Non-coding Class Switch Recombination-related transcription in human normal and pathological immune responses

Helena Kuri-Magaña[1,2]; Leonardo Collado-Torres[3,4]; Andrew E. Jaffe[3,4,5,6]; Humberto Valdovinos-Torres[1]; Marbella Ovilla-Muñoz[1]; Juan M Téllez-Sosa[1]; Laura C Bonifaz Alfonzo[7]; Jesús Martínez-Barnetche[1*]

1.- Centro de Investigación Sobre Enfermedades Infecciosas. Instituto Nacional de Salud Pública. Cuernavaca, Morelos. México.

2.- Programa de Doctorado en Ciencias Biomédicas. Universidad Nacional Autónoma de México.

3.- Lieber Institute for Brain Development. Baltimore, Maryland. USA.

4.- Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland, USA.

5.- Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.

6.- Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.

7.- Unidad de Investigación Médica en Inmunoquímica. Hospital de Especialidades. Centro Médico Nacional Siglo XXI. IMSS. México City. México.

* jmbarnet@insp.mx

Abstract

Background: Antibody class switch recombination (CSR) to IgG, IgA or IgE is a hallmark of adaptive immunity, allowing antibody function diversification beyond IgM. CSR involves a deletion of the IgM/IgD constant region genes placing a new acceptor Constant (C_H) gene, downstream of the VDJ_H exon. CSR depends on non-coding (CSRnc) transcription of donor I_μ and acceptor I_H exons, located 5' upstream of each C_H coding gene. Although our knowledge of the role of CSRnc transcription has advanced greatly, its extension and importance in healthy and diseased humans is scarce.

Methods: We analyzed CSRnc transcription in 70,603 publicly available RNA-seq samples, including GTEx, TCGA and the Sequence Read Archive (SRA) using *recount2*, an online resource consisting of normalized RNA-seq gene and exon counts, as well as coverage *BigWig* files that can be programmatically accessed through R. CSRnc transcription was validated with a qRT-PCR assay for I_μ, $I_{\gamma 3}$ and $I_{\gamma 1}$ in humans in response to vaccination.

Results: We mapped I_H transcription for the human IgH locus, including the less understood *IGHD* gene. CSRnc transcription was restricted to B cells and is widely distributed in normal adult tissues, but predominant in blood, spleen, MALT-containing tissues, visceral adipose tissue and some so-called "immune privileged" tissues. However, significant $I_{\gamma 4}$ expression was found even in non-lymphoid fetal tissues. CSRnc expression in cancer tissues mimicked the expression of their normal counterparts, with notable pattern changes in some common cancer subsets. CSRnc transcription in tumors appears to result from tumor infiltration by B cells, since CSRnc transcription was not detected in corresponding tumor-derived immortal cell lines. Additionally, significantly increased I_δ transcription in ileal mucosa in Crohn's disease with ulceration was found.

Conclusions: CSRnc transcription occurs in multiple anatomical locations beyond classical secondary lymphoid organs, representing a potentially useful marker of effector B cell responses in normal and pathological immune responses. The pattern of I_H exon expression may reveal clues of the local immune response (i.e. cytokine milieu) in health and disease. This is a great example of how the public *recount2* data can be used to further our understanding of transcription, including regions outside the known transcriptome.

Keywords: Ig class switch recombination, B cells, Germinal Center, non-coding transcription, RNA-seq, data mining, tumor microenvironment, antibody, Vaccination, Crohn's disease.

Background

The hallmark of the humoral adaptive immune response is the production of high affinity class-switched antibodies from relatively low-affinity IgM^+ naive precursors. Affinity maturation and class switch recombination (CSR) are tightly regulated, molecular processes that occur in a specialized microenvironment within secondary lymphoid organs known as the germinal center (GC). Upon T-dependent antigen stimulation, IgM^+ naive B cells re-localize into the B cell follicle, undergo clonal expansion within the GC. The GC reaction is an iterative process of mutation-selection that leads to the generation of antigen-specific high affinity, class-switched memory B cells and long-lived antibody secreting plasma cells (reviewed in [1]). Antibody class switching from IgM to IgG, IgA or IgE allows effector function diversification. Both memory B cells and long-lived plasma cells are critical determinants of vaccine efficacy [2].

