OPEN

Robust shifts in *S100a9* expression with aging: A novel mechanism for chronic inflammation

William R. Swindell, Andrew Johnston, Xianying Xing, Andrew Little, Patrick Robichaud, John J. Voorhees, Gary Fisher & Johann E. Gudjonsson

Department of Dermatology, University of Michigan School of Medicine, Ann Arbor, MI 48109-2200.

Correspondence and
requests for materials
should be addressed to
W.R.S. (wswindel@
umich.edu)

The *S100a8* and *S100a9* genes encode a pro-inflammatory protein (calgranulin) that has been implicated in multiple diseases. However, involvement of *S100a8/a9* in the basic mechanisms of intrinsic aging has not been established. In this study, we show that shifts in the abundance of *S100a8* and *S100a9* mRNA are a robust feature of aging in mammalian tissues, involving a range of cell types including the central nervous system. To identify transcription factors that control *S100a9* expression, we performed a large-scale transcriptome analysis of 62 mouse and human cell types. We identified cell type-specific trends, as well as robust associations linking *S100a9* coexpression to elevated frequency of ETS family motifs, and in particular, to motifs recognized by the transcription factor SPI/PU.1. Sparse occurrence of SATB1 motifs was also a strong predictor of *S100a9* coexpression. These findings offer support for a novel mechanism by which a SPI1/PU.1-S100a9 axis sustains chronic inflammation during aging.

Aging is commonly associated with a state of chronic inflammation that contributes to DNA damage, atherosclerosis, stem cell senescence and cognitive decline[1–3]. Although inflammation does not occur uniformly or within all mammalian tissues due to increased age[4], in some tissues aging leads to activation of immune-response pathways and the formation of lymphoid aggregates, particularly within perivascular regions[5–10]. Mechanisms that underpin these processes are not well understood, however, and further work is needed to identify "hubs" within inflammatory and cytokine networks that drive these events. In recent years, two low molecular weight proteins, *S100a8* (calgranulin A) and *S100a9* (calgranulin B), have emerged as central inflammatory regulators capable of driving and responding to inflammation signals[11–13]. On the one hand, *S100a8/a9* mRNA and protein levels are markedly increased in response to cytokine stimuli[14,15]. At the same time, *S100a8/a9* reinforce inflammatory cascades by serving as leukocyte chemoattractants[16], inducing the expression of pro-inflammatory cytokines[17], triggering activation of NF-κB[17,18], and by serving as ligands that interact with and stimulate receptor for advanced glycation end products (RAGE)[19]. *S100a8/a9* therefore play a unique part within inflammatory networks -- acting as inflammation-responsive proteins but yet amplifying inflammatory signals -- ultimately contributing to a positive feedback cycle conducive to unchecked inflammation responses. Such activity may contribute to the development of age-related disease. Recent studies, for instance, support a role for *S100a8/a9* in Alzheimer's disease[20–23], and knockdown of *S100a9* expression was shown to reduce amyloid plaque abundance and improve water maze performance in an Alzheimer's mouse model[21].

Inflammatory processes depend, in part, upon activity within transcriptional regulatory networks, in which key transcription factors (TFs) drive the expression of pro-inflammatory target genes. Such mechanisms reinforce the feed-forward inflammatory circuits that sustain chronic inflammation, potentially in a cell-type-specific fashion. *S100a8* and *S100a9* are, for instance, co-expressed in some cell types and may thus share transcriptional regulatory mechanisms, consistent with instability of *S100a8* or *S100a9* monomers *in vivo* and formation of the noncovalent S100a8-S100a9 heterocomplex (i.e., calprotectin)[24,25]. However, *S100a8* and *S100a9* are not always co-expressed[26,27], suggesting a co-expression network that varies by cell type or according to environmental signals. Multiple DNA-binding factors have been identified as regulators of *S100a9* expression, although collectively, experimental studies have focused upon a heterogeneous set of tissues and transformed cell lines. The transcription factors STAT3 and NF-κB were independently identified as activators of *S100a9* expression in both human and mouse cell types[28–30]. Other TFs have also been identified as *S100a9* regulators, although it remains

unclear whether their regulatory role is limited to humans alone, mice alone, or to a single cell type. Such factors include putative *S100a9* activators C/EBPα and C/EBPβ[31,32], HIF-1α[33], GLI1[34] and SPI1/PU.1[35], in addition to putative *S100a9* repressors BRCA1[36], AP-1[37], SATB1[38] and Arnt/HIF-1β[39]. In the context of aging or disease, such DNA-binding factors may drive or suppress *S100a9* expression, thereby modulating the intensity of age-dependent inflammation. This is one avenue towards development of therapeutic approaches, and indeed, pharmacological inhibition of *S100a9* activators (e.g., NF-κB) has been linked to slowed tumor growth and delayed accumulation of senescent cells with aging[40,41].

We here show that aging leads to shifts in the abundance of *S100a8/a9* in humans and mice, which involve multiple tissues and robust trends across mouse genotypes. These findings demonstrate that shifts in *S100a8/a9* abundance are a feature of normal aging, suggesting a role for *S100a8/a9* in age-associated inflammation that extends beyond their involvement in specific diseases (e.g., Alzheimer's disease). We have further investigated mechanisms of *S100a9* transcription using a large-scale integrative transcriptomics strategy, which has allowed us to systematically screen DNA-binding factors for association with *S100a9* expression across many cell types and transformed cell lines (30 mouse cell types and 32 human cell types). Our approach provides objective statistical assessment of evidence for those TFs with known DNA binding affinities, and offers a means to distinguish cell type-specific patterns from more robust trends supported in multiple cell types. We thus illustrate a strategy that is generally useful for *in silico* study of mammalian gene regulation, and in the current application, our results provide systems-level insight into TFs and pathways that govern *S100a9* transcription. Our findings, moreover, point towards a new mechanism for age-related chronic inflammation, in which over-production of S100a9, enforced by key transcription factors, triggers a feed-forward cycle that sustains a pro-inflammatory microenvironment with increasing age.

Results

Shifts in *S100a8/a9* abundance are a robust feature of normal aging in mouse and human tissues. We used Affymetrix DNA oligonucleotide microarrays to evaluate gene expression in tail skin from young (5 months) and old (30 months) CB6F1 mice (*n* = 5 per age group and sex). Two of the top ten age-increased genes encoded S100 proteins, including *S100a8* an *S100a9* (data not shown). We confirmed this pattern using RT-PCR and showed that *S100a8/a9* expression increased late in the lifespan between 17 and 30 months of age, with no significant change between 5 and 17 months of age (Figures S1A and S1D). Consistent with this, immunostaining did not detect S100a9 in young mice, but did detect S100a9 widely distributed in the epidermis, dermis and subcutaneous layers in old mouse skin (Figure S2). We expected to observe a similar pattern in human skin. However, analysis of a large microarray dataset indicated that *S100A8/A9* expression decreased with age by 1–2% per year (Figures S1B and S1C). We thus used RT-PCR to analyze an independent set of skin biopsies obtained from young (20–40 years old; *n* = 11 males and *n* = 6 females) and elderly human subjects (> 80 years old; *n* = 5 males and *n* = 10 females). This confirmed that *S100A8/A9* expression decreased with age in human sun-protected skin (Figure S1E), in contrast to the trend observed in tail skin from CB6F1 mice (Figures S1A and S1D).

We further investigated the effects of aging on S100 expression in other mouse and human tissues. The AGEMAP study was a large-scale effort designed to identify age-associated gene expression patterns across 16 tissues from C57BL/6 mice (1, 6, 16 and 24 months of age)[42]. However, inspection of the AGEMAP *S100a9* expression pattern revealed no trend towards elevated expression across 16 tissues (Figure S3). Since AGEMAP was limited to a single inbred genotype (C57BL/6) and used cDNA arrays that lacked probes to detect several S100 genes (e.g., *S100a8*), we assembled datasets that used whole-genome Affymetrix arrays to evaluate gene expression in young and old tissues (Figure 1). These data revealed that, in both mice and humans, all S100 gene family members were significantly altered by aging in at least one tissue (P < 0.05), with the strongest and most consistent effects observed for *S100a8* and *S100a9* (Figure 1). These results were further supported by our studies of multiple organs from CB6F1 mice, which showed an increase in S100a9 expression with age (heart, liver, lung and kidney) (Figure S4). Additionally, staining of tissues using an anti-S100a9 antibody revealed a trend towards
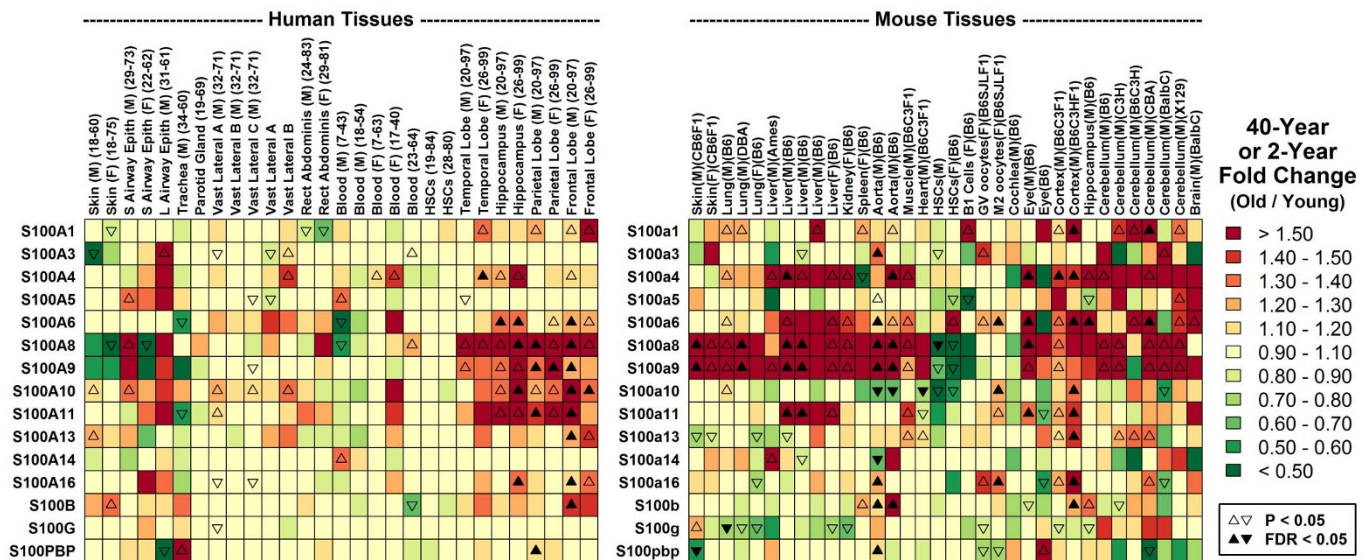
Figure 1 | Shifts in the expression of *S100a8* and *S100a9* are a robust feature of aging in human and mouse tissues (meta-analysis of microarray data). The effects of aging on the expression of S100-encoding mRNAs were evaluated in human (left) and mouse (right) tissues. All human data was generated using the Affymetrix Human Genome U133 Plus 2.0 array platform (29 experiments), and all mouse data was generated using the Affymetrix Mouse Genome 430 2.0 array platform (34 experiments). Data were obtained from the Gene Expression Omnibus or ArrayExpress databases. Colors denote the estimated fold-change over 40 years in humans (left; old/young) or over 2 years in mice (right; old/young). Triangles denote significant effects of aging on the expression of the listed gene (row) with respect to the indicated tissue (column) (P < 0.05 or FDR < 0.05).

increased S100a9 protein in liver (males and females, Figures S5 and S6), kidney (males only; Figure S7) and lung (males and females, Figure S8).

