	18.31	+	8.71E-06	4.95E-03
↳camera-type eye development	317	8	.59	13.59	+	1.21E-07	2.40E-04
↳eye development	357	8	.66	12.06	+	2.93E-07	3.89E-04
↳visual system development	361	9	.67	13.42	+	1.95E-08	7.75E-05
↳sensory system development	366	9	.68	13.24	+	2.18E-08	6.96E-05
↳sensory organ development	555	10	1.03	9.70	+	5.69E-08	1.29E-04
↳anatomical structure morphogenesis	2184	15	4.06	3.70	+	4.06E-06	2.69E-03
spliceosomal tri-snRNP complex assembly	12	3	.02	> 100	+	2.64E-06	1.91E-03
↳spliceosomal snRNP assembly	38	3	.07	42.50	+	5.98E-05	2.03E-02
↳protein-containing complex subunit organization	1656	11	3.08	3.58	+	1.57E-04	4.63E-02
↳cellular component organization	5646	26	10.49	2.48	+	2.55E-07	3.70E-04

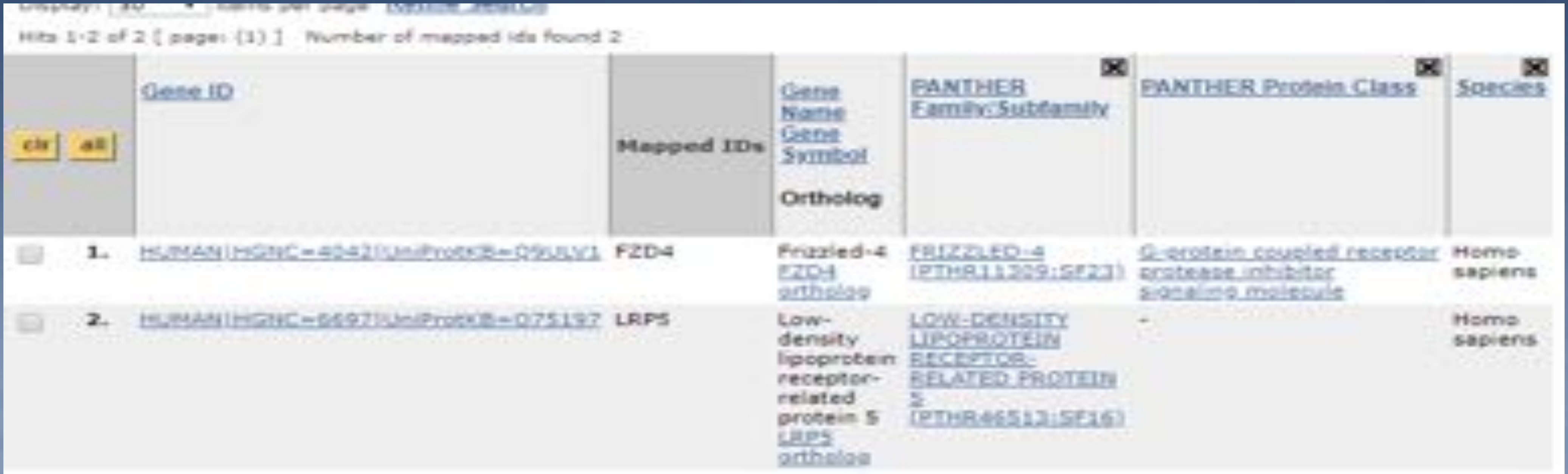


visual perception	220	11	.41	26.92	+	2.82E-13	4.49E-09
↳sensory perception of light stimulus	223	11	.41	26.56	+	3.25E-13	2.59E-09
↳sensory perception	970	12	1.80	6.66	+	1.22E-07	2.15E-04
↳nervous system process	1392	12	2.59	4.64	+	5.38E-06	3.42E-03
↳system process	1974	14	3.67	3.82	+	6.71E-06	3.95E-03
camera-type eye morphogenesis	122	5	.23	22.06	+	3.62E-06	2.51E-03
↳eye morphogenesis	151	6	.28	21.39	+	4.18E-07	4.75E-04
↳sensory organ morphogenesis	265	8	.49	16.25	+	3.15E-08	8.37E-05
↳animal organ morphogenesis	977	9	1.81	4.96	+	6.26E-05	2.08E-02
ossification	263	5	.49	10.23	+	1.31E-04	4.00E-02
regulation of neurogenesis	835	8	1.55	5.16	+	1.31E-04	3.94E-02
↳regulation of developmental process	2625	14	4.88	2.87	+	1.62E-04	4.69E-02
↳regulation of multicellular organismal development	2067	13	3.84	3.39	+	5.73E-05	2.03E-02
↳regulation of multicellular organismal process	3185	17	5.92	2.87	+	2.13E-05	1.06E-02
↳generation of neurons	1565	12	2.91	4.13	+	1.75E-05	9.01E-03
↳neurogenesis	1567	13	3.10	4.20	+	5.79E-06	3.55E-03
↳nervous system development	2374	16	4.41	3.63	+	2.12E-06	1.61E-03
↳cell differentiation	3738	18	6.94	2.59	+	4.37E-05	1.74E-02
↳cellular developmental process	3831	19	7.12	2.67	+	1.47E-05	8.08E-03
positive regulation of cell differentiation	990	9	1.84	4.89	+	6.92E-05	2.25E-02



positive regulation of cell differentiation	990	9	1.84	4.89	+	6.92E-05	2.25E-02
↳positive regulation of developmental process	1390	11	2.58	4.26	+	3.25E-05	1.40E-02
positive regulation of transcription, DNA-templated	1564	12	2.91	4.13	+	1.74E-05	9.25E-03
↳positive regulation of gene expression	2034	13	3.78	3.44	+	4.84E-05	1.84E-02
↳positive regulation of nucleic acid-templated transcription	1660	12	3.08	3.89	+	3.15E-05	1.43E-02
↳positive regulation of RNA biosynthetic process	1661	12	3.09	3.89	+	3.17E-05	1.40E-02
↳positive regulation of cellular biosynthetic process	2023	13	3.76	3.46	+	4.58E-05	1.78E-02
↳positive regulation of biosynthetic process	2055	13	3.82	3.41	+	5.39E-05	1.95E-02
↳positive regulation of RNA metabolic process	1748	12	3.25	3.70	+	5.22E-05	1.93E-02
↳positive regulation of nucleobase-containing compound metabolic process	1912	13	3.55	3.66	+	2.53E-05	1.22E-02
↳positive regulation of macromolecule biosynthetic process	1932	13	3.59	3.62	+	2.82E-05	1.32E-02
positive regulation of multicellular organismal process	1771	12	3.29	3.65	+	5.93E-05	2.05E-02
macromolecule localization	2554	14	4.74	2.95	+	1.21E-04	3.76E-02

Norrin signal(l)ing



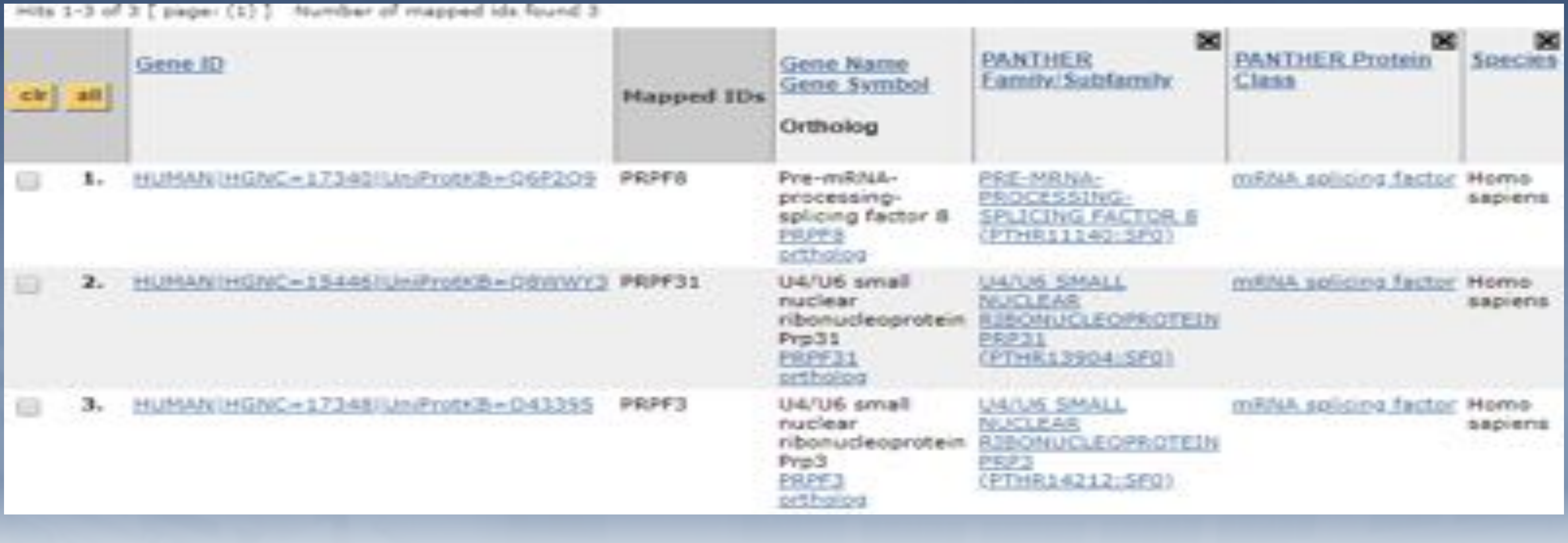
Display: 10 items per page Refine Search

Hits 1-2 of 2 [page: (1)] Number of mapped ids found 2

clr all		Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species
□	1.	HUMAN|HGNC=4042|UniProtKB=Q9ULV1	FZD4	Frizzled-4 FZD4 ortholog	FRIZZLED-4 (PTHR11309:SF23)	G-protein coupled receptor protease inhibitor signaling molecule	Homo sapiens
□	2.	HUMAN|HGNC=6697|UniProtKB=O75197	LRP5	Low-density lipoprotein receptor-related protein 5 LRP5 ortholog	LOW-DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 5 (PTHR46513:SF16)	-	Homo sapiens

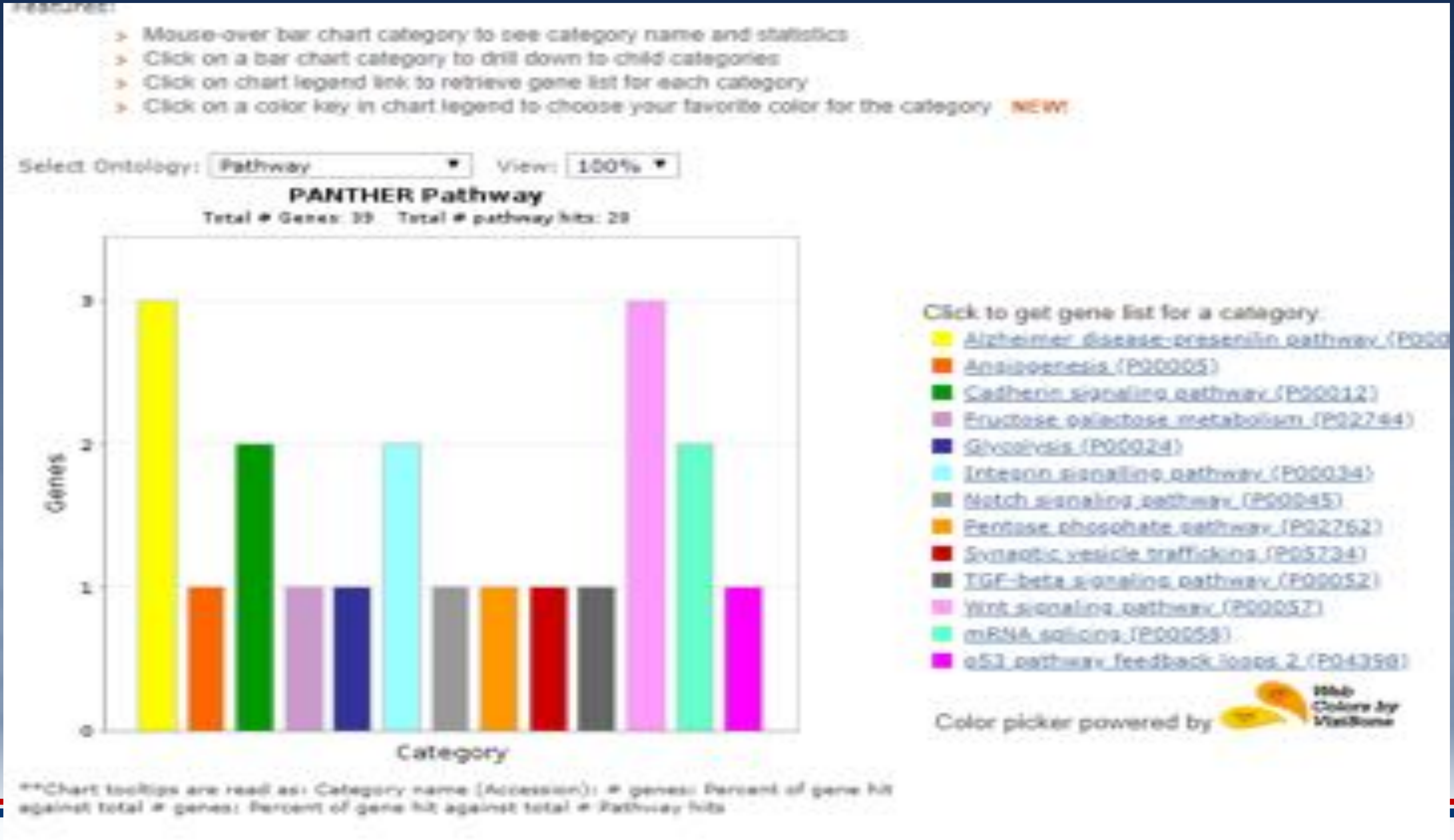
WillsEye Hospital

Spliceosomal complex assembly



Hits 1-3 of 3 [page: (1)]   Number of mapped ids found 3

clr all		Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species
□	1.	HUMAN|HGNC=17340|UniProtKB=Q6P2Q9	PRPF8	Pre-mRNA-processing-splicing factor 8 PRPF8 ortholog	PRE-MRNA-PROCESSING-SPLICING FACTOR 8 (PTHR11140:SF0)	mRNA splicing factor	Homo sapiens
□	2.	HUMAN|HGNC=15446|UniProtKB=Q8WWY3	PRPF31	U4/U6 small nuclear ribonucleoprotein Prp31 PRPF31 ortholog	U4/U6 SMALL NUCLEAR RIBONUCLEOPROTEIN PRP31 (PTHR13904:SF0)	mRNA splicing factor	Homo sapiens
□	3.	HUMAN|HGNC=17348|UniProtKB=O43395	PRPF3	U4/U6 small nuclear ribonucleoprotein Prp3 PRPF3 ortholog	U4/U6 SMALL NUCLEAR RIBONUCLEOPROTEIN PRP3 (PTHR14212:SF0)	mRNA splicing factor	Homo sapiens



Features:

- Mouse-over bar chart category to see category name and statistics
- Click on a bar chart category to drill down to child categories
- Click on chart legend link to retrieve gene list for each category
- Click on a color key in chart legend to choose your favorite color for the category NEW!

Select Ontology: Pathway ▼ View: 100% ▼

PANTHER Pathway

Total # Genes: 39 Total # pathway hits: 28

Click to get gene list for a category:

- 🟨 Alzheimer disease-presenilin pathway (P000
- 🟧 Angiogenesis (P00005)
- 🟩 Cadherin signaling pathway (P00012)
- 🟪 Fructose galactose metabolism (P02744)
- 🟦 Glycolysis (P00024)
- 🟦 Integrin signalling pathway (P00034)
- ⬜ Notch signaling pathway (P00045)
- 🟧 Pentose phosphate pathway (P02762)
- 🟥 Synaptic vesicle trafficking (P05734)
- ⬛ TGF-beta signaling pathway (P00052)
- 🟪 Wnt signaling pathway (P00057)
- 🟩 mRNA splicing (P00058)
- 🟪 p53 pathway feedback loops 2 (P04398)

Color picker powered by

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Pathway hits

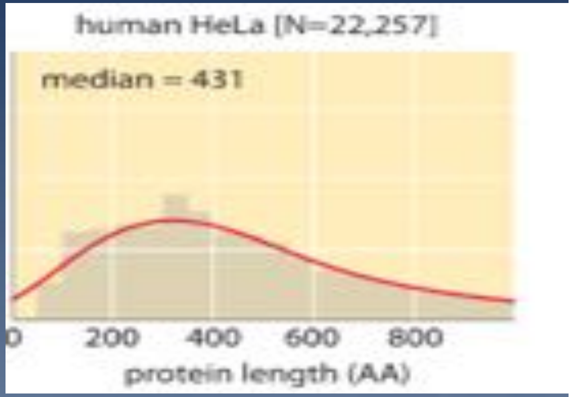
X-linked-8 so +6 from GO=14

- **RPGR**
- **RS**
- **CHM**
- **RP2**
- **OPN1LW**
- **OFD1**
- **PRPS1**
- **DMD**

Large genes >1200kb, >400 aa

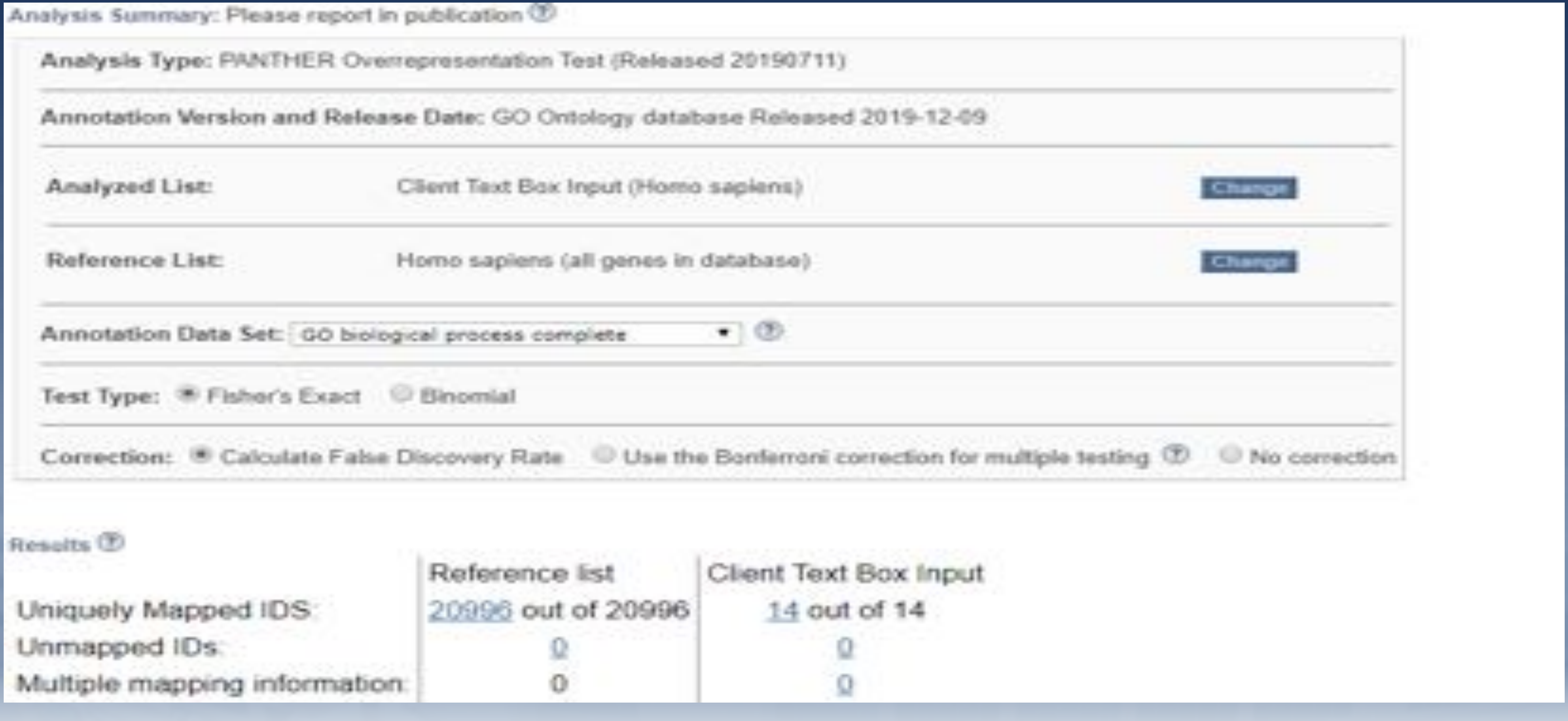
- **OPA1-700aa**
- **EFEMP1-493aa**
- **KIF11-1093aa**
- **COL11A1-1806aa**
- **TOPORS-1045aa**
- **SNRNP200-2136aa**
- **COL2A1-1487aa**
- **ATXN7-945aa**
- **RIMS1-1692aa**
- **JAG1-1218aa**
- **PRPF8-2335aa**

RPGR
RS1
CHM
PRPF31
OPA1
RP2
EFEMP1
PRPF8
KIF11
LRP5*
SNRNP200
PRPF3
FZD4
COL11A1
TOPORS
COL2A1
OPN1LW*
ATXN7*
RIMS1*
JAG1
TEAD1
PITPNM3
OTX2
CCT2
GDF6
CTNNA1
VCAN
MFN2
NR2F1
HK1
PRPF4
AHR
OFD1
PRPS1
ZNF423
RB1
DMD
C3
FBLN5



gnomAD MS>2.99 so o/e<1 so selected against or protected from mutation- 14

PRPF31
PRPF8
KIF11
SNRNP200
PRPF3
KLHL7
PNPLA6
COL2A1*
JAG1*
CTNNA1
NR2F1
HK1
PRPF6
PRPS1



Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (Released 20190711)

Annotation Version and Release Date: GO Ontology database Released 2019-12-09

Analyzed List: Client Text Box Input (Homo sapiens)

Reference List: Homo sapiens (all genes in database)

Annotation Data Set: GO biological process complete

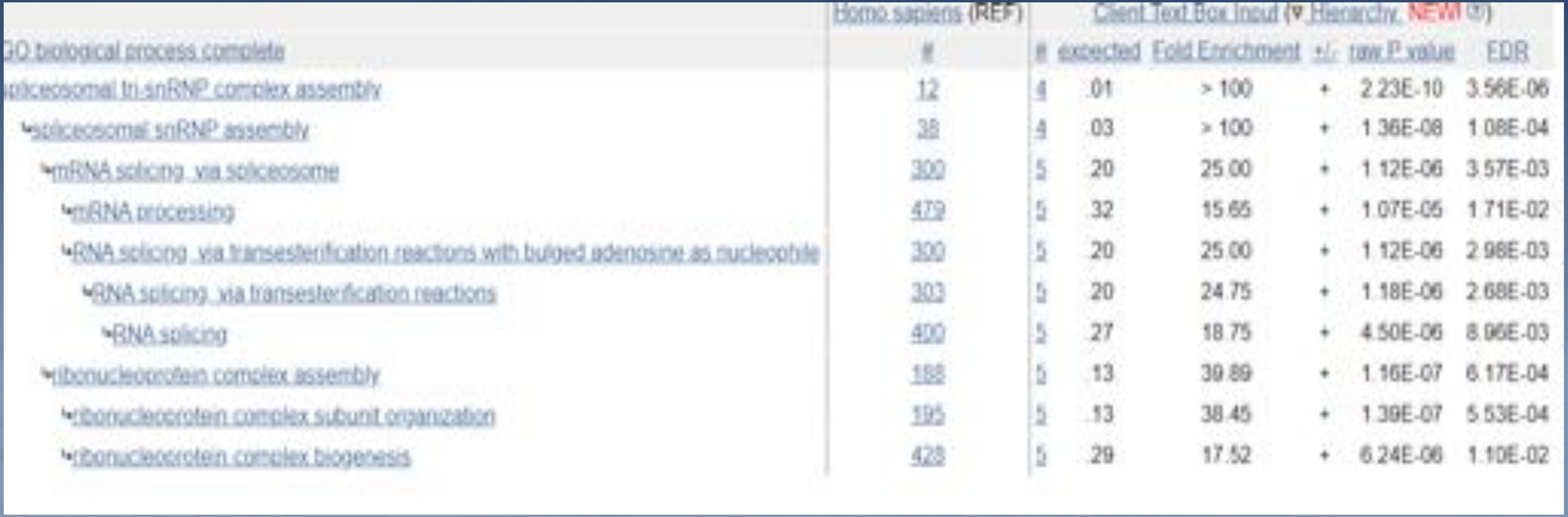
Test Type: ● Fisher's Exact ○ Binomial

Correction: ● Calculate False Discovery Rate ○ Use the Bonferroni correction for multiple testing ○ No correction

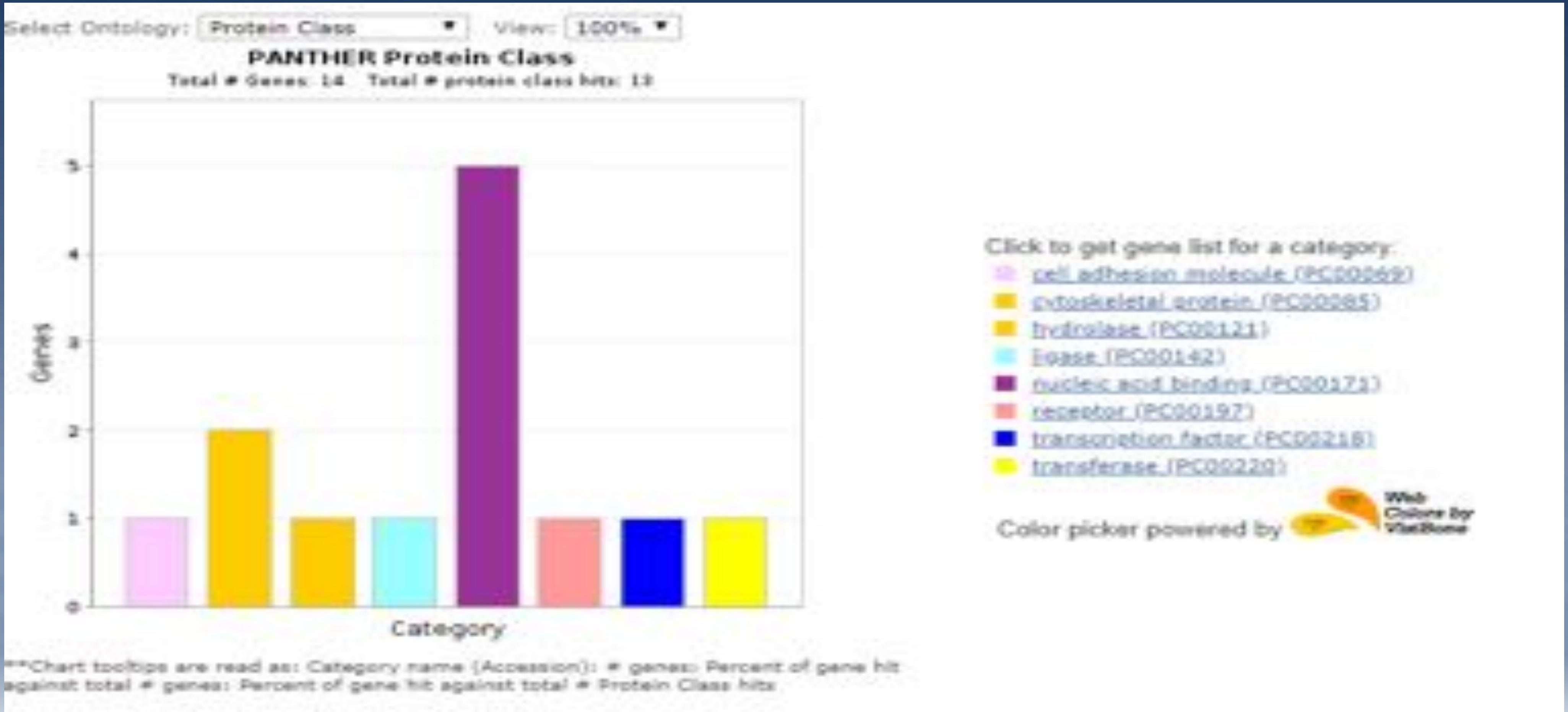
Results

	Reference list	Client Text Box Input
Uniquely Mapped IDS:	20996 out of 20996	14 out of 14
Unmapped IDs:	0	0
Multiple mapping information:	0	0

Spliceosome pathway



	Homo sapiens (REF)	Client Text Box Input (▼ Hierarchy NEW! (?))					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
spliceosomal tri-snRNP complex assembly	12	4	.01	> 100	+	2.23E-10	3.56E-06
↳spliceosomal snRNP assembly	38	4	.03	> 100	+	1.36E-08	1.08E-04
↳mRNA splicing, via spliceosome	300	5	.20	25.00	+	1.12E-06	3.57E-03
↳mRNA processing	479	5	.32	15.65	+	1.07E-05	1.71E-02
↳RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	300	5	.20	25.00	+	1.12E-06	2.98E-03
↳RNA splicing, via transesterification reactions	303	5	.20	24.75	+	1.18E-06	2.68E-03
↳RNA splicing	400	5	.27	18.75	+	4.50E-06	8.96E-03
↳ribonucleoprotein complex assembly	188	5	.13	39.89	+	1.16E-07	6.17E-04
↳ribonucleoprotein complex subunit organization	195	5	.13	38.45	+	1.39E-07	5.53E-04
↳ribonucleoprotein complex biogenesis	428	5	.29	17.52	+	6.24E-06	1.10E-02



Select Ontology: Protein Class ▾ View: 100% ▾

PANTHER Protein Class

Total # Genes: 14 Total # protein class hits: 13

Click to get gene list for a category:

- ■ cell adhesion molecule (PC00069)
- ■ cytoskeletal protein (PC00085)
- ■ hydrolase (PC00121)
- ■ ligase (PC00142)
- ■ nucleic acid binding (PC00171)
- ■ receptor (PC00197)
- ■ transcription factor (PC00218)
- ■ transferase (PC00220)

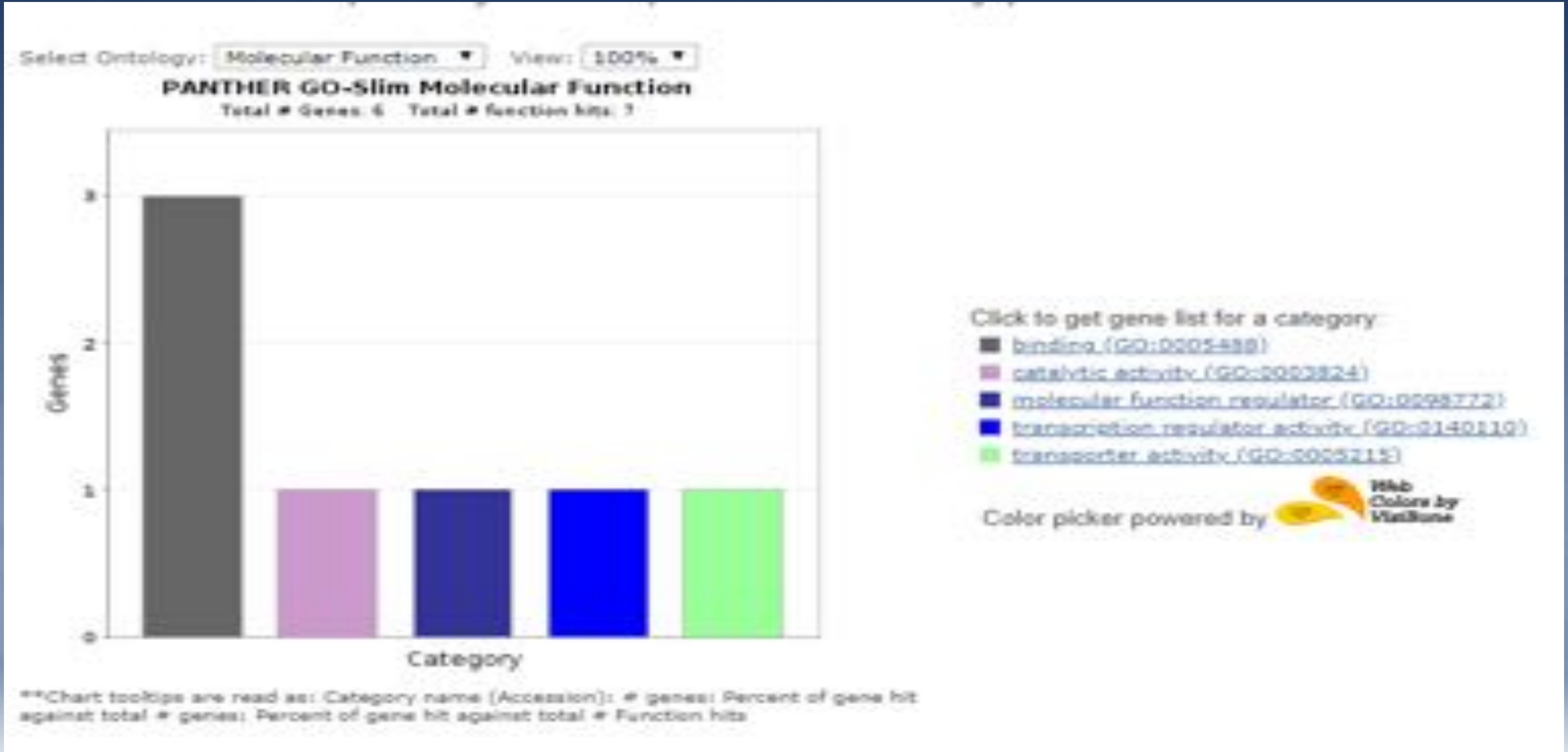
Color picker powered by Web Colors by VisiBone

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Protein Class hits

gnomAD MS Z<-2.99 so o/e >1 so more mutations than chance alone-6

KCNV2
RP1L1*ms*s
ALMS1*
ADAMTS18
WFS1*MS
SAMD11

No enrichment pathway



Select Ontology: Molecular Function ▼ View: 100% ▼

PANTHER GO-Slim Molecular Function

Total # Genes: 6 Total # function hits: 7

Click to get gene list for a category

- ■ binding (GO:0005488)
- ■ catalytic activity (GO:0003824)
- ■ molecular function regulator (GO:0098772)
- ■ transcription regulator activity (GO:0140110)
- ■ transporter activity (GO:0005215)

Color picker powered by Web Colors by VisiBone

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Function hits

Questions

- **What makes these genes statistically very different than other IRD genes?**
- **GO states that spliceosome pathway is over-represented in the loss of function group and in the underrepresented MS group-these are selected against**
- **Norrin signal(l)inggenes are also over-represented in the loss of function group-these are selected against**
- **There are IRD genes that are selectively over or underrepresented-some are basic pathway genes and these are underrepresented-why the other genes are over or under-represented needs to be further evaluated- not purely related to size and X chromosome**
- **Differences with ocular tumor genes, anterior segment morphogenesis, cataract and glaucoma genes are being evaluated**