iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37045

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Bioinformatics methods for identifying Hirschsprung disease genes

Sudheer Menon[1], Vincent Chi Hang Lui[2], Paul Kwong Hang Tam[3]
[1, 2, 3]*Department of Surgery, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Hong Kong.*
[2, 3]*Dr. Li Dak-Sum Research Centre, The University of Hong Kong – Karolinska Institutet Collaboration in Regenerative Medicine, the University of Hong Kong, Hong Kong.*

Abstract: Hirschsprung is a birth defect of Enteric Nervous System (ENS) which is characterized by the absence of enteric neurons along the length of intestine. Hirschsprung is one of the complex diseases which has become a important topic of human genetics. In this article we have focused on RET gene mutation that is most common cause of HSCR disease. Out of seven mutations in RET gene, one mutation S339L is found to be tolerated and have no effect on protein function.

I. INTRODUCTION

Hirschsprung disease (HSCR) is a congenital disorder which is caused due to the lack of enteric neurons down the length of intestine. The occurrence of this disease is different in different ethnic groups averaging almost 1 in 5000 live births (Parisi & Kapur, 2000). HSCR can be categorized into three different types on the basis of length segment which lacks ganglion. It can be total colonic aganglionic (TCA), long segment HSCR and short segment HSCR (Lantieri et al., 2006). Genetically HSCR is quite similar to the other complex diseases with almost 80% sporadic and 20% familial cases. Moreover, in 30% it is found associated to the different congenital disorders, but it is diagnosed as syndromic in very few of these cases (Amiel et al., 2008; Moore, 2006). The mode of inheritance in sporadic cases is believed to be non-Mendelian while the familial cases showed the Mendelian mode of inheritance (Badner et al., 1990). From past few years much attention is being paid to find the genes responsible for HSCR in multi-generational HSCR families. Many studies have reported the genetic mutation involved in the pathgenesis of HSCR i:e *PLD1, BACE2, RET, EDNRB, SOX10, KIAA1279* (Lecerf et al., 2014; Sanchez-Mejias et al., 2010; Tomuschat & Puri, 2015).

The process of identification of gene involved in complex disease is known as gene finding. It is very important process of analyzing the genome of an organism. In past process of gene finding use to be costly and time taking as well as it required a lot of in vivo experimentation. But due to the recent advancement in bioinformatics and statistical tools, gene finding has become easier. The use of many different bioinformatics tools such as M-CAP, SIFT-indel, Polyphen, SIFT and MutationTaster have been reported to identify the variants causing HSCR(I. Adzhubei et al., 2013; Hu & Ng, 2013; Jagadeesh et al., 2016; Ng & Henikoff, 2003; Schwarz et al., 2010; Zhao et al., 2013). In a study conducted by Wu et al seven different genes (*RET, CALN1, FANC1, NPHP3, DRD5, GLYCTK, CAPN9*) have been declared pathogenic by using bioinformatics approaches.

II. OBJECTIVES

The objectives of this study are to:

A. Identification and naming of mutations
B. Verification of mutations via *in silico* analysis
C. Determining the effect of mutation in pathogenicity
D. Enlist the bioinformatics approaches used to identify pathogenic variants

III. MATERIAL AND METHODS

A. BLASTx (for Mutation identification)

Poluphe-2

1) *PolyPhen-2* is a bioinformatics tool that can automatically anticipate probable effect of amino acid substitution on protein's function as well as its structure. The prediction relies on different features consisting of the structural information, sequence and phylogenetic.

For particular amino acid substitution in a human protein, this tool pulls out different sequences and attributes of site of substitution based on structure and then supply them to to the probabilistic classifier.