An *in silico* procedure for identifying transcription factors and motifs that regulate *S100a9* expression in 30 mouse and 32 human cell types. Our data identified shifts in the abundance of *S100a9* mRNA and protein as a robust feature of intrinsic mammalian aging. Given that expression of *S100a9* was often elevated with age (Figure 1), we hypothesized that such an increase could facilitate age-associated inflammation and the formation of lymphocyte aggregates in older tissues, as has been described in prior work[5–9]. To identify mechanisms that may mediate increased *S100a9* expression with aging, therefore, we developed an *in silico* procedure to identify transcription factor binding sites that predict *S100a9* coexpression in mouse and human cell types (Figure 2). For this purpose, we assembled manually curated datasets for each of 30 mouse cell types and 32 human cell types. Our strategy was to identify *S100a9* coexpressed genes for each mouse and human cell type (Figures 2A and 2B), yielding the cell type-specific sub-network surrounding *S100a9* within the genome-wide coexpression network (Figure 2C). We then used generalized additive logistic models (GAM) to identify motifs for which motif frequency in the 2 KB upstream region predicted *S100a9* coexpression within each cell type (Figure 2D). For these analyses, we screened a dictionary consisting of 1209 motifs, where each motif corresponded to a mouse or human DNA-binding protein (see Methods). In addition to cell type-specific analyses, we analyzed a "composite" *S100a9* coexpression network, which was generated by integrating trends across all cell types in mouse or human, respectively (see Methods).

Genes coexpressed with *S100a9* in mouse cells are located in regions enriched with binding sites for ETS transcription factors (e.g., SPI1/PU.1) but with sparse occurrence of SATB1 sites (AATTTT). Genes coexpressed with *S100a9* in the composite network were enriched with DNA motifs bound by members of the ETS transcription factor family (Figures 3 and S9). Of the ten most significant motifs in the composite network, nine were recognized by ETS family members (e.g., SPI1, ETS1, ETS2), and four of these were associated with the ETS member SPI1/PU.1 (Figure 3). The top-ranked motif in the composite network, SPI1|AGGAAGT|MA0080, was significantly enriched among *S100a9* coexpressed genes in 14 of 30 cell types (P < 0.05), although we identified another SPI1/PU.1 motif that was significantly enriched among *S100a9* coexpressed genes in 25 of 30 cell types (SPI1|GGAAG|MA0080; Figure 3). Additionally, we identified five cell types for which the most significantly enriched motif was associated with SPI1/PU.1 (hematopoietic stem cells, B-cells, thymocytes, CD8+ T-cells and CD4+ T-cells; Figure 3). These analyses were based upon 2KB upstream regions, but further analyses of other regions also supported an association between ETS family motifs and *S100a9* coexpression (e.g., conserved regions 2 KB upstream, intronic regions and all non-coding intergenic regions; Figures S9A – S9C). Additionally, we performed an alternative analysis to determine which *k*-mers (of length 2, 3, 4 or 5) best distinguished enriched (P < 0.05 with Z > 0) from non-enriched (P > 0.05) motifs (2 KB upstream region). In the composite network, enriched motifs differed from non-enriched motifs in that they matched the 5-mer AGGAA|TTCCT (Figure S9D). This 5-mer matches the GGA(A/T) core motif recognized by ETS family members[43] and resembles the consensus sequence for several SPI1/PU.1 binding sites (Figures 3 and S9).

Cross-validation was used to determine how well single-motif regression models could predict *S100a9* coexpression. This showed that the "best model" (Akaike information criterion, AIC) was able to predict *S100a9* coexpression for the composite network and 20 of 30 mouse cell types (i.e., observed and null AUC distributions at least 1

SD apart) (data not shown). Performance improved, however, when we analyzed two-motif models (Figure 4). For bivariate motif models, there was strong separation between observed and null AUC distributions for 26 of 30 cell types (Figure 4). The best-performing bivariate models for each cell type often related increased abundance of ETS-associated motifs to increased probability of *S100a9* coexpression (Figure 4). Surprisingly, however, for 6 of 30 cell types, decreased occurrence of SATB1 binding sites was predictive of *S100a9* coexpression (V$SATB1_Q5|AATTTT; splenocytes, keratinocytes, lymphatic endothelial cells, pancreatic islets, 3T3-L1 cells and RAW264 cells) (Figure 4). This association was robustly observed across 29 of 30 cell types (upstream 2 KB region; P < 0.05 with Z < 0), indicating that genes coexpressed with *S100a9* in mouse cells are associated with genomic regions that are deficient with respect to SATB1 binding sites.

The region 150–250 BP upstream of the mouse *S100a9* gene contains the highest concentration of transcription factor binding sites that predict *S100a9* coexpression. The complete statistical profile of all transcription factor binding sites we evaluated provides a tool for detecting loci with cell type-specific regulatory potential. We thus scanned the 2 KB region upstream of the mouse *S100a9* gene to determine which region contained the greatest concentration of motifs enriched among *S100a9* coexpressed genes (Figure 5). For the composite network, and all 30 cell types, sliding window analyses identified a high-scoring locus 150–250 BP upstream (mm9, chr3, 90499762–90499862) (Figure 5). We identified 19 motifs within this region for which increased occurrence was associated with *S100a9* coexpression in the composite network (Figure S10; P < 0.047 with Z > 0). Further inspection uncovered a 13 base element (5-ACTTCCC-TTCCAT-3; mm9, chr3, 90499826–90499838) with binding sites for several significant motifs, including multiple matches to motifs recognized by SPI1/PU.1 and STAT transcription factors (Figure S10).

