


Pathway-based variant enrichment analysis on the example of dilated cardiomyopathy

Christina Backes¹ · Benjamin Meder^{2,3,4} · Alan Lai^{2,3} · Monika Stoll⁵ · Frank Rühle⁵ · Hugo A. Katus^{2,3,4} · Andreas Keller¹ 

Received: 18 July 2015 / Accepted: 30 October 2015 / Published online: 7 November 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Genome-wide association (GWA) studies have significantly contributed to the understanding of human genetic variation and its impact on clinical traits. Frequently only a limited number of highly significant associations were considered as biologically relevant. Increasingly, network analysis of affected genes is used to explore the potential role of the genetic background on disease mechanisms. Instead of first determining affected genes or calculating scores for genes and performing pathway analysis on the gene level, we integrated both steps and directly calculated enrichment on the genetic variant level. The respective approach has been tested on dilated cardiomyopathy (DCM) GWA data as showcase.

To compute significance values, 5000 permutation tests were carried out and p values were adjusted for multiple testing. For 282 KEGG pathways, we computed variant enrichment scores and significance values. Of these, 65 were significant. Surprisingly, we discovered the “nucleotide excision repair” and “tuberculosis” pathways to be most significantly associated with DCM ($p = 10^{-9}$). The latter pathway is driven by genes of the HLA-D antigen group, a finding that closely resembles previous discoveries made by expression quantitative trait locus analysis in the context of DCM–GWA. Next, we implemented a sub-network-based analysis, which searches for affected parts of KEGG, however, independent on the pre-defined pathways. Here, proteins of the contractile apparatus of cardiac cells as well as the FAS sub-network were found to be affected by common polymorphisms in DCM. In this work, we performed enrichment analysis directly on variants, leveraging the potential to discover biological information in thousands of published GWA studies. The applied approach is cutoff free and considers a ranked list of genetic variants as input.

The authors take responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

C. Backes and B. Meder: equal contribution as first authors.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-015-1609-7) contains supplementary material, which is available to authorized users.

✉ Andreas Keller
andreas.keller@ccb.uni-saarland.de

¹ Chair for Clinical Bioinformatics, Saarland University, Building E2.1, 66123 Saarbrücken, Germany

² Department of Internal Medicine III, University Hospital Heidelberg, Heidelberg, Germany

³ German Center for Cardiovascular Research (DZHK), Heidelberg, Germany

⁴ Klaus Tschira Institute for Integrative Computational Cardiology, Heidelberg, Germany

⁵ Department of Genetic Epidemiology, Institute of Human Genetics, University of Münster, Münster, Germany

Introduction

For most frequent and many infrequent human diseases, genome-wide association (GWA) studies have been performed to dissect the influence of common genetic variants. Altogether, more than 2000 GWA studies have been published (e.g. Klein et al. 2005; Haines et al. 2005), building a valuable resource to further investigate the role of human genetic variation (Hindorff et al. 2009). In the so-called GWAS catalogue those studies with at least 100,000 SNPs in the initial stage and SNP-trait associations with p values $<1.0 \times 10^{-5}$ are listed, resulting in nearly 28,000 gene-trait

associations (April 10th, 2015). These associations cover almost 10,000 different human genes.

Besides single polymorphisms in non-coding or coding regions and their association to investigate clinical or biological trait, accumulations of variants in certain functional pathways are likewise interesting. Such categories include ontologies as the Gene Ontology (Ashburner et al. 2000) or biochemical pathways as collected in the KEGG database (Kanehisa and Goto 2000). To discover an accumulation in respective categories, different enrichment approaches have been implemented. Frequently, respective algorithms first aggregate p values of variants and calculate a gene level statistic before the gene level statistic is aggregated into a category level statistics.

The first mapping step is essential, since many databases with a priori-available biological data have lower resolution as compared to GWA data, as discussed by Khatri et al. (2012). Almost all knowledge bases [popular examples include KEGG (Kanehisa and Goto 2000), MetaCyc (Karp et al. 2002) or Reactome (Joshi-Tope et al. 2003)] contain information of genes or gene complexes but not directly on single genetic variants. For the aggregation of variants to genes, different approaches exist. Depending on the gene set analysis method, which is carried out as a second step, these approaches either calculate scores for genes or just determine whether a gene is significantly affected or not. Respective methods range from basic mathematic operations (mean/median value of variants, minimum of variants p values, number of significant variants per gene) to complex scoring functions. In the “Pathways of Distinction Analysis” [PoDA, (Braun and Buetow 2011)], e.g. the most significant SNP of each gene is taken into account to get a single reference per gene. Another straightforward approach treats genes as significant as soon as a single significant variant in the gene has been discovered. An example for a far more comprehensive scoring approach is SPOT (Saccone et al. 2010).

For aggregating genes into a category or pathway level statistics, different classes of algorithms have been developed. These are focused either on sets of affected genes or on ordered list of genes. The first algorithms frequently relied on over-representation analysis (ORA). Here, the affected genes on a particular biochemical category are compared to a background distribution; for example, the screened genes in the study on the same category. Next, significance values are calculated by standard hypothesis testing such as the Chi-squared test. An example of an ORA-like analysis in GWA studies has been presented by Holmans et al. (2009). Beyond such approaches, Functional Class Scoring (FCS) methods have been implemented (Khatri et al. 2012). Among the most applied approaches is the Gene Set Enrichment Analysis (Subramanian et al. 2005) (GSEA), which relies on a

Kolmogorov–Smirnov-like test statistic. To calculate significance values, permutation tests or exact solutions can be derived using dynamic programming (Keller et al. 2007). Although these approaches were primarily developed for transcriptomics applications, GSEA has already been successfully applied in GWA studies (Wang et al. 2007). Reviews and comparisons describing the pros and cons of the different tools have been published in the past years (da Huang et al. 2009; Gui et al. 2011; Liu et al. 2012).

