From Single Variants to Protein Cascades

MULTISCALE MODELING OF SINGLE NUCLEOTIDE VARIANT SETS IN GENETIC DISORDERS*

Received for publication, September 29, 2015, and in revised form, November 12, 2015. Published, JBC Papers in Press, November 24, 2015. DOI 10.1074/jbc.M115.695247

Sabine C. Mueller‡1, Björn Sommer‡, Christina Backes‡, Jan Haas‡‡‡, Benjamin Meder‡‡‡, Eckart Meese‡§, and Andreas Keller‡†

From the ‡Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, ‡‡Department of Human Genetics, Saarland University, 66421 Homburg, Germany, the ‡Bio-/Medical Informatics Department, Faculty of Technology, Bielefeld University, 33501 Bielefeld, Germany, **Clayton School of Information Technology, Faculty of Information Technology, Monash University, Melbourne 3800, Australia, the ***Department of Internal Medicine III, Heidelberg University, 69120 Heidelberg, Germany, and the §§‡‡‡DZHK (German Centre for Cardiovascular Research), 69120 Heidelberg, Germany

Understanding the role of genetics in disease has become a central part of medical research. Non-synonymous single nucleotide variants (nsSNVs) in coding regions of human genes frequently lead to pathological phenotypes. Beyond single variations, the individual combination of nsSNVs may add to pathogenic processes. We developed a multiscale pipeline to systematically analyze the existence of quantitative effects of multiple nsSNVs and gene combinations in single individuals on pathogenicity. Based on this pipeline, we detected in a data set of 842 nsSNVs discovered in 76 genes related to cardiomyopathies, associated nsSNV combinations in seven genes present in at least 70% of all 639 patient samples, but not in a control cohort of healthy humans. Structural analyses of these revealed primarily an influence on the protein stability. For amino acid substitutions located at the protein surface, we generally observed a proximity to putative binding pockets. To computationally analyze cumulative effects and their impact, pathogenicity methods are currently being developed. Our approach supports this process, as shown on the example of a cardiac phenotype but can be likewise applied to other diseases such as cancer.

Genetic alterations such as non-synonymous variants play a critical role in human diseases (1). Non-synonymous single nucleotide variants (nsSNVs) refer to single base changes in DNA coding regions altering the amino acid sequence of a protein. A pathogenic phenotype may arise when an amino acid substitution affects structurally important residues and sites relevant for function, such as residues in catalytic sites of enzymes.

In Mendelian disorders, where variation in a single gene is responsible for the phenotypic consequence, thousand causative genes have been identified already (2). Detecting the causative genes in common diseases such as hypertension or diabetes, however, still remains a challenge. In general, these diseases are caused by a varying number of genetic alterations and are influenced by environmental factors that modulate the severity and type of disease-related phenotypes (3). Several nsSNVs in a single gene may exert a quantitative effect on the phenotypic observation, whereas a phenotype can also result from the combined action of nsSNVs in many genes. Because the experimental analysis of the pathogenic potential of variants is laborious and time consuming, several methods to predict the biological impact of nsSNVs on the corresponding proteins and their function have been developed in recent years. Most of these prediction methods are based on evolutionary information and/or combine functional and structural parameters as well as information derived from multiple sequence alignments. According to the obtained features, nsSNVs are classified into benign or pathogenic using different machine learning methods such as neural networks, random forests, support vector machines or Bayesian methods and mathematical operations (4–6). However, none of these methods actually considers the influence of several nsSNVs or mutual effects. In fact, a human individual usually has more than one nsSNV within interacting genes or even within one single gene. From a medical point of view, especially the individual combination of nsSNVs may play a crucial role in clinical diagnostics regarding personalized medicine, because genetic variations, for example, have been identified to influence selection, dosing, and adverse events of medical drugs (7).