CSR involves a deletion of a genomic segment from the *IGHM/IGHD* coding interval (C_μ-C_δ) to the upstream flank of one of the *IGHG*, *IGHA* or *IGHE* genes in the telomeric region of human chromosome 14. Activated GC B cells upregulate the Activation Induced Cytidine Deaminase (AID), which deaminates cytidines in the G-rich Switch (S) regions located upstream of each immunoglobulin constant coding gene (C_H). Cytidine deamination induces DNA damage response, which eventually leads to double stranded DNA breaks in both donor (S_μ) and the corresponding acceptor S region. The chromosomal ends are rejoined and the C_μ-C_δ-encoding intervening DNA segment is re-circularized in a non-replicating episome by non-homologous end joining (reviewed in [3]).

The initiation of CSR depends on non-coding transcription of I_H exons, known as germline or "sterile" transcripts (referred hereafter as CSRnc transcription). I_H exons are located in the 5' region of each S-C_H gene module. Non-coding transcription of I_H exons extends to the S and C_H region, is coupled to chromatin remodeling and is dependent on splicing [4, 5]. CSRnc transcripts form an R-loop in the corresponding S region, which

recruits AID to target S region deamination and CSR (reviewed in [6]). The precise mechanism of AID targeting to the S_H region remains elusive, and off-target AID activity is implicated in the genesis of B cell malignancies [6, 7].

CSR is a complex cellular process that occurs in specialized microenvironments in secondary and tertiary lymphoid organs. The cellular choice of which I_H to transcribe, and consequently the Ig class to switch to, is influenced by the availability of certain cytokines such as IL-4, IFNγ, TGFβ and PAMP's, among others. Such environmental cues are thought to trigger specific signals that promote selective transcription of a given I_H exon, guiding CRS according to a particular microenvironment or pathogenic insult [3]. CSRnc transcription patterns may reflect distinctive immunological events, such as the dependence of T cell help and other micro-environmental signals. Thus, CSRnc transcription quantitation during normal and pathological human immune responses could uncover novel pathogenic mechanisms and transcriptional signatures with potential clinical value. In addition, despite CSRnc transcription is biologically linked to B cells, its expression in other cell types has not been ruled out.

The recent explosion in the generation of public genomic data, and in particular transcriptome-wide profiling with RNA sequencing (RNA-seq) provides a unique opportunity to explore previously unannotated features in the human genome. To characterize CSRnc transcription in normal and pathological conditions, we tested CSRnc transcription in human vaccination and analyzed the transcriptional landscape of the human IgH locus using more than 70,000 publicly available human RNA-seq samples from a wide variety of research projects, including the Genotype Tissue Expression project (GTEx) [8, 9], The Cancer Genome Atlas (TCGA) [10, 11], and more than 2,000 projects from the Sequence Read Archive (SRA) using *recount2* [12].

Results

CSRnc transcription is B cell-specific and its boundaries are diffuse: Overall, *recount2* [12] comprises a highly heterogeneous catalog of RNA-seq experiments belonging to 2,036 independent studies and comprising 70,603 samples. Each study is composed of an average of 34 samples. However, TCGA [10, 11] and GTEx (SRP012682) [8, 9] are two projects (studies) with the largest number of sequencing samples (11,284 and 9,661 respectively), and represent 29 % of our dataset (**Additional file 1: Figure S1**).

Although human CSRnc transcription has been evaluated by RT-PCR [5, 13, 14], the precise transcription boundaries, including alternate transcription initiation sites and splicing variants remain undefined. We selected RNA-seq samples derived from normal FACSorted B cells to map CSRnc transcription and to define transcriptional boundaries for read count quantitation for further analysis (**Additional file 2: Table S1**) [15-21]. We found that in normal adult B cells, CSRnc boundaries are less sharply defined than coding transcripts, and as expected, extend into switch regions (**Figure 1**) [5]. Projects SRP045500 [18] and SRP051688 [20] describing the transcriptome of isolated peripheral blood (PB) immune cells in healthy adults, including primary neutrophils, monocytes, myeloid dendritic cells, B, NK and T cells revealed that among terminally differentiated hematopoietic lineage-derived cells, CSRnc transcription is restricted to B cells (**Figure 1A, Additional file 1: Figure S2**).