In a previous work, we systematically explored the principle strategy and number of permutation tests for GWAS pathway analysis (Backes et al. 2014). The results of over-representation analysis indicated that already few 1000 permutations (<5000) are sufficient to achieve stable significance values. Moreover, the results suggested that permutation of association p values showed comparable results to permutations of the case–control status.

Here, we set to integrate both statistical approaches. Instead of starting with significance values of variants, calculating significance values for genes and then significance values for pathways, we omit the first step and directly calculate enrichment of variants on pathways. This procedure works especially well for FCS-like methods, thus we performed all calculations analogously to unweighted gene set enrichment analysis. Finally, the performance of the approach has been evaluated using a data set of patients suffering from dilated cardiomyopathy (DCM).

Results and discussion

Based on a genome-wide association study including 909 individuals with dilated cardiomyopathies and 2120 controls, we computed significance scores of around 280,000 human variants as detailed in the “Methods” section and sorted them in increasing order. These have been used as input for our variant enrichment analysis approach. Specifically, we searched among 282 KEGG pathways for those that contained a significantly enriched amount of variants with lower p values, i.e. variants on the respective pathway accumulating on top of the sorted list. For each pathway, permutation testing has been applied to assess a p value. Then, these p values were adjusted for multiple testing using the Benjamini–Hochberg approach. This procedure revealed a total of 65 KEGG networks with adjusted p values below 0.05.

Potential bias by length of genes or pathway size

Frequently, GWA study-based pathway analyses show a bias towards networks containing longer genes, e.g. if just the minimal p value of all variants in a gene is considered.

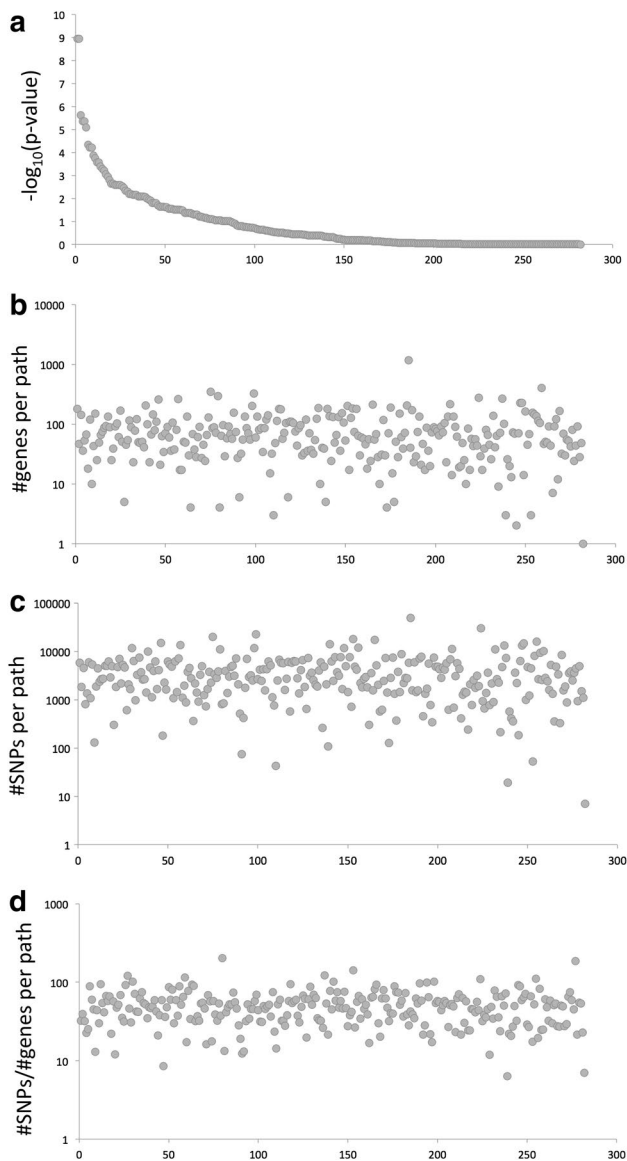


Fig. 1 **a** Negative decade logarithm of the p values for all KEGG paths in decreasing order (i.e. starting with the most significant pathway). In **b–d**, other characteristics of the pathways are plotted to discover a potential bias, e.g. for small or large paths. **b** Number of genes per pathway in the same order as in **a**. **c** Number of variants per pathway in the same order as in **a**. **d** Average number of variants per gene for each pathway in the same order as in **a**

Another potential bias is the pathway size itself. To check for such biases, we first sorted all pathways with respect to the adjusted significance value in increasing order. The negative decade logarithm of these p values is presented in Fig. 1a, starting from 9 ($p = 10^{-9}$) and monotonously decreasing towards 0 ($p = 1$). Next, we considered the number of genes on each pathway in the same order. The respective gene count is presented in Fig. 1b, demonstrating that there is no tendency for larger or smaller pathways depending on the significance value. Likewise, Fig. 1c

provides evidence that we also do not observe a bias towards the number of variants on each pathway, which is around one to two orders of magnitude above the gene count. Consequently, we also did not detect an accumulation of genes with many variants on significant pathways, as Fig. 1d details.

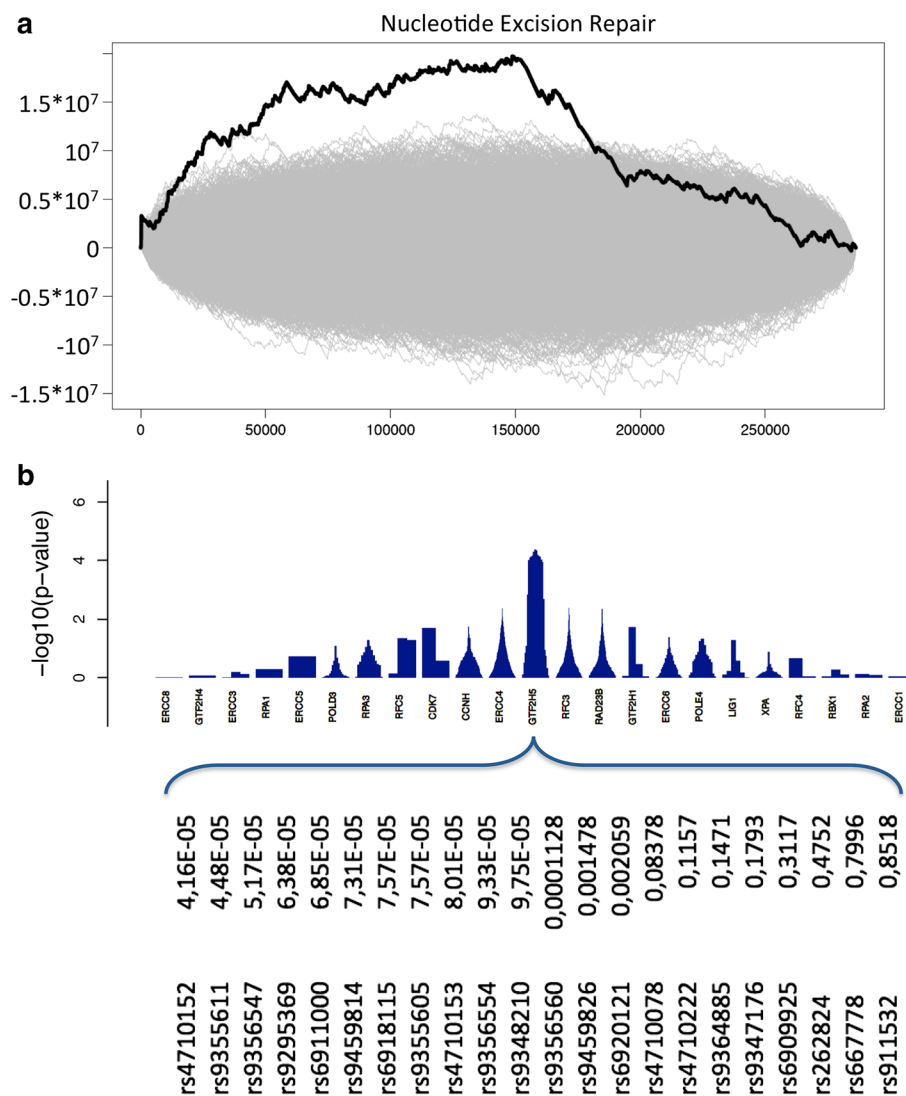
Significant KEGG pathways

Our calculations resulted in a total of 65 pathways that were significantly associated with SNPs from the DCM GWA study. The highest significance was found for “Nucleotide excision repair” and “Tuberculosis” pathways with adjusted p values of 10^{-9} . These two pathways consisted of 47 and 179 genes, respectively, carrying 1847 and 5784 variants from the GWA SNPs set. A complete list of all significant pathways with the number of genes on the pathways, the number of SNPs on pathways, the average number of SNPs per gene, z scores, p values and adjusted p values is presented in Supplemental Table 1.

For nucleotide excision repair, ERCC4, GTF2H5, RFC3 as well as RAD23B showed largest accumulation of variants. The running sum of this path is presented in Fig. 2a, along with the 5000 randomly generated distributions, which were all below the original running sum value. This figure presents the genes on the path with at least one variant. The gene with the largest number of significant variants was GTF2H5. For this gene, the variant ids are presented on the bottom of the figure along with the significance values. Of 22 variants in this gene on chromosome 6, which is overlapping with the TULP4 gene, 11 have p values $<10^{-4}$.

In the tuberculosis pathway, genes belonging to the HLA family dominated the enrichment, including HLA-DRB1, HLA-DQB1, HLA-DQA2 and HLA-DQA1. HLA-D are class II major histocompatibility surface receptors, which play an important role in immune surveillance by presenting peptide sequences to cells of the immune system. In recent years, diverse candidate association studies have linked distinct alleles of HLA-DQB1 to DCM (Li et al. 2012), while only recently the eQTL analysis of the GWA study being basis for this analysis has additionally linked HLA-C to the susceptibility to DCM. These findings may reiterate that DCM has not only a rich genetic basis (Haas et al. 2015), but also may arise in part from chronic autoimmune processes and inflammatory mechanisms due to abnormal immune regulation (Pankuweit et al. 2013). In line, we observed further pathways with adjusted p values below 10^{-4} . These contained multiple non-cardiac but inflammatory pathways, such as “Inflammatory Bowel Disease”. As described, we discovered a significant number of hits in inflammatory or autoimmune genes and pathways. Since an exclusion criterion for this study was (post)myocarditis and inflammatory cardiomyopathy, we hypothesise

Fig. 2 **a** Running sum statistics for the pathway “nucleotide excision repair”. The *black curve* presents the actual distribution. All variants in the data set are processed starting with the most significant one. Whenever a variant belonging to the nucleotide excision repair is discovered the running sum is increased, otherwise it is decreased. The *grey* distribution in the background presents 5000 permutation tests for the same category. **b** Manhattan plot-like representation of p value of variants on that pathway. Genes with lowest significance value are presented in the *middle* of the plot (e.g. most affected genes) while significance decreases towards the *left* and *right* edge. For each variant in the gene, one *bar* with the respective negative decade logarithm of p values is presented. Below the *chart*, the significance values for the most affected gene GTF2H5 are presented



that even in idiopathic/sporadic DCM cases, genetic mechanism modify the immune system and its reaction to several stressors and cardiomyocyte death. For instance, one finds a variety of autoantibodies against cardiac structures and proteins in idiopathic DCM (Xiao et al. 2011; Jane-wit et al. 2007; Korkmaz et al. 2013).

Example KEGG networks that can be related to pivotal cardiomyopathy pathways include the “oxidative phosphorylation” (adjusted $p = 0.001$), which may influence the cardiac energy metabolism. Beyond, the oxidative phosphorylation we also observed the KEGG network “cardiac muscle contraction” (adjusted $p = 0.007$). Moreover, the “vascular smooth muscle contraction” was likewise significant (adjusted $p = 0.007$). This pathway contains several genes that have been described in the literature with respect to DCM, such as RAF1 (Dhandapany et al. 2014). Likewise, adrenergic receptors that are members of the G protein-coupled receptor family

have been described to be of relevance in DCM (Shi et al. 2013). For the latter pathway, the running sum with permutation tests and Manhattan-like pathway plots are shown in Fig. 3.

While we observed several pathways that are described in the context of DCM or cardiomyopathies in general, just a part of all known DCM genes was discovered in our network analysis. This result is partially expected and in line with the general idea of the pathway analysis, i.e. to highlight less obvious molecular mechanisms. These may contribute to, e.g. the disease severity in addition to the established DCM genes.

In the previous analysis, we focused on pathways with large z scores and as a consequence low right-tailed p values. In contrast to these, we also discovered networks with very low z scores, i.e. pathways where almost no significant variants were contained at all. The lowest z score was found for Fatty Acid Metabolism (z score of -1.85 corresponding

Fig. 3 **a** Analogously to Fig. 2a, the running sum for vascular smooth muscle contraction is presented. **b** Analogously to Fig. 2b the Manhattan-like pathway representation for vascular smooth muscle contraction is shown

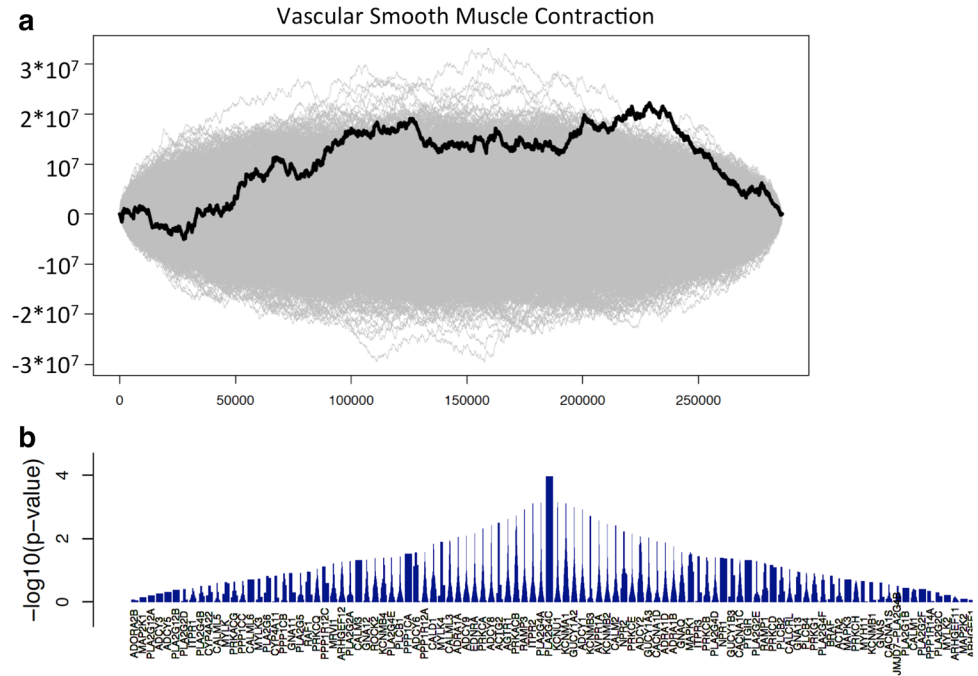
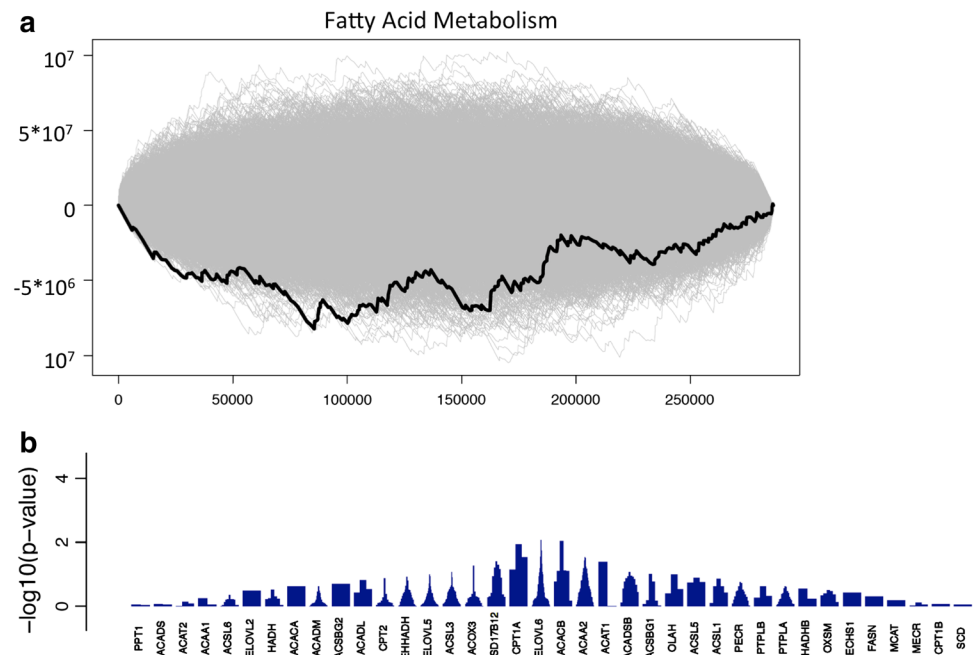


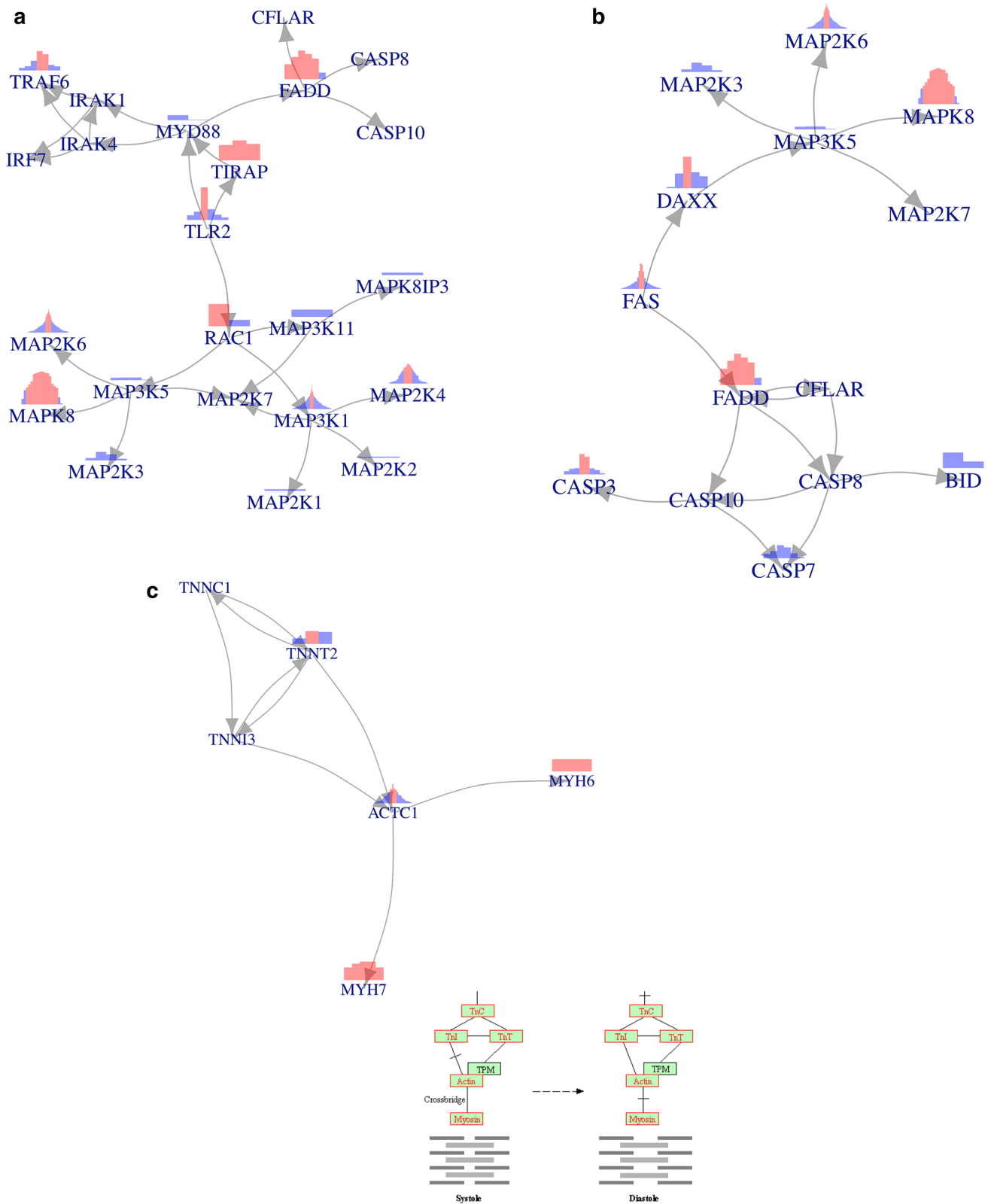
Fig. 4 **a** Running sum for fatty acid metabolism, a pathway that does not contain accumulation of significant variants but has a tendency to harbour variants that have higher p values. **b** Manhattan-like plot for the fatty acid metabolism. As compared to Figs. 2b and 3b, significance values are much larger (bar heights are much lower)



to right-tailed p value of 0.967). This example is presented with running sum and genes/variants in Fig. 4a, b.

As mentioned before, we previously systematically investigated different permutation test strategies for GWAS pathway analysis. The procedure in this study was, however, substantially different. Genes carrying one mutation were considered significantly affected and over-representation analysis has been carried out. Nevertheless, we compared the results of both approaches. While a large

number of pathways were discovered to be significant in both studies, including the focal adhesion, vascular smooth muscle contraction, the MAPK signalling pathway and also Parkinson disease, we observed also differences. One example is “cardiac muscle contraction”, which actually overlaps the vascular smooth muscle contraction by just three genes carrying variants in our study. In the present study, both pathways were significant while in the original study a much higher p value was observed for the cardiac



muscle contraction. This can be explained by the fact that in this category multiple genes carry variants missing the significance threshold slightly. By the hard cutoff of 0.05,

these genes were not considered in our initial analysis. Vice versa, the Dilated Cardiomyopathy pathway contains multiple genes that just contain one variant below the 0.05

◀ **Fig. 5 a** Connected component starting from gene TLR2. Each *bar* corresponds to a variant in the gene. *Red bars* are variants below the 0.05 *p* value threshold. TLR2 regulates TIRAP, which has an influence on the highly affected gene FADD. Likewise, TLR2 regulates RAC1, which in turn has an influence of several affected MAP kinases. **b** Related to the sub-network in Fig. 5a, FAS regulates DAXX, which in turn influences the same MAP kinases as RAC1. Moreover, FAS regulates FADD, which shows 5 of 6 significant variations. c TNNT2 and ACTC1 act together on MYH/and MYH6, belonging to the myosin filaments. The respective KEGG representation, a sub-part of the cardiac muscle contraction pathway, is presented on the *right lower* part of the figure

threshold. While all these genes were included in the initial analysis, the present analysis focusing on accumulation of variants does not lead to a high running sum.

Sub-network analysis

In the above-described pathway analysis, we found a substantial number of 65 KEGG networks (see details in Supplemental Table 1). This astonishing large number may be explained in that KEGG pathways are not independent from each other. Different genes can be part of different pathways. Thus, we asked for each gene on how many of the significant pathways the gene is located. Thereby we found several key players contributing to a magnitude of KEGG networks, including MAPK1 and MAPK3 (21 of 66 pathways), 7 different protein kinases (17 of 66 pathways) or TNF (16 of 66 pathways). These multiple hits could potentially bias the analysis. Beyond this, interactions that pass the pre-defined “borders” of KEGG networks are not considered. This analysis aims at finding the most affected sub-network independently of pre-defined KEGG networks. For each sub-network, we calculated the variants on this sub-network and applied the same calculations as for the pre-defined KEGG pathways. Following adjustment for multiple testing, 64 sub-networks with at least 4 genes remained significant. The maximal number of genes was as high as 42, while on average 12 genes were in each sub-network. Correspondingly, we found up to 1360 variants in a significant sub-network, the average number of variants was 253. On average, the largest fraction of variants being associated significantly with DCM was 17 %.

Focusing on sub-networks of a minimum size of at least 3 genes with a high fraction of significant variants as well as overall sub-network significance, we specifically focused on three core components, starting from FAS, TLR2 and TNNC1. Lowest significance values were obtained for the TLR2 component ($p = 0.0001$), consisting of 23 genes, 279 variants of which 52 (19 %) were significant. This component contains among others also FADD and TIRAP as well as multiple MAP kinases. Related to this sub-network, the FAS cascade with 14 genes, 145 variants of which 34 were significant (23 %) was found with a *p* value of 0.0004. The

FAS protein and parts of the cascade components have been described in the context of cardiomyopathies previously (Lu et al. 2010). FAS plays an important role in the death receptor pathway by binding to FADD, which can activate caspase-8, resulting in the initiation of an apoptotic signal. Cardiomyocyte apoptosis is involved in the progression of different cardiomyopathies (Communal et al. 2002; Purevjav et al. 2007). Finally, we found the TNNC1 component to be significant ($p = 0.0006$), although no variant was included in TNNC1. This component contains also TNNT2, ACTC1, MYH6 and MYH7. Parts of this cascade have been described by Hershberger et al. (2010).

With this analysis the above-mentioned bias in the KEGG pathway analysis that genes contribute in more than one network is not completely removed. Also in this analysis genes can be part of more than one sub-network. To further remove this bias the discovered significant parts can be combined to a consensus network, however. Even more, the algorithm can be called with different sub-network size criteria and the results of those can be combined to build one consensus network.

Another challenge with the sub-network analysis is the number of permutation tests. As described in the “Methods” part 500 permutations have been performed per sub-network based on calculations corresponding to our previous work (Backes et al. 2014). While for the presented showcase of DCM with the distance parameter set to three, 500 permutations were sufficient other cases may require more permutations to reach significant results other applications may require more permutations. This does not only relate to other input data sets but also to changed parameters such as the distance parameter in sub-networks. Thus, for each application the number of permutations should be carefully selected.

Conclusions

In this work, we presented a variant enrichment analysis for pre-defined KEGG networks as well as for specific sub-networks. Our approach has two key characteristics. First, we do not calculate gene level statistics but directly carry out the enrichment analysis on the variant level. Secondly, we perform the enrichment analysis not on a subset of variants but on an ordered list.

Our analysis of dilated cardiomyopathy (DCM) variants as showcase highlights the functionality of the approach. We observe KEGG networks that are closely related to DCM, which have also been discovered in our previous study. Moreover, our sub-network approach presents several core sub-networks that are enriched for DCM variants. The flexible implementation of the approach allows for easy adoption to other traits and is not restricted to DCM.

The sub-network approach has three advantages compared to the analysis, which just focuses on a priori-defined KEGG networks. First, genes can be located on multiple pathways, leading to false-positive hits. Although genes can be also located on different sub-networks, this bias can be reduced in building a consensus network. Second, signalling cascades could “cross the border” between different pre-defined networks. While both pathways alone could be non-significant, the respective cascade could be. Third, a small sub-network of a large network could be significant. Just by focusing on the large pathways, the respective small signalling cascade could be missed in the statistical analysis.

The presented method is thought to deliver complementary information to the standard GWAS and pathway analysis approaches that are frequently tailored to detect the most relevant SNP or gene. In future developments it is, for example, possible to include the information on significance of single variants and to carry out weighted gene set enrichment analysis. Beyond the present approach, dynamic programming, brute force search or global optimization-based techniques could be employed to calculate the paths in networks that are most significantly affected. Respective algorithms bear the potential for an even more precise discovery of affected molecular pathways.

Methods

Genome-wide association study

Our analysis is based on data obtained from Meder et al. (2014). In brief, 909 individuals of European descent with DCM that have been recruited between 2005 and 2008 in Heidelberg were included as well as 2120 controls individuals.

As DCM cases we recruited patients with idiopathic and sporadic DCM and excluded DCM families, since this would potentially affect association statistics. All patients of the GWA screening cohort but one are Caucasians (except one proband with Asian ethnicity). The control individuals have been provided by PopGen and KORA. PopGene is a biobank and databank that contains individuals from the northern part of Schleswig–Holstein and has been set up with the intention to understand the genetic mechanisms of complex disorders. Overall, PopGene contains data and samples of over 10,000 individuals. Details on PopGene are available here: <http://www.uni-kiel.de/medinfo/institut/profil/popgen/>. Similarly, KORA (Cooperative Health Research in the Region Augsburg, established in 1996) is a regional platform for population-based surveys and subsequent follow-up studies. Areas of interest include the fields of epidemiology, health economics, and

health care research. The KORA infrastructure and further details such as data management and quality control aspects are described by Holle et al. (2005).

Case–control association tests were carried out by the PLINK software package version 1.07. Additionally, SNPs with minor allele frequencies of $<3\%$ or call rates $\leq 95\%$ were removed. Furthermore, all SNPs not fulfilling Hardy–Weinberg equilibrium (considering significance level of 0.05 for controls and 0.001 for cases) were discarded such that around 280,000 of all tested SNPs on the array have been included in the analysis. For mapping SNPs to genes information provided by the array vendor Affymetrix have been used. All pathway analyses have been carried out on the full set of genes without pre-selecting gene subsets (e.g. selecting genes that are described with the considered disease already). The significance values were sex and age-corrected values that have been derived by logistic regression. The median of all SNPs divided by the median of a Chi-square distribution with 1° of freedom was calculated to estimate the genomic inflation factor. Thereby, p values were corrected for genomic control (GC) to effectively correct for population stratification (Devlin and Roeder 1999). Details on the data processing are also provided in Meder et al. (2014).

Enrichment analysis

Frequently, GWA study-based enrichment analysis is carried out using two different statistics. First level statistics aggregates variants to genes while second level statistics aggregates genes to pathways. Thus, two mathematical functions are applied. In our approach, an adopted variant of the second function, similar to gene set enrichment analysis, is calculated while no aggregation of variants to genes occurs. Specifically, we first match variants to pathways. Whenever a variant is located at least in a single gene on that pathway, the variant is defined to participate in the pathway. This in turn means that variants located in multiple genes on the same pathway are counted exactly once. Next, variants are ordered according to their significance value with respect to the case control status as described in the previous section. Then, analyses similar to unweighted gene set enrichment (GSEA) are carried out. Here, however, not a sorted list of approx. 20,000 genes is used as input but the sorted list of approx. $m = 280,000$ variants. This list is processed for each pathway p from top to bottom. Assuming that l variants are located on p , a running sum is calculated by processing the sorted list from lowest p values to highest. In each step, the running sum that is initialized by 0 is decreased by l . Whenever a variant is hit that is located on p , the running sum is increased by $m - l$. Thus, after m steps the running sum equals $0 + l \times (m - l) - (m - l) \times l = 0$. As database we used the

Kyoto Encyclopedia of Genes and Genomes KEGG (Kanehisa and Goto 2000), one of the most widely used pathway resources in genomics. From KEGG, we extracted for 282 pathways all genes participating in these pathways and all variants that are located in these genes.

Permutation tests

To validate the significance of results from pathway analyses, re-sampling approaches are commonly applied. We previously explored the influence of different permutation test strategies and found that permuting the case–control status gives similar results as permuting the variants p values. We furthermore found that around 1000 permutations were sufficient to estimate ORA p values in a stable manner. For the present analysis we carried out 5000 permutations of the variant significance values and calculated the running sum as described in the previous section. We then calculated the mean value and standard deviation of the maximum running sum value of the 5000 permutations per pathway. From these two values and the original maximal running sum value we calculated p values using a normal distribution as approximation. The respective p values per pathway were adjusted for multiple testing using the Benjamini–Hochberg approach.

Sub-network analysis

Since genes frequently overlap between KEGG pathways and interacting genes can be located on different pre-defined KEGG pathways we implemented a sub-network-based analysis. We build a directed graph where the genes are the nodes and interactions between genes are the edges. Gene complexes and gene families were not included as nodes. For each node we searched for a connected component with a maximal distance of n interactions. Testing different parameters for n (from 2 to 5) we finally used a maximal distance from the current node of three. The value of three has been chosen because of the average size of the sub-networks. While a value of 2 leads to very narrow sub-networks frequently consisting of only three genes, a value of four often yielded very large sub-networks exceeding 100 genes. With a value of three the average sub-network size was 12. The connected component calculation has been carried out not only on the actual KEGG interaction graph but also on an undirected version of the graph with similar results. Following this approach, for each gene a sub-network was constructed with the corresponding node as source/central node and all nodes in the sub-network were at most three interactions away from this central node. For each sub-network we calculated the variants on this sub-network and applied the same calculations as for the pre-defined KEGG pathways. For this analysis,

500 permutation tests were carried out and z scores have been calculated in the same manner as for the pre-defined KEGG paths. The number of permutations has been decreased from the original analysis by a factor of ten for different reasons, including runtime considerations. Nonetheless, a sensitivity analysis corresponding to the approach in Backes et al. (2014) has demonstrated that 500 permutations were in the case of the sub-network analysis on our data set and with a distance parameter of 3 sufficient. From the 500 permutations, p values have been derived using a normal distribution and the significance values were adjusted for multiple testing using the Benjamini–Hochberg method.

Acknowledgments The work of AK, BM is supported by the European Union FP7 (BestAgeing). BM is grateful for support from the German Center for Cardiovascular Research (DZHK).

Compliance with ethical standards

Conflict of interest The author(s) declare that they have no conflict of interests.

References

- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25(1):25–29
- Backes C, Ruhle F, Stoll M, Haas J, Frese K, Franke A, Lieb W, Wichmann HE, Weis T, Kloos W et al (2014) Systematic permutation testing in GWAS pathway analyses: identification of genetic networks in dilated cardiomyopathy and ulcerative colitis. *BMC Genom* 15:622
- Braun R, Buetow K (2011) Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. *PLoS Genet* 7(6):e1002101
- Communal C, Sumandea M, de Tombe P, Narula J, Solaro RJ, Hajjar RJ (2002) Functional consequences of caspase activation in cardiac myocytes. *Proc Natl Acad Sci USA* 99(9):6252–6256
- da Huang W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1):1–13
- Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55(4):997–1004
- Dhandapany PS, Razzaque MA, Muthusami U, Kunnoth S, Edwards JJ, Mulero-Navarro S, Riess I, Pardo S, Sheng J, Rani DS et al (2014) RAF1 mutations in childhood-onset dilated cardiomyopathy. *Nat Genet* 46(6):635–639
- Gui H, Li M, Sham PC, Cherny SS (2011) Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's disease dataset. *BMC Res Notes* 4:386
- Haas J, Frese KS, Peil B, Kloos W, Keller A, Nietsch R, Feng Z, Muller S, Kayvanpour E, Vogel B et al (2015) Atlas of the clinical genetics of human dilated cardiomyopathy. *Eur Heart J* 36(18):1123a–1135a
- Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR et al (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308(5720):419–421

- Hershberger RE, Norton N, Morales A, Li D, Siegfried JD, Gonzalez-Quintana J (2010) Coding sequence rare variants identified in MYBPC3, MYH6, TPM1, TNNC1, and TNNI3 from 312 patients with familial or idiopathic dilated cardiomyopathy. *Circ Cardiovasc Genet* 3(2):155–161
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106(23):9362–9367
- Holle R, Happich M, Lowel H, Wichmann HE, Group MKS (2005) KORA—a research platform for population based health research. *Gesundheitswesen* 67(Suppl 1):S19–S25
- Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* 85(1):13–24
- Jane-wit D, Altuntas CZ, Johnson JM, Yong S, Wickley PJ, Clark P, Wang Q, Popovic ZB, Penn MS, Damron DS et al (2007) Beta 1-adrenergic receptor autoantibodies mediate dilated cardiomyopathy by agonistically inducing cardiomyocyte apoptosis. *Circulation* 116(4):399–410
- Joshi-Tope G, Vastrik I, Gopinath GR, Matthews L, Schmidt E, Gillespie M, D'Eustachio P, Jassal B, Lewis S, Wu G et al (2003) The Genome Knowledgebase: a resource for biologists and bioinformaticists. *Cold Spring Harb Symp Quant Biol* 68:237–243
- Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30
- Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc database. *Nucleic Acids Res* 30(1):59–61
- Keller A, Backes C, Lenhof HP (2007) Computation of significance scores of unweighted Gene Set Enrichment Analyses. *BMC Bioinform* 8:290
- Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 8(2):e1002375
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308(5720):385–389
- Korkmaz S, Zitron E, Bangert A, Seyler C, Li S, Hegedus P, Scherer D, Li J, Fink T, Schweizer PA et al (2013) Provocation of an autoimmune response to cardiac voltage-gated sodium channel NaV1.5 induces cardiac conduction defects in rats. *J Am Coll Cardiol* 62(4):340–349
- Li X, Luo R, Jiang R, Chen R, Hua W (2012) Human leukocyte antigen-DQ beta 1 chain (DQB1) gene polymorphisms are associated with dilated cardiomyopathy: a systematic review and meta-analysis. *Heart Lung J Crit Care* 41(4):360–367
- Liu G, Jiang Y, Wang P, Feng R, Jiang N, Chen X, Song H, Chen Z (2012) Cell adhesion molecules contribute to Alzheimer's disease: multiple pathway analyses of two genome-wide association studies. *J Neurochem* 120(1):190–198
- Lu D, Lian H, Zhang X, Shao H, Huang L, Qin C, Zhang L (2010) LMNA E82K mutation activates FAS and mitochondrial pathways of apoptosis in heart tissue specific transgenic mice. *PLoS One* 5(12):e15167
- Meder B, Rühle F, Weis T, Homuth G, Keller A, Franke J, Peil B, Lorenzo Bermejo J, Frese K, Hüge A et al (2014) A genome-wide association study identifies 6p21 as novel risk locus for dilated cardiomyopathy. *Eur Heart J* 35(16):1069–1077
- Pankuweit S, Ruppert V, Jonsdottir T, Müller HH, Meyer T, German Competence Network of Heart F (2013) The HLA class II allele DQB1 0309 is associated with dilated cardiomyopathy. *Gene* 531(2):180–183
- Purevjav E, Nelson DP, Varela JJ, Jimenez S, Kearney DL, Sanchez XV, DeFreitas G, Carabello B, Taylor MD, Vatta M et al (2007) Myocardial Fas ligand expression increases susceptibility to AZT-induced cardiomyopathy. *Cardiovasc Toxicol* 7(4):255–263
- Saccone SF, Bolze R, Thomas P, Quan J, Mehta G, Deelman E, Tischfield JA, Rice JP (2010) SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res* 38(Web Server issue):W201–W209
- Shi T, Moravec CS, Perez DM (2013) Novel proteins associated with human dilated cardiomyopathy: selective reduction in alpha(1A)-adrenergic receptors and increased desensitization proteins. *J Recept Signal Transduct Res* 33(2):96–106
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102(43):15545–15550
- Wang K, Li M, Bucan M (2007) Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 81(6):1278–1283
- Xiao H, Wang M, Du Y, Yuan J, Cheng X, Chen Z, Zou A, Wei F, Zhao G, Liao YH (2011) Arrhythmogenic autoantibodies against calcium channel lead to sudden death in idiopathic dilated cardiomyopathy. *Eur J Heart Fail* 13(3):264–270