Recent studies analyzed the existence of compensatory mutations canceling the damaging effect of deleterious mutations (8). A compensatory mutation occurs when the loss caused by one mutation is corrected by its epistatic interaction with a second mutation at a different site in the genome. In contrast, Westphal et al. (9) identified a mild polymorphism in dolichyl pyrophosphate Man9GlcNAc2 α1,3-glycosyltransferase (ALG6) that putatively exacerbates the phenotypic effect caused by dysfunction of the phospho-mannomutase2 (PMM2). The interaction of different genes is generally referred to epistasis, however, a current review claims this term to be confusingly and even conflictively used in available literature (10). To not further add to this confusion, we speak of interacting genes or proteins, respectively. Due to the existence of compensatory mutations, there might also exist a cumulative effect turning benign mutations in an ensemble, regardless whether in one or multiple interacting genes, into a pathogenic or even more pathogenic phenotype.
Traditional approaches predict the influence of nsSNVs on pathogenicity for single variants (11). However, the impact of nsSNVs on the phenotype of the patient can arise from multiple factors (12). Accumulating nsSNVs may have a quantitative effect on the dysfunction of the gene, as well as nsSNVs in interacting genes may aggravate or attenuate a pathological effect. In consequence, a multiscale analysis of nsSNVs, outlined in Fig. 1, which includes a three-dimensional context, interaction information, and functional cascades, is required to capture the whole range of nsSNV impact.

To address this issue, we determined putative nsSNV associations in a high-quality clinical data set of cardiomyopathy patients and studied the possible accumulation events affecting the phenotypes of the patient. We further analyzed available protein three-dimensional structures within our data set and the impact of nsSNVs to gain insights in pathogenic molecular mechanisms. Because subcellular protein localization can also influence phenotypic behavior (13), we analyzed the subcellular localization of the proteins encoded by the identified genes featuring associated nsSNVs.

Materials and Methods

Data Sets—We analyzed a data set comprising 842 nsSNVs in 76 genes that are clinically relevant for dilated cardiomyopathy (DCM) (known causes and likely candidate genes for DCM), found by studying the genetics in 639 unrelated patients with sporadic (51%) or familial (49%) DCM (14). In the investigated region, about 99.1% of the targeted genomic region is covered at least 50-fold and each patient carried an average of 32 nsSNVs.

In addition, we generated a control set based on the general population of the 1000 genomes project, to be able to evaluate detected putatively DCM-related nsSNV patterns. In detail, we downloaded the BAM files from 445 samples and applied the same variant calling and filtering algorithms, as described for the analyzed DCM cohort (14). To match the European INHERITANCE cohort, we only considered individuals with a European descent: Utah residents with northern and western European ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian populations in Spain (IBS), and Toscani in Italy (TSI).

For the nsSNVs in the DCM data set, we collected available information deposited in the SwissProt databases (15), dbSNP (dbSNP build 138) (16) and HGMD (July 2014) (17). SwissProt provides a collection of human polymorphisms and disease mutations (HUMSAVAR) assigned according to literature reports on probable disease association (18). Tightly coupled with dbSNP, ClinVar accessions report human variations and interpretations of the relationship of these variations to human health (19). Entries are labeled according to clinical significance. Moreover, the HGMD collates known (published) gene lesions responsible for human inherited disease. Next, we built...
Multiscale Analysis of nsSNVs

A test set from the DCM data including only nsSNVs with at least one annotation in these three databases as well as benign or disease-linked information. Although about 60% are deposited in dbSNP with an rs ID (reference SNP cluster ID), only 45% have pathogenicity information available. The neutral-labeled set comprises 192 nsSNVs and the disease-associated set 147. A total of 55% nsSNVs in the DCM data set have no available clinical significance information, and even about 40% have neither an rs ID nor other known identifiers and annotations.

Association Rule Learning—Frequently, human individuals carry more than one single nsSNV even within one gene. Beyond this, genes and their encoded proteins interact with each other.