CSRnc and C_H coding transcription in isolated PB $CD19^+$ B cells was analyzed. High relative transcription of I_μ and C_μ, intermediate relative transcription of I_δ, $I_{\gamma3}$, $I_{\gamma1}$, $I_{\alpha1.2}$, $I_{\gamma2}$, $I_{\gamma4}$ and $I_{\alpha2}$, and low transcription of I_ε and $I_{\alpha1.1}$ were characteristic of PB B cells (**Additional file 1: Figure S3A-C**) [15, 18]. Furthermore, CSRnc transcription was analyzed in tonsillar naive and germinal center B cells from project SRP021509 [16]. CSRnc transcription for most I_H but not I_ε, was relatively high in both naive and GC B cells (**Additional file 1. Figure S3D**). These findings indicate that CSRnc transcription is not exclusive of activated B cells, and agrees with previous findings demonstrating constitutive CSRnc transcription [5, 14].

A transcriptionally active 309 base-pair (bp) region within the *IGHM-IGHD* intron was identified (referred hereafter as I_{δ}) and was included for further analysis (**Figure 1B**). This region is homologous to I_{μ}, overlaps with a previously described repeat termed Σμ, implicated in μ-δ CSR in IgD^{+} myelomas [22, 23]. For the *IGHA1* gene, we identified two potential I_H exons ($I_{\alpha 1.1}$ and $I_{\alpha 1.2}$) that were selected for downstream analysis (**Figure 1C**). $I_{\gamma 2}$ overlapped with a previously annotated lincRNA, ENST00000497397.1 (AL928742.1) at ENSEMBL [24]. The genome coordinates of each I_H exon identified and further analyzed are in **Table 1**. Navigation across the IgH locus using the ENSEMBL Genome Browser [24] allowed to confirm that the predicted I_H exons include regions of RNApolII, H3K36 and H3K4 trimethylation enrichment in peripheral blood B cells and EBV-transformed B cells generated by the Roadmap Epigenomics and ENCODE projects [25, 26], indicating active transcription (**Additional file 1: Figure S4**). Overall, these results agree with the current model of CSRnc transcription [5]; however, the identification of novel transcribed elements ads complexity to the transcriptional regulation of the IgH locus.

CSRnc transcription in peripheral blood is modified upon vaccination and does not depend on circulating plasmablasts: CSRnc transcription increases upon B cell activation [5]. To validate that the predicted CSRnc transcripts were indeed induced upon activation, normal human B cells were stimulated with agents mimicking T-dependent activation (CD40 ligand, IL-21 and CpG) and T independent activation (CpG, Pokeweed mitogen and SAC) *in vitro* for 3 and 6 days. Total RNA was obtained and I_{μ}, $I_{\gamma 3}$ and $I_{\gamma 1}$ CSRnc and AID transcripts were quantified by qRT-PCR (**Figure 2A-D**). CSRnc transcripts were detected in B cells activated by both T-dependent and T-independent activators, but transcription levels were significantly higher for T-dependent like activation. In both types of activation, CSRnc transcription at 6 days post-activation was higher than 3 days post-activation (**Figure 2A-D**). The highest transcription was for I_{μ} (3-fold higher than $I_{\gamma 1}$ and $I_{\gamma 3}$). Transcription of AID showed the same pattern as CSRnc transcripts (**Figure 2D**).

Immunization promotes an antigen-specific mobilization of plasmablasts to peripheral blood around day 7 post-challenge [27, 28]. This plasmablast wave is thought to

167 derive from germinal centers [29]. Thus, we hypothesized that the level of CSRnc
168 transcription in peripheral blood correlates with the amount of plasmablasts. We assessed
169 CSRnc transcription and plasmablast proportion in peripheral blood of healthy subjects
170 before, 7 and 14 days post vaccination with either hepatitis B (HB) alone or in combination
171 with tetanus/diphtheria vaccine (TT/Dp) (**Figure 2E-K**). Regarding pre-immune levels, I_{μ}
172 transcription was not affected at day 7 (mean 0.95 ± 0.5, p = 0.76) or at day 14 (mean 2.04
173 ± 2, p = 0.073), however at day 14 was higher than at day 7 (p = 0.01) (Figure 2E). $I_{\gamma 1}$
174 transcription was neither affected at day 7 (mean 2.9 ± 3.2, p = 0.057) nor at day 14 (mean
175 1.02 ± 0.9. p = 0.57) (Figure 2F). Contrastingly, $I_{\gamma 3}$ transcription regarding pre-immune
176 levels was reduced at day 7 (mean 0.76 ± 0.3, p = 0.016) and increased at day 14 (mean
177 1.73 ± 1.1, p = 0.021). As expected, $I_{\gamma 3}$ transcription at day 14 was higher than at day 7 (p =
178 0.0002) (Figure 2G). No changes were detected in AID expression (Figure 2H).