GC-rich motifs and bindings sites for ETS transcription factors (e.g., SPI1/PU.1) predict *S100A9* coexpression in diverse human cell types. We repeated our screen of 1209 transcription factor binding sites (Figure 2), except we attempted to identify motifs predictive of *S100A9* coexpression in 32 human cell types (2 KB upstream regions). A motif-by-motif comparison of *Z* statistics from 13 corresponding cell types in mouse and human revealed good correspondence in 12 cases ($0.21 \leq r_s \leq 0.64$; P < 0.001), and overall, correspondence between results from mouse and human composite networks was convincing ($r_s = 0.43$; P = 10^{-49}) (Figure S11). In line with this, several motifs supported in our analysis of mouse cells were bolstered by trends detected in human cells (Figures 6 and S12). For instance, two motifs associated with ETS transcription factors were consistently enriched among *S100A9* coexpressed genes (V$ETS2_Q6|CTTCCTG and SPI1|GGAAG|MA0080), with significant trends detected for 28 of 32 human cell types including the human composite network (2 KB upstream regions; P < 0.05 and Z > 0) (Figure 6). This pattern was also supported by evaluation of conserved 2 KB upstream sequences, intronic sequences, and complete non-coding intergenic sequences (Figures S12A–S12C). Interestingly, a strong trend in human cells was enrichment of GC-rich motifs in genomic regions associated with *S100a9* coexpressed genes (e.g., V$MOVOB_01|GGGGG and V$SP1_Q2_01|CCCCGCCCC) (Figure 6). Consistent with this, we searched for features that distinguished significantly enriched motifs (P < 0.05 with Z > 0) from all other motifs, and found that enriched motifs more closely matched the 4-mer CCCC|GGGG (Figure S12D). This trend was further supported by cross-validation analyses in which we identified bivariate models that best predicted *S100A9* coexpression (AIC criterion; see Figure 7). For 7 of 32 cell types, increased abundance of an ETS transcription factor motif was
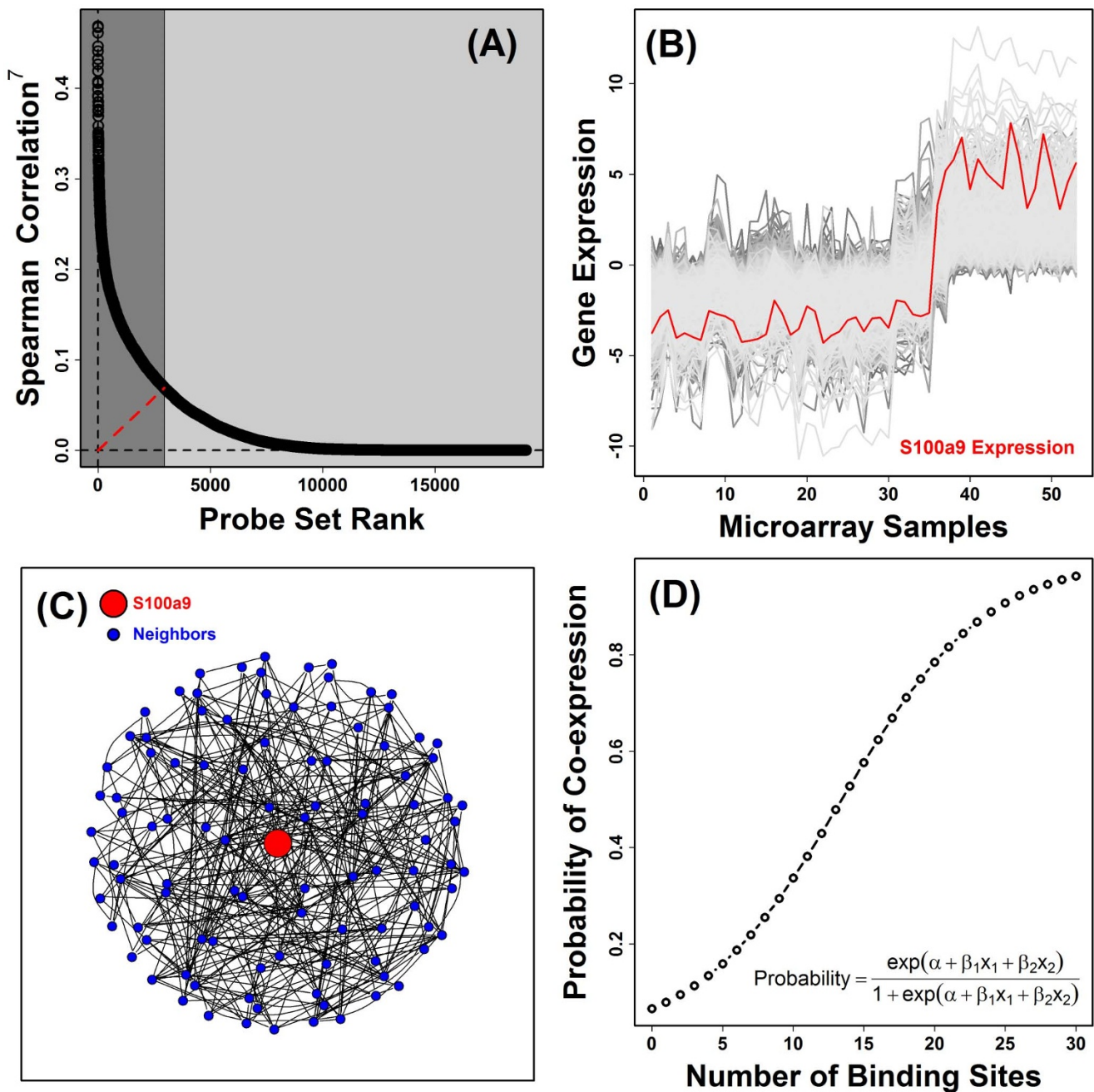
Figure 2 | *In silico* strategy for identifying *cis*-regulatory mechanisms controlling *S100a9* expression. The figure illustrates a general procedure for identifying a cluster of *S100a9*-coexpressed genes (parts A – C), which can then be evaluated to identify TF binding sites that occur at disproportionately high frequency within associated genomic sequences (part D). The procedure is here illustrated for a single cell type (mouse chondrocytes), but we have applied the methodology across a broader panel of 30 mouse and 32 human cell types. In the first step (A), a foreground set of *S100a9*-coexpressed transcripts is identified. This is done by calculating the Spearman rank correlation (r_s) between each transcript and *S100a9*, and then ranking all transcripts by the magnitude of r_s^7. The dashed red line shown in (A) represents the segment with minimal distance between the origin (lower left corner) and the curve shown in the figure. This red line serves to define the foreground set of *S100a9*-coexpressed genes (dark grey region). In part (B), this *S100a9*-coexpression cluster is illustrated with respect to the 53 microarray samples used to calculate Spearman rank correlations shown in (A), where each microarray sample was generated by hybridization with cDNA derived from mouse chondrocytes. The foreground set of *S100a9*-coexpressed genes can thus be viewed as the local sub-network that surrounds *S100a9*, as illustrated in (C). In the final step (D), a generalized additive logistic model (GAM) is used identify significant associations between *S100a9* coexpression and the number of TF binding sites present within the 2 KB region upstream of the transcription start site (or other genomic regions). In GAM models, the probability of *S100a9* coexpression is modeled (vertical axis) as a function of two variables x_1 and x_2, where x_1 is the length of unmasked sequence scanned for a given gene and x_2 is the number of TF binding sites identified in the upstream region. GAM models were fit for each of 1209 TF binding sites, and a significant association between *S100a9* coexpression and binding site occurrence was evaluated based upon significance of the coefficient β_2.
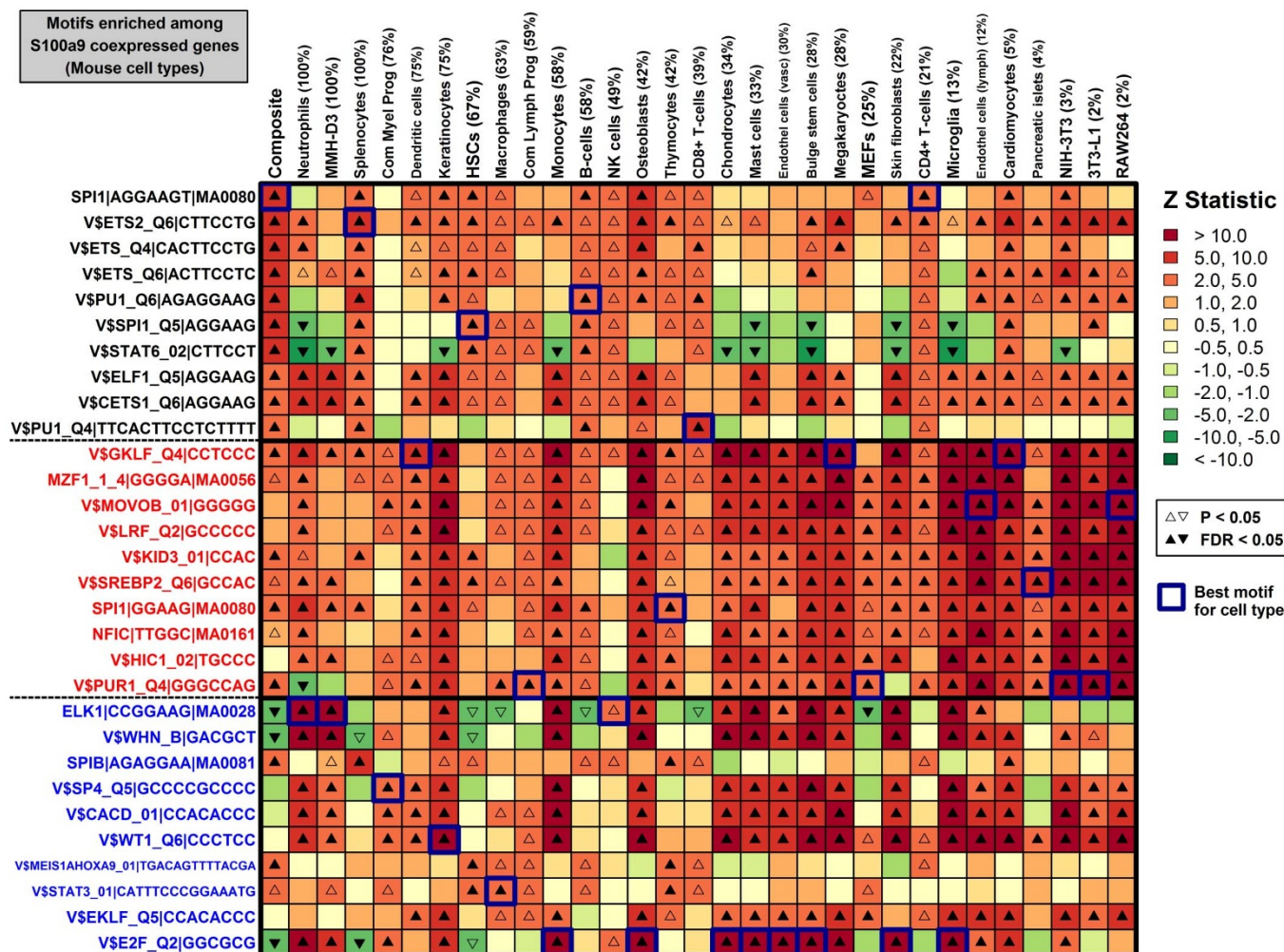
Figure 3 | Top-ranked transcription factor motifs that predict *S100a9* coexpression (30 mouse cell types). Top-ranked motifs are listed in the left margin and were selected based upon three criteria. First, we identified those motifs for which an increased number of occurrences significantly increased the probability of *S100a9* coexpression in the composite network (i.e., lowest p-values in composite with $Z > 0$; left margin labels with black font). Second, we identified those motifs for which an increased number of occurrences significantly increased the probability of *S100a9* coexpression across the largest total number of cell types (i.e., largest number of up-triangles per row; left margin labels with red font). Third, for the 10 cell types that most consistently expressed *S100a9* above background in microarray samples (i.e., neutrophils, …, monocytes), we identified the single motif most significantly associated with *S100a9* coexpression (i.e., lowest p-value for each cell type with $Z > 0$; left margin labels with blue font). Positive Z statistics (red heatmap colors) indicate that *increased* motif occurrence within 2 KB upstream regions was associated with increased probability of *S100a9* coexpression. Negative Z statistics (green heatmap colors) indicate that *decreased* motif occurrence 2 KB upstream was associated with increased probability of *S100a9* coexpression. For each cell type (columns), the percentage shown in parentheses is the fraction of microarray samples for which *S100a9* was expressed above background.
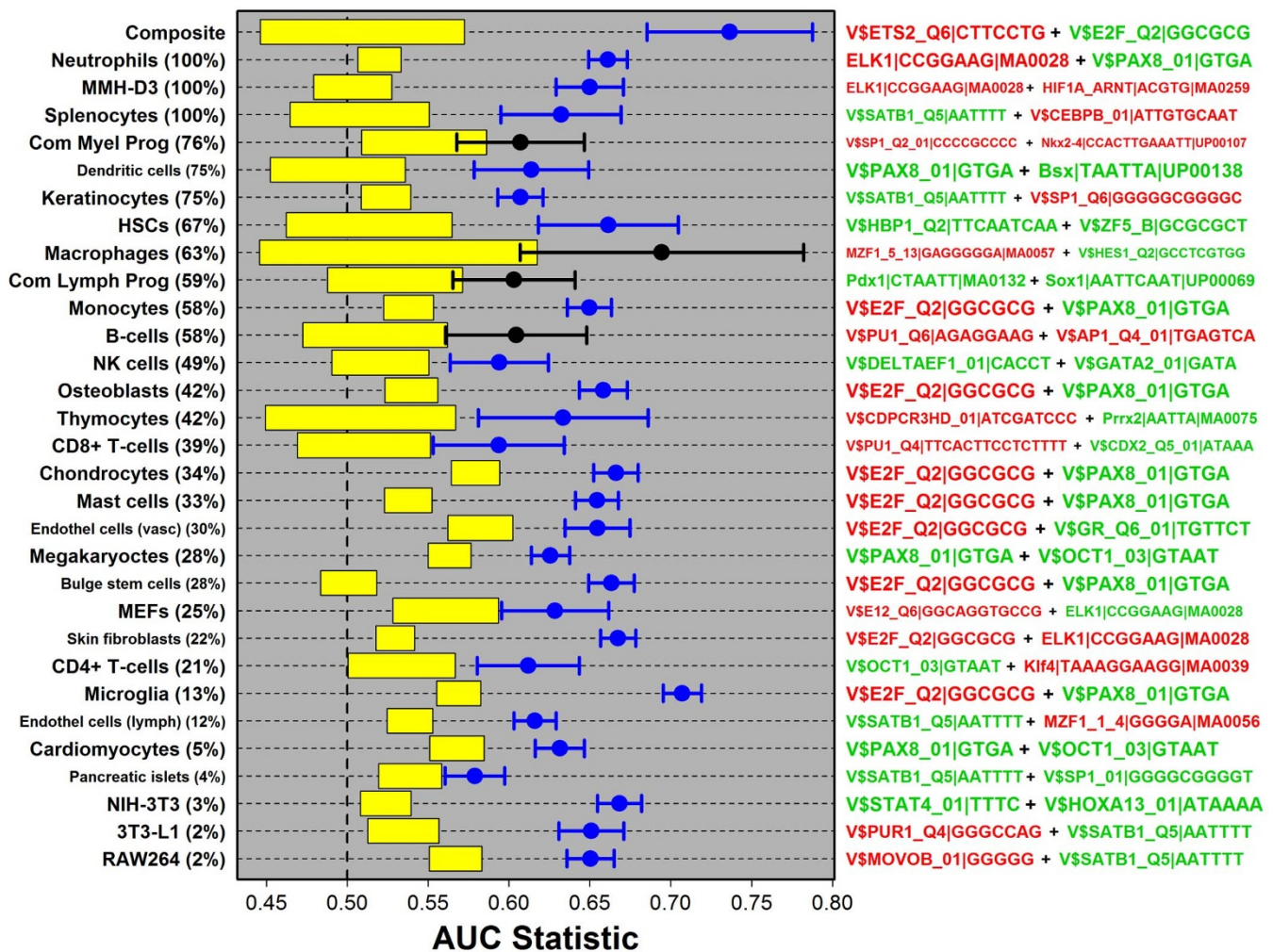
predictive of *S100A9* coexpression in the best-performing bivariate model (e.g., SPI1/PU.1, ETS1, ETS2 and ELK1). However, for approximately half of the 32 cell types, bivariate models related increased frequency of a GC-rich motif to increased probability of *S100A9* coexpression (Figure 7).

A region 930 - 1030 BP upstream of the human *S100A9* gene contains the highest concentration of TF binding sites that predict *S100A9* coexpression. Using results from our *in silico* screen of transcription factor binding sites in 32 human cell types, we scanned the 2KB region upstream of *S100A9* to identify loci containing a high density of motifs enriched among *S100A9* coexpressed genes. For most cell types, the first 500 BP upstream contained abundant matches to significantly enriched motifs (Figure S13). Surprisingly, however, the region 930–1030 BP upstream of *S100A9* contained the greatest concentration of binding sites for high-scoring motifs (hg19, chr1, 153329299 – 153329399; Figure S13), with matches to 43 motifs

significantly associated with *S100A9* coexpression in the composite network (P < 0.05 with $Z > 0$; Figure S14). More than half of these motifs (24 of 43) matched a 19 base pair GC-rich sequence within this region (5-GCCGTGGGGGCGGGCAGGA-3; hg19, chr1, 153329306–153329324).