To discover strong relationships between variables in large data sources, association rule learning is generally applied. Association rule learning uncovers hidden relationships within the tested data by formulating association rules, whereas using different measures of interest to quantify the quality of the generated rules (20). The statistical significance of an association rule is measured by its support and confidence, where support refers to a frequency constraint determining the quantitative applicability of a rule and confidence measures its reliability.

Via association rule learning, we studied whether there are frequent combinations of nsSNVs within our DCM data set and the healthy control cohort. In fact, we identified combinations of nsSNVs occurring within one single gene as well as combinations of nsSNVs in one gene accumulating with combinations in other genes.

We applied the R package a rules (21) using the implemented a priori algorithm (22). The confidence threshold was set to 0.8 and different levels of support, starting with at least 0.5, were tested.

Network Analysis—Due to the growing availability of high throughput biological data, the analysis of molecular networks gained significant interest. To determine the biological and functional connections of the detected associated genes with nsSNV combinations, we used several information sources: the STRING database (23), the UniProt-GOA database (24), and the KEGG PATHWAY database (25).

Besides the biological connections among the associated nsSNVs, we also investigated their topological characteristics within the STRING human interaction network. To detect putative interaction hubs, we determined betweenness and degree for each node in the human STRING network using the R package igraph. The degree of a node identifies the number of edges connected to the node, whereas the node betweenness is an indicator of the centrality of the node in the network.

Structural Location of Amino Acid Substitutions Introduced by nsSNVs—The effect of a nsSNV critically depends on the structural location of the mutated residue, especially if it is buried in the hydrophobic core or exposed on the protein surface (27). We selected proteins within our DCM data set, with a complete Protein Data Bank (PDB) (28) three-dimensional structure available. Next, we calculated solvent accessibility using Naccess (29) to determine whether mutated residues tend to accumulate on the protein surface or are rather buried inside the protein. Because previous studies identified the majority of pathogenic nsSNVs to destabilize the structure of a protein (30), we also analyzed protein stability changes upon mutation based on I-Mutant2.0 predictions (31). The predicted free energy change in I-Mutant2.0 is calculated from the unfolded Gibbs free energy change of the mutated protein minus the unfolding free energy value of the native protein. Predictions were based on default values for temperature (25) and pH (7).

We also predicted possible binding pockets via LIGSITEcsc (32). LIGSITEcsc automatically identifies binding pockets on the surface of a protein based on the Connolly surface and the degree of conservation. We used the default parameter settings of 1-Å grid space and a probe radius of 5 Å.

Subcellular Localization of Mutated Proteins—For visualizing and analyzing the localization of the proteins encoded by genes with nsSNVs, we used the CELlmicrocosmos 4.2 Pathway-Integration (CmPI) (33). CmPI is connected to DAWIS-MD, a data warehouse containing a number of databases (34). In the context of this work, the following databases were queried: BRENDA (35), GO (36), Reactome (37), and UniProt (18). Because each protein usually can obtain different localization entries from databases, CmPI applies a context-based localization prioritization. The cell component acquiring the most localization hits is assigned to the queried protein and verified by cellular topology analysis: connected proteins should not receive distant localizations (e.g., nucleus and extracellular matrix).

Results

Association Rule Learning and Network Analysis—To analyze whether nsSNVs occur accumulated within one gene or within interacting genes, we applied association rule learning. In the DCM data, we identified nsSNV combinations accumulating in the single genes MYPN (Myopalladin), CACNA1C (voltage-dependent L-type calcium channel subunit α1C), DMD (Dystrophin), ADRB2 (β2-adrenergic receptor), and RBM20 (RNA-binding protein 20) with high support and significant confidence values. Table 1 lists the detected nsSNVs significantly associated within one gene. The nsSNV combinations in RBM20 and CACNA1C are even found in at least 90% of all patients.