179 Plasmablast levels peaked at day 7 post-vaccination (p = 0.011), and returned to pre-
180 vaccine levels 14 days post-vaccination, as previously described [27, 28]. There was no
181 correlation between plasmablasts and $I_{\gamma 1}$ transcription increase at day 7 (LTS method.
182 Adjusted R^2 = -0.1265, F-statistic: 0.1013 on 1 and 7 DF, p = 0.75). However, plasmablast
183 and I_{μ} transcription correlated at day 7 (**Figure 2J)** (LTS method. Adjusted R^2 = 0.53; F-
184 statistic = 10.17 on 1 and 7 DF, p = 0.015). Interestingly, plasmablast fold-change at day 7
185 negatively correlated with $I_{\gamma 3}$ transcription (**Figure 2K)** (LTS method. Adjusted R^2 = 0.63;
186 F-statistic = 13.04 on 1 and 6 DF, p-value: 0.011). No changes in I_{μ}, $I_{\gamma 1}$ and $I_{\gamma 3}$ transcription
187 was detected in response to influenza vaccination (**Figure 2L-O**). Overall, these results
188 suggest that an increase in CSRnc transcription in peripheral blood upon vaccination is
189 dependent on vaccine type, and that the contribution of vaccine-mobilized plasmablasts to
190 CSRnc transcription is negligible.

191 **CSRnc transcription is prevalent in a large fraction of the *recount2* dataset, with**
192 **predominant transcription of I_{μ}:** Normalized counts (RPKM) obtained with *recount2*
193 were used to asses CSRnc transcription in the SRA, TCGA and GTEx datasets. We found
194 no CSRnc transcription (RPKM = 0) in any of the 10 I_H exons in a substantial fraction of
195 the whole dataset (n = 26,512 samples; 37%) (**Table 2**). The I_H $\log_2$-transformed RPKM
196 average (2.65) for each non-zero RPKM sample (n = 44,091; 62.4 %) was used to define an

expression cutoff. "High" CSRnc transcription was defined as a mean $\log_2$ RPKM of 2.65 or higher. Only 29.8 % of RNA-seq samples (n = 21,017) were above this expression cutoff. "Low" CSRnc transcription was defined as a mean $\log_2$ RPKM < 2.65 (**Figure 3A. Table 2**), which accounted for the remaining 32.7 % of the samples (n = 23,074). The CSRnc transcription levels varied according to I_H. Higher transcription ($\log_2$ RPKM), as well as a more widespread transcription (proportion of samples) was found for I_μ in all datasets (**Additional file 1: Figure S5**).

The *recount2* dataset is partitioned in distinctive CSRnc transcription profiles: We further de-convoluted CSRnc transcription according to I_H relative transcription profile. To do so, the entire non-zero dataset including GTEx, TCGA and SRA Z-scores was clustered in 10 groups using *k*-means clustering (**Figure 3B. Additional file 1: Figure S6**). Using Z0 as reference, each of I_H showed a distinctive expression pattern in the remaining Z clusters ($\log_2$ RPKM regression analysis. F-statistic: p-value: < 2.2e-16). Clusters Z0, Z2 and Z8 were characterized by low (mean $\log_2$ RPKM < 2.65 or Z score < 0) CSRnc transcription in all I_H classes. Clusters Z4, Z6 and Z7 were characterized by "high" expression of I_μ only (mean $\log_2$RPKM > 2.65). Cluster Z3 showed "high" expression in I_μ and $I_{\alpha 2}$. Cluster Z1 showed high expression in I_μ, $I_{\gamma 1}$ and $I_{\gamma 4}$. Finally, clusters Z5 and Z9 showed "high" expression in all I_H's (**Figure 3B. Additional file 1: Figure S6**).