2) *Polyphen-2* is the advancement of the older **Polyphen** tool which can annotate nonsynonymous SNPs. Some important features of this new edition are:
a) Machine learning method that is base for probabilistic classifier
b) High throughput analysis
c) High quality alignment pipeline for multiple sequence(I. Adzhubei et al., 2013; I. A. Adzhubei et al., 2010)

B. Provean

Provean is used for the prediction of impact of sequence variations on protein function, including small indels and substitution of single amino acid. The basis of the prediction is the change in the similarity of query sequence resulted by the given variation to set of sequences of its related proteins. The principle for this prediction is to figure out a pairwise sequence alignment score among each of the related sequences and the query sequence. The development of Provean was done to predict whether a variation in protein sequence affects its function.
This tool is can predict all kinds of variations in protein sequence which includes amino acid substitution, insertions and deletions either single or multiple.
For all probable amino acid substitution, insertions and maximum of 10 amino acid deletion, a database has been generated by the use of algorithm which contains already computed prediction scores. (Choi, 2012; Choi et al., 2012)

C. MUPro

MUpro is another bioinformatics tool which comprises of machine learning programs that predict the affects of single site amino acid variation on stability of protein. Neural Networks and support vector machines are the machine learning techniques that have been developed. Each of these methods contains a huge mutation database. These methods show 84% accuracy passing through cross validation of 20 folds and are better than various other methods found in the literature. These methods have a major advantage over other methods because they do not require tertiary structures of protein for the prediction of its stability. This can be justified with the experimental results which showed that the accuracy of prediction by the use of sequence analysis alone is as good as using tertiary structures. Hence with the use of this server one can predict accurately even without the availability of tertiary structures. But, if provided with tertiary structures these methods will provide even better results of prediction. A mutation is said to be decreasing stability of protein when the score is less than 0 and a protein stability increases with score bigger than zero. The more the score is smaller or bigger than 0 the more confident prediction is (Cheng et al., 2006).

D. I-Mutation

I-Mutation is based on a set of support vector machines which predicts the affects single point mutation. It is incorporated in a distinct web server. It can automatically predict the change in protein stability due to the point mutation by using only protein sequence or its structure if accessible. In addition it can also predict the human deleterious SNPs. One can choose among the three predictors:
First one is an SVM based predictor starts from structural information and predict changes in stability of protein by single-site mutation. Second one is the I-Mutant-DDG, a SVM based predictors which use sequence information to predict the changes brought by point mutation in stability of protein. Third is the I-Mutant-Disease which is based on SVM and makes use of sequence information to predict human SNPs.(E Capriotti et al., 2006; Emidio Capriotti et al., 2005)

E. PhD.SNP

PhD.SNP is an SVM-based classifier which is used to predict human Deleterious SNPs. In its new version a single SVM driven predictor is developed that test profile information and protein sequence. The input of PHD-SNP SVM is formed by following steps: The substitution from the wild type to the mutant residue for a given mutation is encoded in 20 elements vector which contain at 18 positions, 1 in the position relative to mutant residues and -1 at position relative to mutant residues. Another vector is formed which consist of 20 elements that encodes for the sequence environment. This second vector reports the frequency residues in a window of 19 residues around variant residue. From the above discuss procedure both the occurrence of the mutated (Fi(MUT)) and wild type residue (Fi(WT)) at location (E Capriotti et al., 2006).

F. SIFT

Sorting intolerant from tolerant (SIFT) is used to anticipate whether an amino acid substitution is deleterious and by doing that it helps in bridging the gap amongst phenotypic variations and mutations. The use of SIFT has been done in genetic studies, mutation and disease. The procedure for the use of SIFT has been in *Nature Protocols* previously. (Vaser et al., 2016)The projects of random mutagenesis and SNP studies help in identifying the amino acid substitution in the coding regions of protein. All the substitutions have the ability to affect the function of proteins. SIFT is useful for its users in prioritizing their further on the basis prediction whether an amino acid substitution affects the function of protein. It has been reported that mutagenesis and human polymorphism studies by SIFT can differentiate between amino acid changes which maybe deleterious or functionally neutral (Ng & Henikoff, 2003).