Amplification of a SPI/PU.1-S100a9/A9 inflammatory axis in aged tissues and a model for S100a9/A9-mediated chronic inflammation with aging. Our findings, based upon an integration of transcriptome data from a diversity of mammalian cell types, implicate the ETS transcription factor SPI1/PU.1 as a driver of *S100a9/A9* expression. We investigated whether accumulation of S100a9 with aging in CB6F1 mice was associated with elevation of SPI1/PU.1 (Figure 8). In skin, there was no significant elevation of SPI1/PU.1 mRNA with aging (Figure 8A), although histochemical analysis of old skin identified cells staining positive for both SPI1/PU.1 and S100a9 (Figure S15). For other tissues (liver, kidney and lung), RT-PCR

Figure 4 | Bivariate motif combinations that best predict *S100a9* coexpression in 30 mouse cell types. For each cell type, we identified a set of TF motifs for which the number of binding sites in the region 2 KB upstream of transcription start sites was a significant predictor of *S100a9* coexpression ($P < 0.05$). Among these motifs, all possible pairwise combinations were evaluated as predictors of *S100a9* coexpression within logistic regression models. The right margin lists the best bivariate model identified for each cell type (Akaike information criterion). For motifs shown in red font, *increased* motif occurrence was associated with increased probability of *S100a9* coexpression ($Z > 0$), while conversely, for motifs shown in green font, *decreased* motif occurrence in was associated with increased probability of *S100a9* coexpression ($Z < 0$). 10000 cross-validation simulations were performed to assess the ability of each model to predict *S100a9* coexpression. The average AUC among all simulations is plotted in the figure (filled circles), with error bars spanning $\pm$ one standard deviation. Yellow boxes for each cell type outline the range of AUC statistics obtained for a univariate null model in which unmasked sequence length was the only predictor variable (mean AUC $\pm$ 1 standard deviation). Blue symbols represent cases for which AUC statistic distributions for the null and full models do not overlap, indicating that the combined frequencies of the two motifs yielded a sensitive and specific model for prediction of *S100a9* coexpression.

analyses revealed significant elevation of SPI1/PU.1 mRNA with aging in males (Figures 8B–8D), in correspondence with tissue-specific trends observed for *S100a9* mRNA (Figure S4).

Further histochemical analysis of lung tissue demonstrated increased SPI1/PU.1 protein in older mice and concordance between SPI1/PU.1 and S100a9 staining patterns, with increased SPI1/PU.1 in nuclei and increased S100a9 in cytoplasmic regions (Figure 8E). We hypothesized that the SPI1/PU.1+S100a9+ cells were predominantly leukocyte foci; however, although we could detect macrophages (F4/80), neutrophils and T-cells (CD3) in old lung, stains for these cell types showed only minimal overlap with the *S100a9* distribution (Figures S16, S17 and S18). This suggested additional sources for *S100a9* in old lung, potentially including senescent epithelial cells or other inflammatory cell types (e.g., B-cells or NK-cells). Our findings, taken together, lend support to a model for the development of chronic inflammation and leukocyte cluster formation during aging[5–10], in which inflammatory reactions are

sustained by continued over-production of *S100a9*, which is driven by the action of SPI1/PU.1 and other TFs in resident and/or infiltrating cell types (Figure 8F).

Discussion

The S100A8-S100A9 heterocomplex (calgranulin) was recently identified as a factor contributing to amyloid plaque accumulation and poor cognitive performance in an Alzheimer's mouse model[21]. This finding has called attention to the significance of S100A8-S100A9 in cognitive aging, but a broader role for S100A8-S100A9 in intrinsic aging has not been established. In this study, we have shown that shifts in the abundance of *S100A8* and *S100A9* are a normal feature of intrinsic aging in mammalian tissues. In mice, increased expression of *S100a8/S100a9* mRNA with aging occurs in multiple tissues (e.g., skin, lung, liver), while in humans, there is prominent elevation of *S100A8/S100A9* expression in the central nervous system (Figure 1). These trends are likely to have consequences for development of
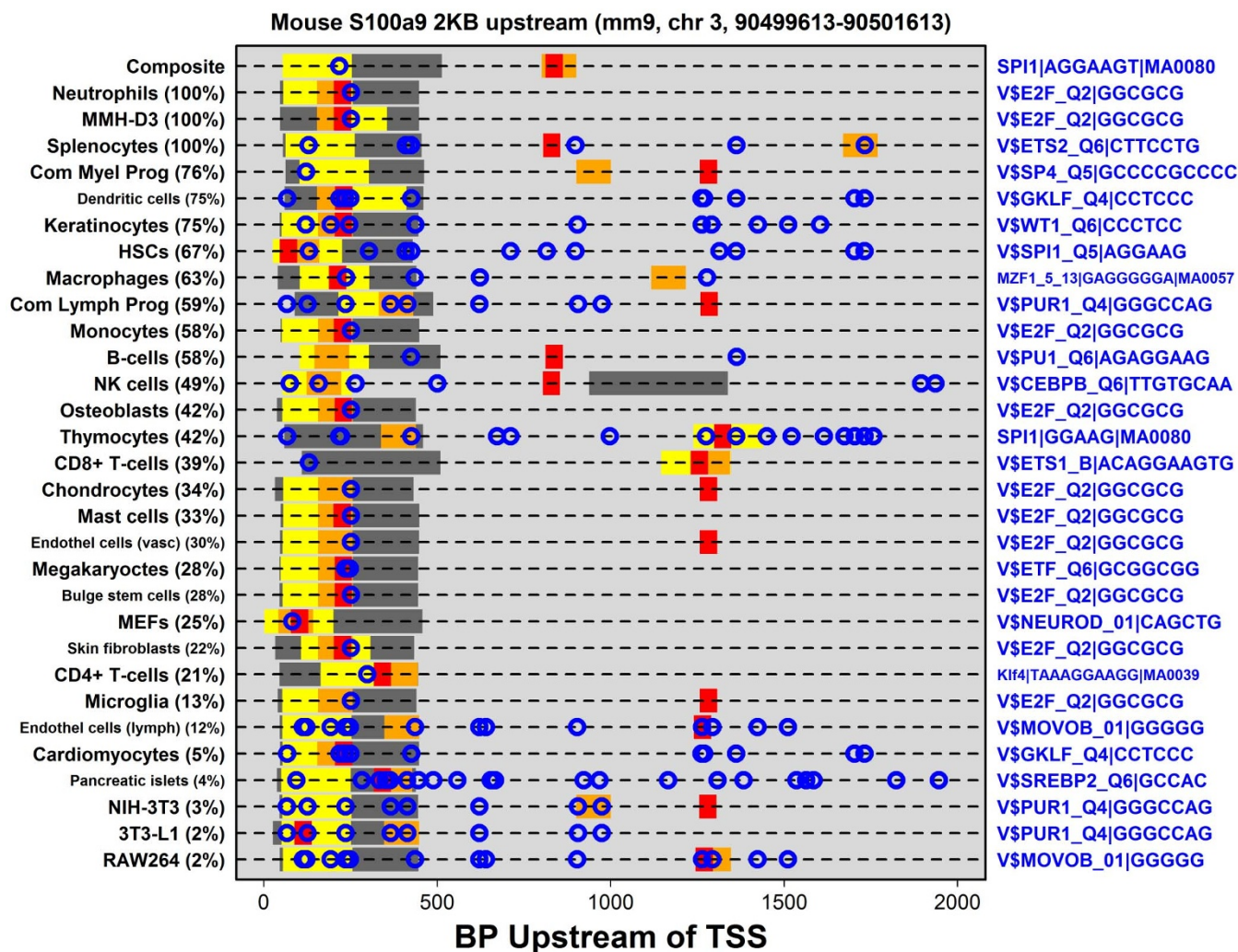
Figure 5 | The region 150–250 BP upstream of the mouse *S100a9* gene (mm9, chr3, 90499762–90499862) contains the highest concentration of TF binding sites that predict *S100a9* coexpression. For each cell type, the 2 KB region upstream of the *S100a9* transcription start site was scanned for matches to the TF binding sites within our dictionary of 1209 motifs. If a match was identified, the matching region was assigned a cell-type-specific score (proportional to the *Z* statistic calculated for that motif and cell type), which quantified the degree to which increased motif occurrence increased the probability of *S100a9* coexpression. If more than one motif matched at a given position, the highest score was assigned. A sliding window analysis was then used to identify regions with greatest concentration of high-scoring base pairs (dark grey = best 400 BP window; yellow = best 200 BP window; orange = best 100 BP window; red = best 50 BP window). The right margin (blue symbols) lists the individual motif, with at least one match 2 KB upstream, for which increased motif occurrence was most significantly associated with *S100a9* coexpression (i.e., lowest p-value with $Z > 0$).

chronic inflammation and formation of lymphocyte foci during aging[13]. It was therefore of interest to identify transcription factors that may control *S100a9* expression, since targeting such factors could provide a basis for effective therapies to treat persistent inflammation, providing a long-term strategy to encourage healthy aging[1–3]. Laboratory investigations of *S100a9* transcriptional regulatory mechanisms, however, have been limited to only a few cell types, and large-scale experimental study of such mechanisms across many cell types is not yet feasible. We therefore developed an integrative transcriptomics strategy to systematically identify candidate *cis*-regulatory mechanisms driving *S100a9* expression in 62 mouse and human cell types (Figure 2). Our approach provides an unsupervised and scalable strategy for identifying transcription factors associated with *S100a9*, and more broadly, with the cell type-specific coexpression networks in which *S100a9* is embedded (Figure 2C).

In previous work, STAT3[28,29], NF-κB[30], C/EBPα and C/EBPβ[31,32], HIF-1α[33], GLI1[34] and SPI1/PU.1[35] have been identified as activators of *S100a9* expression, while BRCA1[36], AP-1[37], SATB1[38] and Arnt/ HIF-1β[39] have been identified as *S100a9* repressors. In addition, multiple lines of evidence support the chromatin-associated enzyme

PARP-1 as an activator of *S100a9* transcription[41,44]. For each of these transcription factors and DNA-associated proteins, it was possible to identify at least one cell type in which an associated motif was significantly enriched among *S100a9* coexpressed genes, and at least one cell type in which an associated motif was significantly underrepresented among *S100a9* coexpressed genes (see Figure S19 summary). We thus observed opposite though nonetheless significant associations between motif abundance and *S100a9* coexpression, depending upon the cell type evaluated. Computational identification of transcriptional regulatory mechanisms is challenging, and for any one cell type, *in silico* methods are not expected to yield a 100% true positive rate and a 0% false positive rate. Nevertheless, detection of consistent trends across cell types, based upon distinct sets of data analyzed independently, serves to highlight robust associations and bolsters confidence in their biological significance. Along these lines, we found that motifs associated with SPI1/PU.1 and SATB1 were repeatedly associated with *S100a9* coexpression in mouse and human cell types (Figures 3, 4, 6, 7, S9 and S12). Our findings, therefore, lend the greatest support to SPI1/PU.1 and SATB1 as potential regulators of *S100a9* expression, since these factors received
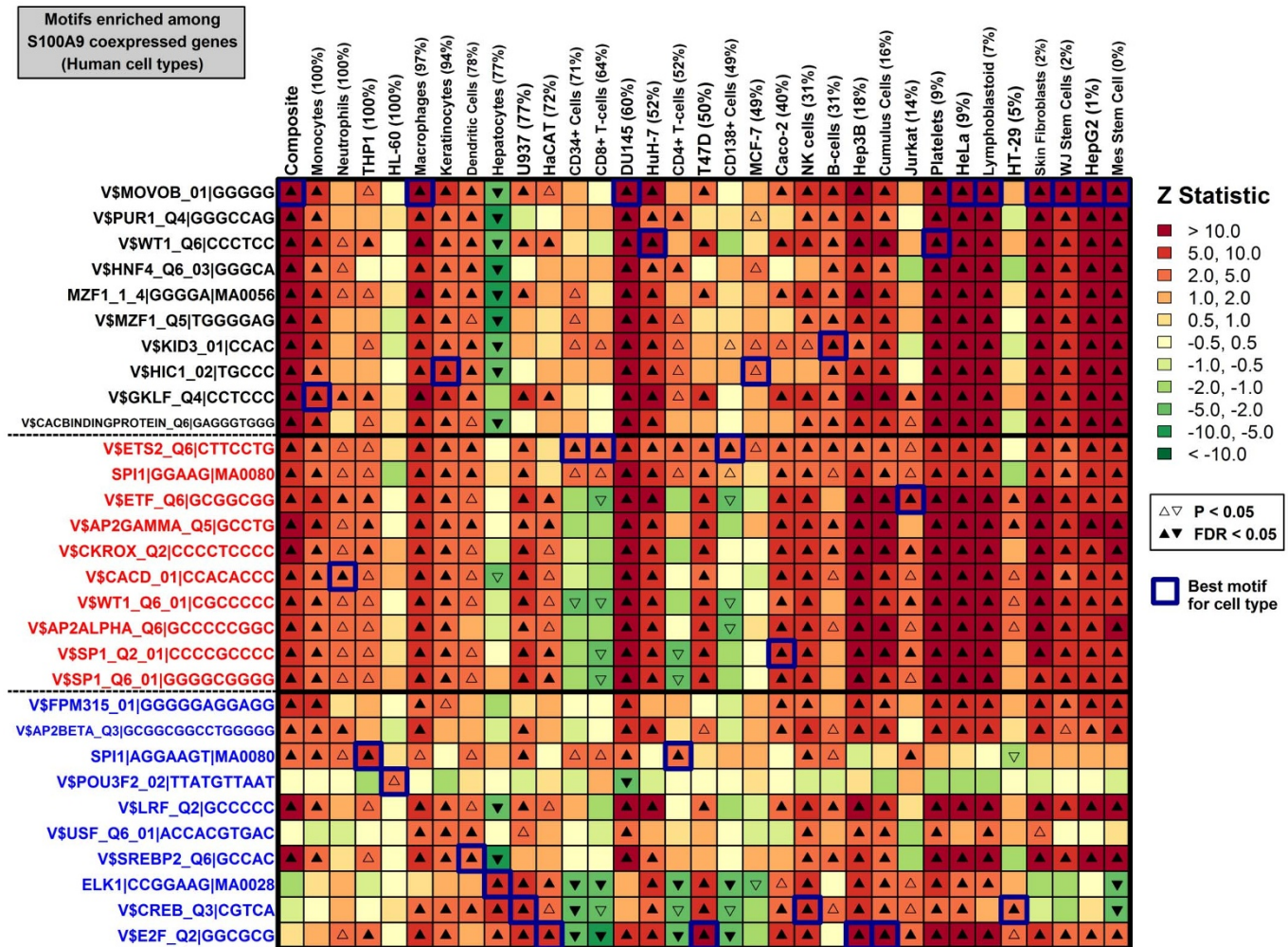
Figure 6 | Top-ranked transcription factor motifs that predict *S100A9* coexpression (32 human cell types). Motifs listed in the left margin were selected based upon the three criteria described in the Figure 3 legend. Positive *Z* statistics (red heatmap colors) indicate that *increased* motif occurrence within 2 KB upstream regions was associated with increased probability of *S100a9* coexpression. Negative *Z* statistics (green heatmap colors) indicate that *decreased* motif occurrence 2 KB upstream was associated with increased probability of *S100a9* coexpression. For each cell type (columns), the percentage shown in parentheses is the fraction of microarray samples for which *S100a9* was expressed above background.

