Gene 519 (2013) 34–40

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methods paper

Roadmap to determine the point mutations involved in cardiomyopathy disorder: A Bayesian approach

Ambuj Kumar [a,1], Vidya Rajendran [a,1], Rao Sethumadhavan [a], Rituraj Purohit [a,b,*]

[a] Bioinformatics Division, School of Bio Sciences and Technology, Vellore Institute of Technology University, Vellore 632014, Tamil Nadu, India
[b] Human Genetics Foundation, Torino, via Nizza 52, I-10126 Torino, Italy

ARTICLE INFO

Article history:
Accepted 27 January 2013
Available online 9 February 2013

Keywords:
MyH7
Ebstein's anomaly
Hypertrophic cardiomyopathy
SNP analysis
Amyloid propensity
Bayes theorem

ABSTRACT

Determining the deleterious non-synonymous single nucleotide polymorphisms (nsSNPs), that might be involved in inducing disease-associated phenomena, is now among the most important field of computational genomic research. The rapid evolution in sequencing technologies has now outranged the limit of available sequence databases and has out-fledged the amount of SNP data that are yet to be characterized. In this article we have performed a comprehensive analysis of deleterious nsSNPs in MyH7 gene associated with cardiomyopathy cases using a set of computational platforms. We implemented a set of computational SNP analysis platforms along with the Bayesian calculations in order to filter the most likely mutation that might be associated with cardiomyopathy associated disorders. The Bayesian calculation depicted 27 fold rises in the likelihood score for causing cardiomyopathy disorder when MyH7 gene mutations were compiled. Furthermore, we reported E466Q mutation in MyH7 motor domain that showed increase in the amyloid propensity of protein, as well as a significant level of pathogenicity was also observed. The prediction roadmap followed in this article has showed a notable range of accuracy and can be used for determining cardiomyopathy associated nsSNPs for other candidate genes.

1. Introduction

Myosins are among the most abundant protein in the human body (Oldfors, 2007). However, identified human diseases caused by mutations in these proteins are rare. Hereditary myosin myopathies have emerged as a new group of muscle diseases with highly variable clinical features at the onset of fetal development, childhood or adulthood, and are mainly caused by mutations in skeletal muscle myosin heavy chain (MyHC) genes (Oldfors, 2007). Novel missense mutation in cardiac beta myosin heavy chain gene was reported in Japanese patient with hypertrophic cardiomyopathy (Nishi et al., 1992) and has been further identified in several other cases (Fan et al., 2011). Several mutations in the beta-MyHC are associated with early onset and high incidence of sudden cardiac death.

Mutations in encoded gene of the human beta myosin heavy chain (MyHC), MyH7, which is expressed in cardiomyocytes and type 1 striated muscle fibers, have been identified in cardiac and skeletal muscle diseases. MyH7 has 40 exons and encodes the beta myosin heavy chain, a protein of 1935 amino acids featuring an N-terminal globular head, a central rod and a C-terminal tail domain (Cullup et al., 2012). Cardiac diseases caused by MyH7 mutations are more frequent than skeletal muscle ones and include familial hypertrophic/dilated cardiomyopathy and cases of Ebstein's anomaly, often associated with left ventricular non-compaction (van Engelen et al., 2011). Sudden death is part of the recognized spectrum associated with MyH7-related cardiomyopathies in humans (Cullup et al., 2012). Muscle diseases due to MyH7 mutations have been classified into two subgroups, according to clinical and pathological findings, laing distal myopathy (LDM) and hyaline body myopathy (or myosin storage myopathy, MSM). The vast majority of MyH7 mutations described till date are dominant. A gross distinction can be made while mutations responsible for myopathies reside in the last exons of the tail domain (exons 32–38 for LDM, 37–40 for MSM). Mutations accounting for the heart phenotype are mostly located in the globular head, but can be spread along the gene (Tasca et al., 2012). Tail domain mutations are supposed to disrupt either myosin dimerization or interactions with other sarcomeric proteins, or both (Meredith et al., 2004). Cases of myopathy together with heart involvement of mutations in the globular head have been widely reported (Darin et al., 2007; Homayoun et al., 2011; Overeem et al., 2007; Tasca et al., 2012). More than 200 of the MyH7 mutations identified till date are heterozygous missense mutations in the globular head domain, giving rise to hypertrophic cardiomyopathy (Cullup et al., 2012). The pathogenic p.Cys905Arg mutation in MyH7 was detected in father and postnatally in the proband in case of noncompaction

Abbreviations: nsSNP, non-synonymous single nucleotide polymorphism; MyH7, myosin heavy chain 7; SVM, support vector machine.

* Corresponding author at: Bioinformatics Division, School of Bio Sciences and Technology, Vellore Institute of Technology University, Vellore 632014, Tamil Nadu, India. Tel.: +91 9944649073.
E-mail address: riturajpurohit@gmail.com (R. Purohit).

[1] The authors wish it to be known that, in their opinion, the first Ambuj Kumar and Vidya Rajendran should be regarded as joint First Authors.

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.01.056

cardiomyopathy (Hoedemaekers et al., 2012). Although different MyH7 mutations have been proven to cause distinct structural, functional and thermodynamic effects in vitro (Armel and Leinwand, 2010), it is hard to delineate tight genotype–phenotype correlations in vivo because of the small number of patients affected by these diseases and the reported intrafamilial variability, which likely implies a role for genetic or environmental modifiers. Thus an efficient computational platform to detect the consequences of these associated mutations will be highly significant and will facilitate in target based drug discovery.