G. Mutation Taster

Mutation Taster is used to assess disease causing ability of variants of DNA sequence. It carries out an array of in-silico tests to anticipate the effect of variant on the product of gene or protein. The tests of Mutation Taster are done on both DNA and protein level hence this tool is not bound to single amino acid substitution but also able to handle variants of introns and synonymous variants (Schwarz et al., 2010, 2014)

Mutation Taster incorporate information from different databases and makes use of conventional tools for analysis. Analysis covers splice-site variations, deficit of protein attributes and variations that can affect the mRNA amount. The results obtained are then analyzed by Bayes classifier2 which can anticipate the potential of disease. Usually a query is solved in 0.3 seconds or less(Schwarz et al., 2010)

H. M-Cap

M-Cap is used to classify clinical pathogenicity by accurately dismissing 60% of rare missense mutations with tentative significance at 95% sensitivity in a genome. Classifiers like MetaLR, CADD, PolyPhen-2 and SIFT help in anticipating of many rare missense variants by deprioritizing some variants in a typical genome. These extensively employed methods misclassify 25 to 50% of identified pathogenic mutations which can eventually direct to wrong diagnosis (Jagadeesh et al., 2016)

It has been demonstrated by using individuals from 1000 Genome Project and exomes of solved patients that M-CAP accurately retains 95% of the pathogenic variants while discharge more than 60% of the variants with unknown consequences in a usual exome. Even though M-CAP dismiss 3/5 of the benign alterations, but it saves much time of geneticist to identify the contributing gene or variant in small number of candidates which makes M-CAP important to clinicians (Ionita-Laza et al., 2016)

IV. RESULTS AND DISCUSSION

BLAST-X is used to compare the amino acid sequence of the given FASTA format of RET gene. By using BLAST-X mutations are identified across the similar sequences in the NCBI database. Seven mutations are common R114C, V292M, D300N, R313Q, S316I, S339L, D353Y and R360Q (So et al., 2011). FASTA format of RET gene is mentioned, which is used to analyze the mutations and their effects on protein.

>NP_066124.1 proto-oncogene tyrosine-protein kinase receptor Ret isoform a precursor [Homo sapiens]MAKATSGAAGLRLLLLLLPLLGKVALGLYFSRDAYWEKLYVDQAAGTPLLYVHALRDAPEEVPSFRLGQHLYG
TYRTRLHENNWICIQEDTGLLYLNRSLDHSSWEKLSVRNRGFPLLTVYLKVFLSPTSLREGECQWPGCARVYFSFFNTSFPA
CSSLKPRELCFPETRPSFRIRENRPPGTFHQFRLLPVQFLCPNISVAYRLLEGEGLPFRCAPDSLEVSTRWALDREQREKYEL
VAVCTVHAGAREEVVMVPFPVTVYDEDDSAPTFPAGVDTASAVVEFKRKEDTVVATLRVFDADVVPASGELVRRYTSTL
LPGDTWAQQTFRVEHWPNETSVQANGSFVRATVHDYRLVLNRNLSISENRTMQLAVLVNDSDFQGPGAGVLLLHFNVSV
LPVSLHLPSTYSLSVSRRARRFAQIGKVCVENCQAFSGINVQYKLHSSGANCSTLGVVTSAEDTSGILFVNDTKALRRPKCA
ELHYMVVATDQQTSRQAQAQLLVTVEGSYVAEEAGCPLSCAVSKRRLECEECGGLGSPTGRCEWRQGDGKGITRNFSTC
SPSTKTCPDGHCDVVETQDINICPQDCLRGSIVGGHEPGEPRGIKAGYGTCNCFPEEEKCFCEPEDIQDPLCDELCRTVIAAA
VLFSFIVSVLLSAFCIHCYHKFAHKPPISSAEMTFRRPAQAFPVSYSSSGARRPSLDSMENQVSVDAFKILEDPKWEFPRKNL
VLGKTLGEGEFGKVVKATAFHLKGRAGYTTVAVKMLKENASPSELRDLLSEFNVLKQVNHPHVIKLYGACSQDGPLLLIV
EYAKYGSLRGFLRESRKVGPGYLGSGGSRNSSSLDHPDERALTMGDLISFAWQISQGMQYLAEMKLVHRDLAARNILVAE
GRKMKISDFGLSRDVYEEDSYVKRSQGRIPVKWMAIESLFDHIYTTQSDVWSFGVLLWEIVTLGGNPYPGIPPERLFNLLKT
GHRMERPDNCSEEMYRLMLQCWKQEPDKRPVFADISKDLEKMMVKRRDYLDLAASTPSDSLIYDDGLSEEETPLVDCNN
APLPRALPSTWIENKLYGMSDPNWPGESPVPLTRADGTNTGFPRYPNDSVYANWMLSPSAAKLMDTFDS