According to database entries in HGMD and SwissProt, RBM20 is already related to DCM and CACNA1C to the Timothy and Brugada Syndrome. MYPN, which incorporates the most associated nsSNV accumulations, is linked to different forms of cardiomyopathies (familial, hypertrophic, and dilated). ADRB2 participates in signal transduction and namely in the adrenergic signaling in cardiomyocytes. DMD is involved in several pathways relevant for cardiac diseases such as DCM, hypertrophic cardiomyopathy, arrhythmogenic right ventricular cardiomyopathy, and viral myocarditis. All identified nsSNV associations, however, are annotated as benign nsSNVs.

Furthermore, we detected nsSNVs in seven different genes (CACNA1C, SMYD2 (N-lysine methyltransferase SMYD2), PARVB (β-parvin), KCNE1 (potassium voltage-gated channel subfamily E member 1), RBM20, KCNQ2 (potassium voltage-gated channel subfamily KQT member 2), and JUP (Junction plakoglobin) significantly associated with each other in the
DCM patients. In at least 70% of all DCM patients, these seven genes including the particular nsSNVs revealed strong associations to each other. Using the corresponding association rule setup, these specific nsSNV-gene combinations are not present in the control cohort. Only SMYD2, PARVB, KCNE1, and JUP revealed an association in healthy controls. According to a large-scale analysis of the human transcriptome in 2004, all of the associated genes revealed significant expression in the heart (38). The majority of detected genes with associated nsSNVs is already known in the context of diseases such as Brugada Syndrome, Long QT Syndrome, Naxos Disease, and different stages of DCM. In contrast, all association rule-detected nsSNVs within these genes are annotated as benign, except the N749T mutation in KCNQ2, which has no available annotations. Fig. 2 compares the information annotations of all genetic variants within the DCM data set with the association rule detected. Among the 639 DCM patients, we identified 26 without already known or annotated disease-associated nsSNVs. The 26 DCM patients mainly carry benign and not annotated variants. Interestingly, the intersection of their inherited nsSNVs revealed exactly the detected associated nsSNVs in CACNA1C, SMYD2, PARVB, KCNE1, RBM20, KCNQ2, and JUP. Table 2 lists the detailed nsSNV combinations.

We further analyzed functional interaction and biological intersection of the genes comprising the identified associated nsSNVs and the corresponding proteins including the introduced amino acid substitutions, respectively. Referring to the corresponding GO terms of the genes with associated nsSNV combinations, the majority participates in protein binding, voltage-gated ion channel activity, and transport. A mutation of residues involved in complex interaction networks can critically influence large interaction cascades by spreading the implemented loss across the network. To provide more insights into a putative relationship of the detected genes, we combined and visualized the extracted interaction information from the STRING human network with the available biological knowledge in Fig. 3. We highlighted GO overlaps within the resulting networks using Cytoscape (39). The edges in the network refer to available interactions between their nodes. CACNA1C, JUP, SMYD2, PARVB, and KCNQ2 are directly connected to large hubs within the human network. KCNE1, KCNQ2, and CACNA1C interact functionally with each other (40). KCNE1 attenuates the current amplitude of the KCNQ2 channel subunit and slows its gating kinetics (41). According to the KEGG PATHWAY database, KCNE1 is part of the adrenergic signaling in cardiomyocytes. A perturbation of its channel function by inherited mutations results in increased susceptibility to cardiac arrhythmias. KCNQ2 belongs to the cholinergic synapse and CACNA1C even takes part in both pathways. Interestingly, KCNE1, CACNA1C, and KCNQ2 are already targets of drugs against arrhythmia, atrial fibrillation, congestive heart failure, left ventricular hypertrophy, and isolated systolic hypertension (42). Furthermore, there is recent evidence that the post–transmembrane domain region of KCNE1 interacts with the KCNQ1 channel to modulate the I(K) current amplitude and gating kinetics (43).