Rapid evolution in genome sequencing technologies has led to flooded databases with large amount of exponentially increasing single nucleotide polymorphism data and thus there is a need to characterize them for their corresponding pathogenic property. Determining the deleterious consequences of these SNPs through in vitro and in vivo study is time consuming and laborious. Several computational algorithms have widely been developed for accurate prediction of these uncharacterized alleles for their disease association property. Mutations involved in diseases such as cardiomyopathies are hard to examine using in vivo examinations. Thus an efficient experimental design specific to these diseases are required which will be an important rationale to examine the disease associated physiology of respective SNPs. Computational studies have previously provided an efficient platform for evaluation and analysis of genetic mutations for their pathological consequences and in determining their underlying molecular mechanisms (Kamraj and Purohit, 2013; Kumar and Purohit, 2012a,b,c; Kumar et al., 2012a,b; Purohit and Sethumadhavan, 2009; Purohit et al., 2008, 2011a, b; Rajendran and Sethumadhavan, in press; Rajendran et al., 2012). In this study we have used a set of tools that showed greater accuracy for the prediction of genetic mutations involved in cardiomyopathy associated disorders. Due to large size of MyH7 protein (1935 amino acids) and unavailability of protein 3D structure, here we used a set of computational platforms that utilizes sequence-based conservation profile, homology-based structure profile information and support vector algorithm to examine the deleterious nsSNPs. Further, the molecular mechanism behind the predicted disease-associated alleles was derived. The Bayesian statistics was implemented to examine the likelihood of phenotypic changes that might occur due to the mutation. Combinely these tools provided an efficient roadmap to determine the cardiomyopathy-associated alleles from the large SNP dataset. The overall workflow implemented in this work has been shown in Fig. 1.

2. Materials and methodologies

2.1. SIFT evolutionary conservation score

SIFT is a sequence homology-based program that evaluates the evolutionary conservation scores to predict the effect of amino acid substitutions in the gene coding region (Kumar et al., 2009). It identifies the highly conserved amino acids and predicts to be intolerant to substitution. The prediction carried out by SIFT program is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI- BLAST (Altschul et al., 1997). Further it scans individual positions of the sequence and calculates the conservation probability of a particular residue for all possible substitution which is recorded in a scaled probability matrix. Generally, a highly conserved position is intolerant to most substitutions with SIFT score <0.05, and the poorly conserved position can tolerate most substitutions showing SIFT score >0.05. The prediction accuracy of SIFT program are 88.3–90.6% (Dai and Cogswell, 2003), when tested with different datasets of human variants. A total of 36 naturally occurring non synonymous exonic polymorphisms and the substitution showing SIFT score <0.05 were further analyzed using structure based evaluation programs.

2.2. Prediction of damaging nsSNPs using Polyphen server

The structural simulation algorithm was used to analyze the damaging probability of the nsSNP's using Polyphen program (Ramensky et al., 2002). The annotations are checked from the SWALL database. It uses Coils2 program (Lupas et al., 1991) to predict coiled coil regions and the SignalP program to predict signal peptide regions of the protein sequences (Ramensky et al., 2002). PolyPhen computes the absolute value of the difference between substitution scores of both allele variants in the polymorphic position. Polyphen uses Blast Algorithm to derive the conservation score of the contig nucleotide and their possible functional roles. The predictions are based on four different structure and sequence annotation based parameters including sequence annotation, sequence prediction, multiple alignment and structure. The structural analysis is carried out by buried site prediction and covalent and non-covalent bond formation. The Polyphen program determines the functional impact by analyzing four key points which includes signal peptide, trans-membrane, ligand binding and protein interaction.

2.3. Support vector machine algorithm for disease causing genomic variations

PhD-SNP is a support vector machine (SVM) based classifier that uses supervised learning approach to classify the disease causing point mutations form the given datasets (Capriotti et al., 2006). For a given mutation the substitution forms the wild residue to the mutant which is encoded in 20 elements vector that has -1 in the position relative to the wild-type residue, 1 in the position relative to the mutant residues and 0 in the remaining 18 positions. Second 20 elements vector encoding for the sequence environment is built reporting the occurrence of the residues in a windows of 19 residues around the mutated residue (Capriotti et al., 2006). A total of 28 nsSNP's which were reported to be commonly deleterious by SIFT and Polyphen program were used for the analysis. Twenty nsSNP's were reported to be associated with disease causing phenomena using PhD-SNP program.

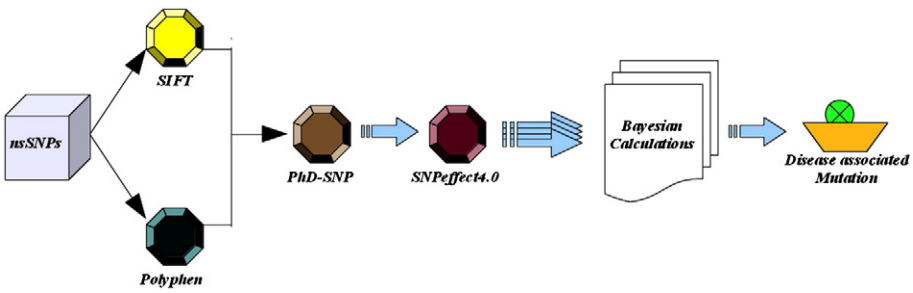

Fig. 1. The workflow model showing the computational pipeline used in this work.