For mutational analysis four bioinformatics tools are used, Polyphen tool gives three types of results, probably damaging, possibly damaging and Benign based on the polyphen score. PhD.SNP results are telling that the mentioned amino acid mutation in RET gene are human deleterious and cause disease. I-Mutant results tell us the stability of the protein after the mutation. And SIFT results indicate whether the amino acid mutation is tolerated and has no effect on protein or is not tolerated and affects the protein function. Table 1 contains the results of all four bioinformatics tools which are used to analyze the mutations and to determine their aftereffects on protein function and pathogenicity.

Table: 1 Results of Polyphen, PhD-SNP, I-Mutant and SIFT, *probably damaging, *possibly damaging, *Predictor of human Deleterious SNPs,

Protein	Functional prediction			
NP_066124.1	Polyphen	PhD-SNP	I-Mutant	SIFT
R114C	PRD	HD-SNPs	Decrease	Tolerated
V292M	PRD	HD-SNPs	Increase	Not Tolerated
D300N	PRD	HD-SNPs	Decrease	Not Tolerated
R313Q	PRD	HD-SNPs	Decrease	Not Tolerated
S316I	PSD	HD-SNPs	Decrease	Not Tolerated
S339L	Benign	HD-SNPs	Increase	Tolerated
D353Y	PRD	HD-SNPs	Decrease	Not Tolerated
R360Q	PSD	HD-SNPs	Decrease	Not Tolerated

V. CONCLUSION

Mutation in RET gene is the most common cause of HSCR disease. Out of hundreds of mutations these seven mutations are reported in different literatures (R114C, V292M, D300N, R313Q, S316I, S339L, D353Y and R360Q) have different effects on RET protein which cause the HSCR. S339L is only mutation which is tolerated and doesn't have any effect on protein stability and function. Rest of the mutations are affecting protein function.

VI. ACKNOWLEDGEMENTS

This study receive funding from "The Hong Kong University Seed Funding for Strategic Interdisciplinary Research Scheme 2019/20"

REFERENCES

[1] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. In Nature methods (Vol. 7, Issue 4, pp. 248–249). https://doi.org/10.1038/nmeth0410-248

[2] Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics, Chapter 7, Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76