**TABLE 1**

Associated nsSNVs in single genes

These genes show significantly associated nsSNVs with confidence of at least 0.8.

| Gene     | Name                        | Transcript | nsSNVs                          | Patients with nsSNV |
|----------|-----------------------------|------------|---------------------------------|---------------------|
| ADRB2    | β2-Adrenergic receptor      | NM-000024  | G16R, E27Q                      | 64%                 |
| CACNA1C  | Voltage-dependent L-type calcium channel subunit α-1C | NM-199460  | M1869V, K1893R, P1868L          | 96%                 |
| DMD      | Dystrophin                  | NM-004009  | D878G, R2933Q                    | 72%                 |
| MYPN     | Myopalladin                 | NM-032578  | S691N, S707N                     | 72%                 |
| MYPT1    | Myopalladin                 | NM-032578  | F628L, S691N, S707N, P1135T      | 70%                 |
| MYPT2    | Myopalladin                 | NM-032578  | F628L, S803R, S691N, S707N, P1135T | 67%                 |
| MYPN     | Myopalladin                 | NM-032578  | S803R, S691N, S707N, P1135T      | 67%                 |
| RBM20    | RNA-binding protein 20      | NM-00113436| E1223Q, W768S                    | 98%                 |

**FIGURE 2. Available information annotations.** Comparison of all genetic variants within the DCM data set and the association rule (AR) detected. Except for one not annotated AR variant, all other AR variants have neutral annotations.

**FIGURE 3.** Structural Location of Amino Acid Substitutions Introduced by nsSNVs—The protein structure reveals interactions between residues that are distant in primary sequence but close in three-dimensional space. Solvent accessibility provides an intuitive and quantitatively reasonable idea of the complexity of the molecular interaction network involved in a residue (44). In consequence, we calculated solvent accessibilities using Naccess (29) for the 8 proteins (comprising 46 amino acid substitutions) in our data set with an available PDB (28) structure to...
analyze whether disease-associated amino acid substitutions cluster on the protein surface or at buried sites. The results confirm the findings of Wang and Moult (45) for single nsSNVs. The majority (89%) of disease-linked mutations introduced by nsSNVs are located inside the protein or probably affecting stability, whereas benign-annotated substitutions mainly cluster on the protein surface (67%). To analyze putative protein stability changes upon mutation, we calculated stability change predictions via I-Mutant2.0. The majority (81%) of substitutions are predicted to decrease protein stability, independent of their location in the three-dimensional protein structure (see Fig. 4).

For five protein structures, we were also able to predict possible binding pockets using LIGSITEcsc. 8 of 10 mutations at the surface of the protein are found close to a predicted binding pocket. Interestingly, two mutations introduced by the detected significantly associated mutation nsSNV combinations, KCNE1 S38G and SMYD2 G165E, are also located close to a possible binding pocket (see Fig. 5) of the corresponding protein. SMYD2 lysine methylates the tumor suppressor TP53, leading to decreased DNA-binding activity and subsequent transcriptional regulation activity of TP53 (46). According to literature, the binding interface of TP53 and SMYD2 is located between the catalytic SET domain (residue 1–282) and the C-terminal domain (47). In addition, SMYD2 has been reported to map primarily to the cytoplasm indicating SMYD2 targets a small subset of histones at specific chromatin loci as well as non-histone substrates (48).

Subcellular Localization of Mutated Proteins—Based on the previously discussed methods, seven genes were identified

### TABLE 2

| Gene     | Name                                | Transcript | Expression | nsSNV combination |
|----------|-------------------------------------|------------|------------|-------------------|
| CACNA1C  | Voltage-dependent L-type calcium channel subunit α-1C | NM-199460  | Heart, brain, ovary, . . . | M1869V, K1893R, P1868L |
| SMYD2    | N-Lysine methyltransferase SMYD2     | NM-020197  | Heart, brain, . . . | G165E |
| PARVB    | β-Parvin                            | NM-001243386 | Heart, skeletal muscle | V6A |
| KCNE1    | Potassium voltage-gated channel subfamily E member 1 | NM-000219  | Heart, liver, . . . | S38G |
| RBM20    | RNA-binding protein 20              | NM-001134363 | Heart | E1223Q, W768S |
| KCNQ2    | Potassium voltage-gated channel subfamily KQT member 2 | NM-172108  | Heart, brain | N749T |
| JUP      | Junction plakoglobin                | NM-002230  | Heart | M697L |

FIGURE 3. GO annotations and interactions of associated nsSNVs. Network is based on the STRING human network. The edges in the network refer to available interactions between their nodes. The associated nsSNV genes show great overlap in their GO annotations. Some are also connected to the top ranked hubs within the STRING human network.