CSRnc transcription is widely distributed in healthy tissues, with particular profiles according to tissue: To gain insight into CSRnc expression patterns in healthy adult human tissues, we used GTEx samples as a reference. Non-zero RPKM per tissue samples were ranked according to their mean $\log_2$ RPKM. Higher average transcription was found in lymphoid tissues such as spleen, EBV-transformed B lymphocytes and whole blood, but also in organs with mucosal-associated lymphoid tissues (MALT) such as terminal ileum, transverse colon, stomach, lung and esophageal mucosa (**Figure 4**). A remarkable difference in average $\log_2$ RPKM is observed in transverse colon (mean 7.5 ± 3.9) and sigmoid colon (mean 2.1 ± 3.2). Interestingly, salivary gland expression was among the tissues with highest transcription, and non-mucosal tissues such as thyroid and pituitary gland showed high average I_H transcription. Another notable difference in average $\log_2$ RPKM was observed between visceral adipose tissue (omentum; mean 4.9 ± 2.3) and

subcutaneous adipose tissue (mean 1.9 ± 1.5). Conversely, samples of tissues such as brain, skeletal and cardiac muscle, skin, as well as chronic myelogenous leukemia cell line K562 and transformed dermal fibroblast cell lines showed the lowest CSRnc transcription levels (mean $\log_2$ RPKM < 2.65) (**Figure 4**).

Using whole blood as reference tissue, we performed a linear regression analysis of each I_H by tissue type. The transcription pattern of each I_H in whole blood was significantly different ($p < 2.2e-16$) (**Figure 5**). I_μ transcription was similar to I_H average transcription (**Figure 4 and 5**), in which higher transcription was observed in spleen, terminal ileum, salivary gland, and transverse colon than blood. Of note, I_μ transcription in testis was particularly low, despite its high average transcription. I_δ transcription was similar to I_μ transcription, but only spleen and terminal ileum were significantly higher. $I_{\gamma 3}$, $I_{\gamma 1}$ and $I_{\gamma 2}$ transcription was similar and was higher in spleen than in blood ($p < 2e-16$) (**Figure 5)**.

For most tissues, transcription of $I_{\alpha 1.1}$ and $I_{\alpha 1.2}$ was highly correlated, however some differences were noted. $I_{\alpha 1.1}$ transcription was higher in terminal ileum and transverse colon than in blood ($p < 0.001$). In contrast, $I_{\alpha 1.2}$ transcription was higher in spleen and salivary gland than in blood ($p < 2e-16$). $I_{\alpha 2}$ transcription was similar to $I_{\alpha 1.2}$ transcription, but also was higher in stomach, esophageal mucosa than in blood ($p < 1.1e-05$) (**Figure 5**). Thus, $I_{\alpha 1}$ and $I_{\alpha 2}$ transcription pattern matches with the fact of IgA as the main immunoglobulin in mucosal tissue. Furthermore, our results suggest that CSR to IgA may involve tissue-dependent alternative transcription initiation and/or splicing in the corresponding CSRnc transcripts.

The most unexpected patterns of CSRnc transcription corresponded to $I_{\gamma 4}$ and I_ε. Transcription for both was higher in spleen, terminal ileum and EBV-transformed B-cells than in blood ($p < 0.01$). However, $I_{\gamma 4}$ transcription was higher in thyroid, visceral adipose tissue (omentum), testis, than in blood ($p < 2e-16$). Similarly, I_ε transcription was higher in in thyroid than in blood ($p < 0.00003$) (**Figure 5**).

Consistently, de-convolution of CSRnc transcription according to I_H relative expression (Z-score) revealed that terminal ileum, spleen, transverse colon, whole blood and salivary gland share a similar expression pattern and are highly enriched (FDR < 0.001)

in clusters Z5 and Z9 (high transcription in all I_H's) (**Figure 6**). Some tissues with MALT such as stomach, esophageal mucosa and lung shared a similar I_H transcription pattern and were enriched in cluster Z5 and Z3 (high I_μ and $I_{\alpha 2}$ transcription) (FDR < 0.001), whereas others such as breast, vagina, liver were enriched in cluster Z3 and Z6 (I_μ only). Testis, thyroid, pituitary and omentum, characterized by the highest $I_{\gamma 4}$ transcription, were enriched in cluster Z1 (I_μ, $I_{\gamma 1}$ and high $I_{\gamma 4}$) (FDR < 0.001). Finally, the remaining tissues such as subcutaneous adipose tissue, arteries, brain, skeletal and cardiac muscle, skin, transformed fibroblasts and K562 cell lines, were enriched in Z0, Z2 and Z8 (low CSRnc transcription in all I_H classes) (FDR < 0.001) (**Figure 6**), and in few cases were enriched in clusters Z4 and Z6 (I_μ only) (FDR < 0.001).

The observed anatomical distribution and abundance of CSRnc transcription suggests that the amount of CSRnc transcription may be dependent on the abundance of B cells present in a given tissue. We used the C_H coding transcript log_2RPKM as a proxy of the amount of B cells in the tissue. In general, C_H transcription was 10 to 100-fold higher than CSRnc transcription. To correlate CSRnc transcription with corresponding C_H transcription, Z-scores were used. As expected for each class, CSRnc and C_H Z-scores where significantly correlated ($p < 1.0e\text{-}16$), suggesting that the higher numbers of B cells or plasma cells in a given tissue, higher CSRnc transcription. However, we have noticed that for most classes, a fraction of samples deviates from the expected orthogonal correlation, indicating higher CSRnc transcription relative to the C_H transcript. This is particularly notable for I_μ, but also, $I_{\gamma 1}$, $I_{\gamma 4}$ and $I_{\alpha 2}$ (**Additional file 1: Figure S7**).

Iμ transcription occurs in early lymphoid progenitors and $I_{\gamma 4}$ is widely expressed in non-lymphoid fetal tissues: CSRnc transcription was addressed in early lymphoid development using data from study SRP058719, which addresses transcription in early lymphoid differentiation prior to and after B and T cell lineage commitment using RNA-seq from FACSorted cells [30]. Interestingly, both B and T lineage precursors expressed I_μ. Enriched hematopoietic stem cells (HSC's $CD34^+CD38^-lin^-$), lymphoid-primed multipotent progenitors (LMPP's, $CD34^+CD38^+CD10^-CD45RA^+lin^-$), common lymphoid progenitors (CLP's, ($CD34^+CD38^+CD10^+CD45RA^+lin^-$), thymic $CD34^+CD7^-$

285 $CD1a^{-}$ $CD4^{-}CD8^{-}$ (Thy1) precursors and fully B cell-committed progenitors (BCPs,
286 $CD34^{+}CD38^{+}CD19^{+}$) expressed Iμ (**Additional file 1: Figures S8 and S9**).

287 CSRnc expression was analyzed in fetal tissue by using project SRP055513 data,
288 which reported an extensive RNA-seq analysis in twenty fetal tissues during gestational
289 weeks 9-22 [31]. Remarkably, a robust $I_{\gamma 4}$ expression was detected in all fetal tissues tested,
290 regardless of the gestational week and in the absence of additional I_H and coding C_H
291 transcription (**Figure 7**). Higher average $I_{\gamma 4}$ expression is in spleen, followed by lung and
292 liver. The latter has lympho-hematopoietic function in the fetal stage. In contrast with their
293 adult counterparts, $I_{\gamma 4}$ was highly transcribed (Z-score > 0.8) in kidney, brain and muscle.

294 **CSRnc transcription in cancer varies according to cancer type and is likely to depend**
295 **on the degree of B cell tumor infiltration:** The TCGA project data was used as a
296 reference to study CSRnc transcription in a wide variety of human cancers. As for GTEx,
297 CSRnc transcription (I_H average $\log_2$ RPKM distribution) was analyzed across 33 cancer
298 types (**Figure 8**). I_H transcription in diffuse large B cell lymphoma (DLBCL) and acute
299 myeloid leukemia (AML) were the highest. In general, CSRnc transcription in neoplastic
300 tissue mimicked its non-neoplastic counterparts, being high in lung, stomach, and testicular
301 germ cell carcinomas. Conversely, CSRnc transcription was low in glial cell and skin
302 cancers (**Figure 8**).

303 A direct comparison between CSRnc transcription in healthy (GTEx) and neoplastic
304 tissue allowed the identification of three distinct patterns (**Figure 9**): **1)** Tumors where
305 average CSRnc transcription is lower than in healthy tissue, such as in DLBCL, prostate,
306 thyroid, liver and colon cancer (**Figure 9 A-E**). **2)** Tumors where average CSRnc
307 transcription was higher than its healthy tissue counterpart, such as breast, rectum,
308 testicular germ cell, pancreas carcinomas, ovarian cystadenoma and skin melanoma
309 (**Figure 9 G-L**). **3)** Tumors with no difference in CSRnc transcription between healthy and
310 neoplastic counterpart, such as stomach and esophagus (**Additional file 1: Figure S10**).
311 Interestingly, using healthy kidney cortex as reference for kidney tumors, CSRnc
312 transcription varied according to cancer type, CSRnc transcription was significantly lower
313 in chromophobe and papillary carcinomas, but not in clear cell carcinoma. (**Figure 9F**). In

adrenal gland, a similar pattern was observed, with lower CSRnc transcription in adrenocortical carcinoma, but not pheochromocytoma (**Additional file 1: Figure S10**).

