

INTRODUCTION

The development of a number of functional impact predictors and annotation tools has been implemented to aid in DNA variant analysis. Many conventional prediction tools, such as SIFT [3] and PolyPhen2 [1], are built for nuclear DNA, and fail to produce accurate predictions for mitochondrial mutations. Newer tools such as APOGEE [2] have addressed the need to annotate mtDNA exonic variants with highconfidence, but are limited when it comes to variants in mitochondrial tRNA and rRNA, which are both common causes of mitochondrial disease. A few papers, such as MitoTIP [4], address the need for tools that predict the pathogenicity of tRNA variants, while no known tools exist for annotating rRNA variants. Our purpose for this experiment is to provide an extensive review of available variant annotation tools that can be used to accurately predict the functional consequence of mtDNA variants.

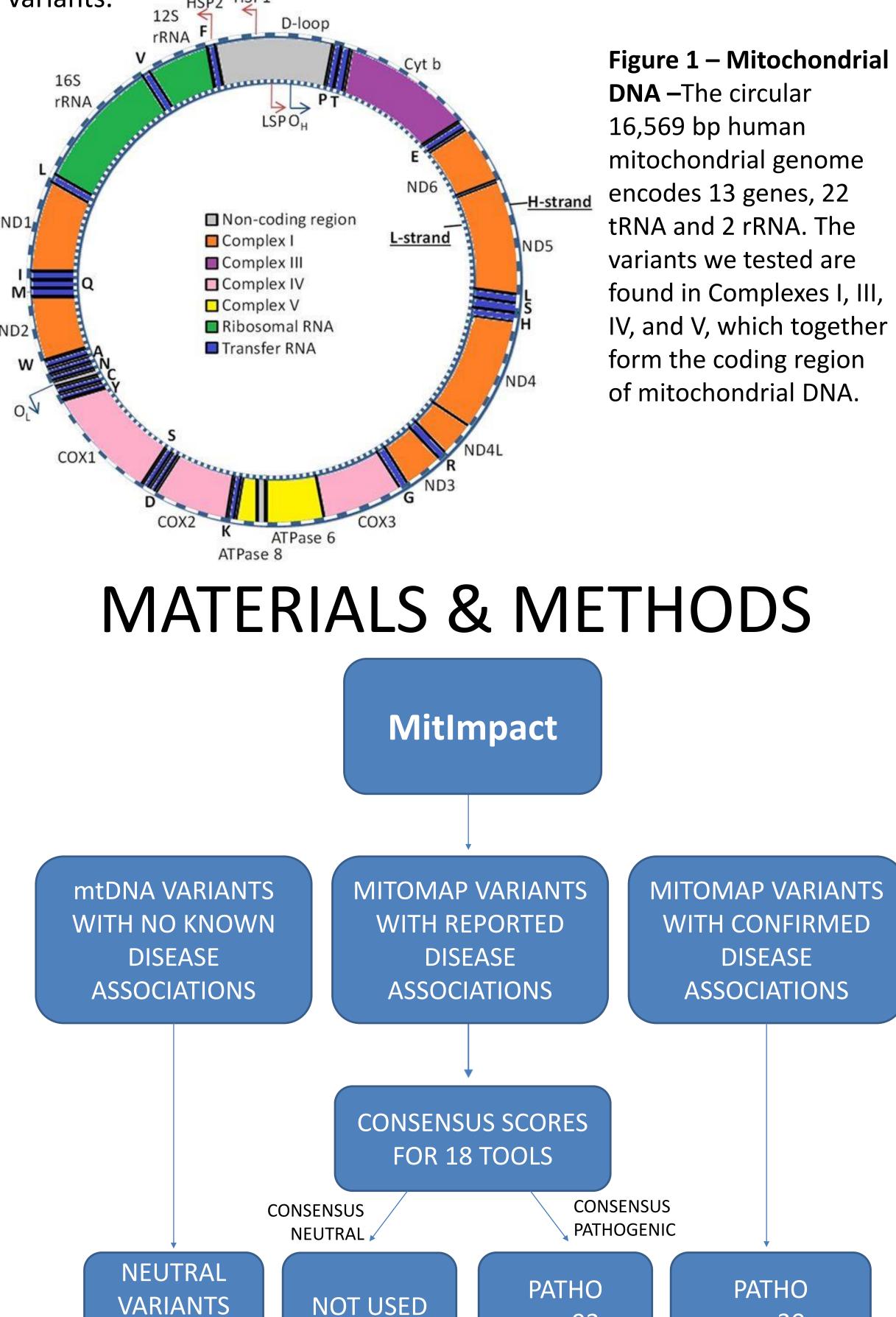


Figure 2 – Curated Test Set of mtDNA variants – Mitochondrial variants with reported, confirmed, and no known disease associations were acquired from MitImpact 2.9[5]. The pathogenic consensus scores from the 18 annotation tools were crosschecked with the reported disease associated variants and combined with the confirmed disease associations and neutral variants with no disease associations. We analyzed the performance of each tool based off of the resulting true or false predictions using a test set of size 1216 mtDNA variants.