**Multiscale Analysis of nsSNVs**
showing specific nsSNVs in more than 70% of all analyzed patients: CACNA1C, JUP, KCNE1, KCNQ2, PARVB, RBM20, and SMYD2. In particular, all of the detected genes have a significant expression in the heart. Using CmPI, Homo sapiens-related potential localizations for these seven genes were acquired using cell component-gene association data from the aforementioned databases. The associated proteins of these genes show five potential localizations: nucleus, cytosol, cell membrane, lysosome, and the extracellular matrix. Moreover, five of them provide multiple potential localizations. An overview and distribution of these locations can be found in Fig. 6. Based on the localization data, the hypothesis can be formulated that these proteins are assembled in a potential cascade starting from the nucleus, through the cytosol, entering the cell membrane, and proceeding to the extracellular matrix, or vice versa. This theory is supported by the fact that RBM20 is exclusively localized at the nucleus and KCNQ2 at the cell membrane, whereas PARVB seems to travel between the extracellular matrix, the cell membrane, and the cytosol. We visualized the connections of these proteins including the assigned subcellular locations in Fig. 7. For the purpose of clarity, Fig. 7 condenses the detected interactions to the identified potential cascade. However, the in silico study requires experimental analysis to validate the formulated cascade hypothesis.

**Discussion**

A human individual usually has more than one nsSNV. From a medical point of view, the individual combination of nsSNVs may play a crucial role in clinical diagnostics regarding personalized medicine (49). Previous studies analyzed the occurrence and characteristics of compensatory mutations (8), although there might also be cumulative effects of mutations, packing single benign effects together to an observable disease phenotype. More precisely, benign-annotated nsSNVs in combination might be responsible for a pathological effect. Westphal et al. (9), for example, studied congenital disorders of glycosylation and identified a mild polymorphism in ALG6 putatively exacerbating an already severe pathogenic phenotype caused by PMM2 dysfunction.

In this study, we developed a multiscale pipeline to systematically analyze the existence of quantitative effects of nsSNV sets. We detected DCM patients without identified disease-associated nsSNVs, but inhering mainly benign-labeled variants. Via association rule learning, we detected associated combinations of nsSNVs within at least 70% of all cardiomyopathy patients in the data set. These specific combinations could not
be identified as associated in the control cohort of healthy humans, which hints to disease relevance and, however, requires further analysis. Due to the lack of prediction tools able to assess a cumulative effect of nsSNVs, a pathogenicity prediction for the identified associated nsSNVs was not possible. Furthermore, a three-dimensional structure of the encoded protein was available for only two of the identified associated nsSNV genes, \textit{KCNE1} and \textit{SMYD2}. For the remaining associated genes, even an adequate template for structural modeling was missing. Interestingly, the associated nsSNVs in \textit{KCNE1} (S38G) and \textit{SMYD2} (G165E) are located at the surface of the encoded protein close to the predicted binding pockets. Stability change predictions based on protein sequence, however, revealed that all of the associated amino acid substitutions decrease protein stability except \textit{SMYD2} G165E. In addition, we studied available pathway information including GO annotations and analyzed the interaction networks. Proteins might act at different stages of the same pathway contributing quantitatively to the progressive dysfunction of the pathway until a disease phenotype is observed. Both genes in the study by Westphal \textit{et al.} (9), for example, encode enzymes involved in a different part of the post-translational modification process without a direct interaction. \textit{KCNE1} and \textit{CACNA1C} as well as \textit{KCNQ2} and \textit{CACNA1C} participate in the same pathways and mainly contribute to voltage-gated ion channel activity and transport. Ion channels are key components in a wide variety of biological processes, such as muscle contraction (e.g., cardiac muscle contraction), epithelial transport of nutrients and ions, or T-cell activation. A number of genetic disorders (e.g., Long QT syndrome, Brugada syndrome) are related to ion channel dysfunctions. \textit{JUP} and \textit{PARVB} are involved in cell junction organization and interact as well as \textit{SMYD2} directly with the top-ranked hubs within the human network. Cell junctions play a major role in communication between neighboring cells and cell stress reduction. The cellular function of the identified genes, however, is largely unclear despite their role in pathological processes. Although, for example, a 2-bp deletion mutation in the junction protein plakoglobin (\textit{JUP}) has been found essential for the molecular genetics of arrhythmogenic right ventricular cardiomyopathy, the molecular basis of reduced \textit{JUP} in the cell membrane of arrhythmogenic right ventricular cardiomyopathy remains unclear (50). It still has to be clarified whether desmosome protein mutations impair desmosome assembly and cause reduced incorporation of \textit{JUP} into the cell membrane.