ample support in our analyses, while both factors have been implicated in *S100a9* regulation by prior experimental work as well[35,38].

SPI1/PU.1 is an ETS family transcription factor that, like other ETS factors, exhibits affinity for sequences featuring a GGA(A/T) core motif[43]. In our study, motifs containing this element were frequently overrepresented among *S100a9* coexpressed genes, indicative of an association between ETS family transcription factors and *S100a9* coexpression, and overall, the strongest trends were associated with motifs recognized by SPI1/PU.1 (see Figure S20 summary). The region upstream of the mouse and human *S100a9* gene contains multiple SPI1/PU.1 motifs, although it is unknown whether SPI1/PU.1 directly interacts with these sites; however, one study reported 10-fold increased expression of *S100a9* following retroviral replacement of SPI1/PU.1 in mouse myeloid cells[35]. Basal expression of SPI1/PU.1 is highest in lymphoid cells, but its expression can also be induced by cytokines such as IFNγ[45]. Within aging tissues, the role of SPI1/PU.1in driving increased *S100a9* expression would most likely occur in the context of local inflammatory reactions, such as those involving microglia in the central nervous system[10], or perivascular formation of leukocyte clusters consisting of infiltrating monocytes, B-cells, T-cells, dendritic cells and macrophages[4,5]. Formation of such lymphoid structures in vascular tissues provides a platform for subsequent expansion of inflammatory cell populations, which is

likely to at least partly account for elevation of *S100a8* and *S100a9* with aging (Figures 1 and S4). Within these contexts, production of S100a9 by resident and/or infiltrating myeloid cells, enforced by SPI1/PU.1, may serve as a feed-forward process that both activates and is responsive to inflammatory signals, ultimately providing a mechanism by which inflammatory reactions become self-sustaining and chronic (Figure 8F). Nevertheless, we expect that the contribution of this mechanism to inflammatory processes will vary among tissues, or potentially, may differ between the sexes[4]. For instance, in kidney and lung, our data show that expression of the gene encoding SPI1/PU.1 is more strongly increased in males as compared to females (Figures 8C and 8D).

SATB1 (special AT-rich sequence binding protein 1) is a chromatin organizing factor that controls gene expression by packaging chromatin into dense loops and by facilitating the aggregation of chromatin remodeling enzymes in specific domains[46]. Absence of SATB1 binding sites (AATTTT) was a strong predictor of *S100a9* coexpression for nearly every cell type we considered (see Figure S21 summary). This was surprising, since initially, we expected that *S100a9* coexpression would be closely associated with increased abundance of certain motifs, rather than their absence. However, absence of SATB1 binding sites in regions surrounding *S100a9* coexpressed genes may allow for an expanded chromatin structure that
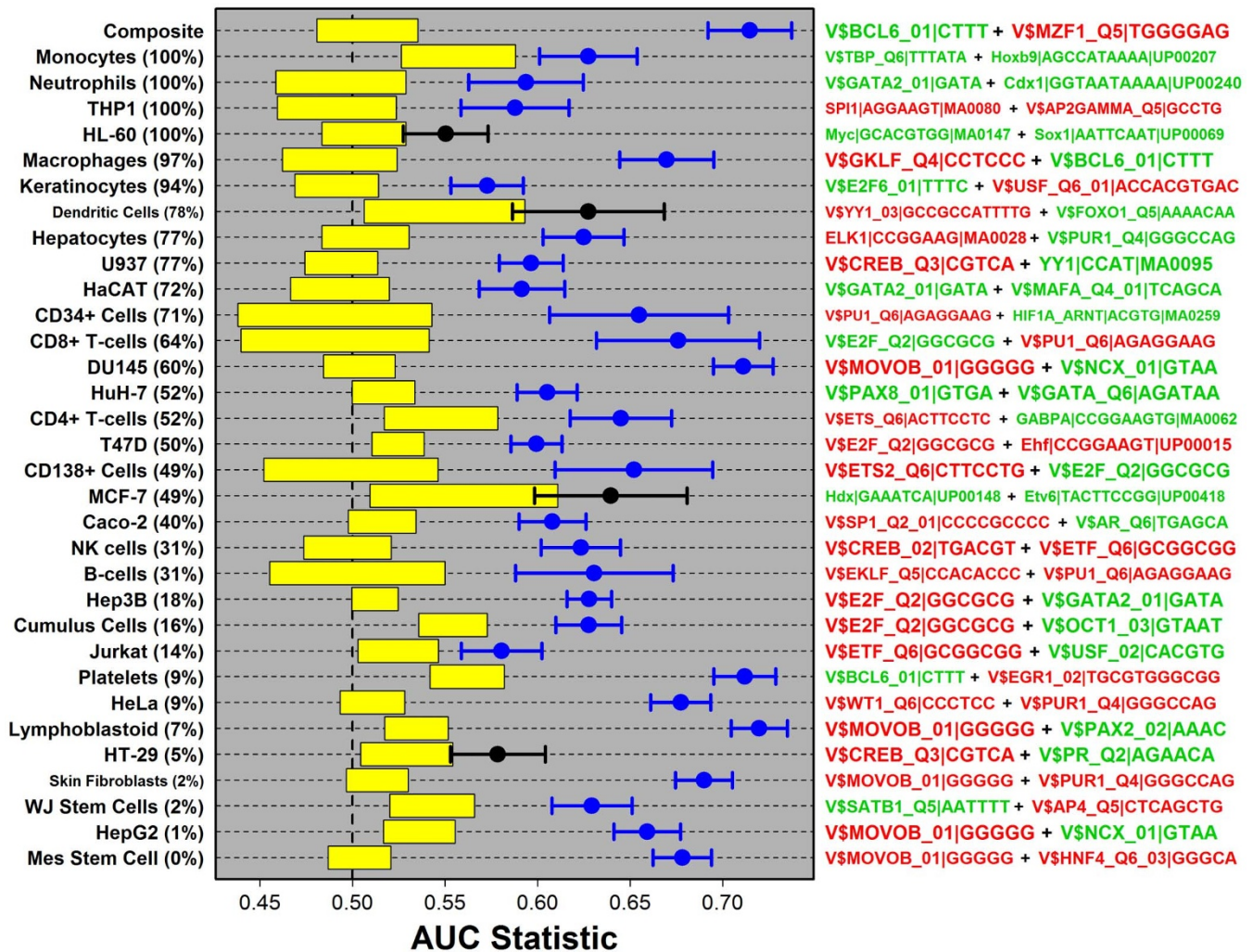
Figure 7 | Bivariate motif combinations that best predict *S100A9* coexpression in 32 human cell types. For each cell type, the right margin lists the bivariate motif model that best predicted *S100A9* coexpression for each cell type (AIC criterion). Bivariate models were chosen and evaluated as described in the Figure 4 legend.

increases accessibility of proximal *cis* regulatory elements[38]. In mouse keratinocytes, SATB1 binding occurs near the epidermal differentiation complex containing the *S100a9* gene, leading to a compressed chromatin conformation[38]. This action is proposed to repress gene expression, since SATB1-null mice exhibit greater than 5-fold elevation of *S100a8* and *S100a9*[38]. In our study, SATB1 motifs were underrepresented in regions proximal to *S100a9* coexpressed genes. One possible explanation for this result is that an expanded chromatin structure, facilitated by a deficiency of SATB1 binding sites, is instrumental for *cis* regulation of *S100a9* and other co-regulated genes. Such indirect regulation of *S100a9* would represent an elegant instance in which a pro-inflammatory transcript is controlled by the balance of two sequence-mediated processes – the higher order control of chromatin packaging by SATB1 – and secondly, by localized interactions between *cis* regulatory elements and their activating transcription factors (e.g., SPI1/PU.1).

Transcriptional regulation can depend upon the cooperative interactions among DNA-binding proteins recruited to a given locus[47], and such interactions are especially common among the ETS family transcription factors[48]. Since an abundance of motifs was significantly associated with *S100a9* coexpression in most cell types, we considered whether such motifs were co-localized in compact sequence regions upstream of the mouse or human *S100a9* gene. In mouse, the region 150–250 BP upstream of *S100a9* contained a high density of motifs significantly enriched among *S100a9* coexpressed genes (Figure 5). Within this region, we have highlighted a 13 BP element with overlapping binding sites for SPI1/PU.1 and STAT3 (Figure S10). The forward strand (5-ATGGAAGGGAAGT-3) features multiple binding sites for SPI1/PU.1 (i.e., AGGAAGT and GGAAG), while the reverse strand (5-ACTTCCCTTCCAT-3) features multiple STAT3 binding sites (i.e., TTCC) (Figure S10). It is interesting to speculate that cooperative interactions between STAT3 and SPI1/PU.1 may occur at this locus, particularly since these factors belong to the same transcriptional regulatory network, in which transcription of the gene encoding SPI1/PU.1 is induced by STAT3[49,50]. With regard to the human *S100A9* gene, regulatory elements have previously been identified less than 600 BP upstream from the transcription start site[30,32,51]. Surprisingly, however, we identified a region 930–1030 BP upstream as most strongly enriched with matches to those motifs overrepresented among *S100A9* coexpressed genes (Figure S13). A key feature of this region was a 19 BP GC-rich element including motifs recognized by OVOL2/MOVOB, WT1 and SP1 (Figure S14). This element was also located 12 BP from a SPI1/PU.1 binding site, raising the possibility of interactions between SPI1/PU.1 and a partner factor with affinity for GC-rich motifs[52].