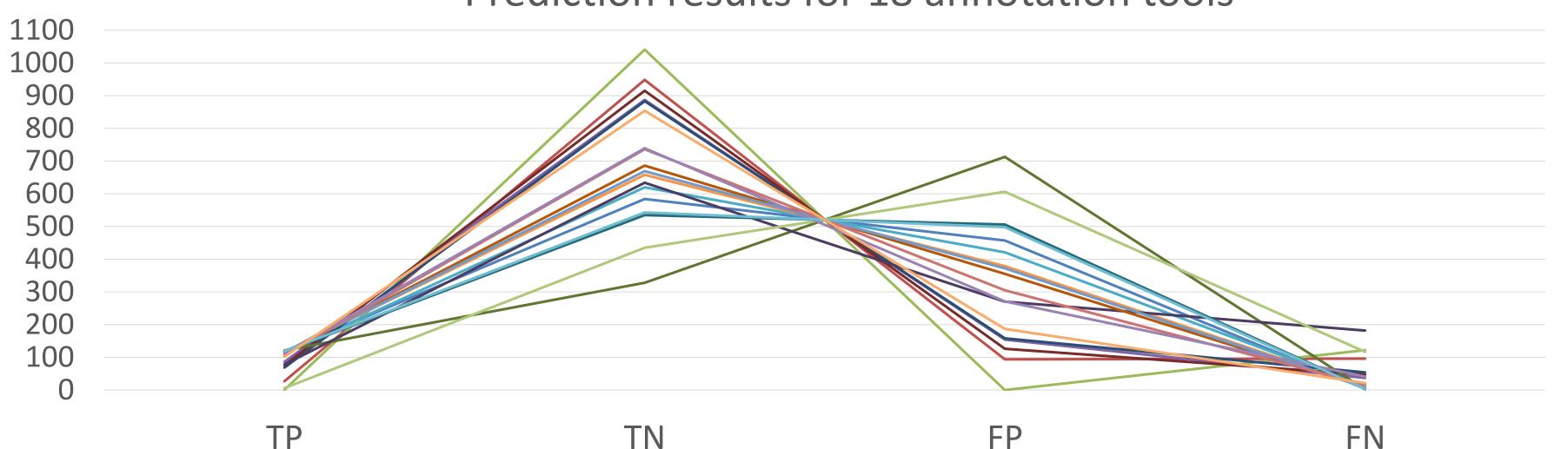
n = 1093

n = 93

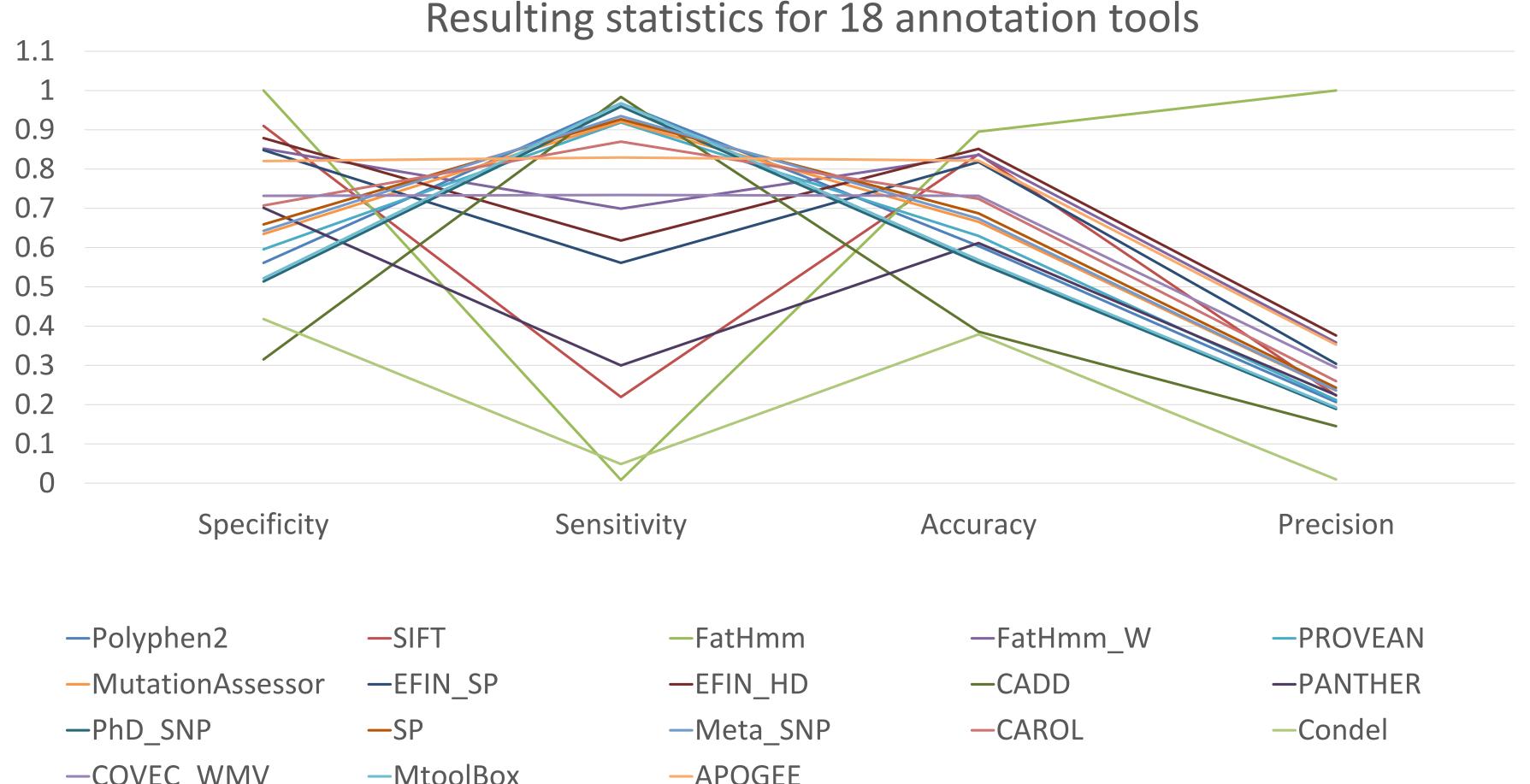
Comparative Analysis of Tools for Predicting the Functional Impact of mtDNA Variants Madeline P. Griffin and Catherine E. Welsh