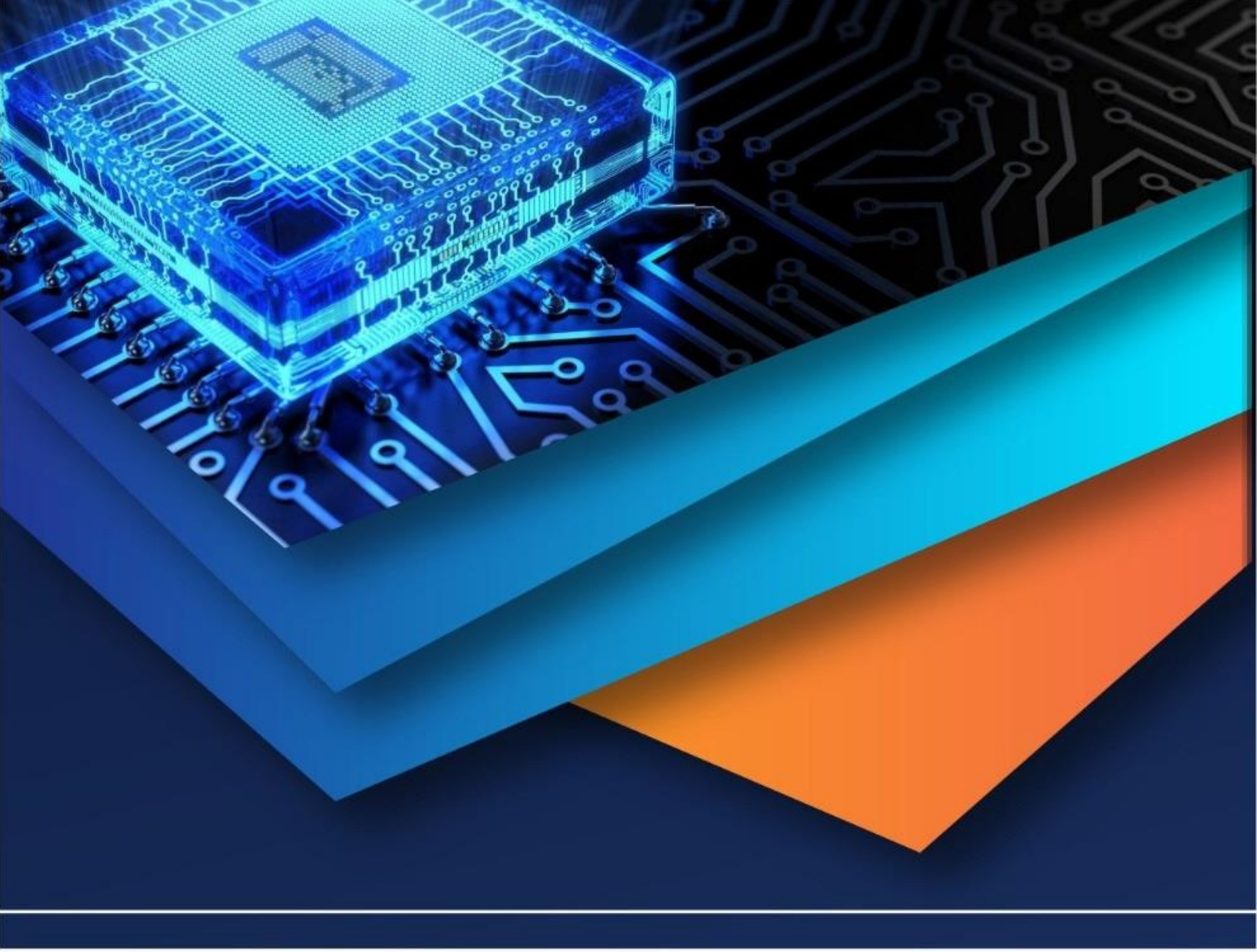
[3] Amiel, J., Sproat-Emison, E., Garcia-Barcelo, M., Lantieri, F., Burzynski, G., Borrego, S., Pelet, A., Arnold, S., Miao, X., Griseri, P., Brooks, A. S., Antinolo, G., de Pontual, L., Clement-Ziza, M., Munnich, A., Kashuk, C., West, K., Wong, K. K.-Y., Lyonnet, S., … Fernandez, R. (2008). Hirschsprung disease, associated syndromes and genetics: a review. Journal of Medical Genetics, 45(1), 1–14. https://doi.org/10.1136/jmg.2007.053959

[4] Badner, J. A., Sieber, W. K., Garver, K. L., & Chakravarti, A. (1990). A genetic study of Hirschsprung disease. American Journal of Human Genetics, 46(3), 568–580.