The importance of the S100a8-S100a9 heterocomplex as a mediator of acute inflammatory reactions is well established[11–13], but the role of S100a8–S100a9 in low grade chronic inflammation during
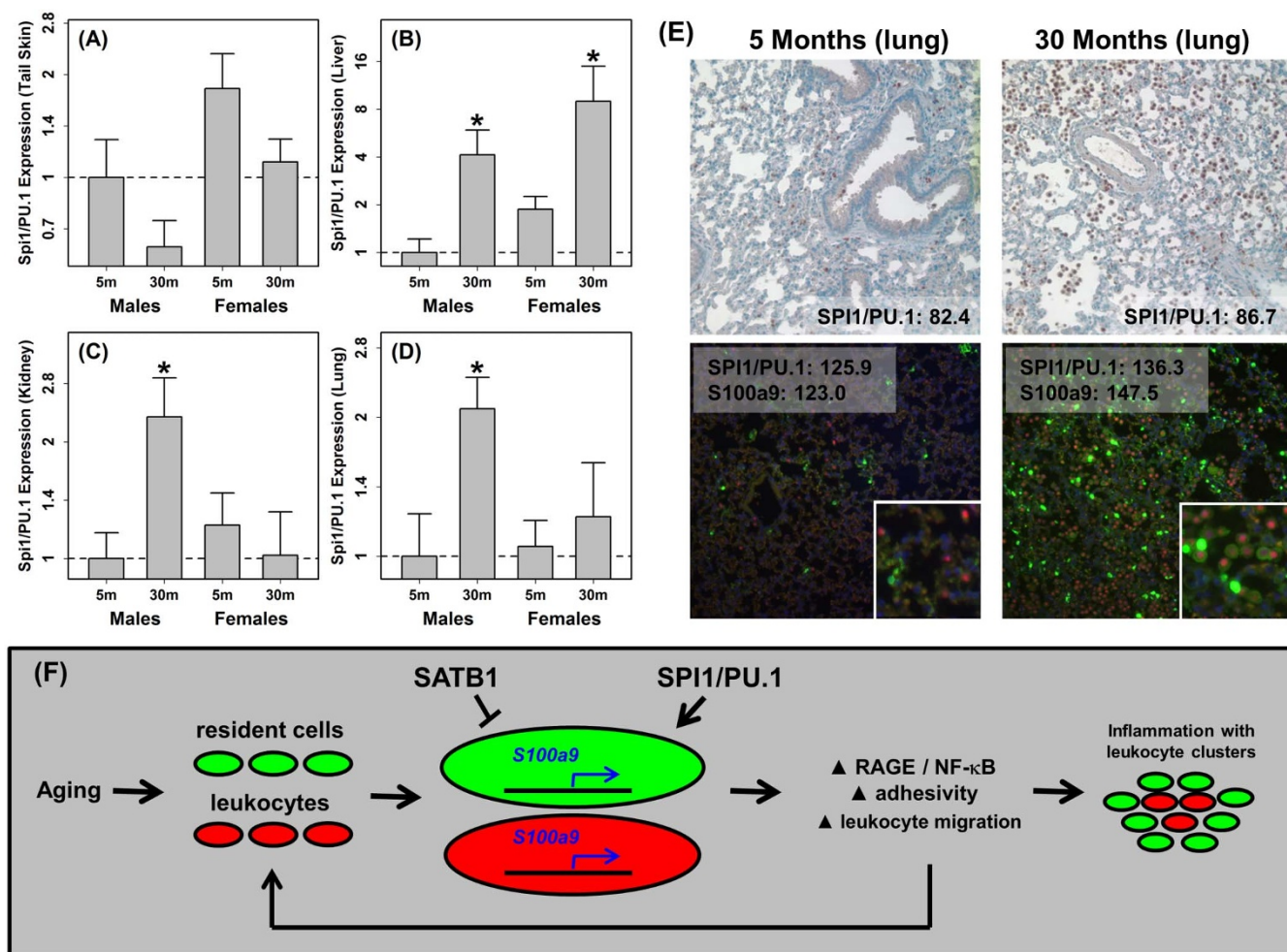
Figure 8 | Aging increases mRNA and protein levels of *SPI1/PU.1* with an overlapping distribution of *SPI1/PU.1* and *S100a9* in older tissues (CB6F1 mice). RT-PCR was used to evaluate the expression of *SPI1/PU.1* mRNA in (A) tail skin, (B) liver (C) kidney and (D) lung of CB6F1 mice scarified at 5 or 30 months of age. Expression was normalized to 18S ribosomal RNA (*Rn18s*). An asterisk symbol denotes a significant difference between young and old mice of the same sex (P < 0.05; two-tailed t-test). In part (E), immunostaining for *SPI1/PU.1* showed increased abundance of SPI1/PU.1 in older tissues (top panels), with increased nuclear *SPI1/PU.1* and elevation of *S100a9* in cytoplasmic regions (bottom panels; red = SPI1/PU.1; green = S100a9; blue = DAPI). (F) Proposed model by which over-production of *S100a9* engenders a pro-inflammatory microenvironment, with sustained activation of the RAGE and NF-κB, leading to recruitment and migration of leukocytes into local tissues. This in turn leads to further infiltration by inflammatory cell types that actively transcribe *S100a9*, driving a self-reinforcing cycle that sustains inflammation and lymphoid aggregation with aging.

aging has not been investigated. Many genes have been reported to increase or decrease expression with aging. However, when data are compared among multiple laboratories, few genes exhibit robust trends similar to those we have described for *S100a8* and *S100a9*[4,53,54]. This paper, therefore, supports a role for S100a8–S100a9 in the process of mammalian aging, and emphasizes the need for further studies of S100a8–S100a9 as a mediator of age-associated inflammation. Within this context, we have developed a strategy for the large scale and cell type-specific identification of *cis* regulatory mechanisms. At present, many transcription factor target gene annotations are generic, with a certain factor *X* recognized as regulating a target gene *Y*, but without any reference to the cell type in which this interaction takes place. Such TF-target relationships, however, may differ among the heterogeneous collections of cell types that interact in most mammalian disease processes[55]. The approach developed here illustrates one strategy for generating cell type-specific hypotheses of mammalian gene regulation. We anticipate that this strategy will become more powerful in future work, as we build a more comprehensive knowledge of binding affinities for mammalian transcription factors, leading to increasingly refined motif dictionaries that can be used to interrogate coexpression networks[56–58]. This can

facilitate the construction of multicellular gene regulatory networks that, by linking inter-cellular interactions with the intracellular mechanisms governing transcription, may better capture the dynamic processes that drive aging and disease in mammals.

Methods

Meta-analysis of age-associated gene expression patterns. The effects of aging on gene expression in mice were evaluated using the Affymetrix Mouse Genome 430 2.0 array, while the effects of aging in humans were evaluated using the Affymetrix Human Genome U133 Plus 2.0 array. Expression analyses of CB6F1 mouse tail skin and human sun-protected skin biopsies (Figure S1) were carried out by our laboratory as described in an earlier publication[4], and raw data are available from Gene Expression Omnibus (GSE35322 and GSE13355). To evaluate aging effects in other tissues (Figure 1), we utilized Affymetrix datasets aggregated and described in an earlier publication[4], although we have updated this collection to include recently released datasets, which here allowed us to also evaluate the effects of aging on gene expression in CB6F1 liver (GSE20426), C57BL/6 aorta (GSE32937), C57BL/6 HSCs

(GSE27686), C57BL/6 B1 cells (GSE28887) and human HSCs (GSE32719 and GSE32725). Effects of aging on gene expression were evaluated using two-sample comparisons (old vs. young) with empirical Bayes methods[59], or if samples were available across a spectrum of individuals varying continuously in age, effects of aging were evaluated using robust regression[4]. All calculations were performed using log2-scaled expression scores generated by Robust Multichip Average (RMA) normalization[60].

Motif dictionary. We have described an *in silico* procedure for screening sets of *S100a9* coexpressed genes in order to identify motifs over- or under-represented among genomic sequences associated with such genes (e.g., 2 KB upstream regions; see Figure 2). An important first step for implementing this procedure was to generate a high-quality dictionary of motifs for mouse and human TFs and other DNA-binding proteins or complexes. For this purpose, an initial set of 2105 probability matrices representing binding preferences of known DNA-binding proteins or complexes was obtained from the Jaspar[56], UniPROBE[57] and TRANSFAC[58] databases, including 145 matrices from the Jaspar CORE vertebrate collection, 295 matrices from UniPROBE, and 1665 matrices from TRANSFAC. Of the 1665 TRANSFAC matrices, we excluded 846 that were not associated with vertebrate species or were already included among those obtained from UniPROBE. This yielded a total of 1259 matrices from all three databases. Each matrix was trimmed to remove columns with low information content at the flanks. Starting at each flank, columns were removed until two consecutive columns with information content greater than 0.25 was encountered. For 9 matrices, this trimming procedure engendered a low information content matrix with fewer than five columns, and such matrices were excluded from further analysis. For the remaining 1250 matrices, a redundancy search was performed to identify very similar matrix models. We identified 41 redundant matrices that shared the same consensus sequence with another matrix, where the difference in base-specific probability estimates differed by less than 0.20 on average. These 41 matrices were removed from consideration, leaving a final set of 1209 matrix models with limited redundancy. These 1209 matrix models were included in our screening procedures. Motif labels shown in our results (e.g., Figures 3 and 6) include the gene symbol for the associated TF or DNA-binding protein, an identifier corresponding to the matrix ID used in the source database (JASPAR, UniPROBE or TRANSFAC), and the representative consensus sequence. JASPAR IDs start with "MA", TRANSFAC IDs start with "V$", and UniPROBE IDs start with "UP".

Genome scans to identify motif matches. Genomic sequences associated with mouse and human genes were scanned for matches to the 1209 matrix models included within our motif dictionary. Sequences were obtained from Bioconductor packages (BSgenome.Mmusculus.UCSC.mm9 and BSgenome.Hsapiens.UCSC.hg19), and coordinates for each gene were defined based upon refGene files from UCSC (mm9 and hg19). Repetitive DNA, assembly gaps and coding regions were masked for all genome scans. In some cases, we also masked less conserved portions of genomic sequences (i.e., Figures S9A and S12A). In such cases, sequences were masked using base-specific PhastCons scores, which range from 0 to 1 for each base, and can be interpreted as the probability that a given base is within a conserved element[61]. For mice, PhastCons scores were derived from the alignment of 29 vertebrate genomes with the mouse genome, while for humans, scores were generated from the alignment of 45 vertebrate genomes with the human genome. For a given sequence, the median PhastCons score across all bases was calculated, and if this median was less than 0.70, we masked out sequences with a PhastCons score less than the median. If the median was greater than 0.70, we masked any sequence with a PhastCons score less than 0.70. For all genome scans, motif matches were identified based upon position weight matrices (PWM), with a match identified only for those loci in which the PWM matching score was greater than 80% of the maximum matching score for that PWM matrix[4,62]. Background frequencies used to construct position weight matrices varied slightly (0.20–0.30), depending upon the genome region scanned, and were calculated by pooling of frequencies across all 2 KB upstream regions, all intronic regions, or all intergenic regions in the mouse or human genomes, respectively[4].

Identification of *S100a9/A9* coexpressed genes in 30 mouse and 32 human cell types. We generated cell type-specific sets of *S100a9/A9* coexpressed genes (Figures 2A and 2B), corresponding to local coexpression networks immediately surrounding *S100a9/A9* for each of 30 mouse and 32 human cell types (Figure 2C). We first assembled 43 datasets, where each dataset contained microarray samples corresponding to a specific cell type (Affymetrix Mouse Genome 430 2.0 platform). Likewise, to identify genes coexpressed with *S100a9* in humans, we assembled 44 datasets, where each dataset contained microarray samples generated from a specific cell type (Affymetrix Human Genome U133 Plus 2.0 platform). The mouse datasets included a combined total of 3555 microarray samples, with a median of 41 samples per cell type (minimum = 9, skin fibroblasts; maximum = 393, macrophages). Human datasets included a combined total of 8572 microarray samples, with a median of 111 samples per cell type (minimum = 10, Raji cells; maximum = 799, CD138+ cells). For each dataset, we calculated the percentage of probe sets for which *S100a9/A9* was expressed significantly above background (mouse probe set id: 1448756_at; human probe set id: 203535_at), based upon presence/absence calls generated by the MAS 5.0 algorithm (Wilcoxon signed rank test)[63]. We identified 13 mouse datasets for which *S100a9* (1448756_at) was not expressed significantly above background on any microarray, as well as 9 human datasets for which *S100a9*

(203535_at) was not expressed significantly above background on any microarray. These datasets were removed and thus we base our analysis on datasets corresponding to 30 mouse cell types and 32 human cell types, where *S100a9/S100A9* was expressed significantly above background for at least one of the microarray samples within each dataset.

Mouse and human datasets for each cell type were normalized using RMA[60]. For each probe set, we subtracted the mean expression across all samples to yield scores reflecting deviation from average expression among all samples. These scores were then scaled to have the same standard deviation across samples within each dataset. We next calculated the Spearman correlation coefficient between scores for all probe sets and scores for the probe set associated with *S100a9* in mouse (1448756_at) or *S100A9* in humans (203535_at). Probe sets were ranked according to this correlation, and we selected a set of *S100a9/A9* coexpressed probe sets based upon the approach illustrated in Figure 2A. The foreground set of *S100a9/A9* coexpressed genes was defined as the non-redundant set of genes associated with those probe sets coexpressed with *S100a9/A9* for a given cell type. Since our ranking procedure included only positive correlation estimates, foreground gene sets are defined based upon the level of positive correlation with the *S100a9/A9* expression pattern ($r_s > 0$), excluding genes with a negative or inverse correlation with the *S100a9/A9* expression pattern ($r_s < 0$).

Statistical analyses required that we define a background gene set for each cell type, where this set contained genes for which there was little evidence of either a positive or negative correlation with the *S100a9/A9* expression pattern. To identify this background set of genes, we defined two gene sets (*A* and *B*), where *A* contained genes positively correlated with the *S100a9/A9* expression pattern and *B* contained genes negatively correlated with the *S100a9/A9* expression pattern. The background gene set was then defined as all genes in the genome excluding any gene within sets *A* and *B*. Set *A* was defined by applying the ranking procedure shown in Figure 2A to all probe sets with a positive correlation estimate (with probe sets ranked according to r_s^2 rather than r_s^7). Similarly, set *B* was defined by applying the ranking procedure shown in Figure 2A to all probe sets with a negative correlation estimate (with probe sets ranked according to r_s^2 rather than r_s^7). Ranking transcripts according to r_s^2 rather than r_s^7 had the effect of creating a less steep curve (Figure 2A) and thus expanding the set of genes included within sets *A* and *B*. This ensured that set *A* was larger than the foreground gene set and ensured that the background gene set did not include genes with even a modest level of association with the *S100a9/A9* expression pattern. This procedure also ensured that foreground and background gene sets would be distinct for each cell type.

Composite sets of *S100a9/A9* coexpressed genes were identified following the same ranking procedure shown in Figure 2A. However, probe sets were ranked according to their (weighted) average correlation with *S100a9/A9* across all 30 mouse cell types or all 32 human cell types. For any one probe set *x*, the Spearman correlation between *x* and *S100a9/A9* (1448756_at/203535_at) was independently calculated for all cell types. A weighted average of correlation estimates was then calculated, with greater weight assigned to correlations estimated from cell types for which more data was available (e.g., macrophages, *n* = 393 samples for mice), with less weight assigned to correlations estimated based upon cell types for which less data was available (e.g., skin fibroblasts, *n* = 9 samples for mice). The procedure was repeated for all probe sets, which were then ranked according to the weighted average correlation (similar to Figure 2A). Because average correlation estimates calculated from many cell types tended to generate a steep curve (Figure 2A), one further modification was that probe sets were ranked according to r_s^3 rather than r_s^7. This increased the size of the foreground gene set such that similar sizes were obtained for both composite and cell type-specific foreground gene sets.

Semiparametric generalized additive logistic models. Semiparametric generalized additive logistic models (GAM) were used to test for a significant association between *S100a9/A9* coexpression and counts reflecting the number of motifs associated with genes in foreground and background gene sets[64]. For each cell type and motif, a 0–1 response variable was generated to denote whether a gene belonged to the foreground gene set (coded 1) or background gene set (coded 0). Two predictor variables were included within each model, including log-transformed length of sequence scanned for a given gene (x_1) and the number of motif sites detected within the sequence (x_2) (Figure 2D). Length of sequence scanned varied among genes due to masking or inherent differences in the amount of intronic sequence or adjacent intergenic sequence. Including x_1 within models was thus necessary to estimate the effect of motif frequency on coexpression probability while controlling for variation in the length of sequence scanned for each gene[65]. For each GAM model, x_1 was included as a non-parametric term with cubic spline smoothing, while x_2 was included as a parametric term. Model fitting was performed using the backfitting procedure with the binomial family error distribution and logit link function[64]. Given these procedures, we assume that the logit of the probability of coexpression is linearly associated with x_2; however, we make no such assumptions involving x_1 but instead estimate this relationship non-parametrically using spline smoothing[64]. In preliminary work, we noted that the largest residuals for fitted models usually occurred for genes with extremely short or long sequences. To improve overall fit, therefore, models were fit after excluding sequences of extreme length (1–5% excluded depending upon the genome region scanned). The association between motif frequency and *S100a9/A9* coexpression probability was evaluated based upon the coefficient estimate associated with x_2. For each coefficient, a *Z* statistic was generated, with $Z > 0$ indicating that increased motif frequency was associated with increased probability of *S100a9/A9* coexpression, and $Z < 0$ indicating that decreased

motif frequency was associated with decreased *S100a9/A9* coexpression. Given the large number of genes included within the procedure, Z statistics were assumed to follow a normal distribution under the null hypothesis of no association between motif frequency and *S100a9/A9* coexpression[66]. P-values were thus calculated for each motif using Z statistics and the standard normal distribution. For each cell type, separate models were fit for all 1209 motifs included within our dictionary, leading us to expect that ~60 significant motifs would be identified by chance for a given cell type ($\alpha = 0.05$). In our results, therefore, we have also reported significance calls based upon a more stringent FDR threshold, which was calculated directly from raw p-values by applying the Benjamini-Hochberg correction[67].

We have highlighted motifs most significantly enriched among *S100a9* coexpressed genes in the composite network, motifs most frequently enriched among *S100a9* coexpressed genes across all mouse or all human cell types, or those motifs most significantly enriched among *S100a9* coexpressed genes in the ten cell types with highest *S100a9* expression (Figures 3, 6, S9 and S12). In some cases, top-ranked motifs shared similar features; for instance, in Figure 6, several motifs have high GC content. We therefore performed secondary analyses to determine which motif features best distinguish enriched motifs (P < 0.05 with Z > 0) from all other non-significant motifs (Figures S9D and S12D). For these analyses, we assigned a score to each motif that was proportional to how well it matched *k*-mers of length 2, 3, 4 or 5. This was done for each motif and each of the 690 possible non-redundant *k*-mers of length 2, 3, 4 or 5. Given column position *j* in a position probability matrix (PPM) ($j \in 1...n$), matrix values from rows corresponding to a given *k*-mer were selected, and the minimum of these values was determined. This minimum value was calculated for each of the PPM columns ($j = 1, …, n$) and the maximum of these minimum values was calculated (m_1). The procedure was repeated for the reverse complement of the *k*-mer (yielding m_2). The *k*-mer score with respect to the PPM was then assigned as the larger of these two maximum values (i.e., $\max(m_1, m_2)$). For each cell type, we evaluated all 690 *k*-mers to identify those with scores significantly greater among motifs enriched among *S100a9* coexpressed genes (P < 0.05 with Z > 0) relative to all other motifs included in our screen (two-sided T-test; see Figures S9D and S12D).

RT-PCR and immunohistochemistry. Tissue samples were obtained from young, middle-aged and old CB6F1 mice that had been fasted 6 – 8 hours prior to sacrifice[4]. Tissue used for subsequent RNA extraction was stored in RNAlater solution (Qiagen cat. no. 76106) and initially placed at 4°C for 24 hours and then transitioned to −20°C for long-term storage. Separate samples used for histochemisty were formalin-fixed and paraffin-embedded. For RNA work, samples were disrupted using a rotor-stator homogenizer and RNA extractions were performed using the Qiagen RNeasy Fibrous Tissue kit (Qiagen cat. no. 74704), with on-column DNAase digestion (Qiagen cat. no. 79254). Quantification of RNA was carried out using the NanoDrop spectrophotometer and quality of RNA was evaluated using the Agilent Bioanalyzer. RT-PCR was performed using commercially available pre-designed primer assays (Applied Biosystems; catalog numbers #Mm00488142_m1 (mouse *SPI1/PU.1*), #Mm00656925_m1 (mouse *S100a9*), and #Hs00610058_m1 (human *S100A9*)). For diaminobenzidine and double fluorescence staining of formalin-fixed paraffin-embedded tissue, we used goat anti-mouse S100A9 polyclonal antibody (0.5 µg/mL; R&D Systems, cat. no. AF2065), SPI1/PU.1 polyclonal antibody (1–2 µg/mL; Cell Signaling Technology, cat. no. 2266S), rat anti-mouse F4/80 (5 ug/ml; eBioscience, cat. no. 14–4801), rat anti-mouse GR1 (2 ug/ml; abcam, cat. no. ab2557), and rabbit anti-mouse/human CD3 (1 : 100 dilution; genetex, cat. no GTX16669).

Image quantification. Diaminobenzidine and fluorescent staining intensities were quantified to facilitate interpretation and objective comparison among immunohistochemical images (Figures 8E, S2, S5, S6, S7, S8, S15, S16, S17 and S18). As an initial step, all pixels from an image were categorized as background or non-background. For each pixel, we calculated the value $X = \min(R, G, B)$, where R is the red channel intensity on the RGB scale (0–255), G is the green channel color intensity on the RGB scale, and B is the blue intensity on the RGB scale. The distribution of X across pixels was generally bimodal, with background pixels belonging to either the lower (black) or upper (white) regions of the distribution. The expectation maximization (EM) algorithm was therefore used to estimate the threshold value of X distinguishing the background and non-background distributions. We utilized a standard implementation of the EM algorithm for mixture of univariate normal distributions (i.e., function "normalmixEM" from R package "mixtools"). After an appropriate threshold was determined for an image, the average red or green intensity of non-background pixels was calculated. The red intensity value for the ith non-background pixel was calculated as $RED_i = R_i – \max(B_i, G_i)$, while the green intensity value for the ith non-background pixel was calculated as $GREEN_i = G_i – \max(R_i, B_i)$. The average red intensity for an image was calculated as the average value of RED among all non-background pixels, while the average green intensity for an image was equal to the average value of $GREEN$ among all non-background pixels.