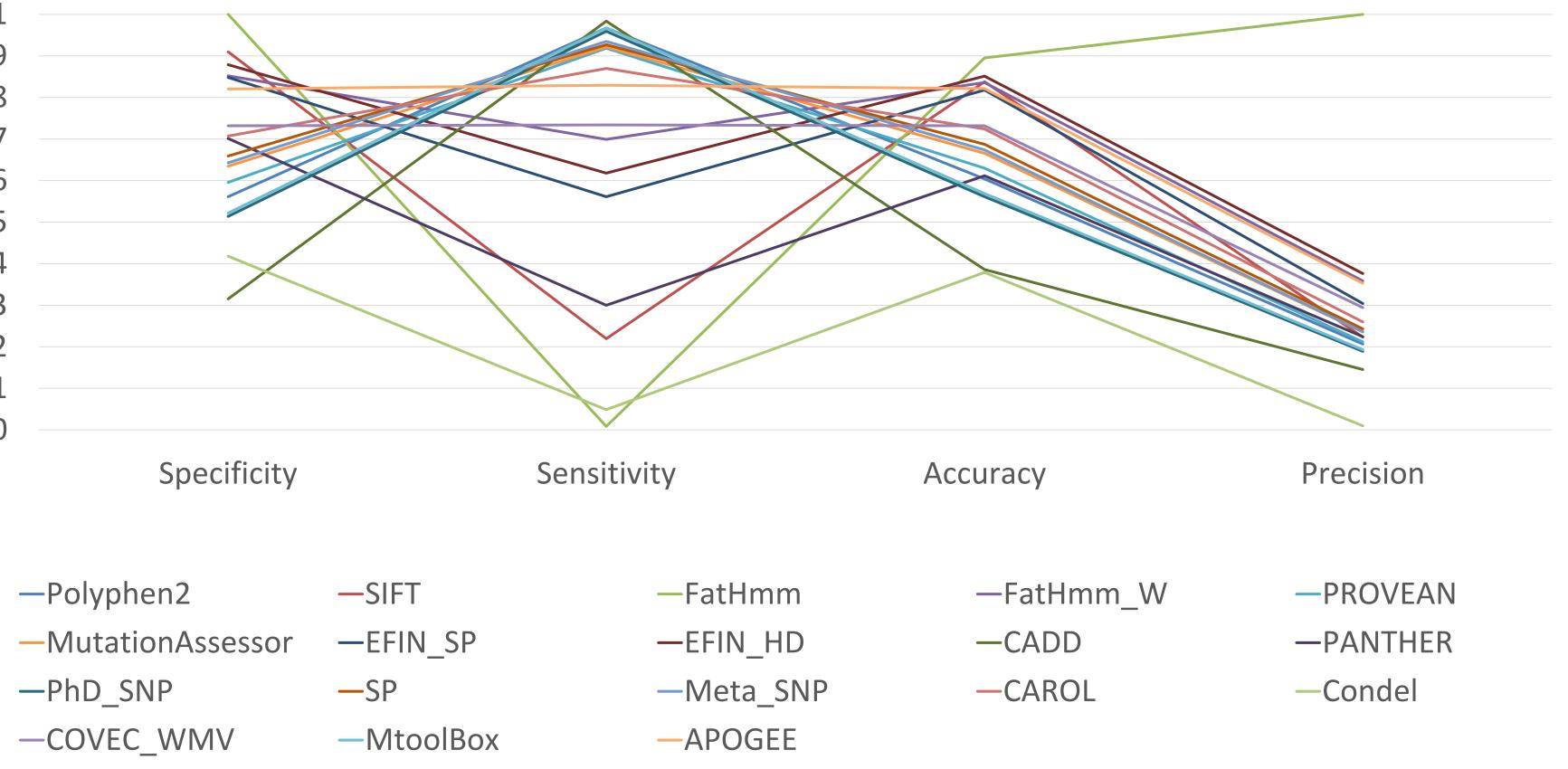
S. F. 1. SIFT 27 948 94 96 0.909789 0.219512 0.837629 0.22314 0.77686 0.045881 0.000841 0 I. 1. FatHmm 1 0 122 1 0.00813 0.895189 1 0 0.002643 0 FatHmm_W 86 887 154 37 0.852065 0.699187 0.835911 0.358333 0.641667 0.208035 0.003576 PROVEAN 113 620 421 10 0.595581 0.918699 0.629725 0.21161 0.78839 0.270804 0.036168 MutationAssessor 117 658 379 10 0.634523 0.92126 0.665808 0.235887 0.764113 0.280322 0.03256 EFIN_SP 69 883 158 54 0.848223 0.560976 0.817869 0.303965 0.696035 0.155033 0.002514 EFIN_HD 76 915 126 47 0.878963 0.617866 0.851375 0.376238 0.623762 0.182086 0.002303 0.02174 0.306271 </th <th></th> <th>MCR</th> <th>MCC</th> <th></th> <th></th> <th>-</th> <th>-</th> <th>• •</th> <th>FN</th> <th>FP</th> <th>TN</th> <th>TP</th> <th>Tools</th> <th></th>		MCR	MCC			-	-	• •	FN	FP	TN	TP	Tools	
1), FatHmm 1 1 1 1 1 1 0 122 1 0.00813 0.895189 1 0 0.002643 0 FatHmm_W 86 87 154 37 0.852065 0.699187 0.835911 0.358333 0.641667 0.208035 0.003576 PROVEAN 113 620 421 10 0.595581 0.91869 0.629725 0.21161 0.78839 0.270804 0.036168 MutationAssessor 117 658 379 10 0.634523 0.92126 0.665808 0.235887 0.764113 0.280322 0.03256 EFIN_SP 69 883 158 54 0.848223 0.560976 0.817869 0.303965 0.696035 0.155033 0.002514 EFIN_HD 76 915 126 47 0.878963 0.617886 0.851375 0.376238 0.623762 0.182086 0.002303 CADD 121 328 713 2 0.315082 0.98374 0.385739 0.145084 0.854916 0.277786 0.306271	The 18	0.098153	0.29413	0.793403	0.206597	0.603952	0.96748	0.560999	4	457	584	119	PolyPhen2	S. 1] <i>,</i>
FatHmm_W 86 87 154 37 0.852065 0.699187 0.835911 0.358333 0.641667 0.208035 0.003576 PROVEAN 113 620 421 10 0.595581 0.918699 0.629725 0.21161 0.78839 0.270804 0.036168 MutationAssessor 117 658 379 10 0.634523 0.92126 0.665808 0.235887 0.764113 0.280322 0.03256 EFIN_SP 69 883 158 54 0.81508 0.617866 0.817869 0.303965 0.696035 0.155033 0.002514 EFIN_SP 69 883 158 54 0.818963 0.617866 0.851375 0.376238 0.623762 0.182086 0.002303 CADD 121 328 713 2 0.315082 0.98374 0.385739 0.145084 0.854916 0.277786 0.306271 PANTHER 78 634 270 182 0.701327 0.3 0.61684	of accu Fathand CAD the Fathand Mise Ann Fath	0.000841	0.045881	0.77686	0.22314	0.837629	0.219512	0.909789	96	94	948	27	SIFT	