The correlation between CSRnc and C_H transcription (**Additional file 1: Figure S7**) suggests that as for healthy tissues, CSRnc transcription derives from tertiary lymphoid infiltrates resulting from tumor-associated inflammation and the corresponding mucosal associated lymphoid tissues [32]. To address this question, we analyzed CSRnc transcription in a wide variety of tumor-derived cell lines in SRA. The majority of the representative tumor derived cell lines tested (i.e., lung cancer A549 cells, breast cancer MCF-7 cells, and colon cancer HCT116 cells), were depleted in samples expressing CSRnc RNA (**Additional file 2: Table S2**), indicating that CSRnc transcription in cancer derives from infiltrating B cells and not the neoplastic cells *per se*. Nevertheless, using this approach, CSRnc transcription in cancer cells *in situ* cannot be ruled out.

CSRnc transcription is altered in certain infectious conditions and a I_δ-$I_{\alpha 2}$ transcriptional signature is associated with pediatric Crohn's disease with deep ulceration: The SRA dataset represents the most diverse collection of data regarding methodological approaches and subjects of interest and represents a useful source of data related to diverse malignancies, as well as infectious and non-infectious inflammatory pathology. Using MetaSRA Disease Ontology annotations [33], we identified significant enrichment of acute AML, breast and lung cancers in SRA samples with CRSnc transcription, confirming our observations derived from TCGA data analysis.

Moreover, the SRA dataset allowed us to identify enriched CSRnc expression in infectious diseases such as diarrhea, brucellosis and malaria. However, in all cases, peripheral blood samples were used for these experiments. To distinguish if the enrichment was due to increased expression, rather than for the inherent enrichment observed in blood (**Figures 4 and 5**), we performed differential expression analysis comparing experimental groups provided by SRA metadata. Significant reduction of CSRnc transcription was detected in enterobacteria, but not in rotavirus diarrheal disease (SRP059039), malaria (SRP032775) [34], brucellosis and leishmaniasis (SRP059172) (**Additional file 2: Table S3**). SRA data from influenza vaccination (SRP020491) [15] was consistent with our own experimental data demonstrating no significant changes in CSRnc transcription 7 days post-

vaccination, regardless the plasmablast migration wave to peripheral blood at that time, indicating that CSRnc transcripts observed in peripheral blood do not derive from recently class-switched plasmablasts (**Figures 2L-O**). CSRnc transcription in response to influenza vaccination contrasted with the observation that natural infection with H7N9 [35] (SRP033696), in which we observed changes in coding and CSRnc transcription (**Additional file 2: Table S3**).

The SRA dataset also showed enrichment for autoimmune disease terms such as systemic lupus erythematous and autoinflammatory diseases like inflammatory bowel disease. No significant differential expression was observed in peripheral blood transcriptomes in systemic lupus erythematosus patients compared with healthy subjects (SRP062966) [36] (**Additional file 2: Table S3)**. However, a study comparing the ileal transcriptome in pediatric Crohn's disease (CD) with and without deep ulceration, and with or without ileal involvement (SRP042228) [37], revealed an increased significant expression for I_{μ}, I_{δ} and $I_{\alpha 2}$ in patients with deep ulcerated CD, regardless the presence of macroscopic or microscopic inflammation (**Figure 10 and Additional file 2: Table S3**).

Discussion

We performed an integral, systematic analysis of human CSRnc transcription using an experimental approach and datamining of a large and diverse public RNA-seq dataset supported by a previously described resource, *recount2* [12]. The precise I-exon boundaries for every class were undefined and are not annotated in current human genome version (GRCh38.p12). Among hematopoietic-derived cells, CSRnc transcription was specific for B cells, and consistent with previous findings was present in naive as well as GC-B cells [5].

The present study has certain limitations. CSRnc transcription is required for CSR, however it does not prove that CSR is actively taking place. Among the most determinant factors influencing the amount of CSRnc transcription observed in an RNA-seq sample is the relative amount of B cells expressing a particular I-exon, as suggested by the high Z-score correlation between CSRnc and its corresponding C_H transcript (**Additional file 1: Figure S7**). We propose that the higher non-coding/coding Z-score ratio indicates a higher

proportion of B cells undergoing a particular CSR event. Other factors influencing the amount of CSRnc transcription is that switch circles are transcriptionally active [38], and our current analysis cannot differentiate if CSRnc transcripts derive from chromosomal or switch circle transcription. Switch circles are non-replicating episomes that decay with B cell division [39-41]. Thus, B cells undergoing cell division at high rate (i.e., GC B cells) should dilute the amount of circle templates in a greater extent than non-dividing B cells.

