



RESEARCH

Open Access



Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types

Simone Ecker^{1,2*} , Lu Chen^{3,4}, Vera Pancaldi¹, Frederik O. Bagger^{4,5,6}, José María Fernández¹, Enrique Carrillo de Santa Pau¹, David Juan¹, Alice L. Mann³, Stephen Watt³, Francesco Paolo Casale⁶, Nikos Sidiropoulos^{7,8,9}, Nicolas Rapin^{7,8,9}, Angelika Merkel¹⁰, BLUEPRINT Consortium, Hendrik G. Stunnenberg¹¹, Oliver Stegle⁶, Mattia Frontini^{4,5,12}, Kate Downes^{4,5}, Tomi Pastinen¹³, Taco W. Kuijpers^{14,15}, Daniel Rico^{1,16†}, Alfonso Valencia^{1†}, Stephan Beck^{2†}, Nicole Soranzo^{3,4*†} and Dirk S. Paul^{2,17*†} 

Abstract

Background: A healthy immune system requires immune cells that adapt rapidly to environmental challenges. This phenotypic plasticity can be mediated by transcriptional and epigenetic variability.

Results: We apply a novel analytical approach to measure and compare transcriptional and epigenetic variability genome-wide across CD14⁺CD16⁻ monocytes, CD66b⁺CD16⁺ neutrophils, and CD4⁺CD45RA⁺ naïve T cells from the same 125 healthy individuals. We discover substantially increased variability in neutrophils compared to monocytes and T cells. In neutrophils, genes with hypervariable expression are found to be implicated in key immune pathways and are associated with cellular properties and environmental exposure. We also observe increased sex-specific gene expression differences in neutrophils. Neutrophil-specific DNA methylation hypervariable sites are enriched at dynamic chromatin regions and active enhancers.

Conclusions: Our data highlight the importance of transcriptional and epigenetic variability for the key role of neutrophils as the first responders to inflammatory stimuli. We provide a resource to enable further functional studies into the plasticity of immune cells, which can be accessed from: <http://blueprint-dev.bioinfo.cnio.es/WP10/hypervariability>.

Keywords: Differential variability, Phenotypic plasticity, Heterogeneity, Immune cells, Monocytes, Neutrophils, T cells, Gene expression, DNA methylation

Background

Phenotypic plasticity is fundamental to human immunity, allowing rapid cellular adaptation in response to changing environmental conditions [1]. Plasticity of immune cells can be influenced by the variability of cellular traits, including gene expression and DNA methylation. The

stochastic nature inherent to cellular processes such as gene regulation gives rise to cell-to-cell variation, enhancing survival under adverse conditions and stress [2–4]. Environmental stimuli, including temperature, hormone levels, and invading pathogens, further affect the expression of genes in a tissue- and temporal-dependent fashion [2, 4, 5].

Rapid and effective response to a stimulus is facilitated and intensified if the cellular trait already exhibits large stochastic fluctuations in the absence of the stimulus [3]. For example, while genes involved in stress response tend to be highly variable [3, 6, 7], genes involved in

* Correspondence: s.ecker@ucl.ac.uk; ns6@sanger.ac.uk; dsp35@medschl.cam.ac.uk

†Equal contributors

¹Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain

³Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK

²UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK

Full list of author information is available at the end of the article

essential cellular functions, such as protein synthesis and metabolism, demonstrate less variable expression levels [8, 9].

B and T cells utilize genetic recombination to generate a highly diverse repertoire of immunoglobulins and T-cell surface receptors, respectively. In addition, immune responses are driven by the variability of key signaling molecules and transcription factors not controlled by genetic factors [10, 11]. Epigenetic states, including DNA methylation, also contribute to plastic gene expression during cell fate commitment, thus enhancing fitness in response to external cues [12, 13].

Transcriptional and epigenetic heterogeneity that is measured across individuals emerges from different origins. While intra-individual variability can relate to different cellular properties in response to external signals, such as cell activation and communication [3, 7, 14], inter-individual variability can relate to differences between the individuals, including genetic makeup, age, sex, and lifestyle. Importantly, it has also been demonstrated that *inter*-individual variability can serve as an appropriate proxy for *intra*-individual variability at the level of single cells [7, 14, 15].

Both transcriptional and epigenetic variability have been shown to strongly correlate with the development and progression of human diseases [12, 16, 17]. For example, gene expression variability has been linked to human immunodeficiency virus (HIV) susceptibility [18], neurological disorders [18, 19], and cancer [20, 21]. Hypervariable DNA methylation loci can be used as biomarkers to predict the risk of neoplastic transformation in stages prior to neoplasia [22, 23].

The extent and functional interpretation of transcriptional and epigenetic variability have not been systematically investigated genome-wide across multiple immune cell types in the general population. Here, we applied a novel analytical approach to measure differential variability of gene expression and DNA methylation in three major immune cell types: CD14⁺CD16⁻ classic monocytes, CD66b⁺CD16⁺ neutrophils, and CD4⁺CD45RA⁺ “phenotypically naïve” T cells. This matched panel of cell types was derived from the same 125 healthy individuals. We show that neutrophils exhibit substantially increased variability of both gene expression and DNA methylation patterns, compared to monocytes and T cells, consistent with these cells’ key role as the first line of host defense. We annotated hypervariable genes (HVGs) and CpGs (HVPs) to known homeostatic and pathogenic immune processes and found subsets of genes correlating with genetic makeup, donor demographic, and lifestyle factors. Our data further reveal potential molecular mechanisms of immune responses to environmental stimuli and provide a resource to enable future functional studies into

the phenotypic plasticity of human immune cells in health and disease.

Results

Deep molecular profiling of immune cells in the BLUEPRINT Human Variation Panel

The analyses described in this study are based on the publicly available resource provided by the BLUEPRINT Human Variation Panel [24]. The resource contains genome-wide molecular profiles of CD14⁺CD16⁻ classic monocytes, CD66b⁺CD16⁺ neutrophils, and CD4⁺CD45RA⁺ naïve T cells. These leukocyte types were chosen due to their important role in mediating immune cell processes, their relative abundance in peripheral blood, allowing for examination of multiple cellular traits, as well as the availability of experimental protocols to prepare cell populations of high purity (>95%). Monocytes and neutrophils are myeloid cells that share the same bone marrow-residing granulocyte-macrophage precursor cell. Monocytes migrate to sites of infection and differentiate into macrophages and dendritic cells to induce an immune response. As part of the innate immune system, neutrophils move within minutes to sites of infection during the acute phase of inflammation. Naïve T cells are lymphoid cells that are part of the adaptive immune system, representing mature helper T cells that have not yet recognized their cognate antigen.

Across an initial cohort of 200 healthy individuals representative of the UK population, purified preparations of these primary cells were probed for gene expression using total RNA sequencing (RNA-seq) and DNA methylation using Illumina Infinium HumanMethylation450 BeadChips (“450 K arrays”). Detailed information about the experimental and analytical strategies for quantifying these cellular traits are provided in the “Methods” section. Additional file 1: Figures S1, S2, and S3 give an overview of the data quality assessment of the gene expression and DNA methylation data sets. All individuals were further profiled for DNA sequence variation using whole-genome sequencing to allow for cell type-dependent, quantitative assessment of the genetic and epigenetic determinants of transcriptional variance [24].

In this study, we exploited this resource, selecting all 125 donors for whom matched gene expression and DNA methylation data sets were available across the three immune cell types. The key analytical advance of the work presented here concerns the measurement and interpretation of differential variability. That is, the identification of loci at which gene expression and DNA methylation levels show significantly greater variation within one cell type compared to the other cell types. An overview of the study design and analytical concept is provided in Fig. 1a.