PROVEAN 113 620 421 10 0.595581 0.918699 0.629725 0.21161 0.78839 0.270804 0.036168 MutationAssessor 117 658 379 10 0.634523 0.92126 0.665808 0.235887 0.764113 0.280322 0.03256 EFIN_SP 69 883 158 54 0.848223 0.560976 0.817869 0.303965 0.696035 0.155033 0.002514 EFIN_HD 76 915 126 47 0.878963 0.617886 0.851375 0.376238 0.623762 0.182086 0.002303 CADD 121 328 713 2 0.315082 0.98374 0.385739 0.145084 0.854916 0.277786 0.306271 PANTHER 78 634 270 182 0.701327 0.3 0.611684 0.224138 0.775862 0.00844 0.001274 NA SP 114 686 355 9 0.658982 0.926829 <th< th=""><td>0</td><td>0.002643</td><td>0</td><td>1</td><td>0.895189</td><td>0.00813</td><td>1</td><td>122</td><td>0</td><td>1041</td><td>1</td><td>FatHmm</td></th<>		0	0.002643	0	1	0.895189	0.00813	1	122	0	1041	1	FatHmm	
MutationAssessor117658379100.6345230.921260.6658080.2358870.7641130.2803220.03256EFIN_SP69883158540.8482230.5609760.8178690.3039650.6960350.1550330.002514EFIN_HD76915126470.8789630.6178860.8513750.3762380.6237620.1820860.002303CADD12132871320.3150820.983740.3857390.1450840.8549160.2777860.306271PANTHER786342701820.7013270.30.6116840.2241380.8108970.2851950.086942NASP11853550650.5139290.959350.5609970.1891030.8108970.2851950.086942NASP11853550650.5139290.9268290.6872850.243070.756930.2811240.033887NASP11468635590.6258920.673540.243070.756930.2811240.033887MetaSNP11566937280.6426510.9349590.673540.236140.763860.2837010.039948ACAROL107736305160.7070120.8699190.7242270.2597090.7402910.2592860.016377ACAROL107736601170.4178670.048780.3788660.009804		0.003576	0.208035	0.641667	0.358333	0.835911	0.699187	0.852065	37	154	887	86	FatHmm_W	
eFin_sp 69 883 158 54 0.848223 0.560976 0.817869 0.303965 0.696035 0.155033 0.002514 eFin_HD 76 915 126 47 0.878963 0.617886 0.851375 0.376238 0.623762 0.182086 0.002303 CADD 121 328 713 2 0.315082 0.98374 0.385739 0.145084 0.854916 0.277786 0.306271 PANTHER 78 634 270 182 0.701327 0.3 0.611684 0.224138 0.775862 0.000844 0.001274 PhD-SNP 118 535 506 5 0.513929 0.95935 0.560997 0.189103 0.810897 0.285195 0.086942 NA SP 114 686 355 9 0.658982 0.926829 0.687285 0.24307 0.75693 0.281124 0.033887 MA SP 114 686 355 9 0.642651 0.934959 0.67354 0.23614 0.76386 0.283701 0.039948 ial		0.036168	0.270804	0.78839	0.21161	0.629725	0.918699	0.595581	10	421	620	113	PROVEAN	
EFIN_HD 76 915 126 47 0.878963 0.617886 0.851375 0.376238 0.623762 0.182086 0.002303 CADD 121 328 713 2 0.315082 0.98374 0.385739 0.145084 0.854916 0.277786 0.306271 to PANTHER 78 634 270 182 0.701327 0.3 0.611684 0.224138 0.775862 0.000844 0.001274 PhD-SNP 118 535 506 5 0.513929 0.95935 0.560997 0.189103 0.810897 0.285195 0.0086942 NA SP 114 686 355 9 0.658982 0.926829 0.687285 0.24307 0.75693 0.281124 0.033887 MataSNP 115 669 372 8 0.642651 0.934959 0.67354 0.23614 0.76386 0.283701 0.039948 MataSNP 115 669 372 8 0.642651 0.934959		0.03256	0.280322	0.764113	0.235887	0.665808	0.92126	0.634523	10	379	658	117	MutationAssessor	~1
CADD 121 328 713 2 0.315082 0.98374 0.385739 0.145084 0.854916 0.277786 0.306271 PANTHER 78 634 270 182 0.701327 0.3 0.611684 0.224138 0.854916 0.277786 0.306271 PANTHER 78 634 270 182 0.701327 0.3 0.611684 0.224138 0.854916 0.277786 0.306271 PANTHER 78 634 270 182 0.701327 0.3 0.611684 0.224138 0.854916 0.277786 0.306271 PANTHER 78 634 270 182 0.701327 0.3 0.560997 0.189103 0.810897 0.285195 0.086942 NA SP 114 686 355 9 0.658982 0.926829 0.687285 0.24307 0.75693 0.281124 0.033887 MetaSNP 115 669 372 8 0.642651 0.934959 0.724227 0.259709 0.740291 0.259286 0.016377 ial CAROL		0.002514	0.155033	0.696035	0.303965	0.817869	0.560976	0.848223	54	158	883	69	EFIN_SP	11
A to PANTHER 78 634 270 182 0.701327 0.3 0.611684 0.224138 0.775862 0.000844 0.001274 PhD-SNP 118 535 506 5 0.513929 0.95935 0.560997 0.189103 0.810897 0.285195 0.086942 NA SP 114 686 355 9 0.658982 0.926829 0.687285 0.24307 0.75693 0.281124 0.033887 MetaSNP 115 669 372 8 0.642651 0.934959 0.67354 0.23614 0.76386 0.283701 0.039948 ial CAROL 107 736 305 16 0.707012 0.869919 0.724227 0.259709 0.740291 0.259286 0.016377 ial Condel 6 435 606 117 0.417867 0.04878 0.378866 0.009804 0.990196 -0.38681 0.00445		0.002303	0.182086	0.623762	0.376238	0.851375	0.617886	0.878963	47	126	915	76	EFIN_HD	