Here, we performed a subcellular localization enabling the formulation of the hypothesis that the seven identified proteins

\[\text{FIGURE 7. Schematic visualization of the subcellular localization.} \] The red gene symbols represent the detected associated genes. Besides the subcellular assignment, interactions between the listed genes are visualized (right side). Red-labeled edges mark direct connections of associated genes. In addition to links within one cell compartment, there are also multiple edges crossing different compartments.

\[\text{Multiscale Analysis of nsSNVs}\]
form a potential cascade: RBM20 is only found in the nucleus, SMYD2 travels between the nucleus and the cytosol, and the other five genes are mostly associated to the cell membrane, where PARVB shows potential localizations between the extracellular matrix, cell membrane, and cytosol. Further experimental studies are required to analyze and validate this cascade hypothesis.

Because the used data set only comprises genes clinically relevant for DCM due to known causes or the proposal of likely candidates, an association analysis might miss genes or nsSNVs not yet identified to have an influence on DCM. Previous studies, however, supposed the inclusion of known disease pathology and prior knowledge to improve analysis results (51). So far, to the best of our knowledge, the methods to analyze the effect of multiple nsSNVs in different genes are limited and suitable approaches have to be developed. In this study, we targeted already detected DCM candidate genes to establish proper analysis strategies.

Finally, all systematic analyses point to a connection of the detected genes featuring associated nsSNVs, in functionality as well as in their contribution to biological pathways. Moreover, the specific nsSNV combinations identified as significantly associated in DCM patients could not be detected associated within the healthy control data. In a next step, further association studies on even larger patient cohorts with cardiomyopathies are required to validate the identified nsSNVs. Additional studies on patients with phenotypes different from cardiomyopathies, in particular, can assess nsSNV specificity. The functional understanding of the identified genes, however, is in many cases limited to their role in physiological or pathological processes leaving many open questions about their cellular role. Without a sufficient understanding of these, the design of experiments, which could be used to address the cumulative effect of specific nsSNVs in several genes, is highly challenging and even near to impossible in many cases. The developed multiscale analysis pipeline has the potential to promote this process.

Conclusion

The impact of nsSNVs in coding genes on the cause and the severity of a disease has become a key task in human health care. A pathogenic phenotype can result from several nsSNVs in one single gene as well as from the combination of nsSNVs in many genes. Single gene dysfunctions linked to diseases have already undergone extensive studies. However, the majority of common diseases such as cardiomyopathy are probably caused by several genetic alterations and environmental influences. In this study, we identified associated nsSNVs in seven genes putatively contributing to cardiomyopathic phenotypes. Due to missing computational methods to analyze cumulative nsSNVs and to assess their impact on pathogenicity, the validation of the clinical relevance is limited. In fact, genetic testing enables predictive diagnosis and can enhance presymptomatic intervention. Future studies, however, focusing on translation of computational findings to applicable mechanisms in clinical routine and capturing diagnostic demands, are highly required.