1. Tuppo, E. E. & Arias, H. R. The role of inflammation in Alzheimer's disease. *Int. J. Biochem. Cell Biol.* **37**, 289–305 (2005).
2. Weber, C. & Noels, H. Atherosclerosis: current pathogenesis and therapeutic options. *Nat. Med.* **17**, 1410–22 (2011).
3. Lepperdinger, G. Inflammation and mesenchymal stem cell aging. *Curr. Opin. Immunol.* **23**, 518–24 (2011).
4. Swindell, W. R. *et al.* Meta-profiles of gene expression during aging: limited similarities between mouse and human and an unexpectedly decreased inflammatory signature. *PLoS One* **7**, e33204 (2012).
5. Singh, P. *et al.* Lymphoid neogenesis and immune infiltration in aged liver. *Hepatology* **47**, 1680–90 (2008).
6. Boyle, A. J. *et al.* Cardiomyopathy of aging in the mammalian heart is characterized by myocardial hypertrophy, fibrosis and a predisposition towards cardiomyocyte apoptosis and autophagy. *Exp Gerontol.* **46**, 549–59 (2011).
7. Ishimoto, Y. *et al.* Age-dependent variation in the proportion and number of intestinal lymphocyte subsets, especially natural killer T cells, double-positive CD4 + CD8 + cells and B220 + T cells, in mice. *Immunology.* **113**, 371–7 (2004).
8. Lipman, R. D., Dallal, G. E. & Bronson, R. T. Lesion biomarkers of aging in B6C3F1 hybrid mice. *J. Gerontol. A. Biol. Sci. Med. Sci.* **54**, B466–77 (1999).
9. Ward, J. M., Goodman, D. G., Squire, R. A., Chu, K. C. & Linhart, M. S. Neoplastic and nonneoplastic lesions in aging (C57BL/6N x C3H/HeN)F1 (B6C3F1) mice. *J. Natl. Cancer Inst.* 1979, **63**, 849–54 (1979).
10. Akiyama, H., Ikeda, K., Katoh, M., McGeer, E. G. & McGeer, P. L. Expression of MRP14, 27E10, interferon-alpha and leukocyte common antigen by reactive microglia in postmortem human brain tissue. *J. Neuroimmunol.* **50**, 195–201 (1994).
11. Thorey, I. S. *et al.* The Ca2 + -binding proteins S100A8 and S100A9 are encoded by novel injury-regulated genes. *J. Biol. Chem.* **276**, 35818–25 (2001).
12. Zwadlo, G., Brüggen, J., Gerhards, G., Schlegel, R. & Sorg, C. Two calcium-binding proteins associated with specific stages of myeloid cell differentiation are expressed by subsets of macrophages in inflammatory tissues. *Clin. Exp. Immunol.* **72**, 510–5 (1988).
13. Gebhardt, C., Németh, J., Angel, P. & Hess, J. S100A8 and S100A9 in inflammation and cancer. *Biochem. Pharmacol.* **72**, 1622–31 (2006).
14. Hsu, K. *et al.* Anti-infective protective properties of S100 calgranulins. *Antiinflamm Antiallergy Agents Med. Chem.* **8**, 290–305 (2009).
15. Liang, S. C. *et al.* Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J. Exp. Med.* **203**, 2271–9 (2006).
16. Lackmann, M. *et al.* Identification of a chemotactic domain of the pro-inflammatory S100 protein CP-10. *J. Immunol.* **150**, 2981–91 (1993).
17. Benedyk, M. *et al.* HaCaT keratinocytes overexpressing the S100 proteins S100A8 and S100A9 show increased NADPH oxidase and NF-kappaB activities. *J. Invest. Dermatol.* **127**, 2001–11 (2007).
18. Hermani, A., De Servi, B., Medunjanin, S., Tessier, P. A. & Mayer, D. S100A8 and S100A9 activate MAP kinase and NF-kappaB signaling pathways and trigger translocation of RAGE in human prostate cancer cells. *Exp. Cell Res.* **312**, 184–97 (2006).
19. Gebhardt, C. *et al.* RAGE signaling sustains inflammation and promotes tumor development. *J. Exp. Med.* **205**, 275–85 (2008).
20. Shepherd, C. E. *et al.* Inflammatory S100A9 and S100A12 proteins in Alzheimer's disease. *Neurobiol. Aging* **27**, 1554–63 (2006).
21. Ha, T. Y. *et al.* S100a9 knockdown decreases the memory impairment and the neuropathology in Tg2576 mice, AD animal model. *PLoS One* **5**, e8840 (2010).
22. Chang, K. A., Kim, H. J. & Suh, Y. H. The role of s100a9 in the pathogenesis of Alzheimer's disease: the therapeutic effects of s100a9 knockdown or knockout. *Neurodegener. Dis.* **10**, 27–9 (2012).
23. Zhang, C., Liu, Y., Gilthorpe, J. & van der Maarel, J. R. MRP14 (S100A9) Protein Interacts with Alzheimer Beta-Amyloid Peptide and Induces Its Fibrillization. *PLoS One* **7**, e32953 (2012).
24. Teigelkamp, S. *et al.* Calcium-dependent complex assembly of the myeloic differentiation proteins MRP-8 and MRP-14. *J. Biol. Chem.* **266**, 13462–7 (1991).
25. Rosenberger, S., Thorey, I. S., Werner, S. & Boukamp, P. A novel regulator of telomerase. S100A8 mediates differentiation-dependent and calcium-induced inhibition of telomerase activity in the human epidermal keratinocyte line HaCaT. *J. Biol. Chem.* **282**, 6126–35 (2007).
26. Grimbaldeston, M. A., Geczy, C. L., Tedla, N., Finlay-Jones, J. J. & Hart, P. H. S100A8 induction in keratinocytes by ultraviolet A irradiation is dependent on reactive oxygen intermediates. *J. Invest. Dermatol.* **121**, 1168–74 (2003).
27. Lian, Z. *et al.* Genomic and proteomic analysis of the myeloid differentiation program. *Blood* **98**, 513–24 (2001).
28. Li, C., Zhang, F., Lin, M. & Liu, J. Induction of S100A9 gene expression by cytokine oncostatin M in breast cancer cells through the STAT3 signaling cascade. *Breast Cancer Res. Treat.* **87**, 123–34 (2004).
29. Cheng, P. *et al.* Inhibition of dendritic cell differentiation and accumulation of myeloid-derived suppressor cells in cancer is regulated by S100A9 protein. *J. Exp. Med.* **205**, 2235–49 (2008).
30. Németh, J. *et al.* S100A8 and S100A9 are novel nuclear factor kappa B target genes during malignant progression of murine and human liver carcinogenesis. *Hepatology* **50**, 1251–62 (2009).
31. Suryono *et al.* Norepinephrine stimulates calprotectin expression in human monocytic cells. *J Periodontal. Res.* **41**, 159–64 (2006).
32. Kuruto-Niwa, R., Nakamura, M., Takeishi, K. & Nozawa, R. Transcriptional regulation by C/EBP alpha and -beta in the expression of the gene for the MRP14 myeloid calcium binding protein. *Cell Struct. Funct.* **23**, 109–18 (1998).
33. Grebhardt, S., Veltkamp, C., Ströbel, P. & Mayer, D. Hypoxia and HIF-1 increase S100A8 and S100A9 expression in prostate cancer. *Int. J. Cancer.* In press (2012).

34. Kasper, M. *et al*. Selective modulation of Hedgehog/GLI target gene expression by epidermal growth factor signaling in human keratinocytes. *Mol. Cell Biol.* **26**, 6283–98 (2006).

35. Henkel, G. W., McKercher, S. R. & Maki, R. A. Identification of three genes up-regulated in PU.1 rescued monocytic precursor cells. *Int. Immunol.* **14**, 723–32 (2002).

36. Kennedy, R. D. *et al*. BRCA1 and c-Myc associate to transcriptionally repress psoriasin, a DNA damage-inducible gene. *Cancer Res.* **65**, 10265–72 (2005).

37. Gebhardt, C. *et al*. Calgranulins S100A8 and S100A9 are negatively regulated by glucocorticoids in a c-Fos-dependent manner and overexpressed throughout skin carcinogenesis. *Oncogene* **21**, 4266–76 (2002).

38. Fessing, M. Y. *et al*. p63 regulates Satb1 to control tissue-specific chromatin remodeling during development of the epidermis. *J. Cell Biol.* **194**, 825–39 (2011).

39. Geng, S. *et al*. Targeted ablation of Arnt in mouse epidermis results in profound defects in desquamation and epidermal barrier function. *J. Cell Sci.* **119**, 4901–12 (2006).

40. Adler, A. S., Kawahara, T. L., Segal, E. & Chang, H. Y. Reversal of aging by NFkappaB blockade. *Cell Cycle* **7**, 556–9 (2008).

41. Martin-Oliva, D. *et al*. Inhibition of poly(ADP-ribose) polymerase modulates tumor-related gene expression, including hypoxia-inducible factor-1 activation, during skin carcinogenesis. *Cancer Res.* **66**, 5744–56 (2006).

42. Zahn, J. M. *et al*. AGEMAP: a gene expression database for aging in mice. *PLoS Genet.* **3**, e201 (2007).

43. Sementchenko, V. I. & Watson, D. K. Ets target genes: past, present and future. *Oncogene* **19**, 6533–48 (2000).

44. Grote, J. *et al*. Identification of poly(ADP-ribose)polymerase-1 and Ku70/Ku80 as transcriptional regulators of S100A9 gene expression. *BMC Mol. Biol.* **7**, 48 (2006).

45. Libregts, S. F. *et al*. Chronic IFN-γ production in mice induces anemia by reducing erythrocyte life span and inhibiting erythropoiesis through an IRF-1/PU.1 axis. *Blood* **118**, 2578–88 (2011).

46. Yasui, D., Miyano, M., Cai, S., Varga-Weisz, P. & Kohwi-Shigematsu, T. SATB1 targets chromatin remodelling to regulate genes over long distances. *Nature* **419**, 641–5 (2002).

47. Wang, D., Rendon, A., Ouwehand, W. & Wernisch, L. Transcription factor co-localization patterns affect human cell type-specific gene expression. *BMC Genomics* **13**, 263 (2012).

48. Verger, A. & Duterque-Coquillaud, M. When Ets transcription factors meet their partners. *Bioessays* **24**, 362–70 (2002).

49. Panopoulos, A. D., Bartos, D., Zhang, L. & Watowich, S. S. Control of myeloid-specific integrin alpha Mbeta 2 (CD11b/CD18) expression by cytokines is regulated by Stat3-dependent activation of PU.1. *J. Biol. Chem.* **277**, 19001–7 (2002).

50. Hegde, S. *et al*. Stat3 promotes the development of erythroleukemia by inducing Pu.1 expression and inhibiting erythroid differentiation. *Oncogene* **28**, 3349–59 (2009).

51. Kerkhoff, C. *et al*. Binding of two nuclear complexes to a novel regulatory element within the human S100A9 promoter drives the S100A9 gene expression. *J. Biol. Chem.* **277**, 41879–87 (2002).

52. Katabami, K. *et al*. Characterization of the promoter for the alpha3 integrin gene in various tumor cell lines: roles of the Ets- and Sp-family of transcription factors. *J. Cell. Biochem.* **97**, 530–43 (2006).

53. de Magalhães, J. P., Curado, J. & Church, G. M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875–81 (2009).

54. Swindell, W. R. Genes and gene expression modules associated with caloric restriction and aging in the laboratory mouse. *BMC Genomics* **10**, 585 (2009).

55. ENCODE Project Consortium *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

56. Portales-Casamar, E. *et al*. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105–10 (2010).

57. Robasky, K. & Bulyk, M. L. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **39**, D124–8 (2011).

58. Matys, V. *et al*. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–10 (2006).

59. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).

60. Irizarry, R. A. *et al*. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).

61. Siepel, A. *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–50 (2005).

62. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–87 (2004).

63. Liu, W. M. *et al*. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18**, 1593–9 (2002).

64. Hastie, T. & Tibshirani, R. *Generalized Additive Models*. Chapman and Hall, London, UK (1990).

65. Aboukhalil, A. & Bulyk, M. L. LOESS correction for length variation in gene set-based genomic sequence analysis. *Bioinformatics* **28**, 1446–54 (2012).

66. Agresti, A. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York, USA (1996).

67. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a powerful and practical approach to multiple testing. *J. Roy. Stat. Soc. B.* **57**, 289–300 (1995).

Acknowledgements

Author contributions

W.R.S., A.J. and J.E.G. conceived the framework, analyzed the data, and wrote the manuscript. X.X., A.L. and P.R. processed tissue samples and performed wet lab procedures. J.J.V. and G.F. participated in drafting the manuscript and provided critical input. All authors read, edited, and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Swindell, W. R. *et al*. Robust shifts in *S100a9* expression with aging: A novel mechanism for chronic inflammation. *Sci. Rep.* **3**, 1215; DOI:10.1038/srep01215 (2013).