NA PhD-SNP 118 535 506 5 0.513929 0.95935 0.560997 0.189103 0.810897 0.285195 0.086942 NA SP 114 686 355 9 0.658982 0.926829 0.687285 0.24307 0.75693 0.281124 0.033887 MetaSNP 115 669 372 8 0.642651 0.934959 0.67354 0.23614 0.76386 0.283701 0.039948 ial CAROL 107 736 305 16 0.707012 0.869919 0.724227 0.259709 0.740291 0.259286 0.00445		0.306271	0.277786	0.854916	0.145084	0.385739	0.98374	0.315082	2	713	328	121	CADD	, ,)
NA SP 114 686 355 9 0.658982 0.926829 0.687285 0.24307 0.75693 0.281124 0.033887 MetaSNP 115 669 372 8 0.642651 0.934959 0.67354 0.23614 0.76386 0.283701 0.039948 ial CAROL 107 736 305 16 0.707012 0.869919 0.724227 0.259709 0.740291 0.259286 0.016377 condel 6 435 606 117 0.417867 0.04878 0.378866 0.009804 0.990196 -0.38681 0.00445		0.001274	0.000844	0.775862	0.224138	0.611684	0.3	0.701327	182	270	634	78	PANTHER	to
MetaSNP 115 669 372 8 0.642651 0.934959 0.67354 0.23614 0.76386 0.283701 0.039948 ial CAROL 107 736 305 16 0.707012 0.869919 0.724227 0.259709 0.740291 0.259286 0.016377 ial Condel 6 435 606 117 0.417867 0.04878 0.378866 0.009804 0.990196 -0.38681 0.00445		0.086942	0.285195	0.810897	0.189103	0.560997	0.95935	0.513929	5	506	535	118	PhD-SNP	
CAROL 107 736 305 16 0.707012 0.869919 0.724227 0.259709 0.740291 0.259286 0.016377 Condel 6 435 606 117 0.417867 0.04878 0.378866 0.009804 0.990196 -0.38681 0.00445		0.033887	0.281124	0.75693	0.24307	0.687285	0.926829	0.658982	9	355	686	114	SP	NA
ial Condel 6 435 606 117 0.417867 0.04878 0.378866 0.009804 0.990196 -0.38681 0.00445		0.039948	0.283701	0.76386	0.23614	0.67354	0.934959	0.642651	8	372	669	115	MetaSNP	
Condel 6 435 606 117 0.417867 0.04878 0.378866 0.009804 0.990196 -0.38681 0.00445		0.016377	0.259286	0.740291	0.259709	0.724227	0.869919	0.707012	16	305	736	107	CAROL	ial
COVEC WMV 113 739 271 41 0.731683 0.733766 0.731959 0.294271 0.705729 0.225178 0.005678		0.00445	-0.38681	0.990196	0.009804	0.378866	0.04878	0.417867	117	606	435	6	Condel	
	Our re relies m genome PROVEA	0.005678	0.225178	0.705729	0.294271	0.731959	0.733766	0.731683	41	271	739	113	COVEC WMV	
$\mathbf{W} = \mathbf{W} = $		0.106959	0.290835	0.807131	0.192869	0.568729	0.96748	0.521614	4	498	543	119	MToolBox)
$\Delta POGFF Bootstran (0.7, 854, 187, 71, 0.870365, 0.879768, 0.871306, 0.357941, 0.647059, 0.754164, 0.00765$		0.00765	0.254164	0.647059	0.352941	0.821306	0.829268	0.820365	21	187	854	102	APOGEE Bootstrap	

Table 1 – Computed statistics for 18 annotation tools – We ran our test set against the available annotations of 18 software predictors, and determined the resulting number of true positive, true negative, false positive, and false negative predictions. We then calculated the specificity, sensitivity, accuracy, and precision of those predictions, as well as the False Discovery Rate, the Misclassification Rate, and Matthew's Correlation Coefficient.









n = 30

Prediction results for 18 annotation tools

- tHmm scored the highest number of true negatives, nd the least amount of false negatives.
- ADD scored the highest number of true positives and e least amount of false negatives.
- tHmm had the highest specificity, accuracy, precision, nd the lowest False Discovery Rate as well as the lowest isclassification Rate.
- nnotation tools that with best performance were tHmm, FatHmm_W, EFIN_SP, EFIN_HD, and APOGEE.

results contradict the current standard for variant annotation that mostly on software predictors built specifically for the nuclear ne. Those particular annotators such as SIFT, PolyPhen, PANTHER, and PROVEAN were retrofitted for mitochondrial DNA, and, based on the results of both our tests and the experiments of others, fall short in a number of statistically significant areas. The APOGEE Bootstrap model also presents convincing results to support the idea that poor scoring annotation tools, such as PolyPhen2, are not capable of accurately predicting the pathogenicity of mitochondrial variants [2]. A number of annotations tools base their predictions on the results of other scoring predictors. For example, CAROL annotation software provides results based almost entirely off the combined predictions of SIFT and Polyphen2. FatHmm, EFIN, and APOGEE algorithms ultimately outperformed conventional annotation tools, and we recommend these tools for mitochondrial annotations and variant prediction.

With these results we intend to build our own annotation tool for mitochondrial DNA specifically, that provides pathogenicity predictions from the tools we have deemed best fit to provide accurate results. We would like to create a software predictor that addresses the prediction of tRNA and rRNA variants, as there are still a limited number of prediction tools that include tRNA variants, and almost none that include rRNA variants.

[1] IA Adzhubei, S Schmidt, L Peshkin, VE Ramensky, A Gerasimova, P Bork, AS Kondrashov, and SR Sunyaev. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* (April 2010). http://dx.doi.org/10.1038/ nmeth0410-248 [2] S Castellana, C Fusilli, G Mazzoccoli, T Biagini, D Capocefalo, M Carella, AL Vescovi, and T Mazza. 2017. High-confidence assessment of functional impact of human mitochondrial non-synonymous genome variations by APOGEE. PLoS Computational Biology (June 2017). http://dx.doi.org/10.1371/journal.pcbi.1005628 [3] P Kumar, S Henikoff, and PC Ng. 2009. Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. *Nature Protocols* (June 2009). <u>http://dx.doi.org/10.1038/nprot.2009.86</u> [4] S Sonney, J Leipzig, MT Lott, S Zhang, V Procaccio, DC Wallace, and N Sondheimer. 2017. Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. *PLoS Computational Biology* (Dec. 2017).<u>http://dx.doi.org/10.1371/ jounral.pcbi.1005867</u> [5] <u>http://mitimpact.css-mendel.it/</u>

RESULTS

L8 annotation tools tested presented varying degrees curacy, precision, sensitivity, and specificity.

DISCUSSION

REFERENCES