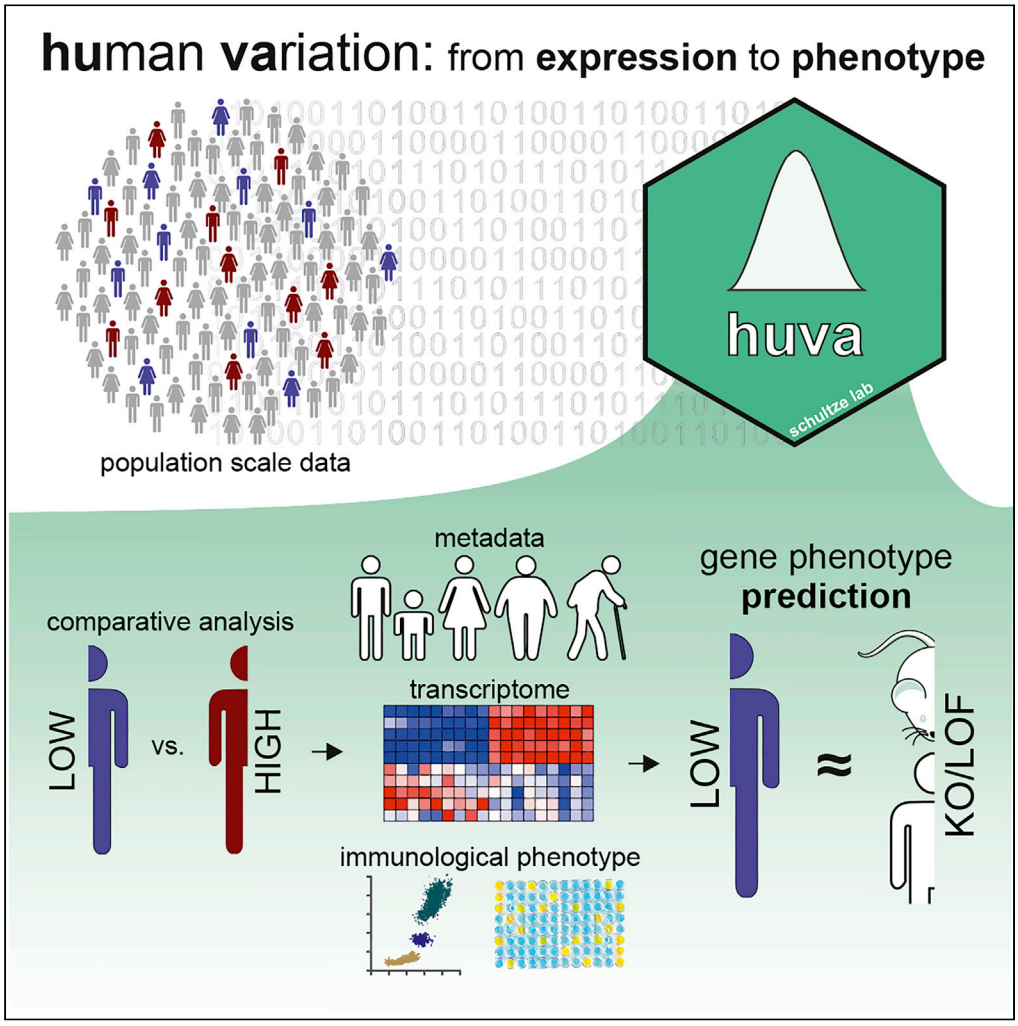
iScience

CellPress
OPEN ACCESS

Article

Human variation in population-wide gene expression data predicts gene perturbation phenotype



Lorenzo Bonaguro, Jonas Schulte-Schrepping, Caterina Carraro, ..., Thomas Ulas, Joachim L. Schultze, Anna C. Aschenbrenner

lorenzobonaguro@uni-bonn.de (L.B.)
anna.aschenbrenner@dzne.de (A.C.A.)

Highlights

Human variation can be exploited to generate gain- or loss-of-function experiments

Huva was used to predict the function of genes central to the immune response

Transcriptome-wide *huva* analysis uncovers the role of *STAT1* in monocytes

Huva is implemented in R as well as accessible via an easy-to-use web interface

Bonaguro et al., iScience *25*, 105328
November 18, 2022
https://doi.org/10.1016/j.isci.2022.105328

iScience

CellPress
OPEN ACCESS

Article

Human variation in population-wide gene expression data predicts gene perturbation phenotype

Lorenzo Bonaguro,[1,2,*] Jonas Schulte-Schrepping,[1,2] Caterina Carraro,[1,3] Laura L. Sun,[2] Benedikt Reiz,[4] Ioanna Gemünd,[1,2,5] Adem Saglam,[1] Souad Rahmouni,[6] Michel Georges,[6] Peer Arts,[7,8] Alexander Hoischen,[7,9] Leo A.B. Joosten,[9,10] Frank L. van de Veerdonk,[9] Mihai G. Netea,[9,11] Kristian Händler,[1,12] Sach Mukherjee,[13,14] Thomas Ulas,[1,2,12] Joachim L. Schultze,[1,2,12,15] and Anna C. Aschenbrenner[1,2,9,15,16,*]

SUMMARY

Population-scale datasets of healthy individuals capture genetic and environmental factors influencing gene expression. The expression variance of a gene of interest (GOI) can be exploited to set up a quasi loss- or gain-of-function "*in population*" experiment. We describe here an approach, *huva (human variation)*, taking advantage of population-scale multi-layered data to infer gene function and relationships between phenotypes and expression. Within a reference dataset, *huva* derives two experimental groups with LOW or HIGH expression of the GOI, enabling the subsequent comparison of their transcriptional profile and functional parameters. We demonstrate that this approach robustly identifies the phenotypic relevance of a GOI allowing the stratification of genes according to biological functions, and we generalize this concept to almost 16,000 genes in the human transcriptome. Additionally, we describe how *huva* predicts monocytes to be the major cell type in the pathophysiology of STAT1 mutations, evidence validated in a clinical cohort.

INTRODUCTION

Any biological parameter is characterized by variation when assessed at the population level. In humans, this is an essential variable when studying diseases in larger cohorts, and it is increasingly recognized when describing healthy populations. Assessment of biological parameters by omics technologies exposes this variation even more owing to the high dimensionality of the data. On the other hand, it encompasses the possibility of comprehensively capturing human variation, especially when assessed in larger cohorts. Genetics has been particularly successful in linking genetic variation to diseases per se but also to physiological phenotypes (GWAS, PheWAS) (Pividori et al., 2020; Tam et al., 2019). More recently genetic variation was directly linked to variation in gene expression (eQTL, expression quantitative trait loci) (GTEx Consortium, 2013, 2020; Kim-Hellmuth et al., 2020; Majewski and Pastinen, 2011; Strunz et al., 2018), or epigenetic variation (epiQTL/hQTL) (Furci et al., 2019; Pelikan et al., 2018), as two examples explaining changes in the transcriptome or gene regulation by variance in the genome. Such genome-wide approaches also allowed for the identification of genetic risk factors for diseases, e.g. APOE for Alzheimer's disease (Kunkle et al., 2019).

Variation of gene expression and regulation is not only determined by the genome. Environmental signals present another important driver of transcriptional variation, resulting in unique individual transcriptional profiles (Favé et al., 2018; Gibson, 2008; Majewski and Pastinen, 2011; Wainberg et al., 2019). From a broader perspective, the breadth of the transcriptional program within a population as a response to environmental challenges is a key evolutionary feature allowing both rapid and long-term adaptation of a species to environmental challenges (López-Maury et al., 2008).

In light of the increasing availability of multi-layered datasets derived from cohorts, including those derived from healthy individuals, it is now possible to move from gene-centric approaches toward data-driven utilization of natural variation. It is evident that genetics and environmental factors impact gene expression,

[1]Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), 53127 Bonn, Germany

[2]Genomics and Immunoregulation, Life and Medical Sciences (LIMES) Institute, University of Bonn, 53113 Bonn, Germany

[3]Department of Pharmaceutical and Pharmacological Sciences, University of Padova, 35131 Padova, Italy

[4]Comma Soft AG, 53229 Bonn, Germany

[5]Department of Microbiology and Immunology, the University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Parkville, 3010 VIC, Australia

[6]Unit of Animal Genomics, GIGA-Institute, University of Liège, 4000 Liège, Belgium

[7]Department of Human Genetics and Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, 6525 Nijmegen, the Netherlands

[8]Department of Genetics and Molecular Pathology, Centre for Cancer Biology, SA Pathology and the University of South Australia, Adelaide, 5000 SA, Australia

[9]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, 6525 Nijmegen, the Netherlands

[10]Department of Medical Genetics, "Iuliu Hatieganu" University of Medicine and Pharmacy, 400012 Cluj-Napoca, Romania

Continued

and we hypothesize that the resulting variation of expression is directly linked to functional phenotypes. For example, inter-individual differences in the expression level of a gene *x* might be linked to the number of a certain cell type in a particular tissue (Newman et al., 2019) or the effect size of a response toward a pathogen (Bossel Ben-Moshe et al., 2019).

A prerequisite to link, for example, gene expression to its biological role is the simultaneous measurement of expression data for each individual in the population under study, e.g. by assessing transcriptomes, and assays considering biological functions and phenotypes, e.g. cell counts in blood, expression of cell surface markers, or induction of soluble mediators by immune cells upon stimulation with immunogenic agents. Such studies have become available only recently and a prime example is the Human Functional Genomics Project (HGFP), in which inter-individual variation of immune responses to prototype pathogens has been studied in several hundred healthy donors in the context of their genetics or microbiomes (Li et al., 2016; Ter Horst et al., 2016). A similar endeavor is currently being pursued by the Milieu Interieur Consortium (Thomas et al., 2015), and further multi-layered datasets from large healthy cohorts have been obtained in the ImmVar project (Raj et al., 2014) as well as the CEDAR study (Momozawa et al., 2018). Exemplary evidence that human gene expression variation can be utilized to define the biological role of a gene emerged when studying Cystein-rich with EGF-like domains 1 (CRELD1) that is involved in T cell homeostasis. Indeed, *CRELD1* gene expression levels in peripheral circulating immune cells were associated with T cell frequencies in three independent human cohorts (Bonaguro et al., 2020).

Here, we present the *huva* (human variation) approach to use the variance of gene expression in human cohort studies as a general and distinct concept to predict the role of individual genes or groups of genes for biological functions by integrating expression, phenotypic, and functional data layers from cohort data. Validity and robustness of *huva* were demonstrated by applying the approach to genes with previously described functions and further illustrated by predicting the phenotype of patients with STAT1-activating mutations (AM). These *huva* predictions were validated by the analysis of the transcriptome of both peripheral blood mononuclear cells (PBMCs) and isolated cell types from STAT1 AM carriers. To facilitate access to our approach, we implemented *huva* as an easy-to-use R-based library as well as an interactive webtool with the built-in datasets of HFGP/500FG, ImmVAR, and CEDAR. *huva* can be applied to any cohort study with available multi-layered data from any organ system for which gene or protein expression as well as functional data are available.

RESULTS

huva allows comprehensive human variation analysis from several cell types

The exploration of human variation to predict and/or define the biological role of individual genes can be best accomplished by exploiting large enough datasets that have become available for example blood, often investigated as an easily accessible surrogate tissue and source of biomarkers for many diseases (Ashton et al., 2020; Li et al., 2016; Momozawa et al., 2018; Rajewsky et al., 2020; Ter Horst et al., 2016). The *huva* approach for the analysis of human natural variation is illustrated by the use of transcriptomics datasets ranging from bulk transcriptome analyses of whole blood to isolated cell types ($CD4^+$ or $CD8^+$ T cells, monocytes, B cells, platelets, and granulocytes) (Figure 1A and Table S1). In the *huva "in population"* approach, two experimental groups are defined according to the value distribution of a parameter of interest, e.g. a gene of interest (GOI). With this initial step, two experimental groups from both ends of the distribution of the gene expression of the GOI, henceforth referred to as LOW or HIGH, are defined (Figure 1A). In principle, depending on the data available for the cohort in the analysis and the specific scientific question, we take as input a gene of interest (*GOI huva experiment*), analyze its variance, and then use its extremes to define the experimental groups (Figure S1A). For the definition of the experimental groups of a *GOI*, it is initially assessed in which of the available datasets the selected gene is present (Figure S1A) followed by analyses of the value distribution of the *GOI* in each of the datasets (Figure S1A). The LOW and HIGH groups of samples taken into the analysis are those falling within the percentile intervals selected for the "*in population*" *huva* experiment and are used for the following comparative analysis (Figure 1B). The LOW group is then compared to the HIGH group for all available parameters of each dataset, including metadata, transcriptome, and immunological or other phenotypes, if available (Figure 1B). We propose that for a specific GOI, this approach can be used as a proxy for the functional characterization of a gene knockout (KO) in a model system (e.g. mice) or a loss-of-function mutation in humans.

[11]Immunology and Metabolism, Life and Medical Sciences (LIMES) Institute, University of Bonn, 53113 Bonn, Germany

[12]Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE and University of Bonn, 53127 Bonn, Germany

[13]Statistics and Machine Learning, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), 53127 Bonn, Germany

[14]MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, UK

[15]These authors contributed equally

[16]Lead contact

*Correspondence: lorenzobonaguro@uni-bonn.de (L.B.), anna.aschenbrenner@dzne.de (A.C.A.)

https://doi.org/10.1016/j.isci.2022.105328

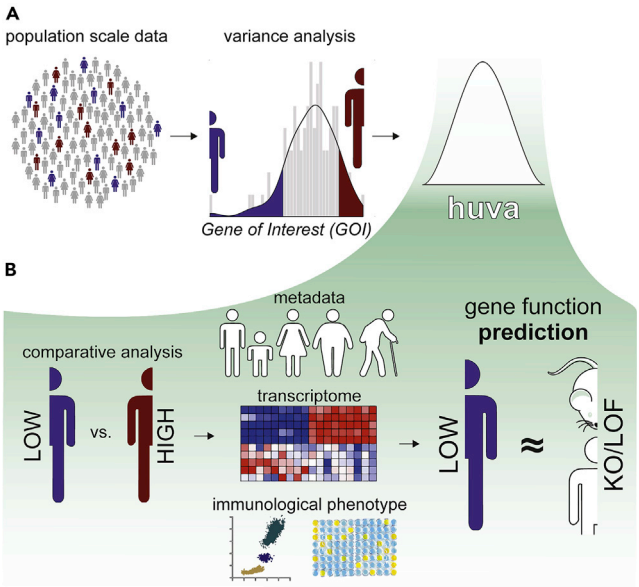


Figure 1. *huva* allows comprehensive human variation analysis from several cell types

(A) Schematic representation of the *huva* framework, starting with the analysis of gene expression across multiple publicly available dataset to (B) the basic workflow of the analysis, the comparison groups of individuals with low and high levels of a parameter of interest. See also Figures S1 and S2; Table S1 (GOI gene of interest).

Computationally, the *huva* experiment runs on standard hardware (see STAR methods for details) within a few seconds for the GOI *huva experiment* calculating the results from around 2400 transcriptomic profiles across 4 datasets and 7 cell types (500FG (Li et al., 2016; Ter Horst et al., 2016), CEDAR (Momozawa et al., 2018), ImmVar (Raj et al., 2014) and PBMC collection (Warnat-Herresthal et al., 2020), Figure S2A). Collectively, this short run-time makes the realization of *huva* analyses broadly applicable. In our R implementation of the approach, user-friendliness and accessibility are provided by a web-based app, which allows easy access to the *huva* framework with no limiting hardware requirements. In addition to ease of use, we also ensured flexibility. For example, users have access to all the datasets already included within *huva* for the comparative analysis and can easily extend the approach also to private/new datasets with minimal programming skills.

huva imputes gene function and phenotypes

To demonstrate the workflow of a *huva* experiment and validate its ability to predict biological differences as a proxy for a potential biological function of a gene within a certain tissue or biosample, *MYD88* was chosen as a paradigm for a well-characterized gene (Figure S3A). This gene was first described in an immunological context in the early 1990s by two independent groups as an adaptor protein that binds the TIR domain of TLRs and consequently mediates NF-κB activation (Lord et al., 1990; Muzio et al., 1997). The role of this protein in signaling transduction downstream TLR activation has been deeply characterized over the last two decades within numerous cellular systems and animal models (Gamrekelashvili et al., 2020; Kaisho and Akira, 2001; Kawai et al., 1999). Furthermore, mutation of *MYD88* has been connected to a number of human diseases, ranging from immunodeficiency to cancer (Platt et al., 2019; von Bernuth et al., 2008; Wang et al., 2014). *MYD88* is highly expressed in immune cells from healthy donors (500FG dataset, Figure 2A) and displays a normally distributed expression across the 95 participants included in the transcriptomics part of the study (Figure S3A; Shapiro-Wilk p = 0.2, Figure S3B). For the *huva experiment*, we defined two experimental groups, HIGH and LOW, according to the expression of *MYD88* (Figure 2B). Contrasting the full transcriptomes of these two groups revealed substantial differences, as depicted by principal component analysis (Figure S3C). Differential expression analysis between the two groups revealed around 400 differentially expressed genes (307 downregulated, 142 upregulated/|FC|>2; FC: fold change; Figures 2C, S3D, and S3E; Table S2). In the differential expression analysis, we compare the LOW vs. HIGH groups as a proxy for an "*in population*" loss-of-function experiment. To clarify this comparison, this means genes defined as downregulated have lower expression in the LOW group compared to

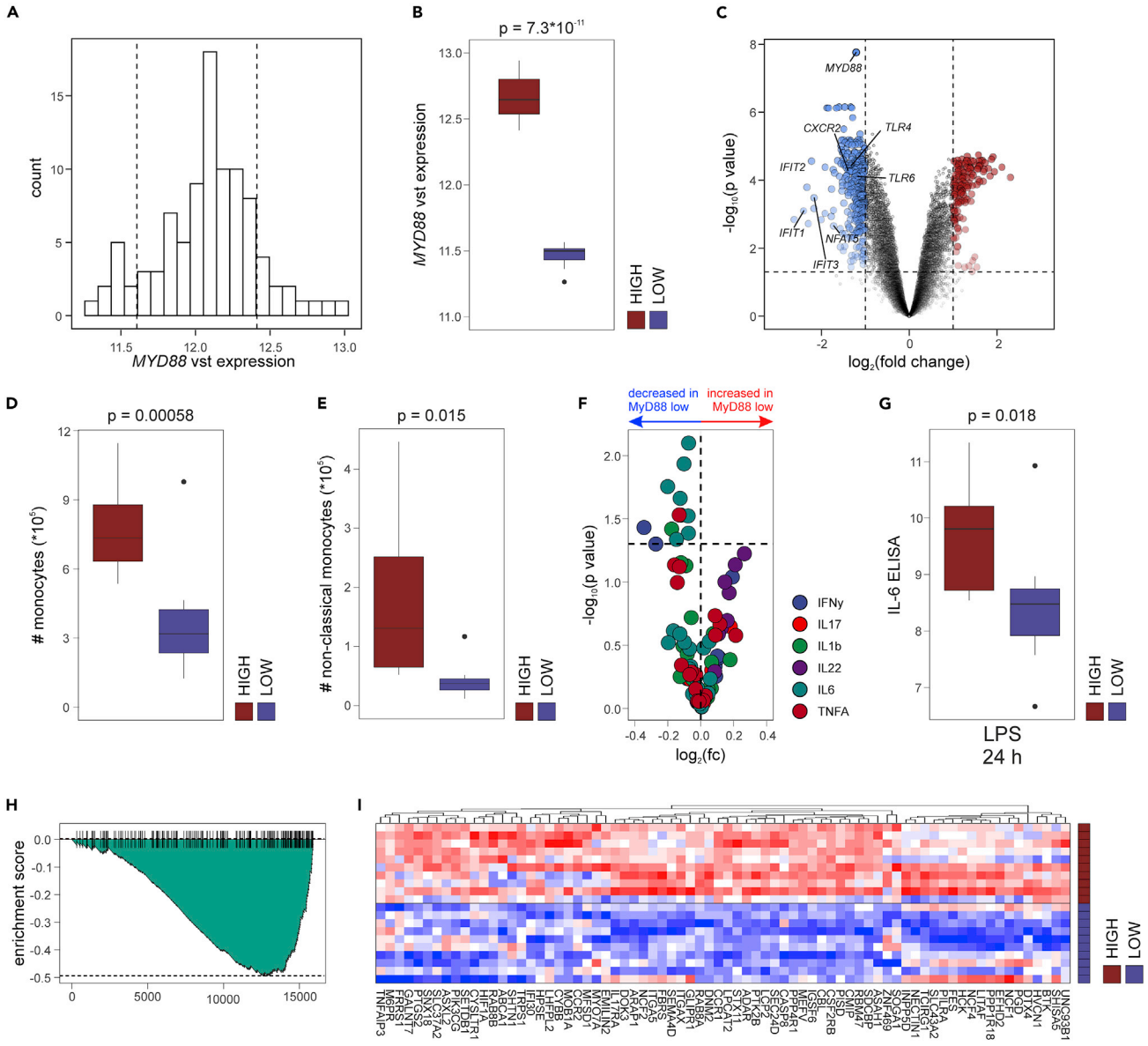


Figure 2. *huva* imputes *MyD88* function and phenotype

(A) *MYD88* vst transformed expression profile in PBMC (500FG).

(B) *MYD88* expression in the HIGH and LOW *huva* experimental groups (n = 10).

(C) Volcano plot of log2 fold change and negative log10 p value of the differential expression between the LOW and HIGH *MYD88 huva* experimental groups.

(D and E) Total cell number of total monocytes (d) and non-classical monocytes ($CD14^{dim}$, $CD16^{++}$; n = 10) (E).

(F) Volcano plot of log2 fold change and negative log10 p value for secreted cytokines after stimulation in the comparison *MYD88* LOW vs. HIGH, colored by measured molecule (ELISA, n = 10).

(G) IL-6 secretion LPS stimulation for 24 h (n = 10).

(H and I) GSEA plot for the GSE22935_WT_VS_MYD88_KO_MACROPHAGE_UP signature on the ranked gene list of the *huva* experiment of PBMC (H) and heatmap of the leading edge of the enrichment (I). Box plots were constructed in the style of Tukey, showing median, 25th and 75th percentiles, exact p value is shown from unpaired two-sided t-test. See also Figures S3 and S4; Tables S2, S3, and S4.

the HIGH group and vice versa for upregulated genes. Among the differentially expressed genes, we found *NFAT5*, known to be important for TLR signaling (Buxadé et al., 2012), *IFIT1/2/3* (Diamond and Farzan, 2013; John et al., 2018), *CXCR2*, another well-known MyD88 target gene (Sabroe et al., 2005), and *TLR4* and *6*, two signaling molecules upstream of MyD88 activation to be downregulated in samples with low *MYD88* expression (Figure 2C).

Finding a notable overlap of differentially expressed genes in dependency of *MYD88* between our human data analysis and the published results from murine loss-of-function models, we investigated the immunological phenotype of the two experimental *huva* groups. Here, we further exploited the 500FG dataset, not only including transcriptome information but also paired data on cell counts for the main circulating immune cell types as well as cytokine secretion profiles upon stimulation with several pathogens (e.g. *Candida albicans*) or pathogen components (e.g. LPS) (Li et al., 2016; Ter Horst et al., 2016). At first, looking at the total cell count in the blood of *MYD88* LOW or HIGH samples, we observed a strong decrease in the number of circulating monocytes, which indeed appeared to be the only cellular population strongly affected by the expression levels of *MYD88* (Figures 2D and S3F; Table S3). Intriguingly, even though the total number of monocytes was strongly affected by the expression level of our gene of interest (Figure 2D), classical monocytes, as the largest fraction of circulating monocytes, showed only a tendency of reduced numbers (Figure S3G), whereas both, non-classical (Figure 2E) and intermediate monocytes (Figure S3H) were most affected. Non-classical monocytes are well-known for their pro-inflammatory phenotype (Kapellos et al., 2019) and their capacity to migrate to inflamed tissues (Kapellos et al., 2019; Randolph et al., 2002). The role of MyD88 as a modulator of cytokine transcription downstream of TLR or IL-1R activation is also well characterized (Akira, 2003; Cohen, 2014). A hallmark cytokine downstream of the MyD88/NF-κB axis is IL-6. Indeed, *Myd88*$^{-/-}$ mice are totally depleted of any circulating IL-6 (Kawai et al., 1999; von Bernuth et al., 2008). Strikingly, when comparing the secretion of cytokines from PBMC of donors in the LOW and HIGH *MYD88* experimental groups, we noticed a general reduction in the secretion of IL-6 upon exposure to several of the used stimulants in the *MYD88* LOW group with almost all other cytokine levels unaltered (Figure 2F and Table S4). Interestingly, the effect on IL-6 production was most prominent upon LPS stimulation (a strong canonical TLR4 antagonist, Figure 2G) (Jin and Lee, 2008) but also downstream of other TLR ligands such as CpG (a TLR9 ligand, Figure S3I) and Pam3Cys (TLR1 agonist, Figure S3J) (Jin and Lee, 2008). Thus, the phenotype we observe "*in population*" with the *huva* framework reflects previous experimental data derived from genetic model systems.

Finally, to provide a more unbiased approach rather than the comparison with single differentially expressed genes, we curated a collection of MyD88-related gene signatures and projected them on the ranked gene list of the differential genes from the *MYD88* LOW vs. HIGH comparison sorted by fold change in the *huva experiment*, with GSEA showing a strong regulation of almost all tested signatures (Figure S3K). Among the most regulated terms, we found the signature "WT VS MYD88 KO MACROPHAGE_UP" (Godec et al., 2016; Qualls et al., 2010). When investigating the leading edge driving the strong negative functional enrichment, in agreement with a high transcriptional similarity between *MyD88* KO cells and the *huva MYD88* LOW group, we found *HIF1A*, *CCR2*, or *CASP8* among the most downregulated genes, all of which are genes reportedly involved in MyD88 signaling (Figures 2H and 2I) (Qualls et al., 2010). Similarly, a published collection of genes found to be co-expressed and/or modulated by MyD88 (Subramanian et al., 2005) (Figures S3L and S3M) was strongly downregulated in the *huva* LOW group as well.

To further strengthen the validity and the biological significance of the *huva* approach, we performed similar analyses for additional genes (*AKT1*, *MAPK3* (ERK), *STAT1*), for which enough experimental evidence for their function in immune cells is available (Hoxhaj and Manning, 2020; Manning and Toker, 2017). For each of these genes, *huva* uncovers - based on the variance within human populations - the biology that was previously determined for these genes in genetic model systems (Figure S4).

Collectively, we provide evidence that human variation analysis by the *huva* approach can predict biological implications and phenotypes for any GOI, generating strong hypotheses that can be further tested for their causal relationship to an observed phenotype.

Statistical validation of the *huva* approach

In essence, *huva* compares, for example, transcriptomes derived within a human population with phenotypic measurements from the same donors focusing on those individuals with high or low levels of parameters of interest, most commonly a particular gene (*GOI huva experiment*). To test the validity of our approach, we performed random permutation sampling (Figure 3A) with 100 random permutations assigning an equal number of random samples to each, the HIGH and LOW group, and compared the fold change of the GOI to the true LOW and HIGH groups (Figures 3B and S5A). We used *SLC12A1* as an example illustrating strong absolute variance and high statistical significance between the LOW and HIGH groups (Figure 3B). In contrast, none of the randomly drawn sample sets came close in terms of absolute fold change differences and significance levels. This was similarly true for genes with low fold change differences, as exemplified for *CRELD1* (Figure S5A). We

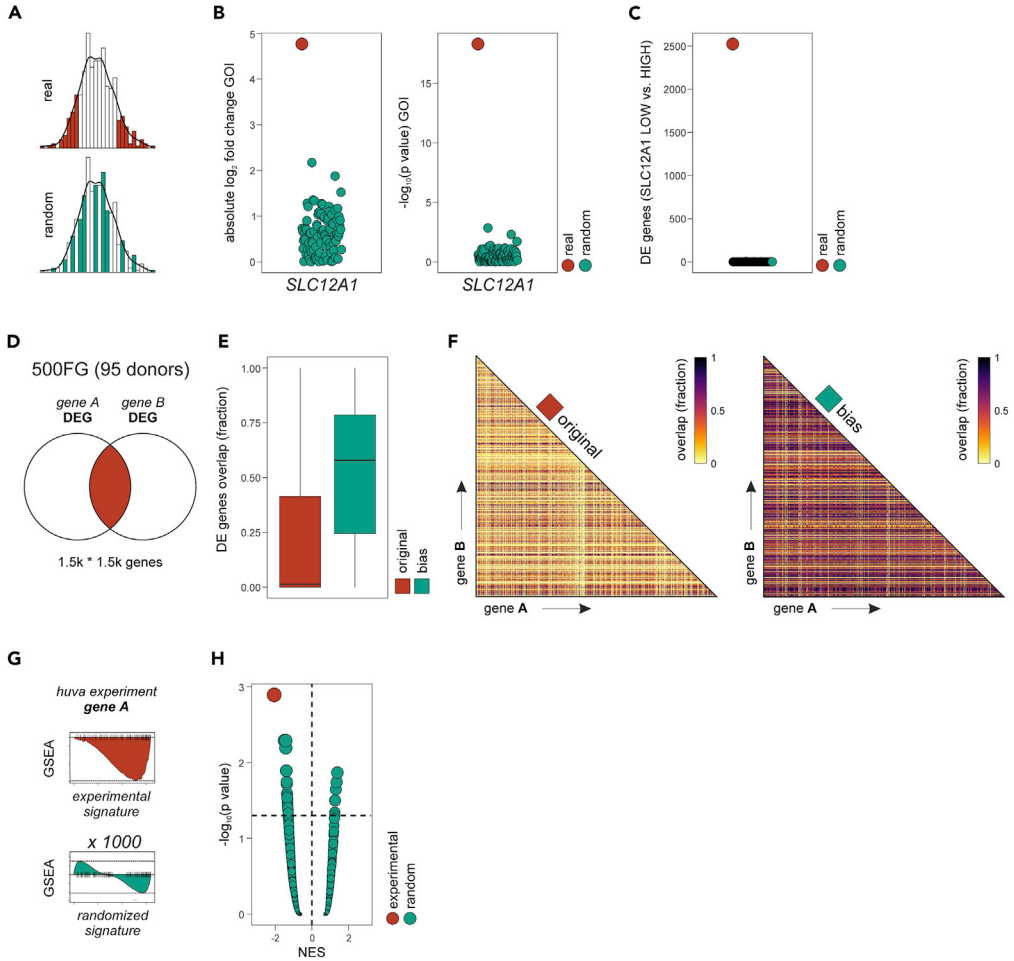


Figure 3. Statistical validation of the *huva* approach

(A) Visual representation of the randomization experiment performed in Figures 3B, 3C, S5A, and S5B.

(B) Absolute fold change and –log10 p value for a selected gene of interest showing high variance across the dataset (*STAT1*). In red the result of the *huva* LOW and HIGH groups, in green the result of random sampling of two experimental groups of equal size (n = 100).

(C) Number of differentially expressed of the *huva* and random sampling experiments shown in b (n = 100).

(D) Schematic view of the validation experiment performed in Figures 3E and 3F.

(E) Fraction of overlapping DE genes from 1,593 *huva GOI experiment* performed on the 500FG dataset and the same dataset with and *in silico* produced data bias.

(F) Heatmap representing the combinatorial overlap of DE genes form the 1,593 *huva experiments* shown in e for both, the original and the biased datasets.

(G) Schematic view of the experiments performed in Figure 3H.

(H) GSEA statistics (NES and –log10 p value) for the "GSE22935 WT VS MYD88 KO MACROPHAGE UP" signature shown if Figure 2G and the result for the enrichment of 1,000 randomly generated signatures of equal length on the same ranked gene list. Box plots were constructed in the style of Tukey, showing median, 25th and 75th percentiles. See also Figures S5–S9 and Table S5.

further tested if the randomization of samples would lead to the identification of differentially expressed (DE) genes (filtered only by p value cut-off). Strikingly, in most of the comparisons, we found no DE genes (Figures 3C and S5B) which further supports the notion that the biology of any given gene within a population can only be revealed by selecting those individuals within a *huva experiment* according to the measured expression level of the GOI being either at the lower or upper end of the expression spectrum.

Intrigued by the stark difference in the number of DE genes between random sampling and expression difference-based sampling, we tested if the sample selection might be influenced by parameters within the

data unrelated to biology, e.g. technical noise leading to unbalanced expression values in some samples, which would result in the selection of the same group of samples, independently from the selected gene or parameter. To test such a potential bias, we generated an artificial dataset based on the original dataset by gradually introducing noise via adding expression values across the complete sample set ranging from 0 to 10% addition of the original expression values for each gene within a sample's dataset and with a fixed added percentage for each individual sample (Figure S6A). The result of introducing such bias is visualized for all donors in the dataset (Figure S6B) and for individual samples (Figure S6C).

We next used the original and the biased datasets to test the influence of such bias. We restricted the analysis to a random selection of 10% of the present genes (n = 1,593, Table S5). We first asked how often a sample falling into the HIGH group for gene A was also included in the HIGH group for gene B, which we termed the sample overlap fraction (Figure S7A). Across the permutation of all genes included, the overlap of samples in the analysis was around 10% for both the LOW and HIGH experimental groups (Figures S7B–S7E), which is essentially the overlap from the random sampling of 10 out of 95 donors in the LOW and HIGH groups, the same percentage of samples of the two *huva* experimental groups in this example, for which the mean is 10% and median 10.52% of overlap (data not shown). The analysis on the same set of randomly selected genes, performed on the biased dataset shows a significant increase in overlap for either the LOW and HIGH groups (Figures S7B–S7E) supporting that our original dataset does not show such bias.

To evaluate whether the introduced bias also has an impact on the biological interpretation as defined by DE gene analysis, we compared the DE gene overlap of the same 1,593 genes in both the original and the biased dataset (Figure 3D). The DE gene overlap fraction was found to be moderate in the original dataset (Figures 3E and 3F). Only for a smaller number of genes, we identified almost complete overlap, strongly arguing for gene co-regulation and shared function as has been demonstrated in previous transcriptome analyses (Tarbier et al., 2020). In contrast, the gene overlap fraction substantially increased in the biased dataset compared to the original data (Figures 3E and 3F) indicating that technical noise needs to be considered and evaluated when working with new population datasets for *huva experiments* to exclude spurious co-regulation. We recommend the user to perform a similar set of tests on the new dataset together with a conventional exploratory data analysis (e.g. PCA and hierarchical clustering) to identify possible aberrant samples (e.g. low library complexity or RNA quality) or unexpected variance in the data that would bias the *huva* experiments.

Next, we compared the original and the biased dataset concerning fold change rank statistics. Here, we performed the *GOI huva experiment* on each of the 1,593 genes and correlated the obtained ranked gene lists based on the fold changes of the gene expression for the LOW vs. HIGH comparison from each experiment against the ranked gene lists of all other experiments (Figure S8A). Within the original data, we obtained an overall mean of 0 for the correlation coefficient between gene ranks as expected for the random ranking of genes while a positive mean was obtained for the biased dataset (Figure S8B).

Collectively, we provide means to test for potential hidden bias in newly generated population data, and we demonstrate that the dataset from the HFGP does not seem to contain any bias influencing the biological interpretation of the *huva experiments* presented here.

To further address that the biological findings for any given gene are not random effects within the dataset, we addressed the biological robustness of the results on the gene level. Here, we used the *MYD88* example and the results from the linked GSEA (Figures 2H and 2I) and compared the enrichment score and p value of the experimental signature to 1,000 randomly generated signatures of equal length (Figure 3G). Strikingly, no randomly generated signature received an equally strong signal enrichment on the fold change ranked gene list and only 4.6% of the random signatures showed p values lower than 0.05 (Figure 3H). This was also true for *STAT1*, *MAPK3* (ERK), and *AKT1* (Figures S9A–S9C), further supporting that the *huva experiment* extracts gene-related biology contained within population-based human variation.

Lastly, we determined the influence of the selected quantile cut-off on the results and biological interpretation of a given *huva experiment* (Figure S9D). We, therefore, performed *huva experiments* with all possible quantile cut-offs for a gene (*MYD88*) and evaluated the fold change and p value of the differential expression of all other genes in the comparison between the resulting experimental groups (Figures S9E

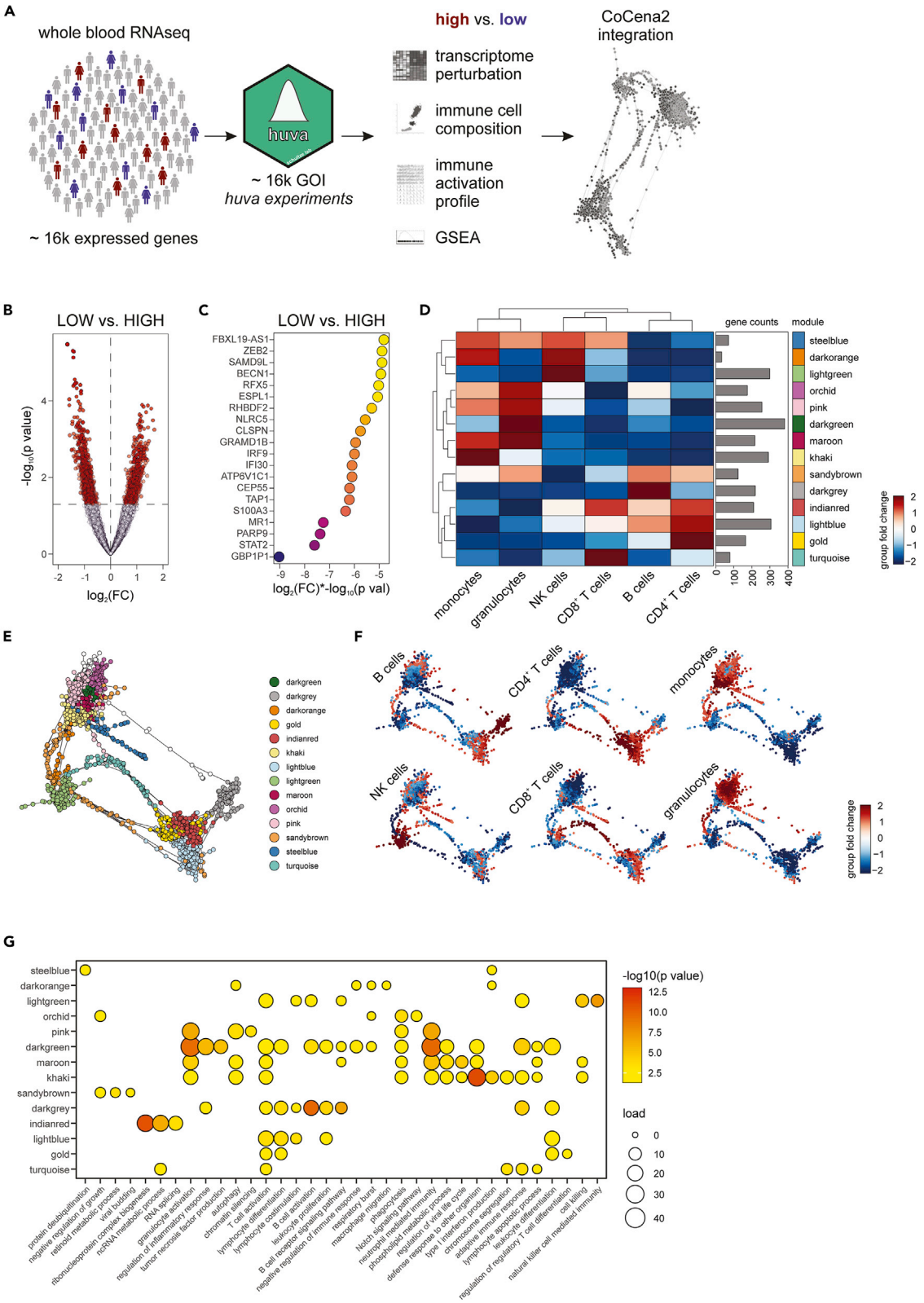

A
whole blood RNAseq
~ 16k expressed genes
huva
~ 16k GOI
huva experiments
high vs. low
transcriptome perturbation
immune cell composition
immune activation profile
GSEA
CoCena2 integration
B
LOW vs. HIGH
-log10(p value)
log2(FC)
C
LOW vs. HIGH
FBXL19-AS1
ZEB2
SAMD9L
BECN1
RFX5
ESPL1
RHBDF2
NLRC5
CLSPN
GRAMD1B
IRF9
IFI30
ATP6V1C1
CEP55
TAP1
S100A3
MR1
PARP9
STAT2
GBP1P1
log2(FC)*-log10(p val)
D
gene counts
module
steelblue
darkorange
lightgreen
orchid
pink
darkgreen
maroon
khaki
sandybrown
darkgrey
indianred
lightblue
gold
turquoise
group fold change
monocytes
granulocytes
NK cells
CD8+ T cells
B cells
CD4+ T cells
E
darkgreen
darkgrey
darkorange
gold
indianred
khaki
lightblue
lightgreen
maroon
orchid
pink
sandybrown
steelblue
turquoise
F
B cells
CD4+ T cells
monocytes
NK cells
CD8+ T cells
granulocytes
group fold change
G
steelblue
darkorange
lightgreen
orchid
pink
darkgreen
maroon
khaki
sandybrown
darkgrey
indianred
lightblue
gold
turquoise
protein deubiquitination
negative regulation of growth
retinoid metabolic process
viral budding
ribonucleoprotein complex biogenesis
ncRNA metabolic process
RNA splicing
granulocyte activation
regulation of inflammatory response
tumor necrosis factor production
autophagy
chromatin silencing
T cell activation
lymphocyte differentiation
lymphocyte costimulation
B cell activation
leukocyte proliferation
B cell receptor signaling pathway
negative regulation of immune response
respiratory burst
macrophage migration
phagocytosis
Notch signaling pathway
neutrophil mediated immunity
phospholipid metabolic process
regulation of viral life cycle
defense response to other organism
type I interferon production
chromosome segregation
adaptive immune response
lymphocyte apoptotic process
leukocyte differentiation
regulation of regulatory T cell differentiation
cell killing
natural killer cell mediated immunity
-log10(p value)
load

Figure 4. Large-scale *huva* analysis revealed a structured phenome

(A) Graphical overview of the analysis of Figures 4, S10, and S11.

(B) Volcano pot visualizing the log2 fold change and negative log10 p value for changes in monocyte cell counts for the transcriptome-wide *huva* analysis.

(C) Top 20 genes most influencing the total number of monocytes according to the analysis in b.

(D) Hierarchical clustering of the GFC (group fold change) for the modules identified in the Co-Cena^2 co-expression network analysis for the changes in cell counts. The number of genes in each module are shown as bar charts.

(E and F) Network visualization of the correlation between *huva experiments* colored according to the defined CeCena^2 modules (E) and GFC of the cellular populations (F).

(G) GOEA of selected gene sets across all Co-Cena^2 modules. See also Figure S10, Tables S6 and S8.

and S9F). The same analysis was performed using randomly selected samples for the two experimental groups (green dots). Here, we observed that altering the quantile setting has a strong impact on both average fold change (Figure S9E) and average p value (Figure S9F) of all genes for the LOW vs. HIGH comparison. Integrating both parameters revealed a bell-shaped distribution with a peak indicating the quantile setting, with maximal fold change and corresponding low p value (Figure S9G). The setting selected with these iterations also leads to the highest number of DE genes (Figure S9G), as evident from the peak of the curve coinciding with the maximum of the DE gene curve at the top side of the graph. Choosing a setting with a high number of DE genes for the gene of interest allows the user to have a more defined and precise picture of the difference between the two experimental groups and thus an easier interpretation of the results and downstream functional analysis. On the other hand, we noticed that the correlation of all ranked gene lists between quantiles appeared almost unchanged (Figure S9H) showing that independently from the cut-off used to define the extreme groups for the analysis, the qualitative difference between gene expression was not changed. This shows that the selection of quantiles will not fundamentally change the *huva* result, but will ease its biological interpretation. Indeed, the *huva* approach, performing a comparison between two defined groups, provides an output easier to implement in the downstream analysis compared to conventional gene-gene correlation analysis in which the difficulty to set a defined cut-off makes the interpretation of the results more challenging. This was not only true for *MYD88*, but also for several other tested genes (e.g. *STAT1*, *CRELD1*, *FOXP3*, data not shown). Of note, the quantile cut-off leading to the highest number of DE genes changes for each tested GOI. This observation was not correlated with the expression variance of the selected gene (data not shown).

Taken together, we demonstrate that the *huva experiment* is a robust approach to identify biological differences based on natural human variation, as supported by the overlap of insight gained from *in population* experiments with experimental data on genes of known function.

Large-scale *huva* analysis revealed a structured phenome

A fundamental characteristic of the *huva "in population" experiment* is the ease of applicability to any gene expressed in a certain cellular population. We here decided to run a *huva experiment* for all 15,927 expressed genes in bulk transcriptomes derived from whole blood immune cells (Li et al., 2016; Ter Horst et al., 2016). For each *GOI huva experiment*, we collected the fold change and p value for the comparison between the LOW and HIGH groups for all genes, cell counts of circulating immune cells, cytokine secretion upon stimulation, and functional enrichment of the hallmark gene sets (Subramanian et al., 2005) (Figures 4A and S10A). We obtained a high dimensional description of the functional connection between the transcriptome and the phenome (15,927 genes, 32,284 parameters) which can now be used to associate any given phenotype to previously undescribed genes or to link any GOI to new and undescribed biological roles for further hypothesis testing (Bonaguro et al., 2020).

Manual examination of the results can give an insight into which genes are mostly related to abundance changes in a certain cell type, for example, monocytes in blood (Figures 4B and 4C). *GOI huva experiments* leading to a negative fold change for the total number of blood monocytes means that having a lower expression of such a gene of interest (LOW group) is associated with a lower total number of monocytes in the blood. We imply in this context that the GOI has a positive contribution to monocyte numbers and their phenotype.

Among the top 20 most influential genes for monocyte abundance were *ZEB2* (Scott and Omilusik, 2019; Wu et al., 2016) and *IRF9* (Lee et al., 2017; Paul et al., 2018; Platanitis et al., 2019), both predominantly expressed in monocytes and known for their crucial function in monocyte activation and differentiation. Further, we noted *PARP9*, described to be involved in monocyte-derived macrophage differentiation (Iwata et al., 2016), and *MR1*, a mediator of ILC activation (Meierovics and Cowley, 2016; Salio et al.,

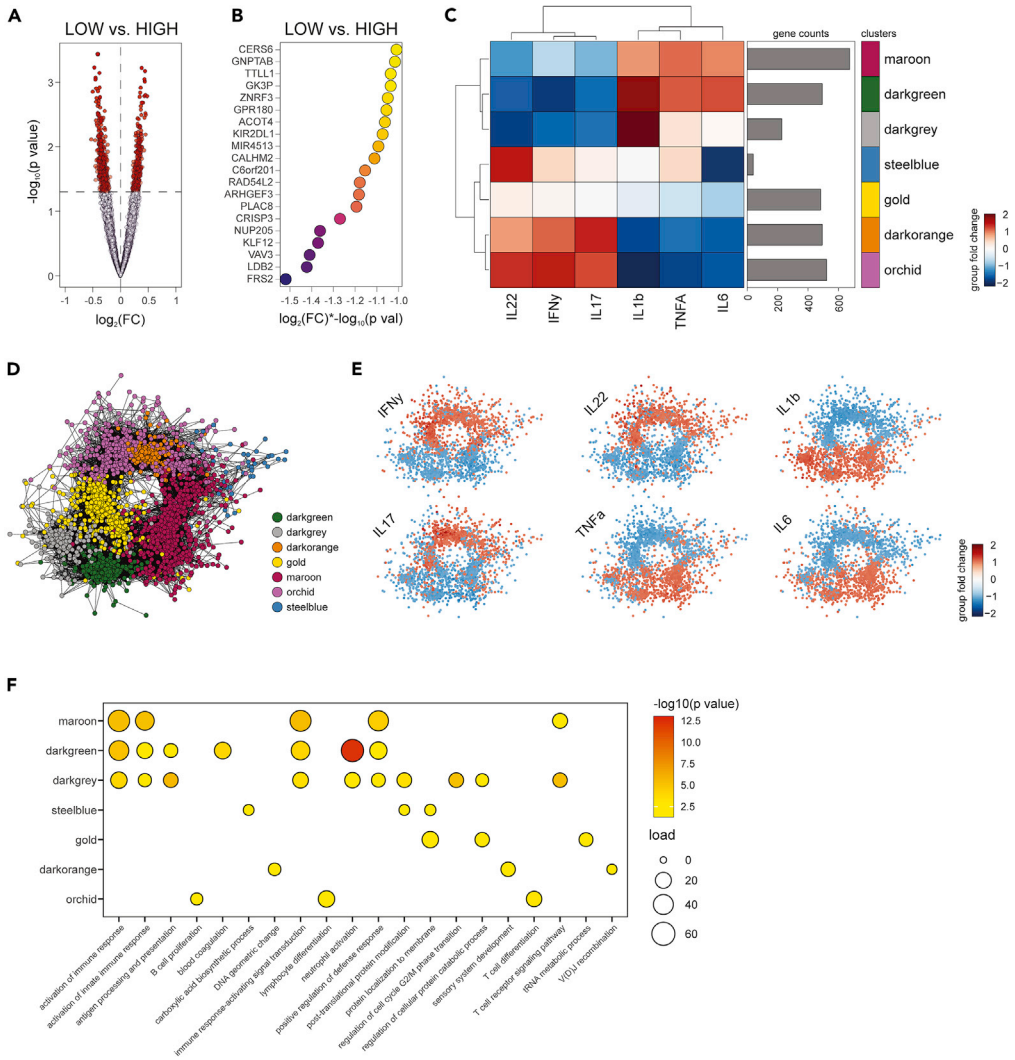


Figure 5. Large-scale *huva* analysis revealed a structured phenome

(A) Volcano plot visualizing the log2 fold change and negative log10 p value for changes in IFN-γ secretion after 48 h of Phytohaemagglutinin P (PHA) stimulation in the transcriptome-wide *huva* analysis.

(B) Top 20 genes most influencing IFN-γ secretion upon 24 h of PHA stimulation according to the analysis in b.

(C) Hierarchical clustering of the GFC (group fold change) for the modules identified in the Co-Cena[2] co-expression network analysis for the changes in cytokine secretion. The number of genes in each module are shown as bar charts.

(D and E) Network visualization of the correlation between *huva experiments* colored according to the defined CeCena[2] modules (D) and GFC of the cellular populations (E).

(F) GOEA of selected gene sets across all Co-Cena[2] modules. See also Figure S10, Tables S7 and S9.

2020; Ussher et al., 2016); both genes that are broadly expressed in immune cells, yet shown here to be regulated in the context of monocyte biology. The same approach can be applied to a functional phenotype of interest. For example, we investigated genes most associated with interferon-gamma secretion levels (Figures 5A and 5B). Here, we observed *KLF12*, a recently reported regulator of NK cell proliferation and IFN-γ production (Lam et al., 2019), or *PLAC8* involved in IFN-γ production in $CD4^+$ T cells (Slade et al., 2020). Interestingly, many genes in this list have previously not been connected to IFN-γ production or T cell activation, indicating that *huva* unveils novel candidate genes influencing distinct cellular functions opening up avenues for future research directions.

The high dimensionality of the results from single *huva experiments* for all expressed genes makes the manual annotation of the outcome challenging. We decided to use an unbiased approach to identify

modules of genes sharing common phenotypes probably involved in similar molecular pathways. To this end, we used the co-expression network analysis pipeline Co-Cena^2 (Aschenbrenner et al., 2021; Oestreich et al., 2022) recently developed for gene co-expression analysis and adapted here to use *huva* results as input. As a parameter for the calculation of the correlation between individual *huva experiments*, we used the product of the negative fold change and the negative logarithmic transformation of the p value (see STAR methods section for details). In order to assign a positive score when a gene is a positive regulator of a parameter of interest allowing for an easier visualization and interpretation of the results, we changed the sign of the fold change.

In this global analysis, we included cellular composition changes of the main immune cell populations in blood (B cells, CD4$^+$ T cells, CD8$^+$ T cells, monocytes, and granulocytes) and the *huva experiments* filtered for the 3,000 GOI showing the highest variance across the analyzed cellular populations. The Co-Cena^2 clustering algorithm identified 14 clusters of "*in population*" *huva experiments* with different patterns of the modulation of the analyzed cellular populations (Figures 4D–4F and Table S6). As expected, we identified clusters specific for the myeloid (orchid, pink and maroon) and lymphoid (indianred) compartments. Interestingly, GOEA for the genes in each cluster in the cell frequency network confirmed the specificity and the importance of each of those for specific cell types or immunological functions (Figure 4G and Table S8). For example, "natural killer cell mediated immunity" was enriched in the lightgreen module that is most prominently associated with a change in NK cell abundance in the blood, "T cell activation" and "lymphocytes differentiation" were enriched in the lightblue module related to B and T cell abundance. Interestingly, genes associated with ncRNA metabolic processes seem particularly important in B and T lymphocytes (indianred and turquoise modules). The connection between ncRNA and lymphocytes biology is still not fully established, showcasing how *huva* can be employed here to shed light on important biological processes and prioritize targets for validation experiments. Similarly, an analogous analysis for the levels of secreted cytokines (instead of cell population abundance) also shows a stark separation of *huva experiments* between lymphoid and myeloid compartment-associated cytokines (Figures 5C–5E and Table S7). Furthermore, dissecting the changes by the different stimulations, we observed modules specific for the stimulation with particular pathogens (e.g. influenza virus infection), disclosing new layers in the analysis and opening up further investigations (Figure S10B). Similarly to the cell count network, GOEA for the genes in each of the cytokine secretion network clusters revealed specific enrichment for terms related to immune cell function (T cell activation/neutrophils activation) but also to common molecular processes (DNA genomic changes/protein localization to the plasma membrane) known to be important for a coordinated immune response (Figure 5F and Table S9).

Taken together, we extended the *huva* framework to the whole transcriptome, integrated this information in a high-dimensional network and provide evidence of how this network enables the further assessment of genes of unknown functions within their functionally connected phenome modules.

huva predicts the predominant monocyte phenotype of patients with STAT1-activating mutations

As the *huva* analysis of the whole transcriptome can identify novel and unexpected roles of well-known genes, we tested this approach for *STAT1* (Figure S11A). Interestingly in our global analysis, STAT1 was part of the khaki module of genes impacting predominantly monocytes (Figure S11B). Being mainly recognized as an important transcription factor in lymphocytes (Mogensen, 2018), the fact that particularly monocytes are affected by the disruption of *STAT1* levels is a surprising novel observation. A closer look at the *GOI huva experiment* for *STAT1* revealed clear changes in the transcriptional program of the *STAT1* LOW and HIGH groups in the 500FG dataset, as visible in the principal component analysis (Figures 6A and S11A), which resulted in a high number of differentially expressed genes (225 down, 207 up/|FC|>2; Figure S11C and Table S10). Functional enrichment for the comparison revealed a strong inflammatory signature in *STAT1* HIGH donors (Figure 6B), with a significant enrichment of terms related to interferon response, inflammatory response, and cellular proliferation (Figures 6B and S11D; Table S11). This strong inflammatory signature is clearly paired with a marked increase in the number of all monocyte subsets, whereas according to the initial observation no changes were observed in any of the T cell populations (Figures 6C and S11E–S11G; Table S12, and data not shown). Investigating the functional changes between the *STAT1* LOW and HIGH groups in *ex vivo* PBMC stimulation experiments showed an increased production of key monocyte-derived cytokines, such as IL-1β (Figure S11H and Table S13). All in all, running the *GOI huva experiment* for *STAT1* highlights a predominant impact in monocyte biology in relation to the level of *STAT1* expression.

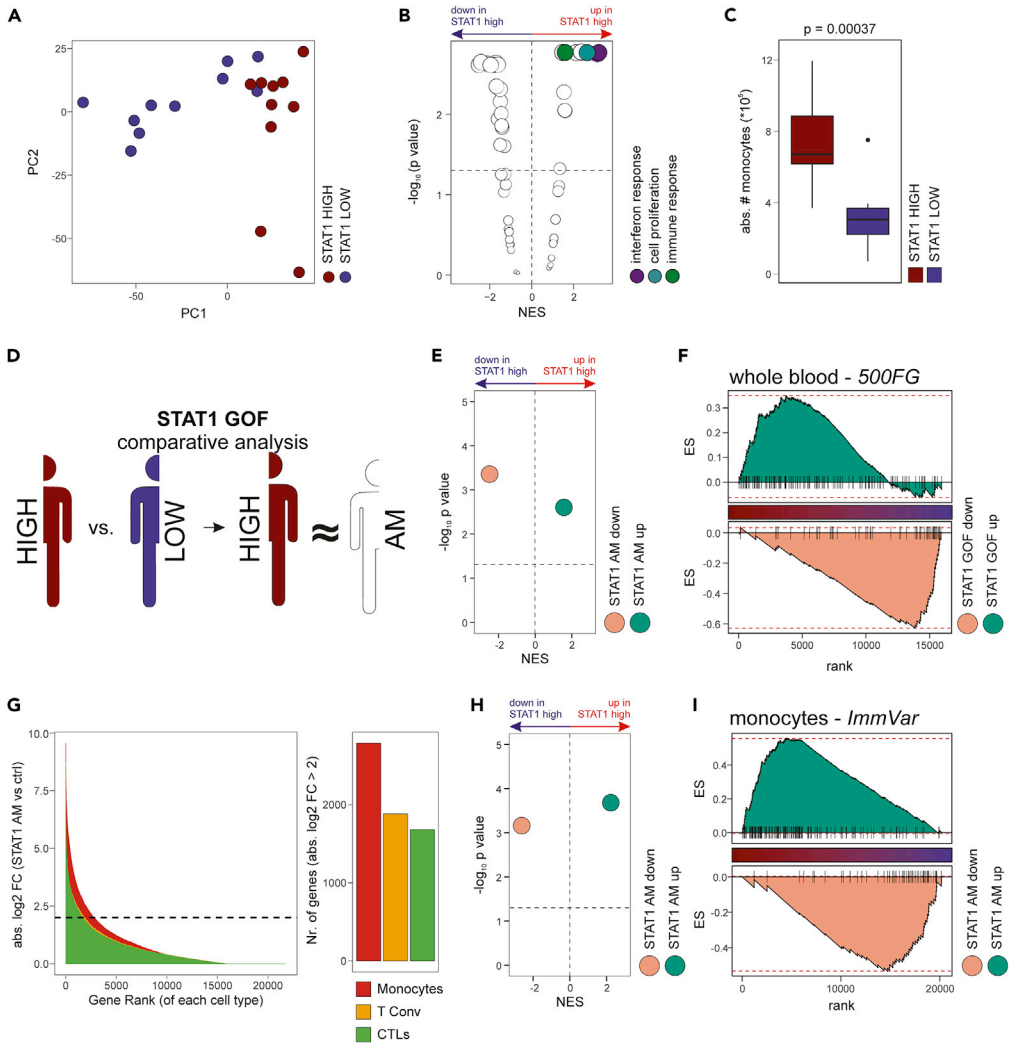


Figure 6. huva predicts predominant monocyte phenotype of patients with STAT1-activating mutations

(A) Principal component analysis for the transcriptome of the *STAT1* HIGH and *STAT1* LOW experimental groups.

(B) Volcano plot of the GSEA output (NES and p value) for the HALLMARK gene sets from the *huva experiment* in the comparison *STAT1* HIGH vs. *STAT1* LOW, colored for categories of interest upregulated in the *STAT1* HIGH group.

(C) Boxplot of the total number of circulating monocytes in the HIGH and LOW *huva* experimental groups.

(D) Schematic representation of the comparative analysis performed between the result of the STAT1 *huva experiment* and the experimental data of patients with STAT1 AM.

(E and F) GSEA of the STAT1 AM patients' PBMC up and down signatures on the ranked gene list of the *STAT1 huva experiment* in the 500FG whole blood dataset, volcano plot of the statistical output (E) and signature mapping (F).

(G) Distribution of the gene fold changes in the comparison of patients with STAT1 AM vs. controls in purified cell types (left) and count of genes with absolute FC higher than 2 (right).

(H and I) GSEA of the STAT1 AM patients' PBMC up and down signatures on the ranked gene list of the *STAT1 huva experiment* on $CD14^{+}$ monocytes from the ImmVar dataset, volcano plot of the statistical output (H) and signature mapping (I) Box plots were constructed in the style of Tukey, showing median, 25^{th} and 75^{th} percentiles, exact p value is shown from unpaired two-sided t-test. See Figures S11–S13; Tables S10, S11, S12, S13, S14, S15, and S16.

Next, by sampling blood from a small patient cohort we studied the actual effect of *STAT1*-activating mutation (AM) on their transcriptome. We conducted a targeted study on three patients carrying an activating point mutation in the *STAT1* gene manifesting as an autosomal dominant chronic mucocutaneous candidiasis (van de Veerdonk et al., 2011) (Table S14). The transcriptomes of PBMCs of the three STAT1 AM donors and three age-matched healthy controls were analyzed by differential expression analysis (Figure S12A) highlighting major differences between STAT1 AM and control samples, as displayed on a global

level by principal component analysis (Figure S12B). Interestingly, the expression of *STAT1* itself appears unchanged showing that no direct or feedback regulation of gene expression takes place owing to the AM mutation (Figure S12C). Conventional differential expression analysis revealed 203 differentially expressed genes (146 up, 57 down/|FC|>2, Figure S12D and Table S15). Typical inflammatory response genes such as *PI3* (Elafin - with antimicrobial function (Simpson et al., 1999)), *CD36* (Park, 2014), or *ALCAM* (CD166, a monocytes/T cell activation marker (Lyck et al., 2017; Nair et al., 2010)) were observed among the top upregulated genes in STAT1 AM PBMCs. Interestingly, also genes with immunosuppressive function, such as *IL1RN*, a scavenger receptor for IL-1α and IL-1β (Perrier et al., 2006) were upregulated, possibly as a result of a negative feedback in the chronic inflammatory state and pointing once more at monocytes as a driving force of the phenotype observed in these patients (Figure S12E). GOEA on the upregulated genes revealed a strong inflammatory signature, as already observed in the *GOI huva experiment* results (Figure S12F and Table S16).

The comparison of the transcriptomes from PBMC of STAT1 AM patients with the results of the *huva "in population" experiment* for *STAT1* was performed next to evaluate the overlap of the two approaches and the predictive potential of the *huva experiment* in the context of a human AM (Figures 6D and S11A). Using the differentially expressed genes from the analysis of patients with AM as a transcriptional signature, we performed GSEA on the ranked gene list from the comparison between *STAT1* LOW and *STAT1* HIGH *huva* groups (Figures 6E and 6F). Here, the downregulated genes in our STAT1 AM data were also among the most downregulated in the PBMC from the *GOI huva experiment* in the *STAT1* HIGH group (Figures 6E and 6F top). Accordingly, the upregulated genes in patients with STAT1 AM are upregulated in the *STAT1* HIGH *huva* experiment group (Figures 6E and 6F bottom). To further confirm the alignment between the *huva* results and the phenotype observed in patients with *STAT1* GOI, we investigated the expression of the *STAT1* target genes *TAP2*, *IRF1*, and *IFIT1* (Figure S13A) showing high agreement between both datasets and cell types.

To pinpoint the difference seen in PBMCs to a cell-intrinsic phenotype of the monocytic compartment, we further analyzed the transcriptomes of purified monocytes, conventional T (T conv) cells and cytotoxic lymphocytes (CTLs). Confirming the result of the *huva experiment*, the highest fold changes were found in monocytes when compared to the other cell types resulting in a total of 2,787 genes with an absolute fold change higher than 2 (compared to 1,882 for T conv, and 1,681 in CTLs) (Figure 6G).

Indeed, also when assessing the *STAT1 huva experiment* in FACS-purified $CD14^{+}$ monocytes from a different healthy population-based dataset (ImmVar CD14) (Figure 6H), both, the patient-derived up- (Figure 6I top) and downregulated (Figure 6I bottom) signatures from STAT1 AM PBMCs were enriched in the *STAT1* HIGH vs. LOW comparison. Thus, in combination with the observed changes in the PBMC data, the results argue for a cell-intrinsic effect of the STAT1 AM mutation in monocytes.

Taken together, our observations show for the first time the central role of monocytes in the pathophysiology of STAT1 AM carriers. The *huva* approach elucidated the mainly affected cell types as well as clinical manifestations of gene perturbation.

DISCUSSION

Human variation is driven by a combination of genetic and epigenetic determinants. The *huva* framework provides a powerful approach to exploit variation as an intrinsic property of any large human cohort to understand phenotype or function linked to a gene of interest. By stratifying gene expression data from the healthy 500FG cohort in a *huva gene of interest (GOI) experiment* for *MYD88*, we contrasted individuals with low vs. high expression with respect to their transcriptomes and cytokine production capacities. This "*in population*" experiment extracted transcriptomic alterations linked to TLR and NF-κB signaling and phenotypic manifestations, such as reduced IL-6 secretion after LPS exposure in *MYD88* LOW individuals, from the human data, which had been previously suggested by complete genetic loss of *MyD88* in murine models (Akira, 2003; Kawai et al., 1999; Qualls et al., 2010). We also applied this variance-based analysis approach to *STAT1*, a well-described transcription factor mediating inflammatory processes, e.g. upon interferon exposure. In PBMCs, *huva* revealed a central role of monocytes as a consequence of STAT1 perturbation. As many clinical genetic variants of *STAT1* have been described, we validated our findings within a cohort of STAT1-activating mutation carriers. We hypothesized that if natural variation in the healthy population can predict the biological role of a gene, our approach not only would uncover LOF phenotypes, but also reveal the phenotypic alterations for a clinically relevant activating mutation

(STAT1 AM). The multi-layered dataset of the 500FG cohort provides the possibility to interrogate if high *STAT1* expression is linked to any alterations in the transcriptome, the abundance of circulating immune cells, or cytokine secretion. Indeed, *STAT1* HIGH individuals exhibited a stronger enrichment of signatures related to inflammatory or interferon response than *STAT1* LOW donors. They produced more IL-1β upon *ex vivo* PBMC stimulation and showed enrichment of those DE-Gs derived from a patient-derived STAT1 AM vs. ctrl transcriptome comparison. Interestingly, the unexpected prediction of monocytes as the most perturbed cell type compared to lymphocytes was also confirmed by analyzing the transcriptome of isolated cell types from a STAT1 AM carrier. This result extends our understanding of the pathophysiology caused by STAT1 AM mutations and may help to tailor better therapeutic strategies.

To accommodate the broad applicability of our approach, we provide access to the *huva* framework for both the data science community and wet-lab scientists. On the one hand, we implemented *huva* in R and compiled predefined environments for *huva* analysis (e.g. Docker containers) giving immediate and versatile access to our framework. On the other hand, to facilitate the usage of *huva* by wet-lab scientists, we designed an interactive easy-to-use interface on the FASTGenomics platform (www.fastgenomics.org) allowing to run *huva experiments* without advanced programming skills (Data S1).

Beyond the technical aspects of the *huva* framework, a further prerequisite for the utilization of *huva* by the research community is the availability of suitable human data with both gene expression data and functional assays. Such studies have only recently become available, which also might explain why the concept of human variation for example to predict loss-of-function or activating mutations of individual genes has not been addressed earlier in population-based human data. The HFGP is certainly a prime example and more such datasets are currently assembled that will allow utilizing the *huva* approach to identify unknown human biology. We provide proof-of-principle how to use *huva* as a conceptually new strategy that does not solely link genetics with gene expression, but includes environmental influences from the beginning.

Using the transcriptome as the net output from the combined effect of genetics and epigenetics on gene expression in order to establish a link to the phenome leads to directly interpretable and exploitable insights into the functional network of any GOI. Thus, multi-layered data (expression data plus phenotypic and functional data) from large human cohort studies are an excellent starting point to propose links between gene expression and expression regulation and function. Other approaches to utilize such data focus on the integration of multi-omic layers within such datasets in an unsupervised fashion. An excellent recent example is MOFA2 (Argelaguet et al., 2018), a factor analysis model which provides a general framework for the integration of multi-omic data with the major aim to determine latent factors across different omics modalities that describe the variation within a given dataset and as results facilitate the identification of major cellular states or disease subgroups within a dataset. *Huva* on the other hand utilizes numerous different (multi-)omic datasets to infer the function of a gene (or a group of genes) based on extreme phenotypes (HIGH/LOW) across different datasets. In a sense, it is orthogonal to recent approaches such as MOFA2.

Genetic variation observed in healthy human cohorts consists of tolerable expression fluctuations. Elimination of a factor, e.g. a GOI, from a system to study its biological role, as it is the case for a gene in a genetic knockout model, will lead to a loss-of-function phenotype, but will also result in effects of compensatory mechanisms (El-Brolosy and Stainier, 2017) by the functional network it is usually embedded in, which may obscure the specific role of the GOI. Instead of complete loss by genetic removal, comparing the extremes of gene expression within its physiological range can allow for a more nuanced description of the connected biological functions and processes. Indeed, gene expression often needs to reach a certain threshold before manifesting in phenotypic changes (Cournac and Sepulchre, 2009; Goldbeter, 2005). Clearly, combined with newer genetic models based on CRISPR technology that can modulate gene expression without necessarily completely deleting a GOI (Qi et al., 2013), approaches such as *huva* starting with observations in the human setting, can define new ways of causally linking gene expression to gene function, complementing conventional approaches to build causality such as genetic models (e.g. KO mice).

The following examples for possible application of *huva* show the broad use of the framework: In the context of basic research, gene-centric scientific approaches will greatly benefit from *huva* and provide a framework to explore functional phenotypes for genes of interest. *Huva* can then help to make informed

decisions, to better design animal experiments and as such help to reduce the number of animal experiments, a major concern for both ethical and economical reasons (Ioannidis et al., 2014; Ter Riet et al., 2012). Data-driven approaches may embrace the concept to employ novel systematic approaches to understand the dynamics and determinants e.g. in all circulating cells of the immune system for which data is available. Third, with respect to medicine, *huva* can help identify pre-determinants in our response to perturbation for example uncovering genes whose expression is phenotypically connected to a more potent reaction to bacterial infection (Bossel Ben-Moshe et al., 2019; Haks et al., 2017).

Currently, *huva* primarily uses information from studies based on circulating immune cells in the blood of healthy cohorts (Li et al., 2016; Momozawa et al., 2018; Raj et al., 2014; Ter Horst et al., 2016). New multi-layered datasets of large cohorts, focusing on other cell types or even specific disease states, are currently being generated - creating the opportunity as well as a need for new avenues to explore this kind of data. We encourage extending *huva* with datasets including at least around 100 donors and one additional phenotypic or functional data layer to the transcriptome. Although one can use *huva* only within the transcriptome layer, the strength of our tool is the seamless integration of the transcriptome with other phenotypic and functional data layers. When including new data, conventional exploratory data analysis should be performed to make sure no technical bias is affecting the data. Furthermore, we provide examples (Figures 3 and S6–S8) and code (GitHub https://github.com/lorenzobonaguro/huva_reproducibility; Zenodo https://doi.org/10.5281/zenodo.7071267) to a series of tests to ensure the consistency of the data.

New questions may be posed, such as what the context/milieu-dependent relationship between the transcriptome and the phenome may be. With this scope in mind, we envision *huva* to be a very helpful tool for the analysis of these new datasets and designed the R implementation to be scalable and capable of quickly implementing new datasets. Future studies will provide more examples of the broad variety of possible applications for the *huva* framework.

In conclusion, *huva* exploits the natural variation found in human populations to infer the relationship between the transcriptome and its phenotypic manifestation. We used *huva* to uncover unknown roles of clinically relevant mutations in STAT1 and provide compelling evidence that *huva* aids basic research in hypothesis generation and experimental design. Given the versatility of the *huva* analysis, its implication in various contexts and the ease of integration with pre-existing workflows, we envision that *huva* will provide important biological insights in many fields.

Limitations of the study

The *huva* approach contrasts individuals with high and low expression of a GOI and we show how this comparison mimics the phenotype of a GOF or LOF setting in human and mouse models. Nevertheless, *huva* cannot infer *per se* a causal link between the GOI and the phenotype, for which further supporting experimental data are required (e.g assessment of human mutations, CRISPR-Cas KO). Further, while we described its application to the circulating immune system, the approach can be easily extended to other organs/tissues and biological questions. Our study offers a starting point for the investigation of human variation in healthy cohorts to explore gene function and as such opens new avenues to directly study the functional basis of genotype-phenotype relationships as well as environmental influences.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Benchmarking *huva* experiments
  - Pre-processing of the publicly available datasets
  - Stepwise description of the *huva* experiment
  - Statistical validation

- Transcriptome-wide *huva experiment* and CoCena[2] co-expression network analysis for *huva* results
- RNA-seq and pre-processing of STAT1 AM PBMC samples
- Analysis of STAT1 AM transcriptome data
- How to include a new dataset to *huva*

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105328.

ACKNOWLEDGMENTS

We thank T. Pecht and T. Kappellos for the critical revision of the *huva* framework and M. Kraut, H. Theis for technical assistance. The work was supported by the German Research Foundation (DFG) to J.L.S. under Germany's Excellence Strategy (EXC2151-390873048) as well as under SCHU 950/8-1; GRK 2168, TP11; SFB704, the BMBF-funded excellence project Diet-Body-Brain (DietBB) and the EU project SYSCID under grant number 733100. A.C.A. was supported by an intramural grant from the Department of Genomics & Immunoregulation at the LIMES Institute. M.G.N. was supported by a Spinoza grant of the Netherlands Organization for Scientific Research and an ERC Advanced Grant (833247).

AUTHOR CONTRIBUTIONS

Conceptualization was by L.B., A.C.A, J.S.-S, and J.L.S. The methodology was devised by L.B., A.C.A, J.S.-S, and J.L.S and C.C., L.L.S. B.R. I.G, and A.S. performed formal analysis. Resources were provided by K.H., T.U., S.R., M.G., P.A., A.H., F.v.d.V, L.A.B.J., and M.G.N. The draft article was written by L.B, A.C.A, and J.L.S. All authors reviewed and edited the article. Visualization was done by L.B. The project was supervised by A.C.A and L.B. Founding acquisition by A.C.A, J.L.S, and M.G.N.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: March 2, 2022
Revised: July 13, 2022
Accepted: October 7, 2022
Published: November 18, 2022

REFERENCES

Akira, S. (2003). Toll-like receptor signaling. J. Biol. Chem. *278*, 38105–38108.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124.

Aschenbrenner, A.C., Mouktaroudi, M., Krämer, B., Oestreich, M., Antonakos, N., Nuesch-Germano, M., Gkizeli, K., Bonaguro, L., Reusch, N., Baßler, K., et al. (2021). Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. Genome Med. *13*, 7.

Ashton, N.J., Hye, A., Rajkumar, A.P., Leuzy, A., Snowden, S., Suárez-Calvet, M., Karikari, T.K., Schöll, M., La Joie, R., Rabinovici, G.D., et al. (2020). An update on blood-based biomarkers for non-Alzheimer neurodegenerative disorders. Nat. Rev. Neurol. *16*, 265–284.

Bigaret, S., Hodgett, R.E., Meyer, P., Mironova, T., and Olteanu, A.-L. (2017). Supporting the multi-criteria decision aiding process: R and the MCDA package. EURO J. Decis. Process. *5*, 169–194.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008.

Bonaguro, L., Köhne, M., Schmidleithner, L., Schulte-Schrepping, J., Warnat-Herresthal, S., Horne, A., Kern, P., Günther, P., Ter Horst, R., Jaeger, M., et al. (2020). CRELD1 modulates homeostasis of the immune system in mice and humans. Nat. Immunol. *21*, 1517–1527.

Bossel Ben-Moshe, N., Hen-Avivi, S., Levitin, N., Yehezkel, D., Oosting, M., Joosten, L.A.B., Netea, M.G., and Avraham, R. (2019). Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human immune cells. Nat. Commun. *10*, 3266.

Buxadé, M., Lunazzi, G., Minguillón, J., Iborra, S., Berga-Bolaños, R., Del Val, M., Aramburu, J., and López-Rodríguez, C. (2012). Gene expression induced by Toll-like receptors in macrophages requires the transcription factor NFAT5. J. Exp. Med. *209*, 379–393.

Cohen, P. (2014). The TLR and IL-1 signalling network at a glance. J. Cell Sci. *127*, 2383–2390.

Cournac, A., and Sepulchre, J.-A. (2009). Simple molecular networks that respond optimally to time-periodic stimulation. BMC Syst. Biol. *3*, 29.

Diamond, M.S., and Farzan, M. (2013). The broad-spectrum antiviral functions of IFIT and IFITM proteins. Nat. Rev. Immunol. *13*, 46–57.

El-Brolosy, M.A., and Stainier, D.Y.R. (2017). Genetic compensation: a phenomenon in search of mechanisms. PLoS Genet. *13*, e1006780.

Favé, M.J., Lamaze, F.C., Soave, D., Hodgkinson, A., Gauvin, H., Bruat, V., Grenier, J.-C., Gbeha, E., Skead, K., Smargiassi, A., et al. (2018). Gene-by-environment interactions in urban populations modulate risk phenotypes. Nat. Commun. *9*, 827.

Furci, L., Jain, R., Stassen, J., Berkowitz, O., Whelan, J., Roquis, D., Baillet, V., Colot, V., Johannes, F., and Ton, J. (2019). Identification and characterisation of hypomethylated DNA loci controlling quantitative resistance in Arabidopsis. Elife *8*, e40655.

Gamrekelashvili, J., Kapanadze, T., Sablotny, S., Ratiu, C., Dastagir, K., Lochner, M., Karbach, S., Wenzel, P., Sitnow, A., Fleig, S., et al. (2020). Notch and TLR signaling coordinate monocyte cell fate and inflammation. Elife *9*, e57007.

Gibson, G. (2008). The environmental contribution to gene expression profiles. Nat. Rev. Genet. *9*, 575–581.

Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A.J., Mesirov, J.P., and Haining, W.N. (2016). Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. Immunity *44*, 194–206.

Goldbeter, A. (2005). Zero-order switches and developmental thresholds. Mol. Syst. Biol. *1*, 2005.0031.

GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

Haks, M.C., Bottazzi, B., Cecchinato, V., De Gregorio, C., Del Giudice, G., Kaufmann, S.H.E., Lanzavecchia, A., Lewis, D.J.M., Maertzdorf, J., Mantovani, A., et al. (2017). Molecular signatures of immunity and immunogenicity in infection and vaccination. Front. Immunol. *8*, 1563.

Hoxhaj, G., and Manning, B.D. (2020). The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. Nat. Rev. Cancer *20*, 74–88.

Ioannidis, J.P.A., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R., Moher, D., Schulz, K.F., and Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. Lancet *383*, 166–175.

Iwata, H., Goettsch, C., Sharma, A., Ricchiuto, P., Goh, W.W.B., Halu, A., Yamada, I., Yoshida, H., Hara, T., Wei, M., et al. (2016). PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. Nat. Commun. *7*, 12849.

Jin, M.S., and Lee, J.-O. (2008). Structures of the toll-like receptor family and its ligand complexes. Immunity *29*, 182–191.

John, S.P., Sun, J., Carlson, R.J., Cao, B., Bradfield, C.J., Song, J., Smelkinson, M., and Fraser, I.D.C. (2018). IFIT1 exerts opposing regulatory effects on the inflammatory and interferon gene programs in LPS-activated human macrophages. Cell Rep. *25*, 95–106.e6.

Kaisho, T., and Akira, S. (2001). Dendritic-cell function in Toll-like receptor- and MyD88-knockout mice. Trends Immunol. *22*, 78–83.

Kapellos, T.S., Bonaguro, L., Gemünd, I., Reusch, N., Saglam, A., Hinkley, E.R., and Schultze, J.L. (2019). Human monocyte subsets and phenotypes in major chronic inflammatory diseases. Front. Immunol. *10*, 2035.

Kawai, T., Adachi, O., Ogawa, T., Takeda, K., and Akira, S. (1999). Unresponsiveness of MyD88-deficient mice to endotoxin. Immunity *11*, 115–122.

Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S.E., Hamel, A.R., Viñuela, A., Roberts, A.L., et al. (2020). Cell type-specific genetic regulation of gene expression across human tissues. Science *369*, eaaz8528.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2016). Fast gene set enrichment analysis. Preprint at bioRxiv. https://doi.org/10.1101/060012.

Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 414–430.

Lam, V.C., Folkersen, L., Aguilar, O.A., and Lanier, L.L. (2019). KLF12 regulates mouse NK cell proliferation. J. Immunol. *203*, 981–989.

Lee, A.J., Chen, B., Chew, M.V., Barra, N.G., Shenouda, M.M., Nham, T., van Rooijen, N., Jordana, M., Mossman, K.L., Schreiber, R.D., et al. (2017). Inflammatory monocytes require type I interferon receptor signaling to activate NK cells via IL-18 during a mucosal viral infection. J. Exp. Med. *214*, 1153–1167.

Li, Y., Oosting, M., Smeekens, S.P., Jaeger, M., Aguirre-Gamboa, R., Le, K.T.T., Deelen, P., Ricaño-Ponce, I., Schoffelen, T., Jansen, A.F.M., et al. (2016). A functional genomics approach to understand variation in cytokine production in humans. Cell *167*, 1099–1110.e14.

Liao, Y., Smyth, G.K., and Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. *47*, e47.

López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat. Rev. Genet. *9*, 583–593.

Lord, K.A., Hoffman-Liebermann, B., and Liebermann, D.A. (1990). Nucleotide sequence and expression of a cDNA encoding MyD88, a novel myeloid differentiation primary response gene induced by IL6. Oncogene *5*, 1095–1097.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and' ' dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Love, M.I., Anders, S., Kim, V., and Huber, W. (2016). RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Res. *4*, 1070.

Lyck, R., Lécuyer, M.A., Abadier, M., Wyss, C.B., Matti, C., Rosito, M., Enzmann, G., Zeis, T., Michel, L., García Martín, A.B., et al. (2017). ALCAM (CD166) is involved in extravasation of monocytes rather than T cells across the blood-brain barrier. J. Cereb. Blood Flow Metab. *37*, 2894–2909.

Majewski, J., and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. *27*, 72–79.

Manning, B.D., and Toker, A. (2017). AKT/PKB signaling: navigating the network. Cell *169*, 381–405.

Meierovics, A.I., and Cowley, S.C. (2016). MAIT cells promote inflammatory monocyte differentiation into dendritic cells during pulmonary intracellular infection. J. Exp. Med. *213*, 2793–2809.

Mogensen, T.H. (2018). IRF and STAT transcription factors - from basic biology to roles in infection, protective immunity, and primary immunodeficiencies. Front. Immunol. *9*, 3047.

Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., Gori, A.-S., et al. (2018). IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nat. Commun. *9*, 2427.

Muzio, M., Ni, J., Feng, P., and Dixit, V.M. (1997). IRAK (Pelle) family member IRAK-2 and MyD88 as proximal mediators of IL-1 signaling. Science *278*, 1612–1615.

Nair, P., Melarkode, R., Rajkumar, D., and Montero, E. (2010). CD6 synergistic co-stimulation promoting proinflammatory response is modulated without interfering with the activated leucocyte cell adhesion molecule interaction. Clin. Exp. Immunol. *162*, 116–130.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.

Oestreich, M., Holsten, L., Agrawal, S., Dahm, K., Koch, P., Jin, H., Becker, M., and Ulas, T. (2022). hCoCena: horizontal integration and analysis of transcriptomics datasets. Bioinformatics *38*, 4727–4734.

Park, Y.M. (2014). CD36, a scavenger receptor implicated in atherosclerosis. Exp. Mol. Med. *46*, e99.

Paul, A., Tang, T.H., and Ng, S.K. (2018). Interferon regulatory factor 9 structure and regulation. Front. Immunol. *9*, 1831.

Pelikan, R.C., Kelly, J.A., Fu, Y., Lareau, C.A., Tessneer, K.L., Wiley, G.B., Wiley, M.M., Glenn, S.B., Harley, J.B., Guthridge, J.M., et al. (2018). Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. Nat. Commun. *9*, 2905.

Perrier, S., Darakhshan, F., and Hajduch, E. (2006). IL-1 receptor antagonist in metabolic diseases: Dr Jekyll or Mr Hyde? FEBS Lett. *580*, 6289–6294.

Pividori, M., Rajagopal, P.S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., Park, Y., Consortium, G., Wen, X., and Im, H.K. (2020). PhenomeXcan: mapping the genome to the phenome through the transcriptome. Sci. Adv. *6*, eaba2083.

Platanitis, E., Demiroz, D., Schneller, A., Fischer, K., Capelle, C., Hartl, M., Gossenreiter, T., Müller, M., Novatchkova, M., and Decker, T. (2019). A molecular switch from STAT2-IRF9 to ISGF3 underlies interferon-induced gene transcription. Nat. Commun. *10*, 2921.

Platt, C.D., Zaman, F., Wallace, J.G., Seleman, M., Chou, J., Al Sukaiti, N., and Geha, R.S. (2019). A novel truncating mutation in MYD88 in a patient with BCG adenitis, neutropenia and delayed umbilical cord separation. Clin. Immunol. *207*, 40–42.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Qualls, J.E., Neale, G., Smith, A.M., Koo, M.-S., DeFreitas, A.A., Zhang, H., Kaplan, G., Watowich, S.S., and Murray, P.J. (2010). Arginine usage in mycobacteria-infected macrophages depends on autocrine-paracrine cytokine signaling. Sci. Signal. *3*, ra62.

Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science *344*, 519–523.

Rajewsky, N., Almouzni, G., Gorski, S.A., Aerts, S., Amit, I., Bertero, M.G., Bock, C., Bredenoord, A.L., Cavalli, G., Chiocca, S., et al. (2020). LifeTime and improving European healthcare through cell-based interceptive medicine. Nature *587*, 377–386.

Randolph, G.J., Sanchez-Schmitz, G., Liebman, R.M., and Schäkel, K. (2002). The CD16(+) (FcgammaRIII(+)) subset of human monocytes preferentially becomes migratory dendritic cells in a model tissue setting. J. Exp. Med. *196*, 517–527.

Ter Riet, G., Korevaar, D.A., Leenaars, M., Sterk, P.J., Van Noorden, C.J.F., Bouter, L.M., Lutter, R., Elferink, R.P.O., and Hooft, L. (2012). Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. PLoS One *7*, e43404.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Sabroe, I., Jones, E.C., Whyte, M.K.B., and Dower, S.K. (2005). Regulation of human neutrophil chemokine receptor expression and function by activation of Toll-like receptors 2 and 4. Immunology *115*, 90–98.

Salio, M., Awad, W., Veerapen, N., Gonzalez-Lopez, C., Kulicke, C., Waithe, D., Martens, A.W.J., Lewinsohn, D.M., Hobrath, J.V., Cox, L.R., et al. (2020). Ligand-dependent downregulation of MR1 cell surface expression. Proc. Natl. Acad. Sci. USA *117*, 10465–10475.

Scott, C.L., and Omilusik, K.D. (2019). Zebs: novel players in immune cell development and function. Trends Immunol. *40*, 431–446.

Simpson, A.J., Maxwell, A.I., Govan, J.R., Haslett, C., and Sallenave, J.M. (1999). Elafin (elastase-specific inhibitor) has anti-microbial activity against gram-positive and gram-negative respiratory pathogens. FEBS Lett. *452*, 309–313.

Slade, C.D., Reagin, K.L., Lakshmanan, H.G., Klonowski, K.D., and Watford, W.T. (2020). Placenta-specific 8 limits IFNγ production by CD4 T cells in vitro and promotes establishment of influenza-specific CD8 T cells in vivo. PLoS One *15*, e0235706.

Smeekens, S.P., Ng, A., Kumar, V., Johnson, M.D., Plantinga, T.S., van Diemen, C., Arts, P., Verwiel, E.T.P., Gresnigt, M.S., Fransen, K., et al. (2013). Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans. Nat. Commun. *4*, 1342.

Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B.H.F. (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. Sci. Rep. *8*, 5865.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484.

Tarbier, M., Mackowiak, S.D., Frade, J., Catuara-Solarz, S., Biryukova, I., Gelali, E., Menéndez, D.B., Zapata, L., Ossowski, S., Bienko, M., et al. (2020). Nuclear gene proximity and protein interactions shape transcript covariations in mammalian single cells. Nat. Commun. *11*, 5445.

Ter Horst, R., Jaeger, M., Smeekens, S.P., Oosting, M., Swertz, M.A., Li, Y., Kumar, V., Diavatopoulos, D.A., Jansen, A.F.M., Lemmers, H., et al. (2016). Host and environmental factors influencing individual human cytokine responses. Cell *167*, 1111–1124.e13.

Thomas, S., Rouilly, V., Patin, E., Alanio, C., Dubois, A., Delval, C., Marquier, L.-G., Fauchoux, N., Sayegrih, S., Vray, M., et al. (2015). The Milieu Intérieur study - an integrative approach for study of human immunological variance. Clin. Immunol. *157*, 277–293.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Ussher, J.E., van Wilgenburg, B., Hannaway, R.F., Ruustal, K., Phalora, P., Kurioka, A., Hansen, T.H., Willberg, C.B., Phillips, R.E., and Klenerman, P. (2016). TLR signaling in human antigen-presenting cells regulates MR1-dependent activation of MAIT cells. Eur. J. Immunol. *46*, 1600–1614.

van de Veerdonk, F.L., Plantinga, T.S., Hoischen, A., Smeekens, S.P., Joosten, L.A.B., Gilissen, C., Arts, P., Rosentul, D.C., Carmichael, A.J., Smits-van der Graaf, C.A.A., et al. (2011). STAT1 mutations in autosomal dominant chronic mucocutaneous candidiasis. N. Engl. J. Med. *365*, 54–61.

von Bernuth, H., Picard, C., Jin, Z., Pankla, R., Xiao, H., Ku, C.-L., Chrabieh, M., Mustapha, I.B., Ghandil, P., Camcioglu, Y., et al. (2008). Pyogenic bacterial infections in humans with MyD88 deficiency. Science *321*, 691–696.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. *51*, 592–599.

Wang, J.Q., Jeelall, Y.S., Ferguson, L.L., and Horikawa, K. (2014). Toll-like receptors and cancer: MYD88 mutation and inflammation. Front. Immunol. *5*, 367.

Warnat-Herresthal, S., Perrakis, K., Taschler, B., Becker, M., Baßler, K., Beyer, M., Günther, P., Schulte-Schrepping, J., Seep, L., Klee, K., et al. (2020). Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics. iScience *23*, 100780.

Wu, X., Briseño, C.G., Grajales-Reyes, G.E., Haldar, M., Iwata, A., Kretzer, N.M., Kc, W., Tussiwand, R., Higashi, Y., Murphy, T.L., et al. (2016). Transcription factor Zeb2 regulates commitment to plasmacytoid dendritic cell and monocyte fate. Proc. Natl. Acad. Sci. USA *113*, 14775–14780.

Xiao, Y., Hsiao, T.-H., Suresh, U., Chen, H.-I.H., Wu, X., Wolf, S.E., and Chen, Y. (2014). A novel significance score for gene selection and ranking. Bioinformatics *30*, 801–807.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ImmVar	GEO	GSE56035
HFGP 500FG	GEO	GSE134080
CEDAR	Array Express	http://cedar-web.giga.ulg.ac.be; E-MTAB-6666 and E-MTAB-6667
STAT1 AM RNAseq	This paper	EGAS00001005041 https://ega-archive.org/studies/EGAS00001005041
Software and algorithms		
Huva v. 0.1.4	This paper	https://github.com/lorenzobonaguro/huva; https://doi.org/10.5281/zenodo.7071267
Huva.db v. 0.1.4	This paper	https://github.com/lorenzobonaguro/huva.db; https://doi.org/10.5281/zenodo.7071267
DEseq2 v. 1.30.1	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Limma v. 3.46.0	Ritchie et al., 2015	https://bioconductor.org/packages/release/bioc/html/limma.html
Fgsea v. 1.12.0	Korotkevich et al., 2016	https://bioconductor.org/packages/release/bioc/html/fgsea.html
ggplot2 v. 3.3.3	R Tidyverse	https://ggplot2.tidyverse.org/
CoCena[2]	Aschenbrenner et al., 2021; Oestreich et al., 2022	https://github.com/MarieOestreich/hCoCena
Igraph - Louvain clustering – v. 1.2.6	Blondel et al., 2008	https://igraph.org/r/
clusterProfiler v. 3.12.0	Yu et al., 2012	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
TopHat v. 2.0.4	Trapnell et al., 2009	https://ccb.jhu.edu/software/tophat/index.shtml
Rsubread v. 2.10.1	Liao et al., 2019	https://bioconductor.org/packages/release/bioc/html/Rsubread.html
Docker desktop v. 20.10.16	https://www.docker.com/products/docker-desktop	RRID:SCR_016445
R v. 4.0.3	http://www.r-project.org/	RRID:SCR_001905
Other		
Code for reproducibility of the analysis	This paper	https://github.com/lorenzobonaguro/huva_reproducibility; https://doi.org/10.5281/zenodo.7071267
Huva web portal	This paper	https://beta.fastgenomics.org/a/huva

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and code should be directed to and will be fulfilled by the lead contact Anna Aschenbrenner (anna.aschenbrenner@dzne.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The STAT1 AM RNA-seq data have been deposited at EGA and are available upon approval of the data access committee as of the date of publication. Accession numbers are listed in the key resources table.
- This paper analyzes existing, publicly available data (ImmVar, 500FG, CEDAR). These accession numbers for the datasets are listed in the key resources table.
- The source code of the *huva* R package has been deposited at Zenodo and GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table. *Huva* is distributed opens source under GPL 3 license.
- The source code of the *huva.db* R package has been deposited at Zenodo and GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table. *Huva.db* is distributed opens source under GPL 3 license.
- Docker images *huva.docker* and *huva.shiny* are available on dockerhub (https://hub.docker.com/u/lorenzobonaguro).
- The code necessary to reproduce the analysis has been deposited at Zenodo and GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table. All the analysis in this manuscript was performed with *huva* v. 0.1.4 with the companion *huva.db* v. 0.1.4 within a dockerized environment.
- *Huva* is also available with a user-friendly interface at FASTGenomics (https://beta.fastgenomics.org/a/huva).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

For the transcriptome analysis of STAT1 AM patients, blood was collected from three patients and three control subjects (4 males, 2 females, age 27–39, for details see Table S14) after informed consent at the Radboud University Nijmegen Medical Centre (RUNMC, Nijmegen, the Netherlands). The study was approved by the Institutional Review Board of the Radboud University Nijmegen Medical Centre.

METHOD DETAILS

Benchmarking *huva* experiments

The evaluation of the performance of the *huva experiments* was performed on a workstation PC (Intel Core I7-8700K @3.70GHz, 32 GB RAM memory, Windows 10 Pro version 2004). The *huva* results for each run were calculated for all available studies (500FG, ImmVar, CEDAR, PBMC collection) and shown as cumulative time in Figure S3. The *huva GOI experiments* were performed on the same set of randomly selected genes expressed in at least one of the included datasets. The performance experiments were performed multiple times with different sampling of genes for the analysis obtaining similar results, one representative result is shown in Figure S3.

Pre-processing of the publicly available datasets

500FG

Row counts, sample annotation and phenotype data (cell count and cytokines secretion) were provided by the HFGP, transcriptome data of this study are available as GEO dataset (GSE134080). RNA-sequencing counts were first normalized according to the DEseq2 (Love et al., 2014) pipeline, filtered for genes with a cumulative number of count higher than 100 and vst transformed. Transformed counts were corrected for the sex of the donor using the function provided in the limma R package (Ritchie et al., 2015). For the use in the *huva* framework ENSEMBL IDs were converted to gene symbols, duplicated gene symbols were filtered choosing the gene showing the highest variance across the samples. The final dataset includes 95 samples and 15,927 genes.

CEDAR

The CEDAR dataset was provided as a log2-transformed pre-processed expression table (http://cedar-web.giga.ulg.ac.be). The dataset was filtered to include healthy donors only and batch corrected for

experimental batch and sex. Duplicated gene symbols derived from translation of Illumina ProbeID were filtered choosing the gene displaying the highest variance across the samples. Each cell type included in the study was pre-processed independently, Finally the dataset includes 242 samples and 10,541 genes including the expression profile of $CD4^+$ T cell (CD4T), $CD8^+$ T cells (CD8T), monocytes (CD14M), B cells (CD19B), granulocytes (CD15G) and platelets (PLA).

ImmVar

The ImmVar dataset was downloaded from GEO as a pre-processed expression table (GSE56035). Duplicated gene symbols were filtered choosing the gene displaying the highest variance across the samples and the expression was log2-transformed to be used in the *huva* framework. The dataset includes expression profiles from $CD4^+$ T cells (CD4T) and $CD14^+$ monocytes (CD14M) including 499 samples and 20,231 expressed genes.

PBMC collection

The three PBMC datasets described in Warnat-Herresthal et al. 2020 (Warnat-Herresthal et al., 2020) were batch corrected for study ID with the appropriate function provided in the limma R package (Ritchie et al., 2015) to remove batch deriving from different experimental settings or covariates. The dataset was filtered to include only PBMC samples from healthy donors for a total of 41 samples and 12,708 genes (dataset 1), 638 samples and 12,708 genes (dataset 2) as well as 61 samples and 12,708 genes (dataset 3).

Stepwise description of the *huva* experiment

The *huva* framework is implemented in R, the analysis in this manuscript was performed using R version 4.0.3. The *huva* package calculates the result of the *huva experiment* consecutively for all datasets included. At first, the quantile parameter, the threshold used to define the HIGH and LOW experimental groups, was selected. If not stated differently, a quantile of 0.1 (10%) was used for the analysis. The enrichment is then used to define the quantiles cut-off and the two experimental groups. Differential expression analysis was performed using the limma R package (Ritchie et al., 2015) using the experimental groups in the design model, p-value correction for multiple testing and fold change cut-off for each experiment are reported in the result section for each experiment separately. GSEA within the *huva* function is performed with the R package fgsea (Korotkevich et al., 2016) with standard setting (1000 random permutations), the gene rank used for GSEA is calculated according to the log2 fold change in the comparison between the LOW and HIGH groups. The results of the *huva experiment* are collected in a *huva_experiment* R object used as input for the provided functions to explore the output for each dataset. Statistical significance for cell count and cytokine secretion was calculated with an unpaired two-sided Student's T test. All graphical output of the *huva experiment* are calculated with built-in functions as ggplot2 objects. The only exceptions are all heat maps calculated with the R package pheatmap within a separate *huva* function.

Statistical validation

Group sample randomization

We performed a *huva experiment* on a gene with high variance (*SLC12A1*) and one with low variance (*CRELD1*), across the 500FG dataset and compared the results of this experiment with the result of the same experiment in which the samples were randomly selected. In this experiment, the samples in the LOW and HIGH groups were randomly selected independently from gene expression. The fold changes and p-values for *SLC12A1* and *CRELD1* were then collected and analysed. For each *huva experiment*, also the number of differentially expressed genes with significant p-value after Benjamini & Hochberg correction for multiple testing is shown. The randomization of the HIGH and LOW groups was permuted 100 times.

Overlap of huva experiments results

With the aim to identify potential bias in the used dataset, we designed a series of experiments to both, investigate this bias in the data used and to simulate the result of a *huva experiment* when this bias was intentionally added to the data (biased dataset). We performed a *huva experiment* for several genes and compared the results between each other on several levels: overlap of samples in the HIGH and LOW group, DE genes, and ranked gene list. The genes used for the analysis were randomly sampled for the 10% of expressed genes in the 500FG dataset. The *huva* result was calculated with both the original *huva* database and an *in silico*-biased dataset. The biased dataset was produced adding randomly to each

sample from 0 to 10% of the original gene expression values. The result of each *huva experiment* was compared to all other results of the selected genes in the analysis.

Randomization of gene signature

To challenge the biological value of the results of the *huva GOI experiment*, we performed GSEA on 1,000 randomly generated signatures of an equal number of genes to the genes of the "GSE22935 WT VS MYD88 KO MACROPHAGE UP" signature expressed in the 500FG dataset (186 genes), the GSEA was performed with the standard setting of the *huva* framework using the fgsea R package (Korotkevich et al., 2016). Similar analysis was performed also for *AKT1*, signature "AKT_UP.V1_UP" (138 genes); *MAPK3*, signature "GO_ERK1_AND_ERK2_CASCADE" (219 genes) and *STAT1*, signature "GSE40666_WT_VS_STAT1_KO_CD8_T-CELL_UP" (195 genes).

Variation of quantile cut-off

To evaluate the impact of the quantile threshold in the results of the *huva* analysis, the *huva experiment* was performed using a quantile setting ranging from 3% to 49% with intervals of 1%. For each experiment, we collected the median $\log_2$ fold change and Benjamini & Hochberg corrected p-value for all present genes in the comparison between the LOW and HIGH *huva* experimental groups. In the calculation of the median value of fold change and p-value, the gene of interest selected as input of the *huva experiment* was removed to avoid bias on the results since the difference in expression of this gene was set by the experimental definition.

Transcriptome-wide *huva experiment* and CoCena^2 co-expression network analysis for *huva* results

For the transcriptome-wide *huva experiment*, we performed a *huva* GOI experiment for all expressed genes in the 500FG dataset. The *huva* results were calculated only for the 500FG datasets and collected in separate tables for fold change and p-value and also for each of the stored parameters (genes, cell count, cytokines secretion and hallmark enrichment). As a metric for the visualization of the results and the calculation of the co-expression network the product of the calculated $\log_2$-transformed fold change and negative $\log_{10}$ transformation of the p-value (Xiao et al., 2014) was used. For the calculation of the correlation network between the *huva* results, we implemented the CoCena^2 co-expression network analysis pipeline (Aschenbrenner et al., 2021; Oestreich et al., 2022). As input for the network calculation we used the table of the product of the calculated $\log_2$-transformed fold changes and negative $\log_{10}$ transformation of the p-values. For the network describing the changes in blood cellular composition, we filtered for changes in B cells, $CD4^+$ and $CD8^+$ T cells, granulocytes, monocytes and NK cells, and the *huva experiments* were filtered for the 3000 most variable ones. For the generation of the correlation network, nodes (*huva experiments*) were connected (edges) with a correlation cut-off of 0.971 for a total of 61,585 edges and 2,894 nodes. Clustering of the *huva experiments* included in the network was performed with the Louvain clustering algorithm (Blondel et al., 2008) with a minimum cluster size of 10 nodes. Group fold changes were calculated for each cell type and overlaid on the *huva experiments* network (Figure 4F) or merged by cluster as heat map (Figure 4D). Similarly, for the network describing the changes in cytokine secretion, the top 3,000 most variable *huva experiments* were used with a correlation cut-off of 0.674 resulting in a network with 159,827 edges and 2,943 nodes. Clustering was performed with the Louvain clustering algorithm (Blondel et al., 2008) with a minimum cluster size of 10 nodes. For each network, the optimal correlation cut-off was calculated according to a weighted sum of the Multicriteria Decision Aiding (MCDA) tabular output favouring a higher R2 and number of edges/nodes but minimising the number of independent networks resulting from the selected cut-off (Bigaret et al., 2017). GOEA was performed for each CoCena^2 cluster independently with the clusterProfiler R package (Yu et al., 2012).

RNA-seq and pre-processing of STAT1 AM PBMC samples

For the transcriptome analysis of STAT1 AM patients, blood was collected after informed consent at the Radboud University Nijmegen Medical Centre (RUNMC, Nijmegen, the Netherlands). The study was approved by the Institutional Review Board of the Radboud University Nijmegen Medical Centre, informed consent was obtained from all subjects. RNA-sequencing experiments were performed on PBMC samples from 3 controls and 3 STAT1 AM patients after 4 hours of *ex vivo* incubation in RPMI. In the analysis of isolated cell types one exemplary STAT1 AM patient and an age-matched control was analysed. The first patient is a 39-year-old male who carries the STAT1 p.R274W mutation and was previously described

(family 1 pt 3) (van de Veerdonk et al., 2011). The other patients have not yet been previously reported; the second patient is a 41-year-old male carrier for the STAT1 p.R274Q variant and the third is a 31-year-old male carrying the STAT1 p.D23V (Table S14). Total RNA was isolated in 800 μL of TRIzol reagent (Invitrogen); RNA integrity was then measured on an Agilent 2100 Bioanalyzer (Agilent) using an Agilent RNA 6000 Pico Chip according to manufacturer's instructions (RIN>7). mRNA was consequently enriched with the Micro-Poly(A) Purist Kit (Ambion) starting from 5μg of total RNA with two rounds of Oligo(dT) selection. Whole transcriptome libraries were prepared using the SOLiD Total RNA-Seq Kit (STaR Kit) following the protocol for low input amounts (Smeekens et al., 2013). Paired-end sequencing (50+25 bases) was performed on a 5500XL sequencer (Life Technologies). Paired-end reads were mapped against the human genome (hg19) using TopHat (v 2.0.4) (Trapnell et al., 2009). The count table was generated starting from TopHat aligned .bam files with the Rsubread tool (v. 1.34.7) (Liao et al., 2019) using standard settings for paired-end reads.

Analysis of STAT1 AM transcriptome data

RNA-sequencing counts were processed according to the DEseq2 differential expression analysis pipeline (Love et al., 2014). Shortly, sequencing reads were normalised and rlog-transformed for visualization. Genes showing a total of less than 100 reads across the dataset were removed due to low expression in the dataset and to increase the calculation speed. Differentially expressed genes were calculated with a significant fold change of 2 and a p-value lower than 0.05 after independent hypothesis weighting (IHW) p-value correction. The design model used for the DE genes estimation includes the genotype (Ctrl, STAT1 AM) and the experimental date to correct for the batch caused by processing the samples in two different days as reported in Table S14. For the visualization of box plots for single genes, heat maps and PCA the rlog counts were corrected using the function included in the limma R package (Ritchie et al., 2015). All differentially expressed genes were used as gene signatures for GSEA (STAT1 AM UP and STAT1 AM DOWN). For the representation of the volcano plot, the log2 fold change and p-value were calculated by the DEseq2 model. GOEA was performed with the clusterProfiler R package (Yu et al., 2012) using the human GO gene list for biological processes as reference (downloaded on 21/11/2018). GSEA was calculated within the *huva* framework using the fgsea R package (Korotkevich et al., 2016). In the analysis of isolated cell types from peripheral blood of control or STAT1 AM donors, kallisto pseudo-aligned counts were normalized according to the DEseq2 pipeline and the fold change was calculated as the log2 transformation of the comparison STAT1 AM vs. ctrl for each cell type (where STAT1 AM is at the nominator and ctrl at the denominator of the comparison). The raw data are available at EGA datasets (EGAS00001005041).

How to include a new dataset to *huva*

We designed the R implementation of *huva* to allow easy integration of new dataset from the user. The *generate_huva_dataset* function takes care to correctly format the transcriptomic data and any further complementary dataset to be used with *huva*. The function also lets the user decide if a standalone dataset should be generated or the new dataset should be merged with those provided with the *huva.db* package.

From our experience, we suggest including datasets with at least 100 donors and one additional phenotypic or functional data layer. When evaluating a new dataset, exploratory data analysis (Love et al., 2016) (eg. PCA, hierarchical clustering) should be performed to identify possible unwanted sources of variation in the data or aberrant samples. We also provide in our Zenodo and GitHub repositories the code to run the same test performed in this manuscript (Figures 3 and S6–S8).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical significance was calculated in R (v. 4.0.3) with an unpaired two-sided t-test if not stated differently. Exact p-values are reported in the figures; a p-value < 0.05 was considered significant. All data were visualized using R (v. 4.0.3) with the packages ggplot2, pheatmap or the built-in functions of *huva* (v. 0.1.4). All box plots were constructed in the style of Tukey, showing median, 25^{th} and 75^{th} percentiles; whisker extends from the hinge to the largest or lowest value no further than 1.5 * IQR from the hinge (where IQR is the interquartile range, or distance between the first and third quartiles); outlier values are depicted individually.