Acute wheeze-specific gene module shows correlation with vitamin D and asthma medication

Shintaro Katayama [1*], Katarina Stenberg Hammar [2,3*], Kaarel Krjutškov [1, 4, 5], Elisabet Einarsdottir [1, 5, 6], Gunilla Hedlin [2, 3], Juha Kere [1, 5, 7], Cilla Söderhäll [1, 3§]

Affiliations

[1] Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden;

[2] Astrid Lindgren Children´s Hospital, Karolinska University Hospital, SE-171 64 Stockholm, Sweden;

[3] Department of Women´s and Children´s Health, Karolinska Institutet, SE-171 77 Stockholm, Sweden;

[4] Competence Centre on Health Technologies, Tartu EE-50410, Estonia

[5] Folkhälsan Institute of Genetics, and Stem Cells and Metabolism Research Program, University of Helsinki, FI-00014 Helsinki, Finland

[6] Current affiliation: SciLifeLab, Department of Gene Technology, KTH-Royal Institute of Technology, SE-171 21 Solna, Sweden

[7] School of Basic & Medical Biosciences, King's College London, Guy's Hospital, London SE1 9RT, UK

*= shared first authorship

§Corresponding author

Cilla Söderhäll, Department of Women´s and Children´s Health, Karolinska Institutet, Karolinska vägen 37A, Q2B:04, SE-171 77 Stockholm, Sweden Tel: +46 852481058.

Email: cilla.soderhall@ki.se

Supplementary methods

Material and methods

Study design and subject enrollment

Children in this study are part of a longitudinal study on preschool children with wheezing enrolled between 2008 and 2012, recruited consecutively when visiting the Paediatric Emergency Department at Astrid Lindgren Children's Hospital, Stockholm, Sweden as a result of acute wheezing. Diagnosis of acute wheeze was based on a clinical diagnosis made by the treating physician at the Pediatric Emergency Department. The enrolment criteria were confirmed by the study doctor. Of children with acute wheeze, 80% were hospitalised for at least 24h [1]. The children came back for a revisit 2-3 months later (median 12 weeks), and thereafter annually to the same paediatrician and allergologist (study doctor KSH) until school-age. This study is still ongoing with follow-ups. The children are well characterized with clinical examinations, standardized questionnaires, and biological sampling at all visits. Guardians and children responded to questions in structured interviews concerning medication, contact with healthcare, days of absence due to illness until first follow up 2-4 mo later [1] and also during the year preceding each visit. They also reported symptoms of allergy and eczema at each visit. Lung function tests at age 7 yrs were performed. For inclusion and exclusion criteria see Table1. Included in this study are the acute visit (transcriptomics and clinical information), the first revisit after 2-4 months (transcriptomics and clinical information), and the annual visit at 7 years of age (clinical information). Age-matched healthy control children were recruited at the Surgical Day-care Ward, Astrid Lindgren Children´s Hospital. For the study design see Figure 1, and for inclusion and exclusion criteria see Table 1. In total, 334 samples were included in the transcriptome study (Acute wheeze (ACW) n=138, revisit (REV) n=114, healthy controls (CTRL) n=83). Details of some of the definitions of clinical parameters are found below, and further explanations are found in Table E1.

Definitions of clinical parameters

For inclusion and exclusion criteria see Table1, and for clinical parameters see TableE1. Diagnosis of acute wheeze was based on a clinical diagnosis made by the treating physician at the Pediatric Emergency Department, whereof 80% were hospitalised for at least 24h. At the follow up at 7 years of age, all children were examined by the study doctor (KSH) and assessed for the diagnosis of asthma. Asthma at 7 years of age (7Y_ASTHMA_GA2LEN) was defined as a positive answer to either the question; Have you had an attack of asthma in the last 12 months?' OR the question "Are you currently taking or have you during the last 12 months taken any medication for asthma, including

short-acting b2-antagonists, inhaled corticosteroids, and montelukast?", modified from[2]. In addition, allergic asthma (7Y_ASTHMA_ALLERGIC) was defined as asthma with allergic sensitization and clinical symptoms of allergy until the age of 7 years. (7Y_FEV1%_FVC_RATIO) the ratio between FEV1%/FVC at the 7 years visit. FEV1% = Percent of the expected forced expiratory volume during 1s, FVC = Forced vital capacity. (7Y_LTRA) Leukotriene receptor antagonist medication the year preceding the 7 years visit. (7Y_ASTHMA_CONTROL_TEST) Self-reported asthma control test at the 7 years visit was assessed using the ACT[3].

Sampling

For the wheezing children blood samples and nasopharyngeal swab samples were obtained at the acute visit as well as at the follow-up visit 2-3 months later (median 12 weeks). For the age-matched healthy control children blood was drawn at the same time as an intravenous line was inserted prior to surgery and anaesthesia. The legal guardian filled out a standardized questionnaire (cases and controls), as detailed previously [1, 4].

Laboratory analysis

Blood samples were analyzed for total blood cell counts at the Karolinska University Hospital Laboratory at all visits. Presence of RV was detected by PCR in the nasopharyngeal samples, as described elsewhere [4]. The levels of bound antigen-specific antibodies against recombinant VP1 proteins from RV2, 16, 89 (RV-A), RV14 (RV-B) and RV-YP (RV -C), were previously analysed in plasma samples at the acute visit and the follow-up visit 2-3 months later (median 12 weeks) as described elsewhere [1, 4]. Vitamin D; The levels of 25-hydroxyvitamin D (25(OH)D) was assessed using direct, competitive chemiluminescence analysis (CLIA; DiaSorin Inc, Stillwater, MN, USA), as described elsewhere [1].

RNA extraction

Total RNA was extracted from white blood cells (buffy coat) using RiboPure-Blood extraction kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions. For RNA extraction, white blood cells were freshly isolated from the blood, immediately put into RNA*later* (Thermo Fisher Scientific) and stored at -20° and -80°C until RNA extraction. RNA quality and quantity were assessed using NanoDrop 8000 (Thermo Fisher Scientific), Qubit Fluorometric Quantitation (Thermo Fisher Scientific) and Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). RIN-value >8 was used as cut-off for inclusion.

RNA sequencing including GlobinLock and statistical analyses

80 ng of total RNA from each individual was added to the library preparation. In total, 334 samples were included (Acute wheeze ACW n=138, healthy controls CTRL n=83, revisit REV n=114) and subdivided into eight 48-plex libraries. Peripheral blood leukocyte RNA samples were first treated by GlobinLock® oligonucleotides [5] to exclude highly abundant globin mRNA molecules from the cDNA synthesis, followed by RNA sequencing library preparation according to the Single-cell Tagged Reverse Transcription (STRT) method [6, 7]. In detail, human globin mRNA alpha 1 and 2 (*HBA,* hereafter α) and beta (*HBB*, β) were first denatured and subsequently locked by specific oligonucleotides to mask the binding site of the anchored oligo-dT primer. As a result of this, the GlobinLock treatment significantly reduces the amount of sequences from *HBA* and *HBB*, making direct whole blood full transcriptome analysis possible from 80 ng input material.

Peripheral blood leukocyte RNA samples were diluted with RNase-DNase-free water to a concentration of 40 ng/µl, and 2 µl was added to 4 µl of GlobinLock buffer. The RNA samples (n =368) were placed randomly in eight 48-plex GlobinLock-STRT reaction plates, and each well was tagged for sequencing with an individual barcode. After mixing GlobinLock and RNA on ice, the RNA was denatured for 30 s at 95°C and incubated for 10 min at 60°C for GlobinLock masking and continued for 60 min at 42°C. Just after the 60°C incubation, the block was cooled to 42°C, and 5 µl of reverse transcriptase (RT) mixture was added to initiate cDNA synthesis. The RT mixture contained 1 M betaine (Sigma), 50 mM Tris (pH 8.0, Sigma), 5 mM DTT (Sigma), 7.5 mM $MgCl_2$ (Sigma), RiboLock (0.7 U/µl, Thermo), 400 nM T30VN and RevertAid Premium reverse transcriptase (7 U/µl, Thermo). The concentrations were calculated for final RT in a volume of 10 µl, including the GlobinLock® buffer. Two microliters of ERCC Mix 1 (Ambion), a 1:500 spike-in dilution with nuclease-free water, were used per whole 48-plex library. After a 60 min RT reaction at 42°C and a 5 min inactivation of RT at 85°C, the contents of all 48 reaction wells (480 µl) were pooled into a low-binding 2.0-ml tube. One hundred microliters of Dynabeads MyOne C1 Streptavidin (Thermo) beads were washed twice and used to capture the cDNA molecules (and free primers) according to instructions. After three rounds of EB buffer (10 mM Tris, pH 8.0) and one round of water washing, the DNA-enriched beads were suspended in 75 µl of water and incubated at 75°C for 3 min to release biotin from the streptavidin beads. The supernatant was used as a template for further full cDNA amplification as described previously [7]. The purified cDNA pool was first amplified using 15 cycles of PCR followed by 15 additional cycles to introduce the complete sets of adapters for Illumina sequencing. The libraries were size-selected (200–400 bp) using the sequential AMPure XP (Beckman Coulter) bead selection protocol described previously [7].

All libraries were quality-controlled by TapeStation HS assay (Agilent) and quantified by KAPA Library Quantification Kit (Kapa Biosystems) in a concentration of 1–10 nM. The amplified libraries were alkaline denatured and diluted to 10 pM library prior to Illumina cluster generation. Single 59 bp reads were sequenced on an Illumina HiSeq2000 instrument using a v3 single read kit. In total, each 48-plex library was sequenced on three HiSeq2000 flow-cell lanes.

The raw sequences were processed, aligned and summarized using the STRTprep pipeline version 3 (branch 3vdev, commit 8cb9974; https://github.com/shka/STRTprep/wiki; [7]). Raw read redundancy was corrected through the use of unique molecular identifiers (UMI [8]). The corrected reads were demultiplexed according to the barcode sequence. The demultiplexed reads were aligned to UCSC hg19 human reference genome, human ribosomal DNA unit [GenBank:U13369] and spike-in sequences by Bowtie v. 1.1.0 [9] and Tophat v.2.0.12 [10] with NCBI RefSeq genes as a transcriptome reference; the aligned reads uniquely within 5'-UTR or the proximal upstream of protein coding genes were counted by genes and by samples; 5'-end capture rates in protein coding genes were also calculated. Library bias in the counts was corrected by an approximation-based approach [11]. After the library bias correction, spike-in based normalization was applied [12]. To select variable genes, significance of variation of gene expression was evaluated by comparison with technical variation in the spike-in RNAs, as described in Supplementary text S1 of [7]. Outlier samples in each of the CTRL, ACW and REV groups were examined by pvclust [13] on the normalized expression levels of variable genes (adjusted $P < 0.05$) in each group, and excluded. Expression of genes, which contribute to similar function, tends to correlate [14]. Moreover, because of non-random topology in the regulatory network [15], mutants of different genes that are involved in the same cellular processes have been shown to display similar expression profiles [16]. Therefore, grouping of co-regulated genes followed by association with phenotypes is another approach on functional genomics, which is supposed to work well for identification of diagnostic/prognostic marker genes as well. WGCNA [17] is one of the packages to perform such correlation analysis, and the "gene module" is a set of co-regulated genes with consideration of the topology. Weighted correlation network analysis (WGCNA) [17] was applied according to the developers' recommendations; in detail, (i) genes which were weakly variable in at least either ACW, RVE or CTRL (adjusted variation p-value < 0.25) were selected, (ii) approximation of scale-free topology, signed network construction and module detection used biweight mid-correlation [18] with maxPOutliers=0.05 on the logged normalized levels of variable genes (adjusted $P < 0.25$) in either CTRL, ACW or REV, (iii) relating modules to binary traits used hybrid robust-Pearson correlation [18], and to quantitative traits used biweight mid-correlation with maxPOutliers=0.05. To investigate similarity of the modules between the three groups, consensus modules, which are set of genes correlating in all the three groups, were defined with the same

parameters, then related to the group-specific modules. Significance of differential expression between the sample groups was evaluated by SAMstrt [12] and STRTprep [7]. In detail, the differentially expressed genes between two groups were those with the variation p-value < 0.05 (to guarantee the significant fold-change; adjusted by BH correction) and the differential expression q-value < 0.05 (to guarantee the significant difference between the groups; estimated by permutation, as described in Li et al [19]); the variation index is gene-to-spike-in ratio on the squared coefficient of variance; same amount of spike-in RNA was added to all samples, to model the technical variation, and the variation p-value was estimated by the technical variation, as described in [7]. Hierarchical clustering was performed using Spearman's correlation distance and Ward's clustering method. Gene set enrichment analysis was performed by EnrichR [20]. Multiple logistic regression analysis was performed using glm function in R.

References

1. Stenberg Hammar, K., et al., *Subnormal levels of vitamin D are associated with acute wheeze in young children.* Acta Paediatr, 2014. **103**(8): p. 856-61.
2. Sundbom, F., et al., *Asthma symptoms and nasal congestion as independent risk factors for insomnia in a general population: results from the GA(2)LEN survey.* Allergy, 2013. **68**(2): p. 213-9.
3. Liu, A.H., et al., *Development and cross-sectional validation of the Childhood Asthma Control Test.* J Allergy Clin Immunol, 2007. **119**(4): p. 817-25.
4. Stenberg-Hammar, K., et al., *Rhinovirus-specific antibody responses in preschool children with acute wheeze reflect severity of respiratory symptoms.* Allergy, 2016. **71**(12): p. 1728-1735.
5. Krjutskov, K., et al., *Globin mRNA reduction for whole-blood transcriptome sequencing.* Sci Rep, 2016. **6**: p. 31584.
6. Islam, S., et al., *Highly multiplexed and strand-specific single-cell RNA 5' end sequencing.* Nat Protoc, 2012. **7**(5): p. 813-28.
7. Krjutskov, K., et al., *Single-cell transcriptome analysis of endometrial tissue.* Hum Reprod, 2016. **31**(4): p. 844-53.
8. Kivioja, T., et al., *Counting absolute numbers of molecules using unique molecular identifiers.* Nat Methods, 2011. **9**(1): p. 72-4.
9. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
10. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.
11. Katayama, S., et al., *Guide for library design and bias correction for large-scale transcriptome studies using highly multiplexed RNAseq methods.* BMC Bioinformatics, 2019. **20**(1): p. 418.
12. Katayama, S., et al., *SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization.* Bioinformatics, 2013. **29**(22): p. 2943-5.
13. Suzuki, R. and H. Shimodaira, *Pvclust: an R package for assessing the uncertainty in hierarchical clustering.* Bioinformatics, 2006. **22**(12): p. 1540-2.
14. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
15. Featherstone, D.E. and K. Broadie, *Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network.* Bioessays, 2002. **24**(3): p. 267-74.
16. Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2000. **102**(1): p. 109-26.
17. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
18. Langfelder, P. and S. Horvath, *Fast R Functions for Robust Correlations and Hierarchical Clustering.* 2012, 2012. **46**(11): p. 17.
19. Li, J. and R. Tibshirani, *Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data.* Stat Methods Med Res, 2013. **22**(5): p. 519-36.
20. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.

Supplementary figures

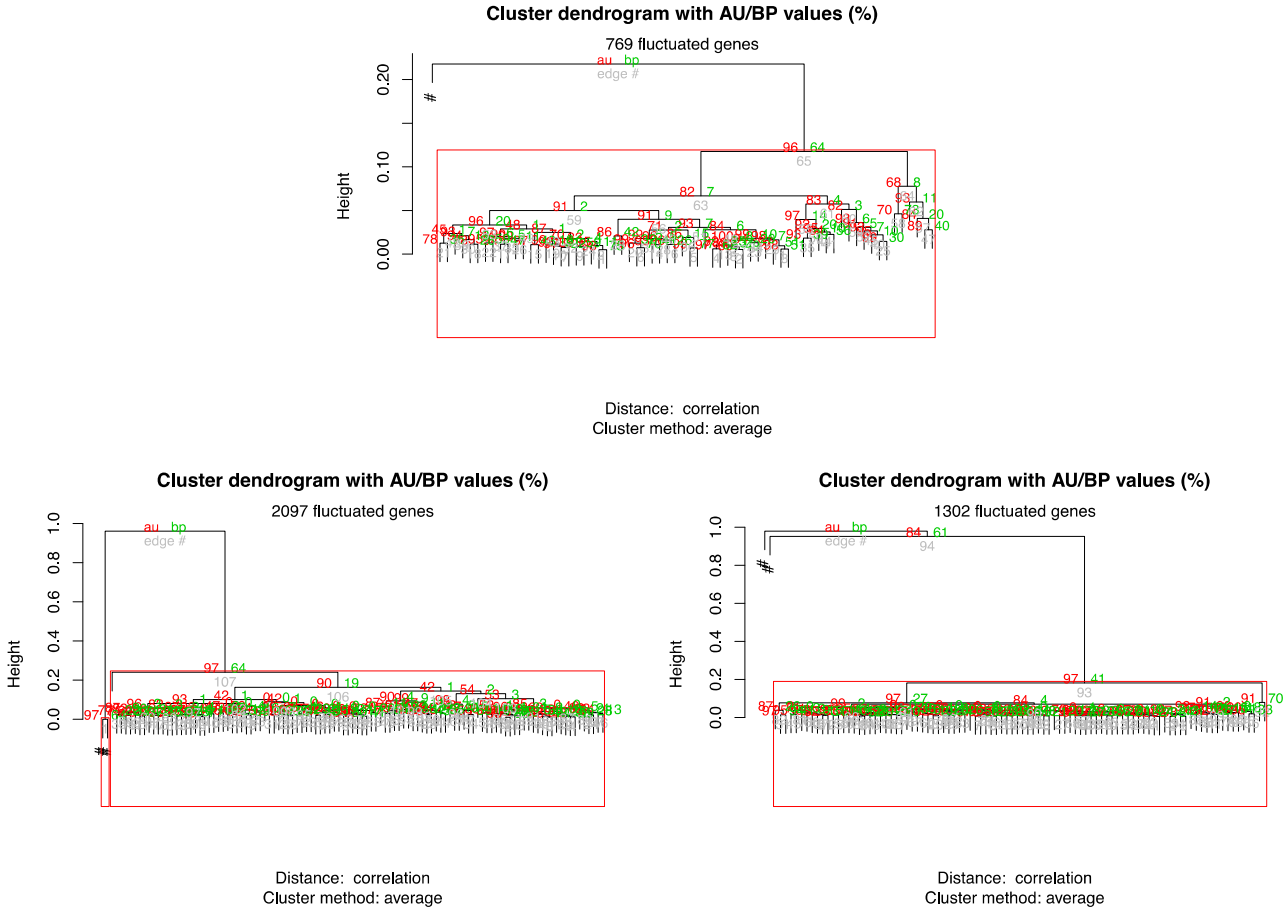

FIG E1. *Outlier check. Each dendrogram elucidates outlier samples on the leukocyte transcriptome profile of the variable protein coding genes (adjusted variation p-value < 0.05) in the healthy controls (top), the cases at the acute visit (bottom left) and the cases at the follow-up visit (bottom right). Red value in each branch is approximately unbiased p-value (AU), and green is bootstrap probability. Clusters with AU ≥ 95% are highlighted by red rectangles, which are strongly supported as certain cluster by normalized expression levels of variable genes in each group; whereas the samples "#" are outliers.*

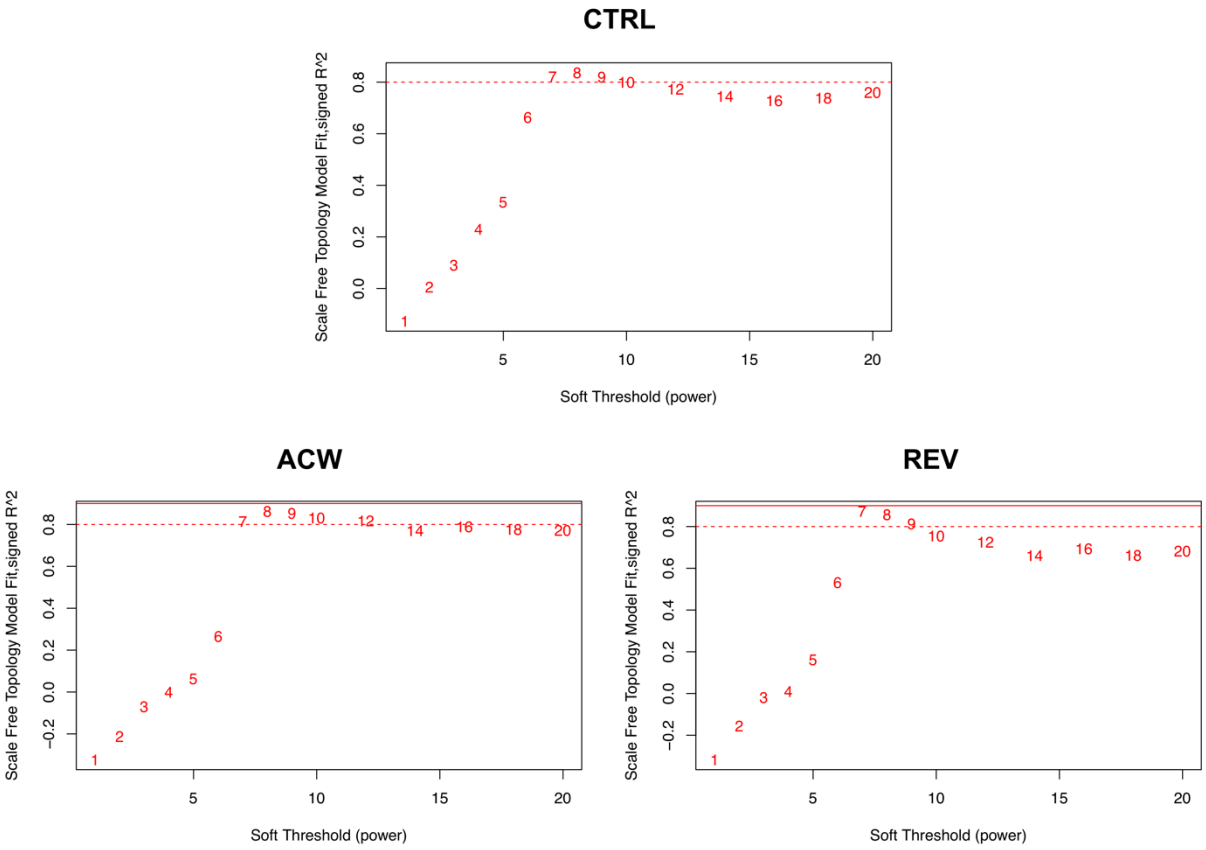

FIG E2. *Analysis of network topology for various soft-thresholding powers. Panels illustrate the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis) on the healthy controls (top), the cases at the acute visit (bottom left) and the cases at the follow-up visit (bottom right). Red horizontal lines are guides of the index at 0.8 (dashed) and 0.9 (solid). At the power=7, the index curve flattened out upon reaching the higher value in all groups; it is a recommended soft-thresholding value by the authors of WGCNA. ACW = acute wheeze, REV= cases at revisit after 2-3 months, CTRL = healthy controls*

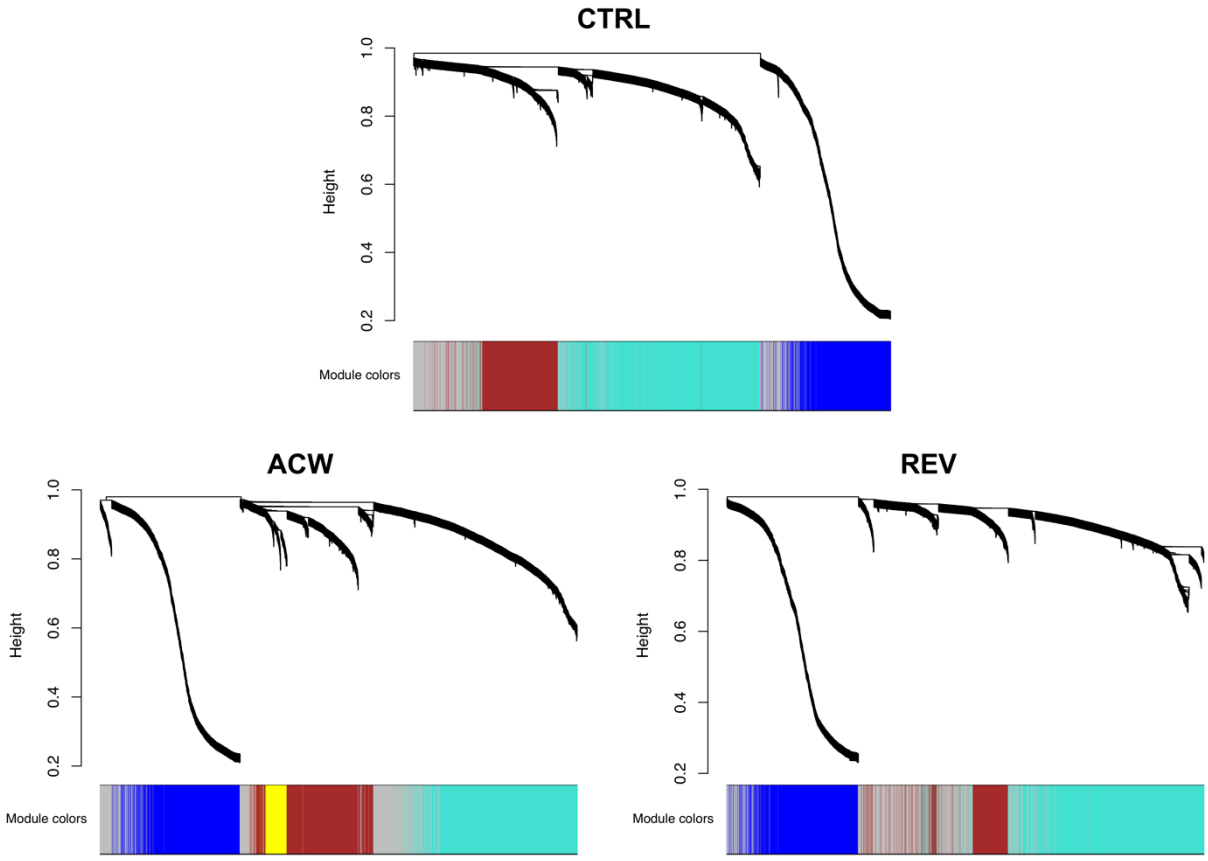

FIG E3. *Hierarchical clustering of logged normalized levels of the variable coding genes and the module assignment. Each dendrogram illustrates similarity on the leukocyte transcriptome profile of the variable protein coding genes (adjusted variation p-value < 0.25) in the healthy controls (top), the cases at the acute visit (bottom left) and the cases at the follow-up visit (bottom right), and the module assignment (bottom of each panel). Gray module color is a reserved one for genes that are not part of any module. Module detection in a block-wise manner by WGCNA does not define identical grouping with the hierarchical clustering, because of its use of the topological overlap measure. ACW = acute wheeze, REV= cases at revisit after 2-3 months, CTRL = healthy controls*

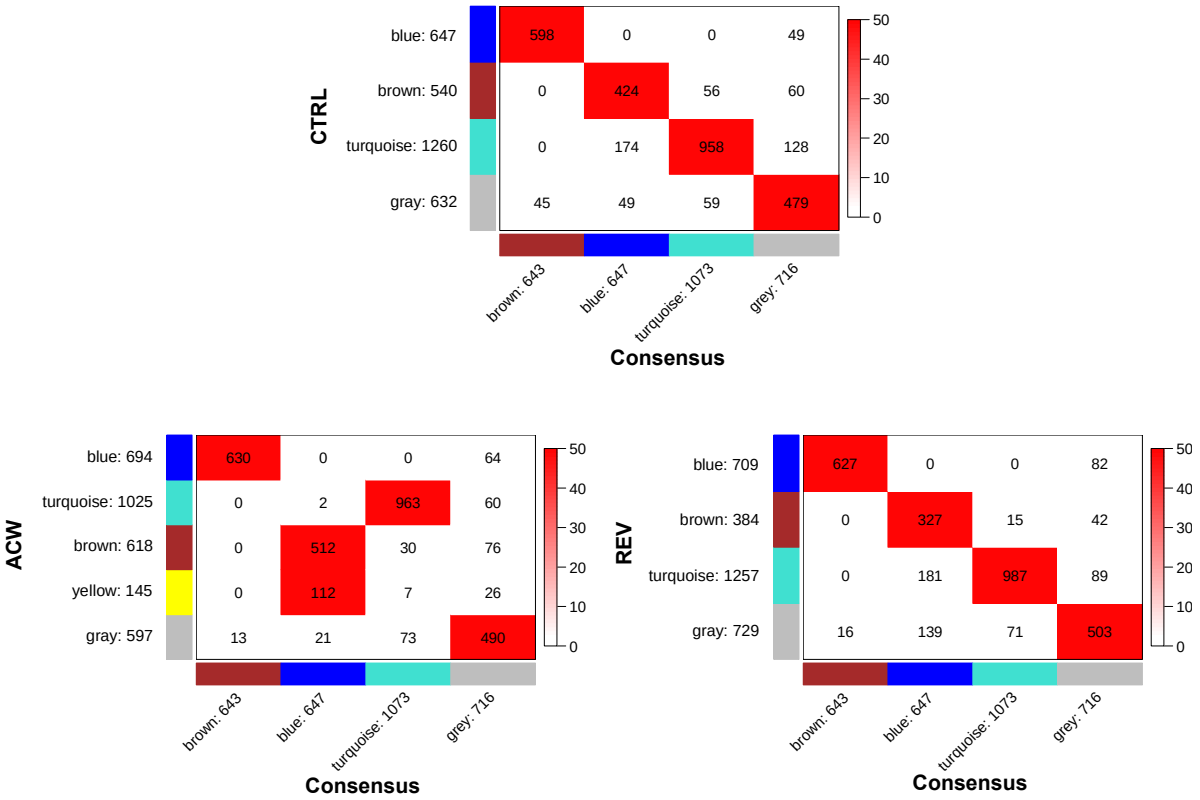

FIG E4. *Gene correspondence between group-specific modules and the consensus modules. Each matrix represents the correspondence between the consensus modules and group-specific modules defined by the healthy controls (top), the cases at the acute visit (bottom left) and the cases at the follow-up visit (bottom right). Each row corresponds to a group-specific module labeled by color (but gray is a reserved color for genes that are not part of any module) and numbers of the member genes, and each column corresponds to one consensus module. Numbers in each cell is number of genes in the intersection of the corresponding modules. Color of each cell is -log(p), where p is by the Fisher's exact test for the overlap of the two modules. ACW = acute wheeze, REV= cases at revisit after 2-3 months, CTRL = healthy controls*

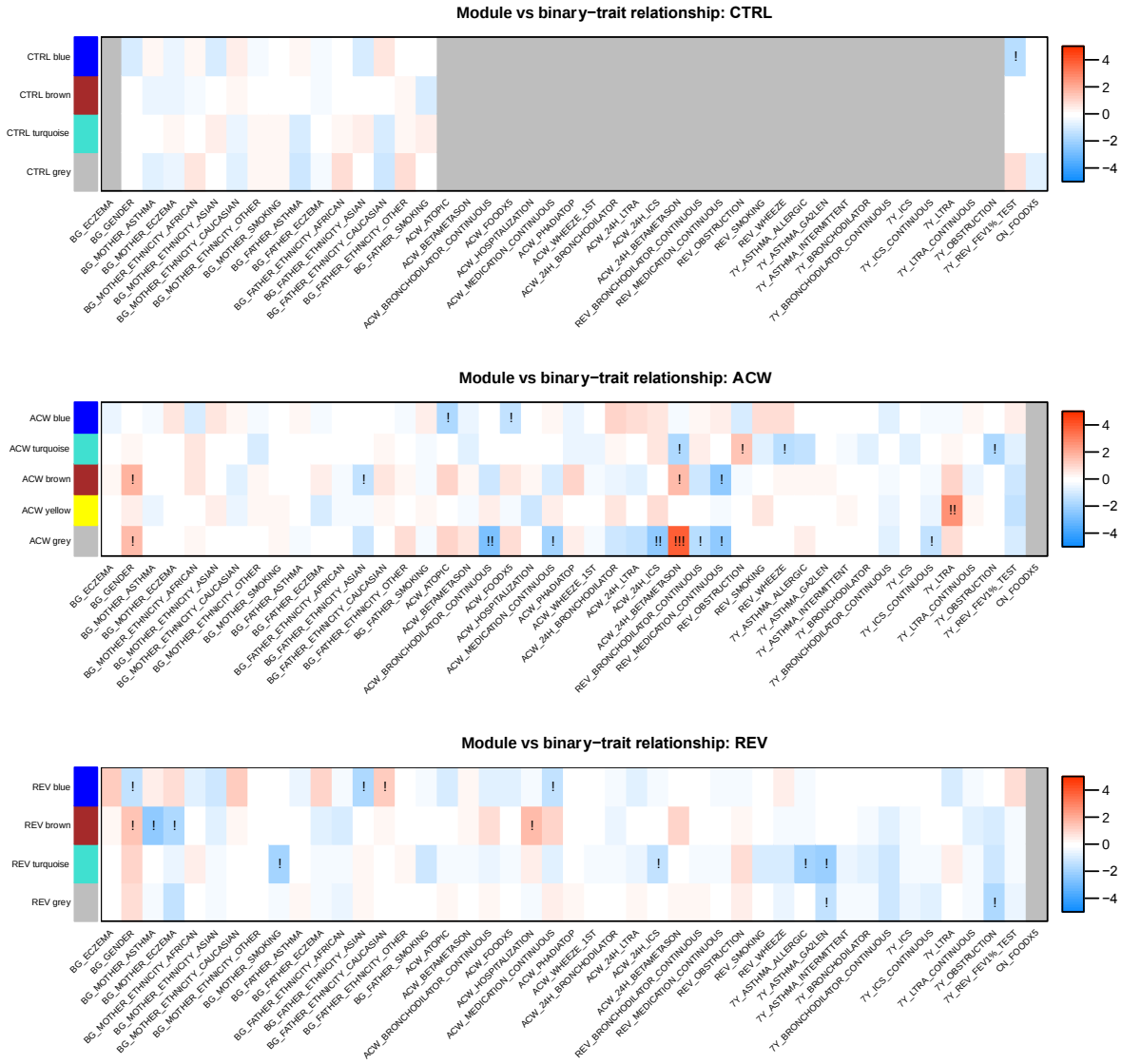

FIG E5. *Associations between the binary traits and the group-specific modules. Each panel illustrates an association between the traits and the modules of the controls (top), the cases at the acute visit (middle) and the cases at the follow-up visit (bottom). Each row is a module labeled by a color (gray is a reserved color for genes that are not part of any module), and each column is a trait. Each cell is colored by -log_{10}(p) x sign(r), where p is p-value of the corresponding correlation, and r is the correlation coefficient. !, !! and !!! in each cell are p < 0.05, 0.005 and 0.0005, respectively. Legend of the trait IDs is Table E1. ACW = acute wheeze, REV= cases at revisit after 2-3 months, CTRL = healthy controls*

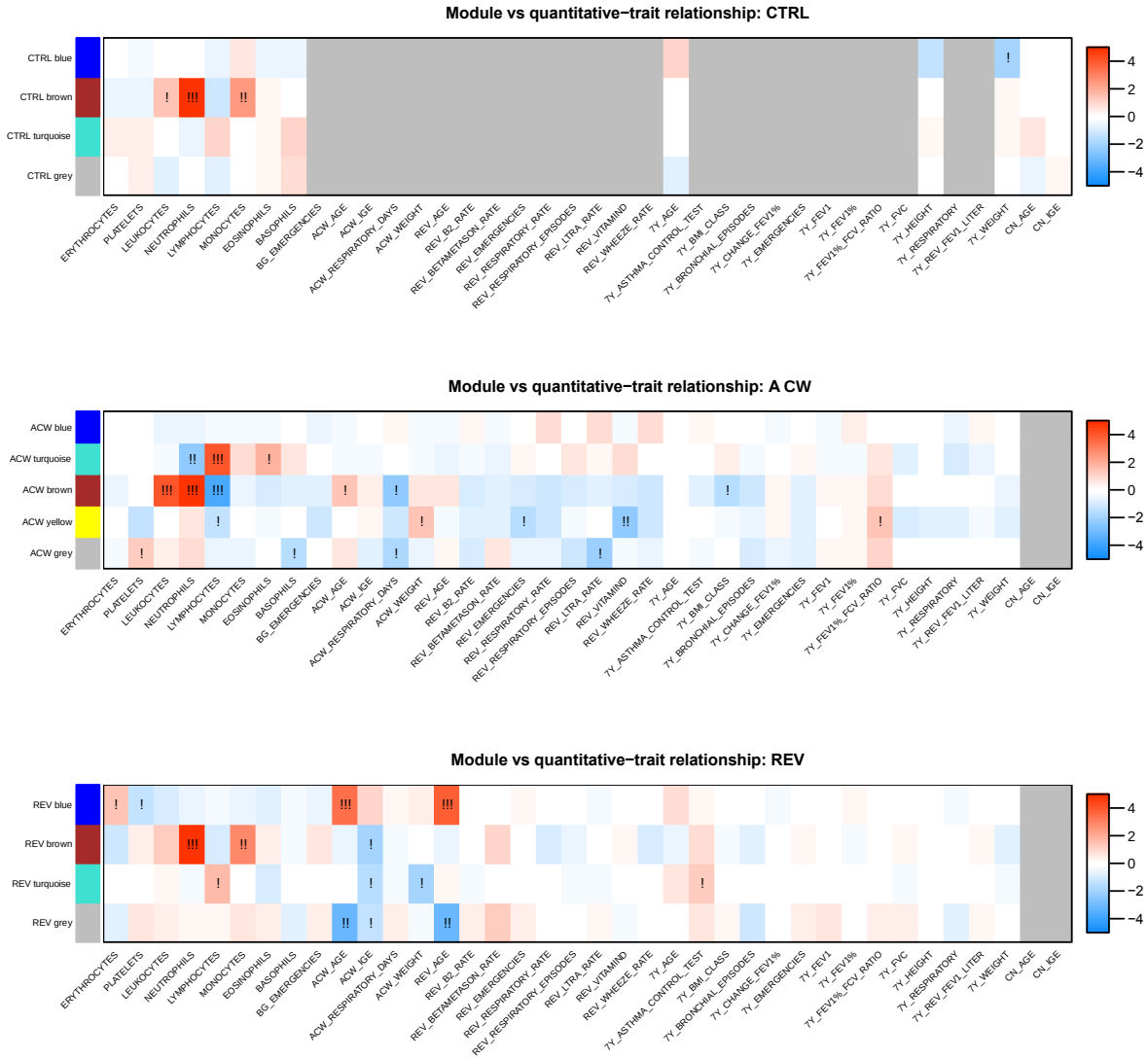

FIG E6. *Associations between the quantitative traits and the group-specific modules. Each panel illustrates an association between the traits and the modules of the controls (top), the cases at the acute visit (middle) and the cases at the follow-up visit (bottom). Each row is a module labeled by a color (gray is a reserved color for genes that are not part of any module), and each column is a trait. Each cell is colored by -log$_{10}$(p) x sign(r), where p is p-value of the corresponding correlation, and r is the correlation coefficient. !, !! and !!! in each cell are p < 0.05, 0.005 and 0.0005, respectively. Legend of the trait IDs is Table E1. ACW = acute wheeze, REV= cases at revisit after 2-3 months, CTRL = healthy controls*

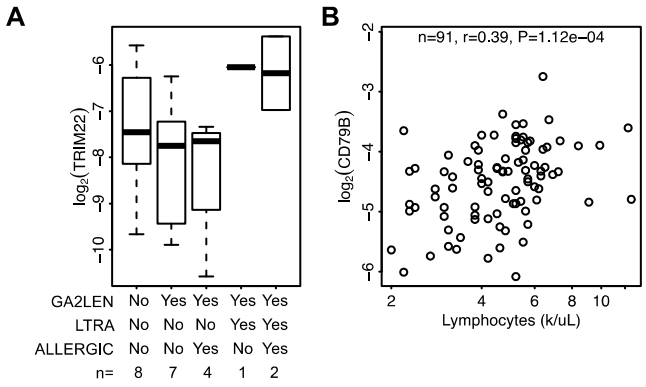

FIG E7. *Differential gene expression of ACW yellow module genes and the association with cilinical traits.*

A, *Differential gene expression of TRIM22 (y-axis) at the acute visit in the cases with the asthma-related traits at 7yrs (x-axis). GA2LEN = (7Y_ASTHMA_GA2LEN) Asthma at 7 years of age (7Y_ASTHMA_GA2LEN) was defined as a positive answer to either the question; Have you had an attack of asthma in the last 12 months?' OR the question "Are you currently taking or have you during the last 12 months taken any medication for asthma, including short-acting b2-antagonists, inhaled corticosteroids, and montelukast?", modified from[2].LTRA = leukotriene receptor antagonists medication the year preceding the 7 year visit (7Y_LTRA), and ALLERGIC = (7Y_ASTHMA_ALLERGIC) Allergic asthma was defined as asthma as above with allergic sensitization and clinical symptoms of allergy.*

B, *Positive correlation between lymphocyte counts (x-axis) and expression of B-cell marker gene (y-axis; CD79B) in the REV turquoise module at the revisit. Biweight midcorrelation coefficient (r) and the significance (P) are labeled.*

Supplementary table legends

TABLE E1. *Legends of the clinical traits. Binary traits have either 0 or 1.*

TABLE E2. *Differentially expressed genes between the groups. Sheet label represents the pair of compared groups, the cases (ACW= acute phase, REV = revisit), the healthy controls (CTRL=controls). These tests were unpaired, except for the comparison between REV vs ACW which were paired. DE.score is a statistic value of the differential expression test between the two groups; in a comparison A vs B, positive DE.score is up-regulation in the group B. DE.qvalue is the false discovery rate on the differential expression test. FL.pvalue is a corrected significance on degree of the variation of all samples in the compared groups. FL.score is the statistic value of the variation test.*

TABLE E3. *Modules and the member genes. Group CTRLACWREV is the consensus module of CTRL, ACW and REV.*

TABLE E4. *Correlation between the LTRA medication in the last year before the revisit at seven years of age and expression of ACW yellow genes. Correlation coefficient (bicor) and the significance (p.adj) were calculated by hybrid biweight mid-correlation with Benjamini and Hochberg correction.*

TABLE E5. *Correlation between the vitamin D concentration at the first revisit and expression of ACW yellow genes. Correlation coefficient (bicor) and the significance (p.adj) were calculated by biweight mid-correlation with Benjamini and Hochberg correction.*