Author Contributions—S. C. M. and C. B. contributed to data analysis. B. S. performed the cell localization studies. B. M. and J. H. were responsible for data generation. S. C. M., B. S., E. M., and A. K. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgment—We thank the INHERITANCE Project Group (EU FP7) for their valuable contribution.

References

1. Bailey, J. N., Pericak-Vance, M. A., and Haines, J. L. (2014) The impact of the human genome project on complex disease. Genes 5, 518–535
2. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437, 1299–1320
3. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slak, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., and Visscher, P. M. (2009) Finding the missing heritability of complex diseases. Nature 461, 747–753
4. Ng, P. C., and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814
5. Capriotti, E., Calabrese, R., Faraelli, P., Altman, R. B., and Casadio, R. (2013) WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics 14, 56
6. Addzahui, I. A., Schmidt, S., Peshkin, L., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249
7. Giacomini, K. M., Brett, C. M., Altman, R. B., Benowitz, N. L., Dolan, M. E., Flockhart, D. A., Johnson, J. A., Hayes, D. F., Klein, T., Krauss, R. M., Kroetz, D. L., McLeod, H. L., Nguyen, A. T., Ratain, M. J., Relling, M. V., Reus, V., Roden, D. M., Schaefer, C. A., Shuldiner, A. R., Skaar, T., Tantisira, K., Tyndale, R. F., Wang, L., Weinsilboum, R. M., Weiss, S. T., Zineh, I., and Pharmacogenics Research Network. (2007) The pharmacogenomics research network: from SNP discovery to clinical drug response. Clin. Pharmacol. Ther. 81, 328–345
8. Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2007) Characterization of compensated mutations in terms of structural and physico-chemical properties. J. Mol. Biol. 365, 249–256
9. Westphal, V., Kjaergaard, S., Scholl, E., Martens, K., Grunewald, S., Schwartz, M., Matthijs, G., and Freeze, H. H. (2002) A frequent mild mutation in ALG6 may exacerbate the clinical severity of patients with congenital disorder of glycosylation Ia (CDG-Ia) caused by phosphomannomutase deficiency. Hum. Mol. Genet. 11, 599–604
10. Cordell, H. J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. 11, 2463–2468
11. Mueller, S. C., Backes, C., Haas, J. T., Katus, H. A., Meder, B., Meese, E., and Keller, A. (2015) Pathogenicity prediction of non-synonymous single nucleotide variants in dilated cardiomyopathy. Brief Bioinform. 16, 769–779
12. Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009) Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. 19, 212–219
13. Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J. Mol. Biol. 315, 771–786
14. Haas, J., Fresé, K. S., Peil, B., et al. (2015) Atlas of the clinical genetics of human dilated cardiomyopathy. Eur. Heart J. 36, 1123–1135A
15. Yin, Y. L., Scheib, H., Diemand, A. V., Gattiker, A., Famiglietti, L. M., Gasteiger, E., and Bairoch, A. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum. Mutat. 23, 464–470
16. Sherry, S. T., Ward, M. H., Kolodov, M., Baker, J., Phan, L., Smigielski,
30. Yue, P., Li, Z., and Moult, J. (2005) Loss of protein structure stability as a
31. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig,
32. Capriotti, E., and Altman, R. B. (2011) Improving the prediction of disease-
33. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M.,
34. Roth, A., Kelley, R., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S.,
35. Martin, M. J., Mazumder, R., O’Donovan, C., Redaschi, N., and Suzek, B.
36. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry,
37. Hunter, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla,
38. Su, A. I., Wiltshire, T., Batalov, S., Batalov, S., Lapp, H., Ching, K. A., Block,
39. Cline, M. S., Smoot, M., Bocchini, C. A., and McKusick, V. A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowl-
40. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick,
41. Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy,