Statistical methods for the analysis of complex omics data

Tissier, R.

4

Gene co-expression network analysis for family studies based on a meta-analytic approach

Abstract

For a better understanding of the biological mechanisms involved in complex traits or diseases, networks are often useful tools in genetic studies: coexpression networks based on pairwise correlations between genes are commonly used. In case of a family-based design, it can be problematic when there is a large between-family variation in expression levels. We propose here a gene coexpression network analysis for family studies. We build a coexpression network for each family and then combine the results. We applied our approach to data provided for analysis in the Genetic Analysis Workshop 19 and compared it to 2 naive approaches-ignoring correlations among the expressions and decorrelating the gene expression by using the residuals of a mixed model and a single-probe analysis. Our approach seemed to better deal with heterogeneity with regard to the naive approaches. The naive approaches did not provide any significant results, while

This chapter has been published as: Renaud Tissier, Hae-Won Uh, Erik van den Akker, Brunilda Balliu, Spiridoula Tsonaka, Jeanine J. Houwing-Dusitermaat (2016). Gene co-expression network analysis for family studies based on a meta-analytic approach. *BMC Proceedings*; 10(Suppl 7): 119-123.

our approach detected genes via indirect effects. It also detected more genes than the single-probe analysis.

4.1 Background

Weighted gene co-expression network is a widely used method for studying biological networks based on pairwise correlations. This method provides more insight in the underlying biological mechanisms and offers a tool for dimension reduction by summarizing identified modules (clusters) of genes (Plaisier et al., 2009; de Jong et al., 2012). How to perform such an analysis for family data is an open question. For family data Kraft et al. (2003) noted that testing association between expression levels and traits without taking into account the family structure can lead to spurious results, especially when the number of families is small and in the presence of large between-family variation. In this paper, we propose a novel strategy for network analyses in a small set of relatively large families. For this family-based approach, we first construct family-specific co-expression networks and test for association between the modules and the traits of interest. Common set of genes for all families were obtained by using the intersection and the union of family specific modules. We compare this family-based approach with two naive approaches: namely, one using the gene expression of the families directly (ignoring correlation) and one that first decorrelates the gene expressions and then applies the standard approach. We also compare our results with single probe analyses.

4.2 Methods

4.2.1 Study sample

The gene-expression dataset is composed of 647 individuals from 17 large families. These samples are from the dataset described in Almasy and Blangero (1998). Here, we focus on the largest 5 families: namely family 2, 5, 6, 8 and 10 with 65, 55, 45, 62 and 49 family members, respectively. The total number of individuals is 276. In total gene expressions of 20634 probes are available. We used the simulated quantitative phenotypes Systolic Blood Pressure (SBP) and the phenotype Q1 at time point 1 as outcome variables. The simulation model of SBP comprises 15 genes and that of Q1 does not contain any of these genes. SBP, Q1 and all probes were corrected for age and sex by regressing out covariates and using residuals.

In order to decorrelate the gene expressions, we fitted for each probe a linear mixed model: $X_{ij} = + u_{ij} + v_i + _{ij}$, with X_{ij} the value of the probe for the individual j in family i, u_{ij} a normally distributed random genetic effect: $u_i j \; N(0, S)$ where $S = 2 * K * s_g$ with K kinship matrix and $s_g$genetic variance, v_i a normally distributed random effect representing shared environmental effects, and $_{ij}$ a normally distributed residual. To obtain the residuals X^*_{ij} of this model we used the function lmekin, which fits linear mixed models with specific structure of the variance-covariance matrix from the package coxme (Therneau, 2018) in R.

4.2.2 Single probe analysis

For the single probe analysis the following mixed model was used:

$$Y_ij = \; + u_ij + v_i + X_ij + {}_ij$$

with Y_{ij} the value of SBP or Q1 and X_{ij} the value of the probe for individual j of family i. The random effects u_{ij}, v_i and ${}_{ij}$ are the genetic effect, the shared environmental effect and residuals respectively. The parameter β represents the effect of the probe on the outcome variable.

4.2.3 Network constructions

Co-expression networks were built on the dataset without correction for family structure based on X_ij (naive approach), the dataset adjusted for family structure based on X_ij^* (naive decorrelated approach), and on the datasets from the five families separately.

We used signed co-expression networks. The adjacency matrix $A = [a_{lk}]$ of each network was computed as follows: $a_{lk} = |0.5 + 0.5cor(x_l, x_k)|^\gamma$, with $cor(x_l, x_k)$ the correlation between x_l the values vector of probe l and x_k the values vector of probe k. The parameter γ is acting as a soft threshold in the adjacency matrix, when we increase the value γ the coefficient of the adjacency matrix will tend to 0 except for values really close to 1. We used the biweight midcorrelation based on the median, which is more robust than the Pearson correlation. The co-expression networks were constructed with the R package WGCNA (Langfelder and Horvath, 2008). For each obtained module, the first principal component (eigengene) was computed.

4.2.4 Phenotype analysis

From all modules and all families, the following models were fitted:

$$Y_j = \; + u_j + \beta eigengene_j^k + {}_j,$$

where Y_j is the outcome, u_j the random genetic effect and $eigengene_j^k$ the value of the eigengene of module k of family member j. Let E_{F2}^M to E_{F10}^M be the most significant eigenvalues of the family specific networks (N_F2 to N_F10) and let E_F^M be the most significant eigenvalue of these five eigenvalues and M_F^M be the corresponding module. Identify the modules of the family-specific networks, which have the highest overlap with M_F^M (denoted as M_{F2}^O to M_{F10}^O). Next, two common sets of genes for all families were obtained by taking the intersection ($M_F = M_{F2}^O \cap M_{F5}^O \cap M_{F6}^O \cap M_{F8}^O \cap M_{F10}^O$) and the union ($M_F = M_{F2}^O \cup M_{F5}^O \cup M_{F6}^O \cup M_{F8}^O \cup M_{F10}^O$) of the family specific modules. The first principal components of the two common sets were computed. The principal component that explained most of the variance of the corresponding set of genes was used as the eigengene EF of the family based approach.

The eigengenes of the naive approach (EN), the naive approach after decorelation (END) and the family-based approach (EF) are tested for association with the two phenotypes SBP and Q1. Here, the following mixed model was used:

$$Y_{ij} = \;+ u_{ij} + v_i + \beta eigengene_{ij}^{k} + {}_{ij}$$

with Y_{ij} the phenotype value for individual j of family i and $eigengene_{ij}^{k}$ the value of eigengene of module k of individual j of family i. And u_{ij}, v_i and ${}_{ij}$ are the genetic effect, the shared environmental effect and residuals respectively. The parameter β represents the effect of the eigengene k on the outcome variable.

Finally since spurious associations are especially expected in the presence of large between family heterogeneity (Kraft et al., 2003) we also performed a network analysis using the subset of 25% most heritable probes when performing the network analysis (n=4911 probes with heritability between 0.33 and 0.88).

To test for significance we used a nominal alpha level of 0.05 and the Bonferroni correction was applied to take into account multiple testing.

4.3 Results

4.3.1 Results obtained with all probes

For per family analysis, the module that showed the highest correlation with the SBP was the magenta module obtained in family 8 (M_{F8}^{M}) (β=2.52, p=0.0011). M_{F8}^{M} comprises 710 genes. For each family, the number of genes of the module with the highest overlap is given in Table 4.1. The intersection and the union of these five family modules, comprises 62 and 1746 probes respectively. The first principal component (eigengene) of the probes in the intersection set explained more than 50% of the variance for each family, while for the union set the eigengene explained only between 23% and 31% of the variance of the expression levels. Therefore the eigengene of the intersection set was used as summary for the family approach (E_F). In Table 4.2, for each family the effect of E_{Fi} on SBP (β of model (2)) is given. For families 2 and 8, the eigengenes (E_{F2} and E_{F8}) were significantly associated with SBP.

When analysing all families together none of the approaches provided significant results. The joint analysis of the families using EF as eigengene in model (3) did not provide a significant association SBP (β=-0.13, p=0.49). For the naive approach, the eigengene of the module magenta (E_N) had the smallest p-value (β=-3.21, p=0.01). For the naive approach using the decorrelated dataset, the eigengene of the module grey60 (E_{ND}) had the smallest p-value (β=-3.03, p=0.0061). After multiple-testing correction (between 43 and 50 modules in each network) none of the results were significant. Finally the single

	M^O_{F2}	M^O_{F5}	M^O_{F6}	M^O_{F8}	M^O_{F10}
Module size	446	694	499	710	446
Size of the overlap with M^M_F	187	308	240	710	372

Table 4.1: Module size of M^O_{F2} to M^O_{F10} and overlap size with M^M_F in the all-probes analysis

	All probes		25% most heritable probes	
	SBP	Q1	SBP	Q1
E_{F2}	-0.57(0.2) (0.02)^a	9.90(4.3)(0.02)	0.27(0.1)(0.07)	-1.62(1.0)(0.11)
E_{F5}	0.34(0.2)(0.21)	14.0(4.7)(3.3e-3)	0.18(0.2)(0.41)	-2.13(1.3)(0.11)
E_{F6}	0.08(0.3)(0.78)	8.90(3.7)(0.02)	0.66(0.3)(0.01)^a	2.49(0.9)(9.6e-3)
E_{F8}	-0.62(0.3)(0.04)^a	10.47(4.2)(0.01)	0.07 (0.2)(0.68)	2.47(1.0)(0.02)
E_{F10}	0.14(0.3)(0.67)	7.55(4.5)(0.09)	0.02(0.2)(0.91)	-2.22(1.2)(0.06)
E_F	-0.13(0.2)(0.49)	-	0.21(0.09)(0.02)^a	-
E_N	-3.21(1.3)(0.01)	2.75(0.7)(5.6e-4)^a	1.93(0.8)(0.01)	-0.96(0.5)(0.06)
E_{ND}	-3.03(1.1)(6.1e-3)	-1.41(0.4)(9.4e-4)^a	1.94(0.7)(5e-3)^a	-0.41(0.2)(0.06)

Table 4.2: Parameter estimates of the association between eigengenes and Q1 and SBP. In parentheses are standard errors and p values, respectively. For E_{F2} to E_{F10} model (2) was used; for E_F, E_N and E_{ND} model (3) was used. For Q1 the association results for E^M_{F2} to E^M_{F10} are presented. ^{a} Denotes significant test after multiple testing corrections.

probe analysis preformed in the five families by using model (1) provided one significantly associated probe with SBP (CRIP2; β= -13.68, p= 1.7e-06).

The intersection module of the family based approach did not contain any of the 15 genes used for the simulation. Also the identified gene of the single probe analysis is not among these 15 genes. We hypothesized that correlation might exist between E_{F2}, E_{F8}, and the gene expression of CRIP2 on one hand and the set of 15 genes on the other hand. Indeed E_{F2} showed significant correlation with PSMD5 (p=0.004) and GTF2IRD1 (p=0.007) and E_{F8} showed significant correlation with ZNF443 (p=5e-05), PSMD5 (p=3e-05) and ABTB1 (p=6e-05). When the presence of these 15 genes in the modules was investigated, it appeared that they were in different modules (see Table 4.3). The gene CRIP2 which was significant in the single probe analysis showed significant correlation with the gene KRTAP11-1 (p= 3.1e-03).

4.3.2 Analysis of Q1

The results of the analysis of Q1 are also given in Table 4.2. For the family approach, none of the modules obtained in family-specific network analysis was significantly asso-

	N_N	N_{ND}	N_{F2}	N_{F5}	N_{F6}	N_{F8}	N_{F10}
MAP4	-	-	-	-	-	7	-
NRF1	-	-	-	1	-	-	-
TNN	11	-	19	5	14	-	-
LEPR	-	1	19	-	-	-	-
FLT3	5	-	-	8	4	-	1
GTF2IRD1	-	4	13	3	-	-	-
FLNB	9	-	16	21	13	-	2
ZNF443	8	-	5	1	23	6	1
GSN	2	15	3	-	-	1	-
CABP2	11	-	-	5	14	2	16
LRP8	-	-	6	-	-	12	-
PSMD5	3	1	18	10	28	17	-
GAB2	20	15	1	3	22	-	5
ABTB1	3	-	4	4	1	2	2
KRTAP11-1	4	19	2	1	18	4	1

Table 4.3: List of the top genes involved in the simulation model and their module number in each network. -, Denotes the grey module in which all nonclustered genes are combined. The different colours represent genes in the same module for a specific network

ciated with Q1 and no common set could be defined. In table 4.2 the estimates of strongest associated modules E_F^M for each family are given. For the naive approach, the module red (β=2.75, p=0.00056) was significant and for the naive approach using the decorrelated data the module green (β=-1.41, p=0.00094) was significantly associated with Q1.

4.3.3 Results obtained with the 25% most heritable probes

For the naive and the family approaches, the results of the network based analyses using only the gene expressions of the 25% most heritable probes (n=4911 probes with heritability between 0.33 and 0.88) are also given in Table 4.2. None of the 15 genes used in the simulation model for SBP was among this set of most heritable probes. For Family 6 the E_{F6} was significantly associated with SBP (p=0.01). The association of E_F in the five families with SBP was also significant (p=0.02). For Q1 none of the approaches provided significant results. With regard to the single probe analysis, no other probes than CRIP2 was significantly associated.

4.4 Discussion

In this paper, we have proposed a novel strategy to perform a co-expression network analysis with family data: building a network for each of the large pedigrees, and defining a common module by taking the intersection of family specific modules. We compared our family-based approach with two naive network approaches and a single probe analysis. All analyses were performed in a small set of five relatively large families. None of the 15 genes in the simulation model was identified in this small dataset. However the family-based approach identified significant associations between the eigengene and SBP in two families. This eigengene was significantly correlated with 4 of the 15 genes. When analyzing all families jointly the family eigengene was not significant. Also the naive network approaches did not provide any significant result. The single probe analysis provided one significant gene which was correlated with one of the 15 genes. To study the performance of the methods with regard to false positive findings, we also analyzed the trait Q1 for which no gene expressions were included in the simulation model. The family approach did not provide a significant finding, while both naive approaches identified a significant module for Q1. The result in the naive approach based on gene expression (X_{ij}) is in line with the findings of Kraft et al. (2003). We did not expect to have a false positive finding when using the decorrelated data (X^*_{ij}) as input for our network analysis. Possible explanations for this finding are the fact that the correlation based on the kinship coefficient might not be appropriate for gene expressions, and randomness. In addition to the set of all probes, also networks were built using only the 25% most heritable probes. Especially for these variables that show large between-family variation spurious associations might occur when the family structure is not taken into account. This was not confirmed in our analysis. More research is needed to study the sensitivity of the methods for between-family variation.

We did not know the answers when we developed the family-based approach and analyzed the data. The simulation model used to create the datasets may not be well suited to pick up the 15 genes directly by network analysis. The 15 genes present in the model were in different pathways: they were not correlated. Moreover our approach was able to identify indirect effects: i.e. the eigengenes were correlated with the 15 genes. Thus the significant association of the family based network approach represented the largest number of genes from the simulation model. We expect that especially in the presence of large between-family variation our approach would perform best. A thorough simulation study is required to investigate the performance of our method further.

Network analysis provides a tool to reduce the number of tests by first summarizing the data in sets of genes with correlated gene expressions and summarizing the gene set by the first principle component. Another obvious reduction step is to only consider the heritable probes for the analysis. It appeared that by using the heritable probes the results across the families were less heterogeneous. The family approach as well as the naive approach using decorrelated data provided significant results for SBP. In this paper we combined the family-specific modules by taking the intersection of the modules which showed most overlap. This approach worked well for the relatively small set of five families. When we applied our method to the six largest families, similar results were obtained (data not shown). However intersection might not be the most appropriate approach to combine modules across families, because the intersection set becomes too small. Alternative approaches have to be developed. For example lasso type of methods can be used to select probes from the union sets. Development of a method for constructing a common set from the family specific modules is a topic for future research. Finally more research is needed to evaluate the performance of our method with regard to false positive and false negative findings in relationship to heterogeneity, family size, the number of families and the heritability of gene expressions.