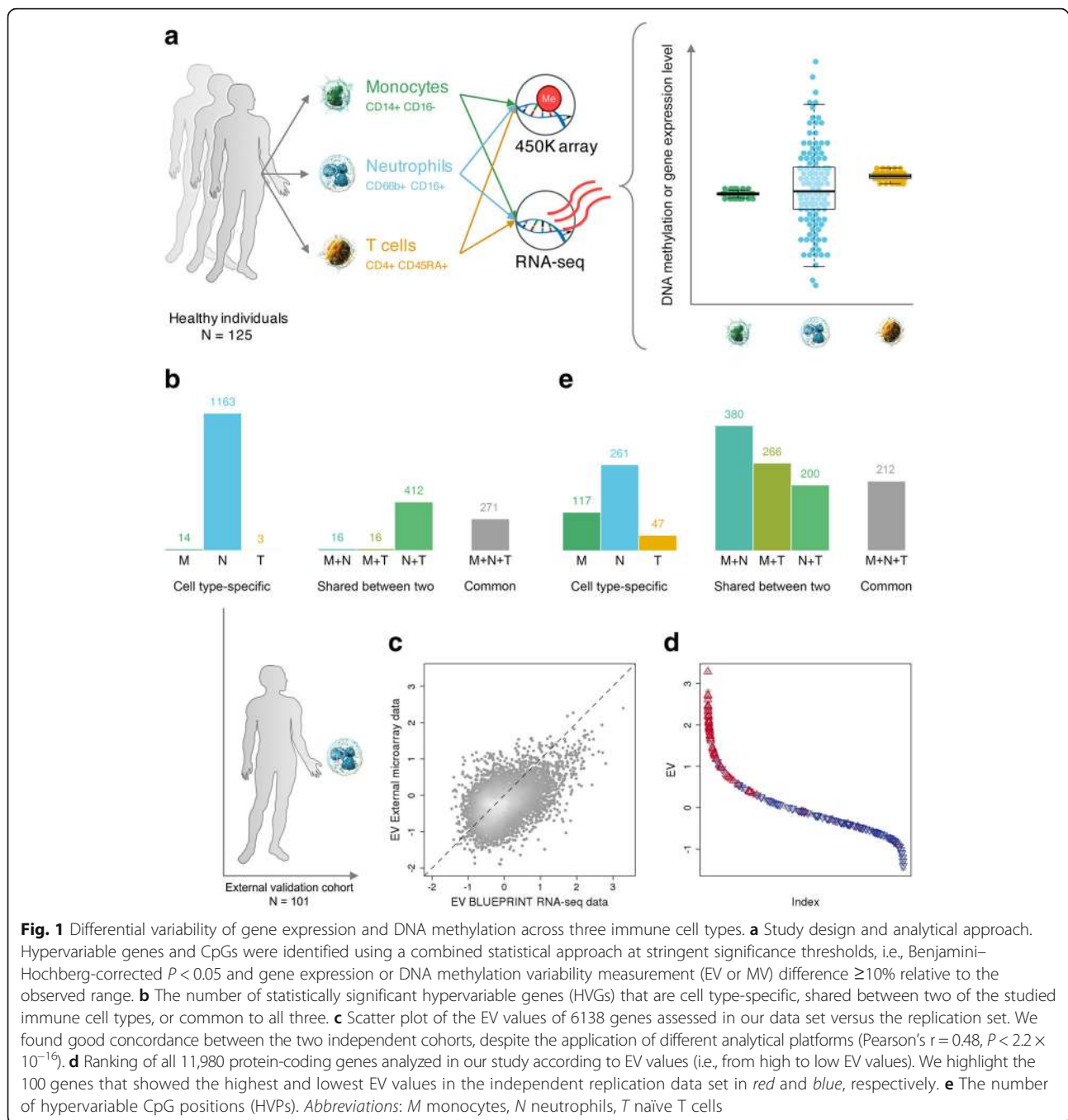


Fig. 1 Differential variability of gene expression and DNA methylation across three immune cell types. **a** Study design and analytical approach. Hypervariable genes and CpGs were identified using a combined statistical approach at stringent significance thresholds, i.e., Benjamini–Hochberg-corrected $P < 0.05$ and gene expression or DNA methylation variability measurement (EV or MV) difference $\geq 10\%$ relative to the observed range. **b** The number of statistically significant hypervariable genes (HVGs) that are cell type-specific, shared between two of the studied immune cell types, or common to all three. **c** Scatter plot of the EV values of 6138 genes assessed in our data set versus the replication set. We found good concordance between the two independent cohorts, despite the application of different analytical platforms (Pearson's $r = 0.48$, $P < 2.2 \times 10^{-16}$). **d** Ranking of all 11,980 protein-coding genes analyzed in our study according to EV values (i.e., from high to low EV values). We highlight the 100 genes that showed the highest and lowest EV values in the independent replication data set in red and blue, respectively. **e** The number of hypervariable CpG positions (HVPs). *Abbreviations: M* monocytes, *N* neutrophils, *T* naïve T cells

Genome-wide patterns of differential gene expression variability across immune cell types

We first assessed inter-individual expression variability of 11,980 protein-coding, autosomal genes that showed robust expression in monocytes, neutrophils, and T cells (“Methods”). We applied an improved analytical approach for the assessment of differential variability (“Methods”), taking into account the strong negative correlation between mean gene expression levels and expression variability (Additional file 1: Figure S4).

Figure 1b gives an overview of the number of identified HVGs that are cell type-specific, shared between two of the studied immune cell types, or common to all three. Neutrophils were found to have the largest number of HVGs overall ($n = 1862$), as well as of cell type-specific HVGs ($n = 1163$). In contrast, we found only a small number of cell type-specific HVGs in monocytes and T cells ($n = 14$ and 3 , respectively). In addition, we identified 271 genes that were highly variable across all three immune cell types using a rank-based approach

(“Methods”). Mature neutrophils (as profiled here) show low proliferative capacity and reduced transcriptional and translational activity [25, 26]. The latter could potentially impede comparable assessment of differential variability if the relationship between variability and mean expression levels was not taken into account. Thus, using our analytical approach, we assessed and confirmed that overall reduced gene expression levels did not technically confound the observed increased variability of gene expression levels in neutrophils (Additional file 1: Figure S4).

We then aimed to replicate the detected HVG levels in an independent sample cohort. We retrieved a gene expression data set generated using Illumina Human HT-12 v4 Expression BeadChips consisting of CD16⁺ neutrophils derived from 101 healthy individuals; these donors were, on average, 34 years of age (range 19–66 years) and 50% were male [27]. Of the 11,023 gene probes assessed on the array platform, 6138 could be assigned to a corresponding gene identifier in our data set. First, we ranked all 11,980 genes analyzed in our study according to gene expression variability (EV) values from high to low. Then, we assessed the position of the top 100 genes with highest and lowest EV values from the independent validation data in this ranking to confirm that the variability patterns are consistent between the two data sets. Neutrophil-specific HVGs measured using RNA-seq were also found to be hypervariable using expression arrays in the independent cohort of healthy individuals (Fig. 1c, d).

In summary, we devised and assessed a novel method for the identification of differential gene expression variability. Overall, we found strongly increased variability of gene expression in neutrophils compared to monocytes and T cells and replicated the detected neutrophil-specific HVG patterns in an external cohort.

Biological significance of differentially variable genes across immune cell types

Next, we explored the characteristics of the identified HVGs. We performed ontology enrichment analysis of gene sets using the Goseq algorithm [28]. This method takes into account the effect of selection bias in RNA-seq data that can arise due to gene length differences [28]. Additional files 2 and 3 summarize the annotation data of all identified HVGs and observed gene ontology enrichment patterns, respectively.

Genes showing expression hypervariability across all three cell types were enriched in biological processes related to chemotaxis, migration, and exocytosis (Additional file 3). For neutrophil-specific HVGs, we found gene ontology enrichment in oxidoreductase activity and cellular processes related to virus response and

parasitism (Additional file 3). Notable genes among those with hypervariable expression values were *CD9* (Fig. 2a), *CAPN2* (Fig. 2b), and *FYN* (Fig. 2c). *CD9* showed increased variability across all three cell types. The gene encodes the CD9 antigen, a member of the tetraspanin family. It functions as cell surface protein that forms complexes with integrins to modulate cell adhesion and migration and mediate signal transduction [29, 30]. The neutrophil-specific HVGs *CAPN2* and *FYN* encode a calcium-activated neutral protease involved in neutrophil chemotaxis [31] and a tyrosine-protein kinase implicated in intracellular signal transduction [32], respectively.

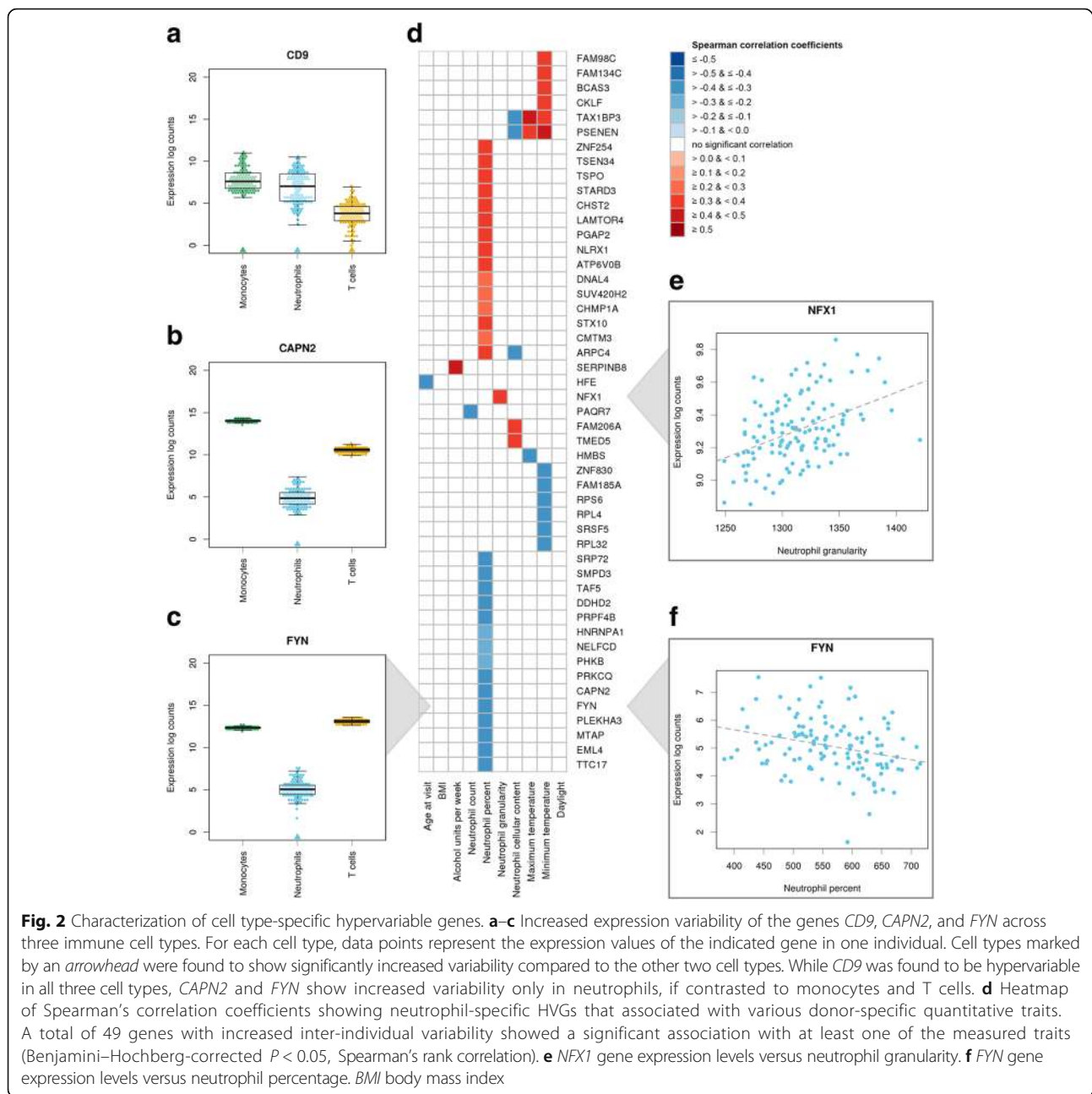
Taken together, functional enrichment of HVG sets revealed that many of the identified HVGs are involved in mediating immune-related processes. This suggests that neutrophils exhibit specific gene loci that are highly adaptable to external cues.

Determinants of inter-individual cell type-specific gene expression variability

Following the discovery and characterization of genes that present hypervariable expression levels between individuals, we next aimed to delineate potential sources of heterogeneity that can be associated with differences between individuals. We hypothesized that these sources mainly relate to genetic variation, age, sex, and lifestyle factors.

First, we determined the subset of cell type-specific HVGs that correlated with genetic variants. We retrieved gene sets with a local (*cis*) genetic component designated by expression quantitative trait locus (eQTL) and variance decomposition analyses, as described in the BLUEPRINT Human Variation Panel (Additional file 1: Figure S5a). In neutrophils, we found that 638 of the 1163 cell-specific HVGs (55%) associate with *cis* genetic variants (Additional file 2), at least partly explaining the observed gene expression variability. These data are consistent with previous reports, highlighting the role of genetic variants in mediating transcriptional variance [33–35].

Second, we correlated cell type-specific HVGs with various quantitative traits measured in individual donors: demographic information (age, body mass index, and alcohol consumption); cellular parameters as assessed by a Sysmex hematology analyzer (e.g., cell count and size); and season (i.e., minimum/maximum temperature and daylight hours of the day on which blood was drawn). The results of this analysis are provided in Additional files 2 and 4. In neutrophils, we identified 49 HVGs that show significant association with at least one of the measured traits (Fig. 2d). For example, we found *NFX1*, a nuclear transcription factor that regulates *HLA-DRA* gene transcription [36], to associate with neutrophil granularity



(Fig. 2e). An increase in neutrophil granularity can be reflective of a potential infection; this parameter is routinely monitored in a clinical setting. *FYN* gene levels (reported above) were negatively correlated with neutrophil percentage (Fig. 2f).

Third, we investigated whether sex was an important source of inter-individual (autosomal) gene expression variability. We found only two of the 1163 neutrophil-specific HVGs, *SEPT4* and *TMEM63C*, to be differentially expressed between sexes (Additional file 1: Figure S6a), and high expression variability was observed for

both sexes in these genes. However, in neutrophils we identified a surprisingly large number of sex-specific differentially expressed genes of small effect size, which corresponded to important immune cell functions. We present a detailed analysis of these genes in the “Sex-specific differential gene expression across immune cell types” section.

In conclusion, we found that genetic makeup is an important determinant of transcriptional variability. Donor demographic and lifestyle factors also contributed towards transcriptional variability.

Neutrophil-specific hypervariable genes not mediated by *cis* genetic effects

Next, we studied in detail the subset of neutrophil-specific genes that showed hypervariable expression but did not associate with local genetic variants ($n = 525$). Although some of these genes could be mediated by distal (*trans*) genetic factors not detected in the BLUEPRINT Human Variation Panel, it is conceivable that expression heterogeneity of this gene set was primarily due to external triggers or stochastic fluctuations.

We generated a correlation matrix of expression levels of the 525 HVGs and identified clusters of correlated genes that may act in concert or be co-regulated. The identified co-expression network contained 259 connected genes and consisted of three distinct gene

modules (Fig. 3). We inferred biological functions corresponding to the three gene modules. All modules were highly enriched for genes with important immune-related functions.

The first and largest gene module ($n = 105$ genes, green in Fig. 3) showed enrichment for inclusion body, receptor signaling, and immune response activation. The second module ($n = 78$ genes, yellow) was enriched in biological processes related to RNA processing and chaperone binding. The third gene module ($n = 33$ genes, red), contained many genes with particularly high variation in their expression patterns. *RSAD2*, an interferon-inducible antiviral protein, showed the highest variability among many other interferon-inducible genes present in module three. These genes are essential in

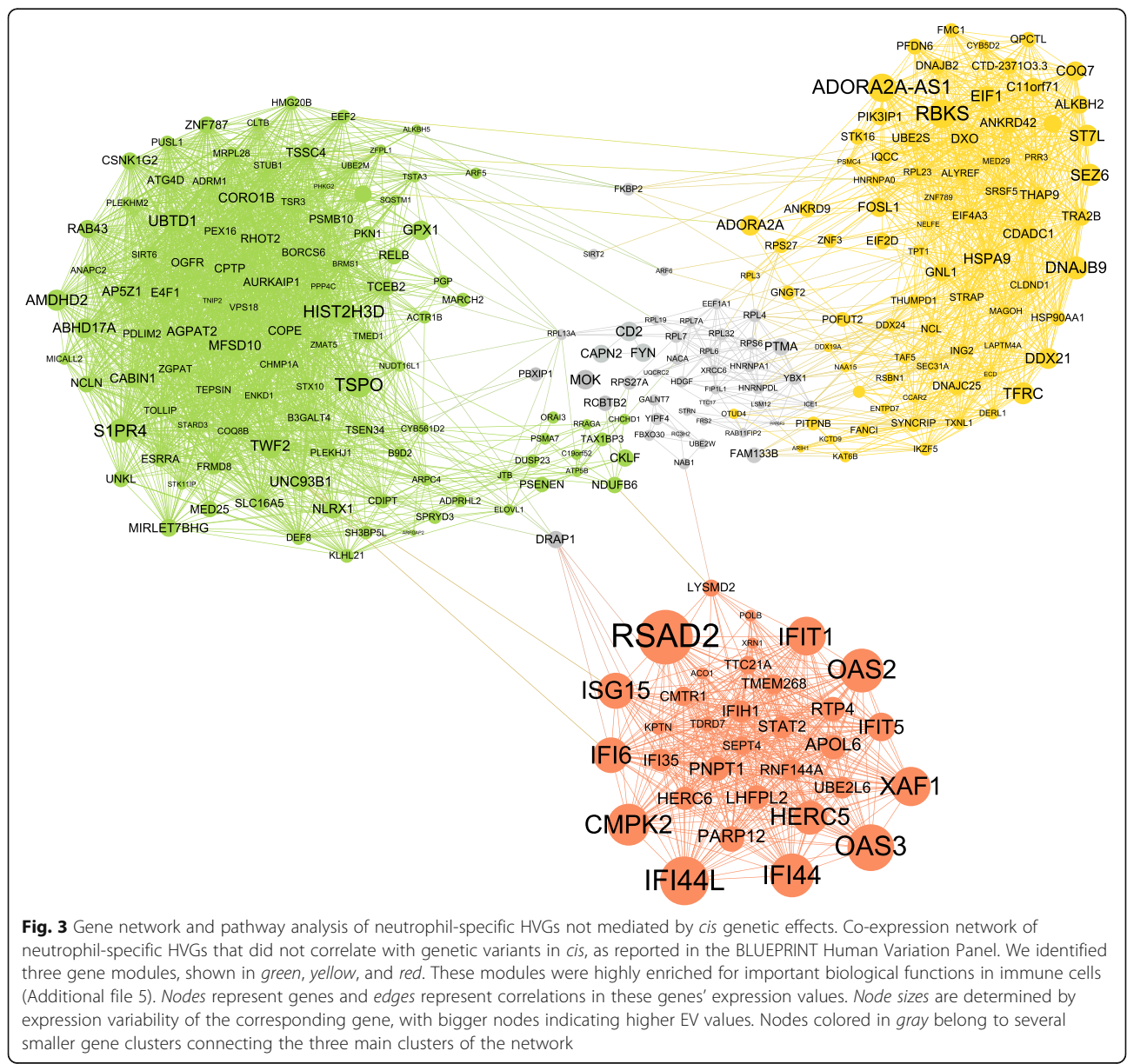


Fig. 3 Gene network and pathway analysis of neutrophil-specific HVGs not mediated by *cis* genetic effects. Co-expression network of neutrophil-specific HVGs that did not correlate with genetic variants in *cis*, as reported in the BLUEPRINT Human Variation Panel. We identified three gene modules, shown in green, yellow, and red. These modules were highly enriched for important biological functions in immune cells (Additional file 5). Nodes represent genes and edges represent correlations in these genes' expression values. Node sizes are determined by expression variability of the corresponding gene, with bigger nodes indicating higher EV values. Nodes colored in gray belong to several smaller gene clusters connecting the three main clusters of the network

innate immune response to viral infections [37]. Gene ontology and pathway analyses of all genes in the network module further showed a strong enrichment for response to type I interferon and several viral disease pathways, including influenza A, herpes simplex, and hepatitis (Additional file 1: Figure S7). A detailed functional annotation of all three network modules is provided in Additional file 5.

Sex-specific differential gene expression across immune cell types

In our analysis, we only detected differences in mean gene expression levels between male and female donors with log-fold change ≥ 1 , for 21 genes in neutrophils, two of which were also found to be HVGs in neutrophils (Additional file 1: Figure S6a). Nonetheless, when no minimum log-fold change criterion was applied, we found that sex-dependent mean expression of autosomal genes (Additional file 1: Figure S6b) was highly abundant in neutrophils ($n = 3357$ genes) compared to T cells ($n = 895$) and monocytes ($n = 64$).

As many autoimmune diseases have a higher incidence in females, and females show generally elevated immune responses compared to males [38], we hypothesized that genes with elevated gene expression levels in females may account for the increased incidence rates. Indeed, genes with higher mean expression levels in neutrophils derived from females ($n = 682$) were enriched in immune response and related pathways (Additional file 6). In contrast, genes with increased mean expression in male donors ($n = 2675$) were enriched in basic cellular processes, such as RNA processing and translation (Additional file 6). In addition, in male donors, genes were strongly enriched in cellular compartments, such as nuclear lumen (Additional file 6).

Genome-wide patterns of differential DNA methylation variability across immune cell types

Following the analyses of differential gene expression variability, we then applied our improved analytical approach to determine the inter-individual variability of DNA methylation levels at 440,905 CpG sites ("Methods"). Again, our method accounted for confounding effects due to the correlation between mean and variability measurements (Additional file 1: Figure S8).

Concordant with our findings for gene expression variability (Fig. 1b), we found that neutrophils had the largest number of hypervariable CpG positions (HVPs) overall ($n = 1053$), as well as cell-specific HVPs ($n = 261$). Neutrophils and monocytes shared a considerable number of HVPs ($n = 380$) in contrast to T cells (Fig. 1e). Finally, we identified 212 HVPs common to all three cell types. An overview of the number of HVPs is shown in Fig. 1e.

Following the discovery of HVPs, we examined whether these sites were overrepresented at particular gene elements and epigenomic features. To this end, we focused on cell type-specific HVPs, correlating their DNA methylation levels with distinct cellular characteristics and molecular pathways. In Additional file 7, we summarize the detailed annotation of all HVPs across the three profiled immune cell types. In neutrophils, we found that cell type-specific HVPs were depleted at CpG islands, which typically occur near transcription start sites ($P = 6.37 \times 10^{-19}$, hypergeometric test; Fig. 4a), and enriched at intergenic regions ($P = 0.03$; Fig. 4b).

We hypothesized that cell type-specific HVPs localize at distal gene regulatory elements such as enhancer sequences, of which many are known to be also cell type-specific [39]. To test this hypothesis, we retrieved reference chromatin state maps of primary human monocytes, neutrophils, and T cells from the data repository provided by the BLUEPRINT Consortium [40]. Chromatin states are defined as spatially coherent and biologically meaningful combinations of multiple chromatin marks [41, 42]. A total of five chromatin states were designated, which correspond to functionally distinct genomic regions, namely active promoters, enhancers, and regions related to transcriptional elongation and polycomb-repression. In addition, a "variable" chromatin state was defined here, indicating frequent changes of local chromatin structure across samples of the same cell type. Indeed, neutrophil-specific HVPs were found to be strongly enriched in the enhancer ($P = 1.32 \times 10^{-12}$, hypergeometric test; Fig. 4c) and variable chromatin states ($P = 3.81 \times 10^{-8}$; Fig. 4c).

Biological significance of immune cell type-specific hypervariable CpGs

To interpret the potential cellular and biological implications of cell type-specific hypervariable CpGs, we annotated the genes in close proximity to each CpG using the Genomic Regions Enrichment of Annotations Tool (GREAT) [43]. This tool is valuable in assigning putative functions to sets of non-coding genomic regions [43].

Overall, we found enrichment in gene ontology terms attributed to genes close to HVPs in a cell type-dependent context (Additional file 8). For example, genes located near neutrophil-specific HVPs were enriched in gene signatures related to acute *Streptococcus pneumoniae* infection and cysteine synthase activity; the latter molecular process is important to hold off infections [44]. Consistent with established neutrophil function, this suggests that the identified HVPs play a

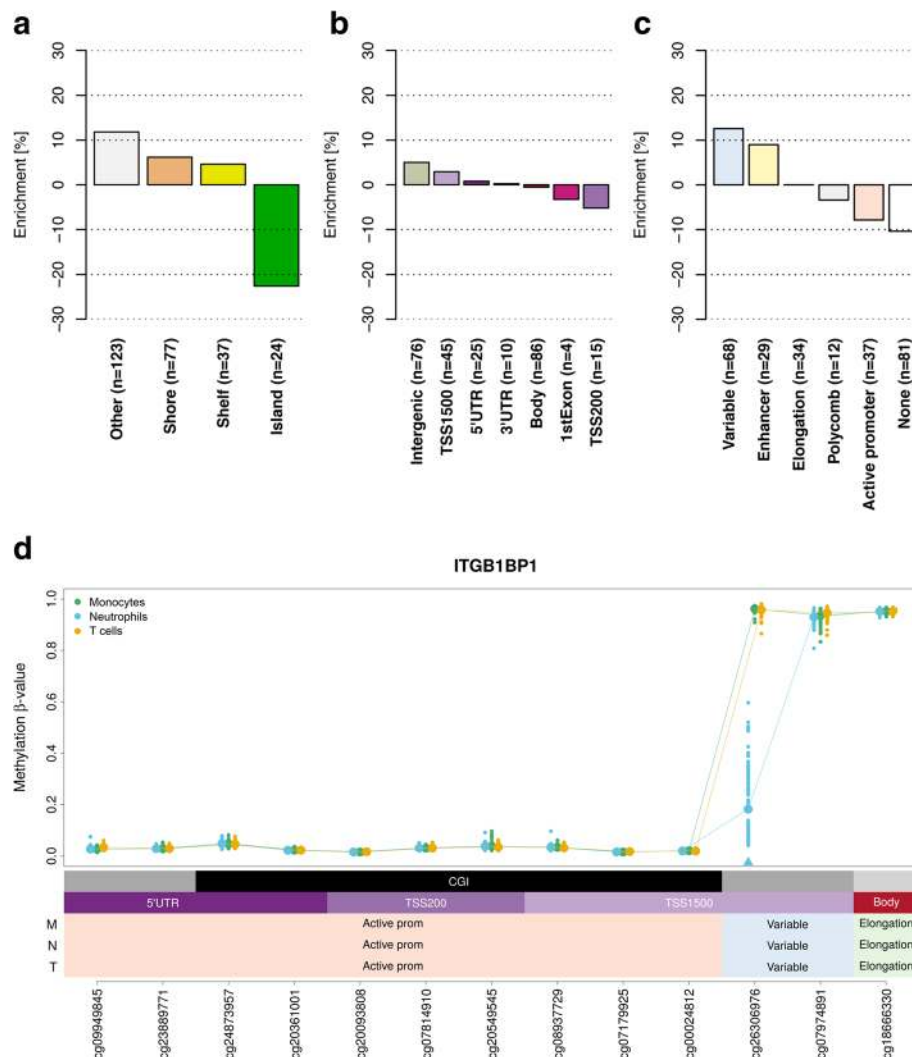


Fig. 4 Functional annotation of neutrophil-specific hypervariable CpG positions. **a** Enrichment of neutrophil-specific HVPs ($n = 261$) at genomic features. We found neutrophil-specific HVPs to be depleted at CpG islands ($P = 6.37 \times 10^{-19}$, hypergeometric test). **b** Enrichment of neutrophil-specific HVPs at gene elements. Neutrophil-specific HVPs were enriched at intergenic regions ($P = 0.03$). **c** Enrichment of neutrophil-specific HVPs at distinct reference chromatin states in neutrophils. The HVPs were enriched at enhancer ($P = 1.32 \times 10^{-12}$) and “variable” ($P = 3.81 \times 10^{-8}$) chromatin states. A variable chromatin state denotes a state that was observed in less than 80% of the biological replicates ($n \geq 5$) within a given cell type and indicates dynamic changes of local chromatin structure. **d** Regional plot of an exemplar neutrophil-specific HVP mapping to the promoter of the *ITGB1BP1* gene, encoding the integrin beta 1 binding protein 1. The statistically significant HVP is indicated with an arrowhead. For each cell type, data points represent the DNA methylation β values (y -axis) at the indicated CpGs (x -axis) in one individual. For each CpG site, we calculated the mean DNA methylation value (indicated with a larger data point). Every CpG site is annotated with regards to genomic feature, gene element, and chromatin state. Abbreviations: *M* monocytes, *N* neutrophils, *T* naïve T cells, *TSS* transcription start site, *CGI* CpG island, *UTR* untranslated region, *prom* promoter

role in regulating the expression of neutrophil-specific genes in response to infection.

In Fig. 4d, we provide an example of a neutrophil-specific HVP at the promoter of the *ITGB1BP1* gene, encoding the integrin beta 1 binding protein 1. Integrins are essential cell adhesion proteins that induce intracellular signaling pathways upon activation by matrix binding [45, 46]. They function as signal transducers allowing

for rapid responses to cell surface signals [46]. Notably, the highlighted HVP mapped to a variable chromatin state at this locus, indicating that it influences local chromatin dynamics upon an internal or external trigger (Fig. 4d).

In conclusion, we show that cell type-specific HVPs clustered in enhancer and dynamic chromatin states at intergenic regions, suggesting they play a role in the

regulation of cell type-specific gene expression programs in response to environmental changes. Genes in proximity to HVPs were enriched in gene sets relevant to important immunological functions.

Determinants of inter-individual cell type-specific DNA methylation variability

Subsequent to the identification and annotation of CpGs with hypervariable DNA methylation levels, we explored potential reasons for the discovered inter-individual DNA methylation heterogeneity.

In agreement with our findings for gene expression variability, we determined that a large proportion of cell type-specific HVPs correlated with *cis* genetic variants reported in the BLUEPRINT Human Variation Panel (Additional file 1: Figure S5b). In neutrophils, we found that 167 of the 261 cell type-specific HVPs (64%) associated with DNA methylation quantitative trait loci (Additional file 7). Our data further revealed that none of the cell type-specific HVPs were differentially methylated between male and female donors. The complete numerical results of all correlation analyses are provided in Additional file 9.

HVPs specific to monocytes showed frequent association with seasonal effects, such as temperature and daylight ($n = 12/117$ HVPs; Additional file 1: Figure S9). This finding is consistent with recent analyses reporting fluctuations of gene expression levels in monocytes depending on the season and circadian rhythm [47]. Many $CD4^+$ T cell-specific HVPs particularly correlated with donor age ($n = 14/46$ HVPs; Additional file 1: Figure S9), in line with previous findings on age-related DNA methylation changes in T cells [48, 49]. These alterations are especially interesting in the context of immunosenescence, for which dysregulation in T-cell function is thought to play a crucial role [50, 51]. Naïve $CD4^+$ T cells have further been reported to become progressively longer-lived with increasing age [52], which possibly also impacts their DNA methylation patterns.

Correlation of DNA methylation variability with transcriptional output

DNA methylation at active gene elements can directly control the regulation of gene expression. While methylated gene promoters usually lead to transcriptional silencing, methylated gene bodies typically lead to transcriptional activation [53]. We next aimed to probe this paradigm in the context of gene expression and DNA methylation variability.

We measured the correlation of DNA methylation variability with transcriptional output at the level of single genes. Specifically, we studied cell type-specific HVPs that map to gene promoters and bodies, correlating their

DNA methylation level with the gene expression level in the same individuals. At promoters, 30.1% (range 23.5–33.3%) of HVPs showed a negative correlation with gene expression (Fig. 5a), in support of the conventional role of DNA methylation in gene repression. At gene bodies, a small subset of HVPs (5.0%; range 0.0–10.8%) showed a positive correlation with gene expression (Fig. 5b). Additional file 10 gives a full account of these genes and numeric results.

An example is provided in Fig. 5c, showing a monocyte-specific HVP at the gene promoter of *MSRI*. At this CpG site, DNA methylation levels were significantly correlated with gene repression (Benjamini–Hochberg (BH)-corrected $P < 2.2 \times 10^{-16}$, Spearman's rank correlation). *MSRI*, encoding the CD204 antigen, is involved in endocytosis of modified low-density lipoproteins.

Relationship between DNA methylation variability and gene expression variability

Finally, we examined global patterns of DNA methylation variability in relation to transcriptional variability. In neutrophils, highly variable gene expression levels were observed at promoters exhibiting highly variable DNA methylation levels, and also at promoters showing very stable DNA methylation levels (Fig. 5d). For DNA methylation variability at gene bodies, this relationship was weaker and showed a linear tendency (Fig. 5e). Importantly, these global patterns were consistent across all three immune cell types (Additional file 1: Figure S10).

To characterize these promoter regions further, we counted the number of transcription factor binding motifs at these regions (“Methods”). We found an accumulation of binding motifs at promoters presenting either highly variable or very stable DNA methylation levels (Fig. 5f; Additional file 1: Figure S8). Next, we explored the properties of the 100 genes that showed both the highest expression variability and the highest DNA methylation variability at their promoters. We found that of the 100 genes in each cell type, 66 were common to all three cell types; in turn, ten of these 66 genes encode transcription factors. For example, in neutrophils this included *ELF1*, a transcriptional regulator of genes involved in immune response signaling pathways [54]. Neutrophil-specific HVPs were also enriched at genes with promoter sequences that contain the consensus binding motif of *ELF1* (BH-corrected $P = 1.2 \times 10^{-5}$; MSigDB analysis).

Taken together, these results provide evidence that DNA methylation variability and gene expression variability could be mediated by the sequence-specific binding of transcription factors, such as *ELF1* in neutrophils. Future studies will be required to further investigate the functional relevance of the observed correlation.

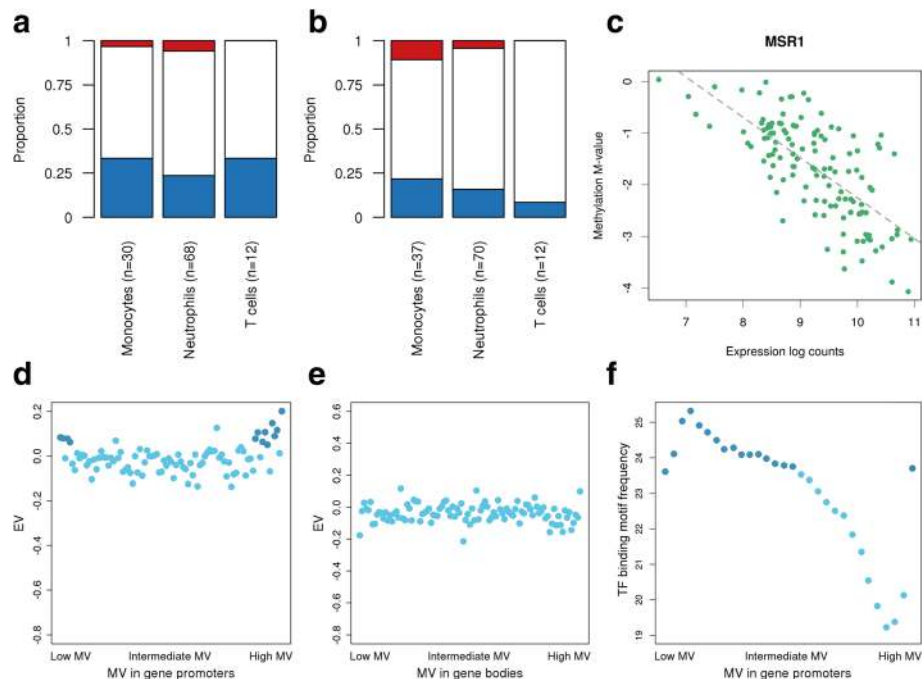


Fig. 5 Relationship between DNA methylation and gene expression. **a** The proportion of cell type-specific HVPs that map to gene promoters and are positively (*red*), negatively (*blue*), or not (*white*) associated with gene expression levels at Benjamini–Hochberg-corrected $P < 0.05$ (Spearman’s rank correlation). We found that around one-third of these HVPs (30.1%; range 23.5–33.3%) are negatively correlated with gene expression. **b** Same as panel **a** but for HVPs that map to gene bodies. **c** The negative correlation of *MSR1* promoter DNA methylation with gene expression in monocytes ($r = -0.70$, $P < 2.2 \times 10^{-16}$; Spearman’s rank correlation). **d** Correlation between DNA methylation variability (MV) and gene expression variability at gene promoters in neutrophils. First, gene-wise MV values were calculated. Then, the values were ordered from low to high MV value, grouped together in bins of 100 genes, and plotted against the EV values, maintaining the ordering by MV values. This binning strategy was applied to reduce the complexity of the data. HVPs at gene promoters were defined as CpG sites annotated to TSS1500, TSS200, 5’ UTR, and first exon, according to the Illumina 450 K array annotation manifest. *Darker data points* indicate the subset of bins that is further discussed in the “Results” section. **e** Same as panel **d** but for HVPs that map to gene bodies. HVPs at gene bodies were defined as CpGs annotated to body and 3’ UTR, according to the 450 K array annotation manifest. **f** The number of consensus transcription factor (TF) binding motifs at promoter regions versus MV values in neutrophils. Promoter regions were defined as ± 500 bp around the transcription start site. *Darker data points* indicate the subset of bins that is further discussed in the “Results” section

Discussion

In this study, we investigated the transcriptional and epigenetic variability that enables immune cells to rapidly adapt to environmental changes. To this end, we devised a novel analytical strategy to assess the inter-individual variability of gene expression and DNA methylation as a measure of functional plasticity across three immune cell types.

Our integrative analyses revealed two key insights. Firstly, neutrophils exhibit substantially increased variability of both gene expression and DNA methylation when directly compared to monocytes and T cells (Additional files 2 and 7). Gene expression variability in monocytes and T cells is either shared with neutrophils or does not reach statistical significance in comparison to neutrophils and/or the other cell type (Fig. 1b). We hypothesized that neutrophils require higher gene expression variability in order to perform their unique biological functions as first responders of the immune

system. Neutrophils have to readily react to changing conditions, which is facilitated by highly variable gene expression patterns. Functional enrichment and network analyses characterizing the neutrophil-specific hypervariability supported this hypothesis (Fig. 3; Additional files 3, 5, 8; Additional file 1: Figure S7). For example, genes with important functions in intracellular signaling, cell adhesion, and motility showed increased variability (Additional files 3 and 8). Such variability is likely mediated or additionally influenced by epigenetic mechanisms. Indeed, a subset of these genes was found to be under sole epigenetic control, such as *RSAD2*, a gene involved in interferon-mediated immune response (Fig. 3). Notably, cell type-specific differential DNA methylation variability was also highest for neutrophils compared to the other cell types (Fig. 1e).

Secondly, neutrophils display an increased number of sex-specific gene expression differences compared to

monocytes and T cells (Additional file 1: Figure S6b). Genes with elevated expression levels in neutrophils derived from females were associated with immune-related processes (Additional file 6). These results suggest a potential mechanistic basis for the higher incidence rates of many autoimmune diseases in females, such as multiple sclerosis, rheumatoid arthritis, and autoimmune hepatitis.

We acknowledge that our study has limitations: The data underlying the BLUEPRINT Human Variation Panel was generated in different laboratories. While the study design using systematic, paired analyses across cell types and individuals, and stringent quality control and statistical approaches reduced possible batch effects (Additional file 1: Figure S1), residual technical effects may still be present. Heterogeneity may also be partly explained by differing stages and rates of cell activation and cell death during experimental processing, as well as unaccounted environmental effects such as circadian rhythm, diet, physical activity, and psychological stress, which could affect one cell type more than the other(s).

Differences in the proportions of cellular subpopulations may contribute to overall elevated variability between individuals. We have thus assessed the expression profiles of a number of genes that identify distinct cellular subpopulations of neutrophils [55]: *CXCR4*, *CD63*, *CD62L* (also known as *SELL*), and *CD49* (also known as *ITGA4*). We did not observe inter-individual gene expression differences of surface markers corresponding to known neutrophil subpopulations, with the exception of *CD49* (Additional file 1: Figure S11). We note that *CD49* gene expression levels did not correlate with neutrophil granularity (BH-corrected $P = 0.89$, Spearman's rank correlation). These data suggest that variation in neutrophil subpopulations is unlikely to be a main determinant of increased inter-individual variability. Future studies are required to corroborate these results and to determine whether uncharacterized cellular subpopulations may contribute to the observed heterogeneity.

Novel transcriptome and epigenome profiling (e.g., scM&T-seq [56] and scWGBS [57]) and computational tools (e.g., single-cell latent variable models (scLVM) [58]) are now available to allow for measurements of gene expression and DNA methylation at the level of single cells. Such approaches have already been successfully used to quantify cell-to-cell expression variation and to identify otherwise undetected subpopulations of primary human immune cells [59–61]. Multi-omics single cell assays that capture not only transcriptomes and epigenomes but also proteomes and metabolomes will be used for the comprehensive functional annotation of single cells [62]. The application of these approaches will facilitate the dissection of cellular subpopulations and

reveal valuable additional information about the functional heterogeneity of neutrophils.

In summary, we provide a novel analytical strategy and comprehensive resource for future research into the plasticity of immune cells. For our analyses, we exploited the unique resource provided by the BLUEPRINT Human Variation Panel, enabling us to conduct the most comprehensive study of differential variability in primary cell types to date. We have prepared all data sets generated in this study as an easily accessible and freely available online resource, comprising all results that showed statistical significance ($n = 3378$) [63]. The portal enables the research community to further characterize the hypervariable gene–phenotype associations (Additional files 4 and 9) using experimental approaches. For example, gene expression and DNA methylation hypervariability could be correlated to pathophysiological triggers of immune responses, such as interferon- γ and lipopolysaccharide [64]. These future studies will help elucidate how increased variability of gene expression and DNA methylation relate to functional diversity and effective adaptability during homeostatic and potentially pathogenic immune processes.

Conclusions

We found that neutrophils show increased variability in both their gene expression and DNA methylation patterns compared to monocytes and T cells. Our data suggest that increased variability in neutrophils may lead to cellular plasticity, enabling rapid adaptation to new or changing environments such as inflammation and pathogen intrusion. A detailed molecular understanding of the role of cellular heterogeneity in the human immune system is crucial to specifically target a pathogenic cellular subset without compromising immunity, ultimately advancing therapeutic design and treatment strategies in hematopoietic and immunological diseases.

Methods

Sample collection and isolation of cell subsets

As part of the BLUEPRINT Human Variation Panel, a total of 200 healthy blood donors were recruited from the NIH Cambridge BioResource [65]. Donors were on average 55 years of age (range 20–75 years) and 46% of donors were male. For all donors, a unit of whole blood (475 ml) was collected in 3.2% sodium citrate, of which an aliquot was collected in EDTA for genomic DNA purification and a full blood count using a Sysmex hematology analyzer. Blood was processed within 4 h of collection. We purified $CD14^+CD16^-$ monocytes, $CD66b^+CD16^+$ neutrophils, and naive $CD4^+CD45RA^+$ T cells using a multi-step purification strategy. In brief, whole blood was diluted 1:1 in a buffer of Dulbecco's

phosphate-buffered saline (PBS, Sigma) containing 13 mM of sodium citrate tribasic dehydrate (Sigma) and 0.2% human serum albumin (HSA, PAA), and then separated using an isotonic Percoll gradient of 1.078 g/ml (Fisher Scientific). Peripheral blood mononuclear cells were collected, washed twice with buffer, diluted to 25 million cells/ml, and separated into a monocyte-rich layer and a lymphocyte-rich layer using a Percoll gradient of 1.066 g/ml. Cells from each layer were washed with PBS containing 13 mM of sodium citrate and 0.2% HSA, and subsets purified using a strategy based on magnetic beads conjugated to highly specific antibodies. First, CD16⁺ cells were depleted from the monocyte-rich layer using CD16 MicroBeads (Miltenyi) according to the manufacturer's instructions. Cells were washed in PBS (13 mM of sodium citrate and 0.2% HSA) and CD14⁺ cells were positively selected using CD14 MicroBeads (Miltenyi). Next, CD4⁺ naïve T cells were negatively selected using an EasySep Human Naive CD4⁺ T Cell Enrichment Kit (StemCell) according to the manufacturer's instructions. Finally, the dense layer of cells from the 1.078 g/ml Percoll separation was lysed twice using an ammonium chloride buffer to remove erythrocytes. The resulting cells (including neutrophils and eosinophils) were washed, and neutrophils positively selected using CD16 MicroBeads (Miltenyi) following the manufacturer's instructions. The purity of each cell preparation was assessed by multi-color fluorescence-activated cell sorting (FACS). The following antibodies were used: CD14 (M4P9, BD Biosciences) and CD16 (B73.1/Leu-11c, BD Biosciences) for monocytes; CD16 (VEP13, MACS, Miltenyi) and CD66b (BIRMA 17C, IBGRL-NHS) for neutrophils; and CD4 (RPA-T4, BD) and CD45RA (HI100, BD) for T cells. Purity was on average 95% for monocytes, 98% for neutrophils, and 93% for T cells. Purified cell aliquots were pelleted, stored at -80 °C, and transported to the processing institutes. Further details about the experimental protocols and quality control assessments are provided by the BLUEPRINT Human Variation Panel.

RNA-sequencing assay and data preprocessing

RNA-seq sample preparation and library creation were performed for monocytes and neutrophils at the Max Planck Institute for Molecular Genetics (Germany), and for T cells at McGill University (Quebec, Canada). Purified cell aliquots were lysed and RNA extracted using TRIZOL reagent (Life Technologies) following the manufacturer's protocol. Sequencing libraries were prepared using a TruSeq Stranded Total RNA Kit with Ribo-Zero Gold (Illumina). Adapter-ligated libraries were amplified and indexed via PCR. Libraries were sequenced using 100-bp single-end reads for monocytes and neutrophils and paired-end reads for T cells. Reads from each

RNA-seq library were assessed for duplication rate and gene coverage using FastQC [66]. Then, PCR and sequencing adapters were trimmed using Trim Galore. Trimmed reads were aligned to the GRCh37 reference genome using STAR [67]. We used GENCODE v15 to define the annotated transcriptome. Read counts of genes and exons were scaled to adjust for differences in total library size using DESeq2 [68]. We adjusted for batch effects related to sequencing center using an empirical Bayesian method, ComBat [69]. Batch effects were assessed using cross-over samples, i.e., identical samples of each cell type per sample batch that were sent to the reciprocal center not processing the cell type. Visual inspection of the results by multidimensional scaling showed a successful reduction of batch effects following the application of ComBat (Additional file 1: Figure S1a). In addition, we calculated the correlation coefficients of all cross-over samples after batch effect correction. We obtained a mean correlation coefficient of $r = 0.96$ ($n = 15$ cross-over samples), indicating data consistency across the processing centers. An overview of the RNA-seq data quality assessment is provided in Additional file 1: Figure S2.

Quantification of gene expression

Analyses on RNA-seq data were performed on exon-based read counts per gene. We omitted all genes not expressed in at least 50% of all samples in each of the three cell types, leaving only genes that were robustly expressed in all three cell types. In addition, we included only protein-coding genes, resulting in a final set of 11,980 genes. RNA-seq read counts were converted into expression log counts by applying the formula $\log_2(x + 1)$.

Illumina Infinium HumanMethylation450 assay and data preprocessing

For monocytes and neutrophils, cell lysis and DNA extraction were performed at the University of Cambridge (UK), followed by bisulfite conversion and DNA methylation profiling at University College London (UK). T cells were processed at McGill University (Quebec, Canada). DNA methylation levels were measured using Infinium HumanMethylation450 assays (Illumina) according to the manufacturer's protocol. All 450 K array data preprocessing steps were carried out using established analytical methods incorporated in the R package minfi [70]. First, we performed background correction and dye-bias normalization using NOOB (normal-exponential convolution using out-of-band probes). The method estimates the background mean intensity using the over 135,000 out-of-band control probes, which provide signals in the opposite fluorescent channel from the probe design. NOOB effectively adjusts for differences in

background distribution and average intensities in the fluorescent channels between samples run on different arrays [71]. Then, we applied SWAN (subset-quantile within array normalization), a within array normalization method that reduces the differences in β -value distribution between Infinium I and II probe types [72]. Next, we filtered out probes based on the following criteria: (1) low detection P value ($P \geq 0.01$) in at least one sample; (2) bead count of less than three in at least 5% of samples; (3) mapping to sex chromosomes; (4) ambiguous genomic locations [73]; (5) non-CG probes; and (6) containing SNPs ($MAF \geq 0.05$) within 2 bp of the probed CG. Finally, we adjusted for batch effects due to processing center and analysis date using an empirical Bayesian framework [69], as implemented in the ComBat function of the R package SVA [74]. Multidimensional scaling analyses following the application of ComBat revealed no apparent batch effects (Additional file 1: Figure S1b). After batch effect correction, the mean correlation coefficient across cross-over samples was $r = 0.99$ ($n = 9$ samples), confirming data consistency across processing centers. An assessment of the DNA methylation data quality is shown in Additional file 1: Figure S3. In parallel, we performed singular value decomposition (SVD) of the DNA methylation data, which determined the components of variation (Additional file 1: Figure S3c).

Quantification of DNA methylation

The final data set that passed quality control consisted of 440,905 CpG sites. DNA methylation values were represented as either M values or β values. The methylation M value is the \log_2 ratio of the intensities of the methylated probe versus the unmethylated probe on the 450 K array, while the β value is the ratio of the methylated probe intensity and the overall intensity. All analyses of DNA methylation data were performed using M values. Due to their easier interpretability (i.e., 0–100% DNA methylation), β values were used for the visualization of DNA methylation data in most figures.

Analysis of differential variability

To assess differential variability across the three cell types, we applied a combined statistical approach based on DiffVar [75], which is embedded in the framework of limma [76, 77]. DiffVar calculates the median absolute deviation (MAD) from the group mean of expression levels of a particular gene, or DNA methylation at a given CpG site, across all individuals for two conditions, e.g., two distinct cell types. Then, a moderated t -test is used to test for a significant increase or decrease in MAD value between the two conditions. However, we found that the MAD variability measurement employed by DiffVar is correlated with mean levels (Additional file 1: Figures S4 and S8), which could potentially confound

the assessment of variability. Therefore, we included an additional measurement of variability that corrects for the dependency of variability measurements on the mean [8], here referred to as EV (gene expression variability value) and MV (DNA methylation variability value). The corresponding algorithm models variance as a function of the mean and then calculates the ratio of the observed variance to expected variance in order to get a variability measurement independent of the mean. Differential variability was tested in three group-wise comparisons. Statistical significance was defined as BH-corrected [78] $P < 0.05$ and EV/MV difference $\geq 10\%$ relative to the observed range of EV/MV values. For each cell type, both contrasts in which the cell type is involved were considered to define statistically significant differential variability. For example, for a gene to be a neutrophil-specific HVG, it must show significantly increased variability in both the comparison versus monocytes and versus T cells. For a gene to be classified as hypervariable across two cell types (shared hypervariability), it must exhibit significantly increased variability in the two corresponding cell types but low variability in the third. Thus, no gene can appear in more than one list. The statistical tests were performed in a paired fashion, taking into account that all three cell types were derived from the same individuals. This procedure corrects for potential differences related to individuals and sample processing.

Analysis of variability common to all three cell types

To identify HVGs common to all three cell types, we applied a rank-based approach. We ordered both MAD and EV values of all genes in the three cell types from high to low variability and then took the top n genes with the highest variability across all three cell types, where n corresponds to the mean number of results obtained for the gene lists of differential variability. Specifically, $n = 271$ for gene expression variability and $n = 212$ for DNA methylation variability.

Gene set enrichment analyses

For HVGs, we applied GOSEq using the default parameters and set 'use_genes_without_cat' = FALSE, thus ignoring genes without an annotated category for the calculation of P values [28]. With regards to HVPs, we analyzed the biological functions of flanking genes with GREAT [43] using the standard parameters: association rule = basal + extension (constitutive 5 kb upstream, 1 kb downstream, up to 1 Mb extension); curated regulatory domains = included. In both analyses, we used the set of analyzed features as background, and the cutoff for statistical significance was set at BH-corrected $P < 0.25$.

Gene co-expression network and pathway analysis

For neutrophil-specific HVGs not associated with *cis* genetic variants in the BLUEPRINT Human Variation Panel, we first constructed a co-regulation network by calculating gene expression correlations. The threshold of gene correlations was set at Pearson's $r > 0.6$. Unconnected genes were removed. The resulting correlation network was then further analyzed using Cytoscape [79]. Clusters were identified by the agglomerative clustering method FAG-EC [80] of the ClusterViz plugin. Enrichment analyses of resulting gene clusters were performed using clueGO [81], setting the Kappa score to 0.4 and the cutoff for statistical significance at BH-corrected $P < 0.05$. All networks were visualized using Gephi [82].

Correlation analyses

Associations between both gene expression and DNA methylation levels with donor-specific quantitative traits, cellular parameters, as well as weather and seasonal effects were assessed by calculating Spearman's rank correlation coefficients (ρ) and their corresponding P values. Results were considered statistically significant at BH-corrected $P < 0.05$. This threshold was also used for the correlation analyses between DNA methylation and gene expression data.

Analyses of seasonal effects

We downloaded historical raw weather data for the minimum and maximum daily temperature in London Heathrow (UK) for the period of data collection from the National Climatic Data Centre (USA) [83]. We applied linear interpolation to account for missing values. Additionally, we downloaded daylight hours for London [84]. The obtained data were then correlated with gene expression and DNA methylation values corresponding to the date of blood donation using Spearman's rank correlation coefficient (see details above).

Analyses of sex-specific differential gene expression

In each cell type, mean gene expression and DNA methylation differences between male and female donors were identified using limma [76, 77]. A moderated t-test was performed and statistical significance defined as BH-corrected $P < 0.05$ and log-fold change ≥ 1 . Results could be driven by differences in menopause status between female donors. Therefore, we performed the same analysis on only the subset of donors who are younger than 50 years and obtained very similar results compared to the complete donor group.

Functional annotation of hypervariable CpGs

For the enrichment analyses with regards to gene elements and epigenomic features, we used the annotation provided by the Illumina 450 K array manifest.

Enrichment was assessed by repeated random sampling ($n = 1000$) using all probes that passed quality control ($n = 440,905$), as previously described [85].

Transcription factor motifs analysis at gene promoter regions

Consensus transcription factor binding motifs were retrieved from the database "JASPAR_CORE_2016_vertbrates.meme" [86]. Using FIMO [87], we scanned for transcription factor binding motifs ($P < 1 \times 10^{-5}$) at promoter regions, defined as ± 500 bp around the transcription start site of genes listed in the reference gene set "UCSC.hg19.knownGene".

Programming language

If not indicated otherwise, analyses were performed using R v3 (R Development Core Team, 2008) and Bioconductor [88].

Additional files

Additional file 1: Supplementary figures. Supplementary document (.pdf) containing all Supplementary figures. (PDF 1165 kb)

Additional file 2: Annotation of genes showing hypervariable gene expression. For cell type-specific hypervariable genes, we provide the results of the correlation analyses of gene expression levels with genetic variation, donor-specific information, cell counts in peripheral blood, as well as seasonal data (corresponding to the date of blood donation). Only correlations with BH-corrected $P < 0.05$ are reported (indicated by "TRUE"). For all traits other than genetic variation the results presented are based on Spearman correlation tests (see Additional file 4 for correlation coefficients and P values). More detailed information, such as the corresponding SNP IDs for expression QTLs (obtained by the BLUEPRINT Human Variation Panel), can be found in our data portal available online at: <http://blueprint-dev.bioinfo.cnio.es/WP10/hypervariability>. (XLSX 467 kb)

Additional file 3: Gene ontology enrichment of genes showing hypervariable expression. Ontology enrichment analysis of genes showing hypervariable expression using GOseq. (XLSX 83 kb)

Additional file 4: Correlation of cell type-specific hypervariable gene expression with donor information. Correlation of cell type-specific hypervariable gene expression with donor information. (XLSX 358 kb)

Additional file 5: Neutrophil network gene ontology enrichment. Ontology enrichment analyses of three modules of a correlation network of neutrophil-specific hypervariable genes not mediated by *cis* genetic effects using ClueGO. (XLSX 68 kb)

Additional file 6: Gene ontology enrichment of sex-specific differentially expressed genes in neutrophils. Ontology enrichment analysis of differentially expressed genes between males and females in neutrophils using GOseq. (XLSX 57 kb)

Additional file 7: Annotation of CpGs showing hypervariable DNA methylation. See also description of Additional file 2 for further details about the information provided. Correlation coefficients and P values can be found in Additional file 9. Corresponding SNP IDs for methylation QTLs can be retrieved from our data portal available online at: <http://blueprint-dev.bioinfo.cnio.es/WP10/hypervariability>. (XLSX 211 kb)

Additional file 8: Gene ontology enrichment of genes showing hypervariable DNA methylation. Ontology enrichment analysis of genes in proximity to CpGs showing hypervariable DNA methylation using GREAT. (XLSX 60 kb)

Additional file 9: Correlation of cell type-specific hypervariable DNA methylation with donor information. Correlation of cell type-specific hypervariable DNA methylation with donor information. (XLSX 118 kb)

Additional file 10: Relationship between DNA methylation and gene expression. Relationship between DNA methylation and gene expression among genes showing cell type-specific DNA methylation hypervariability. (XLSX 29 kb)

Abbreviations

BH: Benjamini–Hochberg; EV: Gene expression variability value; HSA: Human serum albumin; HVG: Hypervariable gene; HVP: Hypervariable CpG position; MAD: Median absolute deviation; MV: DNA methylation variability value; PBS: Phosphate-buffered saline; RNA-seq: RNA sequencing.

Acknowledgments

We would like to thank K. Pearce and M. Kristiansen (UCL Genomics) for processing the Illumina Infinium HumanMethylation450 BeadChips; D. Balzereit, S. Dökel, A. Kovacovics, and M. Linser (Max Planck Institute for Molecular Genetics) for help with generating the RNA-seq data; B. Phipson (Murdoch Childrens Research Institute), H.C. Bravo (University of Maryland), and P. Guilhamon (UCL Cancer Institute) for advice on statistical analyses; C. Bock (CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences) for useful discussions; A. Orozco (University of Costa Rica) for technical support; V. Naranbhai, B. Fairfax, and J. Knight (University of Oxford) for providing access to the neutrophil gene expression data set for replication; and L. Phipps for proofreading the manuscript. We gratefully acknowledge the participation of all NIHR Cambridge BioResource volunteers, and thank the Cambridge BioResource staff for their help with volunteer recruitment. We thank members of the Cambridge BioResource SAB and Management Committee for their support of our study and the NIHR Cambridge Biomedical Research Centre (BRC) for funding. BLUEPRINT Consortium: Cornelis A. Albers (Radboud University), Vyacheslav Amstislavskiy (Max Planck Institute for Molecular Genetics), Sofie Ashford (University of Cambridge), Lorenzo Bombà (Wellcome Trust Sanger Institute), David Bujold (McGill University), Frances Burden (University of Cambridge), Stephan Busche (McGill University), Maxime Caron (McGill University), Shu-Huang Chen (McGill University), Warren A. Cheung (McGill University), Laura Clarke (European Bioinformatics Institute), Irina Colgiu (Wellcome Trust Sanger Institute), Avik Datta (European Bioinformatics Institute), Oliver Delaneau (University of Geneva), Heather Elding (Wellcome Trust Sanger Institute), Samantha Farrow (University of Cambridge), Diego Garrido-Martín (Centre for Genomic Regulation), Bing Ge (McGill University), Roderic Guigo (Centre for Genomic Regulation), Valentina Iotchkova (European Bioinformatics Institute), Kousik Kundu (Wellcome Trust Sanger Institute), Tony Kwan (McGill University), John J. Lambourne (University of Cambridge), Ernesto Lowy (European Bioinformatics Institute), Daniel Mead (Wellcome Trust Sanger Institute), Farzin Pourfarzad (Sanquin Research and Landsteiner Laboratory), Adriana Redensek (McGill University), Karola Rehnstrom (University of Cambridge), Augusto Rendon (University of Cambridge), David Richardson (European Bioinformatics Institute), Thomas Risch (Max Planck Institute for Molecular Genetics), Sophia Rowlston (University of Cambridge), Xiaojian Shao (McGill University), Marie-Michelle Simon (McGill University), Marc Sultan (Max Planck Institute for Molecular Genetics), Klaudia Walter (Wellcome Trust Sanger Institute), Steven P. Wilder (European Bioinformatics Institute), Ying Yan (Wellcome Trust Sanger Institute), Stylianos E. Antonarakis (University of Geneva), Guillaume Bourque (McGill University), Emmanouil T. Dermizakis (University of Geneva), Paul Flicek (European Bioinformatics Institute), Hans Lehrach (Max Planck Institute for Molecular Genetics), Joost H. A. Martens (Radboud University), Marie-Laure Yaspo (Max Planck Institute for Molecular Genetics), Willem H. Ouwehand (University of Cambridge).

Funding

This work is predominantly funded by the EU-FP7 Project BLUEPRINT (HEALTH-F5-2011-282510). S. Ecker is supported by a “la Caixa” pre-doctoral fellowship. V. Pancaldi is supported by a FEBS Long-Term Fellowship. F.O. Bagger is supported by The Lundbeck Foundation. K. Downes is funded as a HSST trainee by NHS Health Education England. M. Frontini is supported by the British Heart Foundation (BHF) Cambridge Centre of Excellence (RE/13/6/30180). S. Beck acknowledges support from the Wellcome Trust (WT99148), a

Royal Society Wolfson Research Merit Award (WM100023), and the UK National Institute for Health Research (NIHR) UCLH Biomedical Research Centre (BRC84/CN/SB/5984). N. Soranzo’s research is supported by the Wellcome Trust (WT098051 and WT091310), EPIGENESYS (257082), and NIHR Cambridge Biomedical Research Centre (BRC). The INB-CNIO Unit is a member of ProteoRed (PRB2-ISCIII) and is supported by PE I + D + i 2013–2016 (PT13/0001), ISCIII, and FEDER. The Cardiovascular Epidemiology Unit is supported by the UK Medical Research Council (G0800270), BHF (SP/09/002), and NIHR Cambridge BRC.

Availability of data and materials

All data sets generated as part of this study are available at the European Genome-phenome Archive (EGA) [89] under the following accession numbers: EGAS00001001456 for 450 K array data; EGAS00001000752 and EGAS00001000327 for RNA-seq data.

Authors’ contributions

SE and DSP designed the study. SE, DSP, LC, VP, FOB, ECdeSP, DJ, NS, NR, and AM analyzed data. All other authors provided samples or analytical tools. DSP and SE wrote the manuscript. DR, AV, SB, NS, and DSP supervised the study. All authors read and approved the final manuscript.

Competing interests

P. Flicek is a member of the Scientific Advisory Board for Omicia, Inc.

Consent for publication

Not applicable.

Ethics approval and consent to participate

All sample material was collected at the NHS Blood and Transplant Centre in Cambridge (UK) with informed consent (REC 12/EE/0040). The experimental methods comply with the Declaration of Helsinki.

Author details

¹Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain. ²UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK. ³Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK. ⁴Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge, Hinxton, UK. ⁵National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK. ⁶European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ⁷The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. ⁸Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. ⁹The Bioinformatics Centre, Department of Biology, Faculty of Natural Sciences, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. ¹⁰National Center for Genomic Analysis (CNAG), Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Carrer Baldri i Reixac 4, 08028 Barcelona, Spain. ¹¹Department of Molecular Biology, Faculty of Science, Radboud University, Nijmegen 6525GA, The Netherlands. ¹²British Heart Foundation Centre of Excellence, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK. ¹³Department of Human Genetics, McGill University, 740 Dr. Penfield, Montreal H3A 0G1, Canada. ¹⁴Blood Cell Research, Sanquin Research and Landsteiner Laboratory, Plesmanlaan 125, Amsterdam 1066CX, The Netherlands. ¹⁵Emma Children’s Hospital, Academic Medical Center (AMC), University of Amsterdam, Location H7-230, Meibergdreef 9, Amsterdam 1105AX, The Netherlands. ¹⁶Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. ¹⁷Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Wort’s Causeway, Cambridge CB1 8RN, UK.

Received: 25 October 2016 Accepted: 17 January 2017

Published online: 26 January 2017

References

- Yosef N, Regev A. Writ large: Genomic Dissection of the Effect of Cellular Environment on Immune Response. *Science*. 2016;354:64–8.
- Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science*. 2002;297:1183–6.
- Lehner B, Kaneko K. Fluctuation and response in biology. *Cell Mol Life Sci*. 2011;68:1005–10.
- Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. *Science*. 2005;309:2010–3.
- Snijder B, Pelkmans L. Origins of regulated cell-to-cell variability. *Nat Rev Mol Cell Biol*. 2011;12:119–25.
- Blake WJ, Balázi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, et al. Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell*. 2006;24:853–65.
- Dong D, Shao X, Deng N, Zhang Z. Gene expression variations are predictive for stochastic noise. *Nucleic Acids Res*. 2011;39:403–13.
- Alemu EY, Carl JW, Corrada Bravo H, Hannehalli S. Determinants of expression variability. *Nucleic Acids Res*. 2014;42:3503–14.
- Basehoar AD, Zanton SJ, Pugh BF. Identification and distinct regulation of yeast TATA box-containing genes. *Cell*. 2004;116:699–709.
- Busslinger M, Tarakhovskiy A. Epigenetic control of immunity. *Cold Spring Harb Perspect Biol*. 2014;6:a019307.
- Paszek P, Ryan S, Ashall L, Sillitoe K, Harper CV, Spiller DG, et al. Population robustness arising from cellular heterogeneity. *Proc Natl Acad Sci U S A*. 2010;107:11644–9.
- Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A*. 2010;107:1757–64.
- Pujadas E, Feinberg AP. Regulated noise in the epigenetic landscape of development and disease. *Cell*. 2012;148:1123–31.
- Choi JK, Kim Y-J. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat Genet*. 2009;41:498–503.
- Brock A, Chang H, Huang S. Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours. *Nat Rev Genet*. 2009;10:336–42.
- Hansen KD, Timp W, Bravo HC, Sabuncian S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*. 2011;43:768–75.
- Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, et al. Locally Disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*. 2014;26:813–25.
- Li J, Liu Y, Kim T, Min R, Zhang Z. Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Comput Biol*. 2010;6:e1000910.
- Mar JC, Matigian NA, Mackay-Sim A, Mellick GD, Sue CM, Silburn PA, et al. Variance of gene expression identifies altered network constraints in neurological disease. *PLoS Genet*. 2011;7:e1002207.
- Bravo HC, Pihur V, McCall M, Irizarry RA, Leek JT. Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinformatics*. 2012;13:272.
- Ecker S, Pancaldi V, Rico D, Valencia A. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. *Genome Med*. 2015;7:8.
- Teschendorff AE, Jones A, Fiegl H, Sargent A, Zhuang JJ, Kitchener HC, et al. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med*. 2012;4:24.
- Teschendorff AE, Liu X, Caren H, Pollard SM, Beck S, Widschwendter M, et al. The dynamics of DNA methylation covariation patterns in carcinogenesis. *PLoS Comput Biol*. 2014;10:e1003709.
- Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*. 2016;167:1398–414.
- Geering B, Simon H-U. Peculiarities of cell death mechanisms in neutrophils. *Cell Death Differ*. 2011;18:1457–69.
- Subrahmanyam YVBK, Yamaga S, Prashar Y, Lee HH, Hoe NP, Kluger Y, et al. RNA expression patterns change dramatically in human neutrophils exposed to bacteria. *Blood*. 2001;97:2457–68.
- Naranbhai V, Fairfax BP, Makino S, Humburg P, Wong D, Ng E, et al. Genomic modulators of gene expression in human neutrophils. *Nat Commun*. 2015;6:7545.
- Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11:R14.
- Berditchevski F. Complexes of tetraspanins with integrins: more than meets the eye. *J Cell Sci*. 2001;114:4143–51.
- Hemler ME. Tetraspanin functions and associated microdomains. *Nat Rev Mol Cell Biol*. 2005;6:801–11.
- Nuzzi P, Senetar M, Huttenlocher A. Asymmetric localization of calpain 2 during neutrophil chemotaxis. *Mol Biol Cell*. 2007;18:795–805.
- Saito YD, Jensen AR, Salgia R, Posadas EM. Fyn: a novel molecular target in cancer. *Cancer*. 2010;116:1629–37.
- GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015;348:648–60.
- Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurivsky A, et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*. 2015;11:e1004958.
- Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*. 2013;342:744–7.
- Song Z, Krishna S, Thanos D, Strominger J, Ono S. A novel cysteine-rich sequence-specific DNA-binding protein interacts with the conserved X-box motif of the human major histocompatibility complex class II genes via a repeated Cys-His domain and functions as a transcriptional repressor. *J Exp Med*. 1994;180:1763–74.
- Schneider WM, Dittmann Chevillotte M, Rice CM. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol*. 2014;32:513–45.
- Fairweather D, Frisancho-Kiss S, Rose NR. Sex differences in autoimmune disease from a pathological perspective. *Am J Pathol*. 2008;173:600–9.
- Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518:317–30.
- Carrillo de Santa Pau E, Juan D, Pancaldi V, Were F, Martin-Subero I, Rico D, et al. Searching for the chromatin determinants of human hematopoiesis. *bioRxiv*. 2016. doi:10.1101/082917
- Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*. 2010;28:817–25.
- Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9:215–6.
- McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*. 2010;28:495–501.
- Dröge W, Holm E. Role of cysteine and glutathione in HIV infection and other diseases associated with muscle wasting and immunological dysfunction. *FASEB J*. 1997;11:1077–89.
- Harburger DS, Calderwood DA. Integrin signalling at a glance. *J Cell Sci*. 2009;122:159–63.
- Miranti CK, Brugge JS. Sensing the environment: a historical perspective on integrin signal transduction. *Nat Cell Biol*. 2002;4:E83–90.
- Dopico XC, Evangelou M, Ferreira RC, Guo H, Pekalski ML, Smyth DJ, et al. Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nat Commun*. 2015;6:7000.
- Golbus J, Palellan TD, BCR A, Arbor A. Quantitative changes in T cell. *Eur J Immunol*. 1990;20:1869–72.
- Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol*. 2016;17:11.
- Maue AC, Yager EJ, Swain SL, Woodland DL, Blackman MA, Haynes L. T-cell immunosenescence: lessons learned from mouse models of aging. *Trends Immunol*. 2009;30:301–5.
- Weng N. Aging of the immune system: how much can the adaptive immune system adapt? *Immunity*. 2006;24:495–9.
- Tsukamoto H, Clise-Dwyer K, Huston GE, Duso DK, Buck AL, Johnson LL, et al. Age-associated increase in lifespan of naive CD4 T cells contributes to T-cell homeostasis but facilitates development of functional defects. *Proc Natl Acad Sci U S A*. 2009;106:18333–8.
- Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012;13:484–92.
- Gallant S, Gilkeson G. ETS transcription factors and regulation of immunity. *Arch Immunol Ther Exp*. 2006;54:149–63.
- Silvestre-Roig C, Hidalgo A, Soehnlein O. Neutrophil heterogeneity: implications for homeostasis and pathogenesis. *Blood*. 2016;127:2173–81.

56. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods*. 2016;13:229–32.
57. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep*. 2015;10:1386–97.
58. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol*. 2015;33:155–60.
59. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science*. 2014;343:776–9.
60. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*. 2015;163:1663–77.
61. Lu Y, Biancotto A, Cheung F, Remmers E, Shah N, McCoy JP, et al. Systematic analysis of cell-to-cell expression variation of T lymphocytes in a human cohort identifies aging and genetic associations. *Immunity*. 2016;45:1162–75.
62. Bock C, Farlik M, Sheffield NC. Multi-omics of single cells: strategies and applications. *Trends Biotechnol*. 2016;34:605–8.
63. BLUEPRINT. WP10 data portal: Hypervariability. 2016. <http://blueprint-dev.bioinfo.cnio.es/WP10/hypervariability>. Accessed 7 Oct 2016.
64. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*. 2014;343:1246949.
65. The Cambridge NIHR. BioResource. 2016. <http://www.cambridgebioresource.org.uk>. Accessed 7 Oct 2016.
66. Andrews S, FastQC A. Quality Control tool for High Throughput Sequence Data. 2014. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 26 Jun 2015.
67. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
68. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol*. 2014;15:550.
69. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118–27.
70. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30:1363–9.
71. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*. 2013;41:e90.
72. Makismovic J, Gordon L, Oshlack A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol*. 2012;13:R44.
73. Nordlund J, Bäcklin CL, Wahlberg P, Busche S, Berglund EC, Eloranta M-L, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol*. 2013;14:r105.
74. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28:882–3.
75. Phipson B, Oshlack A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol*. 2014;15:465.
76. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43:e47.
77. Smyth GK. Limma: Linear Models for Microarray Data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. *Bioinformatics and computational biology solutions using R and Bioconductor*. New York: Springer; 2005. p. 397–420.
78. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*. 1995;289–300.
79. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498–504.
80. Li M, Wang J, Chen J. A fast agglomerate algorithm for mining functional modules in protein interaction networks. *BioMed Eng Informatics*. 2008;1:3–7.
81. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25:1091–3.
82. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. *ICWSM*. 2009;8:361–2.
83. National Climatic Data Center (NCDC). 2016. <http://www.ncdc.noaa.gov/cdo-web/search>. Accessed 8 Apr 2016.
84. Timeanddate.com. 2016. <http://www.timeanddate.com>. Accessed 8 Apr 2016.
85. Guilhamon P, Eskandarpour M, Halai D, Wilson GA, Feber A, Teschendorff AE, et al. Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2. *Nat Commun*. 2013;4:2166.
86. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2014;42:D142–7.
87. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*. 2011;27:1017–8.
88. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5:R80.
89. European Genome-phenome Archive. 2016. <https://www.ebi.ac.uk/ega/>. Accessed 7 Oct 2016.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