[5] Capriotti, E, Calabrese, R., & Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics (Oxford, England), 22(22), 2729–2734. https://doi.org/10.1093/bioinformatics/btl423

[6] Capriotti, Emidio, Fariselli, P., Calabrese, R., & Casadio, R. (2005). Predicting protein stability changes from sequences using support vector machines. Bioinformatics (Oxford, England), 21 Suppl 2, ii54-8. https://doi.org/10.1093/bioinformatics/bti1109

[7] Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. Proteins, 62(4), 1125–1132. https://doi.org/10.1002/prot.20810
[8] Choi, Y. (2012). A Fast Computation of Pairwise Sequence Alignment Scores between a Protein and a Set of Single-Locus Variants of Another Protein. Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, 414–417. https://doi.org/10.1145/2382936.2382989
[9] Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLOS ONE, 7(10), 1–13. https://doi.org/10.1371/journal.pone.0046688
[10] Hu, J., & Ng, P. C. (2013). SIFT Indel: Predictions for the Functional Effects of Amino Acid Insertions/Deletions in Proteins. PLOS ONE, 8(10), null. https://doi.org/10.1371/journal.pone.0077940
[11] Ionita-Laza, I., McCallum, K., Xu, B., & Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics, 48(2), 214–220. https://doi.org/10.1038/ng.3477
[12] Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature Genetics, 48(12), 1581–1586. https://doi.org/10.1038/ng.3703
[13] Lantieri, F., Griseri, P., & Ceccherini, I. (2006). Molecular mechanisms of RET-induced Hirschsprung pathogenesis. Annals of Medicine, 38(1), 11–19. https://doi.org/10.1080/07853890500442758
[14] Lecerf, L., Kavo, A., Ruiz-Ferrer, M., Baral, V., Watanabe, Y., Chaoui, A., Pingault, V., Borrego, S., & Bondurand, N. (2014). An impairment of long distance SOX10 regulatory elements underlies isolated Hirschsprung disease. Human Mutation, 35(3), 303–307. https://doi.org/10.1002/humu.22499
[15] Moore, S. W. (2006). The contribution of associated congenital anomalies in understanding Hirschsprung's disease. Pediatric Surgery International, 22(4), 305–315. https://doi.org/10.1007/s00383-006-1655-2
[16] Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13), 3812–3814. https://doi.org/10.1093/nar/gkg509
[17] Parisi, M. A., & Kapur, R. P. (2000). Genetics of Hirschsprung disease. Current Opinion in Pediatrics, 12(6), 610–617. https://doi.org/10.1097/00008480-200012000-00017
[18] Sanchez-Mejias, A., Watanabe, Y., M Fernández, R., Lopez-Alonso, M., Antiñolo, G., Bondurand, N., & Borrego, S. (2010). Involvement of SOX10 in the pathogenesis of Hirschsprung disease: report of a truncating mutation in an isolated patient. Journal of Molecular Medicine (Berlin, Germany), 88(5), 507–514. https://doi.org/10.1007/s00109-010-0592-7
[19] Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. In Nature methods (Vol. 11, Issue 4, pp. 361–362). https://doi.org/10.1038/nmeth.2890
[20] Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods, 7(8), 575–576. https://doi.org/10.1038/nmeth0810-575
[21] So, M. T., LeonThomas, T. Y. Y., Cheng, G., TangClara, C. S. M., Miao, X. P., Cornes, B. K., Ngo, D. N., Cui, L., NganElly, E. S. W., LuiVincent, V. C. H., Wu, X. Z., Wang, B., Wang, H., Yuan, Z. W., Huang, L. M., Li, L., Xia, H., Zhu, D., Liu, J., … Garcia-Barcelo, M. M. (2011). RET mutational spectrum in Hirschsprung disease: Evaluation of 601 Chinese patients. PLoS ONE, 6(12), 2–6. https://doi.org/10.1371/journal.pone.0028986
[22] Tomuschat, C., & Puri, P. (2015). RET gene is a major risk factor for Hirschsprung's disease: a meta-analysis. Pediatric Surgery International, 31(8), 701–710. https://doi.org/10.1007/s00383-015-3731-y
[23] Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. Nature Protocols, 11(1), 1–9. https://doi.org/10.1038/nprot.2015.123
[24] Zhao, H., Yang, Y., Lin, H., Zhang, X., Mort, M., Cooper, D. N., Liu, Y., & Zhou, Y. (2013). DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biology, 14(3), R23. https://doi.org/10.1186/gb-2013-14-3-r23



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY