Missense Variant Pipeline

Presented by Brady Hoskins

Project Motivation

- Motivation: gain a greater understanding of mutations in ASD that will lead to better diagnostics
- What we know now: mutations in ASD probands are more likely to land in protein interaction domains than in their unaffected siblings
- Knowledge gap: the impact of multiple mutations on disrupting protein interactions has not been studied
- Model deployed and advantages: deep learning models allow for the effect of multiple mutations to be modeled in protein interactions
- Hypothesis: multiple mutations affecting protein interaction will be more common in ASD probands versus their unaffected siblings

Missense Variant Pipeline

OLD:

- Step 1: Pass two gene id strings to top layer function
- Step 2: Solicit/clean gene variant information
- Step 3: Map related proteins
- Step 4: Map protein sequences
- Step 5: Apply variant changes to sequences
- Step 6: Determine proband and sibling variants
- Step 7: Return new data frame for modeling

NEW:

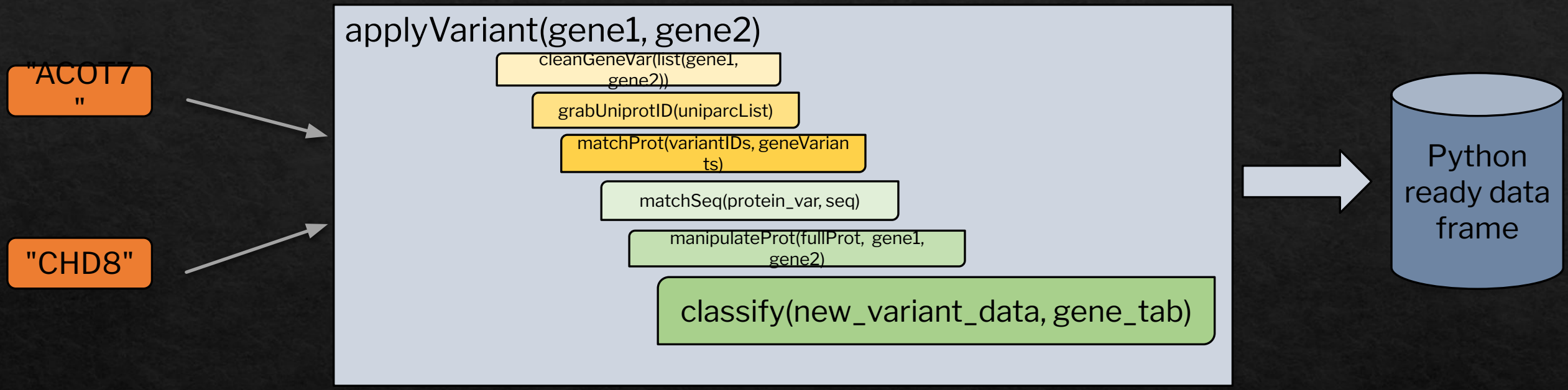
Step 1: Pass two gene id strings to top layer function

- Step 1a: Solicit/clean gene variant information
- Step 1b: Map related proteins
- Step 1c: Map protein sequences
- Step 1d: Apply variant changes to sequences (Keep Sequence string)
- Step 1e: Determine proband and sibling variants by number of occurrences
- Step 1f: Return new data frame for modeling/encoding

Step 2: Read file into python function

- Step 2a: Python function encodes and models data, returns Wilcoxon statistics

Last Time: Pipeline Architecture



Old Output:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Protein	Sequence	uniparc	symbol	gene	feature	biotype	hgvsp	protein_pos	seq_length	Old_aa	New_aa	Old_codon	New_codon	ensp	used_ref	variant_id	seq_matrix	variantSeq_matrix	source
2	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Asn227Ser	227	246	N	S	aAc	aGc	ENSP00000402532	T	36496	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	both
3	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Lys225Asn	225	246	K	N	aaG	aaC	ENSP00000402532	C	36497	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	both
4	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Gly217Arg	217	246	G	R	Ggg	Agg	ENSP00000402532	C	36498	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	proband
5	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Met212Leu	212	246	M	L	Atg	Ctg	ENSP00000402532	T	36501	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	both
6	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Gly205Arg	205	246	G	R	Gga	Aga	ENSP00000402532	C	36610	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	proband
7	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Pro172Leu	172	246	P	L	cCa	cTa	ENSP00000402532	G	36627	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	proband
8	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Asp168His	168	246	D	H	Gac	Cac	ENSP00000402532	C	36628	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	both
9	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Gly167Glu	167	246	G	E	gGg	gAg	ENSP00000402532	C	36629	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	proband
10	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Tyr152Ter	152	246	Y	*	taT	taG	ENSP00000402532	A	36630	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	sib
11	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Arg149Gln	149	246	R	Q	cGg	cAg	ENSP00000402532	C	36633	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	sib
12	O00154	MKLLARALRLCEFG	UPI000002A4AF	ACOT7	ENSG00000097021	ENST00000418124	nonsense_mediat	ENSP00000402532.1:p.Glu146Asp	146	246	E	D	gaG	gaC	ENSP00000402532	C	36634	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	proband

Updated: Pipeline Architecture

"ACOT7"

"CHD8"

applyVariant(gene1, gene2)

cleanGeneVar(list(gene1, gene2))

grabUniprotID(uniparcList)

matchProt(variantIDs, geneVariants)

matchSeq(protein_var, seq)

Code to apply variant & count number of variants occurrences

Code to create row per number of variant occurences, return file

ppiPipeline(datafile)

Encode amino acids

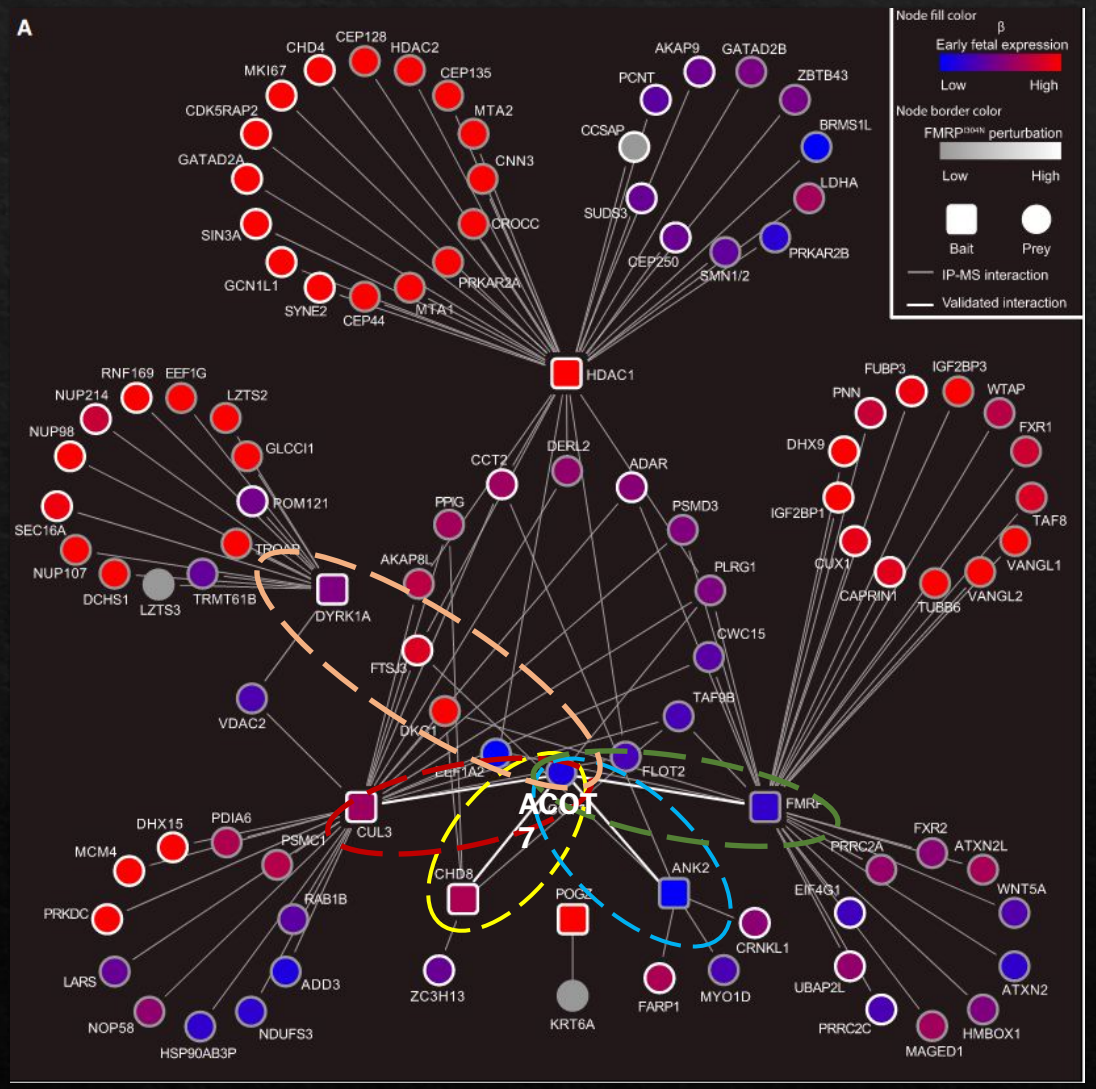
Run data through RNN model

Return Wilcoxon test statistic

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Protein	Sequence	uniparc	symbol	gene	feature	biotype	hgvsp	protein_pos	seq_length	Old_aa	New_aa	Old_codon	New_codon	ensp	used_ref	variant_id	seq_matrix	variantSeq_n	Proband	Sibling
2	O00154	MKLLARALR	UPI000002A	ACOT7	ENSG00000	ENST00000E	protein_codi	ENSP00000	337	338	Q	K	Cag	Aag	ENSP00000	G	36411	MKLLARALR	MKLLARALR	1	0
3	O00154	MKLLARALR	UPI000002A	ACOT7	ENSG00000	ENST00000E	protein_codi	ENSP00000	334	338	A	V	gCg	gTg	ENSP00000	G	36413	MKLLARALR	MKLLARALR	1	0
4	O00154	MKLLARALR	UPI000002A	ACOT7	ENSG00000	ENST00000E	protein_codi	ENSP00000	334	338	A	V	gCg	gTg	ENSP00000	G	36413	MKLLARALR	MKLLARALR	1	0
5	O00154	MKLLARALR	UPI000002A	ACOT7	ENSG00000	ENST00000E	protein_codi	ENSP00000	331	338	Q	R	cAg	cGg	ENSP00000	T	36415	MKLLARALR	MKLLARALR	1	0
6	O00154	MKLLARALR	UPI000002A	ACOT7	ENSG00000	ENST00000E	protein_codi	ENSP00000	331	338	Q	R	cAg	cGg	ENSP00000	T	36415	MKLLARALR	MKLLARALR	1	0

Pipeline Results

"ACOT7" + "CHD8"	applyVariant(gene1, gene2) + ppiPipeline(proteinVariants)	Wilcoxon_proband_vs_sibling = -0.892, p = 0.373	Proband: 0.96758 Sibling: 0.97011
"ACOT7" + "CUL3"	applyVariant(gene1, gene2) + ppiPipeline(proteinVariants)	Wilcoxon_proband_vs_sibling = 1.462, p = 0.144	Proband: 0.15548 Sibling: 0.15342
"ACOT7" + "ANK7"	applyVariant(gene1, gene2) + ppiPipeline(proteinVariants)	Wilcoxon_proband_vs_sibling = -0.522, p = 0.601	Proband: 0.57002 Sibling: 0.57256
"ACOT7" + "FMR1"	applyVariant(gene1, gene2) + ppiPipeline(proteinVariants)	Wilcoxon_proband_vs_sibling = -0.628, p = 0.530	Proband: 0.57010 Sibling: 0.57696
"ACOT7" + "DYRK1A"	applyVariant(gene1, gene2) + ppiPipeline(proteinVariants)	Wilcoxon_proband_vs_sibling = -1.894, p = 0.058	Proband: 0.94380 Sibling: 0.95009



Identification of Human Neuronal Protein Complexes Reveals Biochemical Activities and Convergent Mechanisms of Action in Autism Spectrum Disorders. Li, et al. *Cell Systems*, 2015.