A novel transcribed element (I_δ) located in the *IGHM-IGHD* intergenic region was identified, which overlaps with a previously described repeat Σ_μ region, downstream of an atypical switch region ($\sigma\delta$) involved in non-classical CSR to IgD in mice [42] and humans [43-45]. The $\Sigma\mu$ region was originally described as mediator of μ–δ CSR by homologous recombination in myeloma cell lines [22, 23]. The biological role of $\Sigma\mu$ is uncertain, because IgD^+ EBV-transformed cell lines and human tonsillar B cells undergo μ–δ CSR by non-homologous recombination using $\sigma\delta$ as acceptor switch region [43, 45], in an AID-dependent fashion [44]. Here, we demonstrate that I_δ ($\Sigma\mu$) can be actively transcribed (**Figure 1B**). Nonetheless, I_δ as an acceptor I-exon does not fit into the general model of CSR, because it is located downstream of the $\sigma\delta$ region. Further research is required to elucidate if I_δ transcription is involved in the non-classical μ–δ CSR.

An important motivation for this work was the identification of early transcriptional signatures in blood that correlated with the strength and quality of the humoral response, and in particular with the GC response. Interestingly, Hepatitis B and or Tetanus/Diphteria vaccination induced I_μ, and $I_{\gamma 3}$ transcription in peripheral blood at day 14 post-vaccination, but not when plasmablasts peaked (day 7 post-vaccination), suggesting that CSRnc transcription may not be the result of plasmablast mobilization to peripheral blood. Moreover, we observed increased $I_{\gamma 1}$ and $I_{\alpha 1}$ transcription in natural H7N9 infection, but not upon influenza vaccination (**Figure 2, Additional file 2: Table S3**), or in rotavirus infection (**Additional file 2: Table S3**). The differential response observed upon influenza vaccination and natural rotavirus infection in contrast to HBV/tetanus-diphteria vaccination and H7N9 infection may be the result of the common repeated exposure to seasonal influenza or rotavirus, which would reactivate IgG^+ memory B cells and in the absence of CSR [46].

Although I_H transcription is highly inducible upon activation (**Figure 2A-D**), GTEx and SRA mining revealed that I_μ transcription was higher than other I_H. This is consistent with the current model of CSR, in which I_μ is constitutively transcribed under the control of the μ intronic enhancer (E_μ), and its transcription is required for CSR regardless of the acceptor class [47]. E_μ participates in chromatin remodeling of the IgH locus during lineage commitment and VDJ recombination [48, 49], and its deletion impairs B cell development, I_μ transcription and CSR [50].

We identified distinctive CSRnc transcriptional patterns related to known immunological functions, such as I_μ and $I_{\alpha 2}$ co-expression in MALT-rich organs, where active IgA secretion takes place. Of particular interest is cluster Z1, defined by an $I_{\gamma 1}$, $I_{\gamma 2}$ and $I_{\gamma 4}$ transcriptional signature. Z1 cluster was the only cluster with higher expression of $I_{\gamma 4}$, and was characteristic of the testis, thyroid gland and visceral adipose tissue (omentum), but not subcutaneous adipose tissue. Visceral adipose lymphoid B cells are immunologically active cells implicated in adipose tissue homeostasis that may play important pro-inflammatory roles associated with metabolic syndrome and obesity [51]. The expression of $I_{\gamma 1}$ and $I_{\gamma 4}$ transcriptional signature in testis and thyroid is unexpected, because they are regarded as immune-privileged sites devoid of secondary/tertiary lymphoid organs [52]. Similarly, higher I_ε transcription in the thyroid gland is a striking finding worthy of further research due to the common association of atopic disease with autoimmune thyroiditis [53]. At present, we do not know if CSR to IgG_4 or IgE is taking place, however further research is required given that IgG_4 is an atypical Ig that lacks Fc-mediated effector functions [54] and IgE could be implicated in autoimmunity. Furthermore, $I_{\gamma 4}$ transcription in different fetal tissues suggests that its transcription is not limited to lymphoid tissue, may be a common feature of multipotency and is not necessarily coupled with C_H transcription, suggesting additional functions beyond CSR.

A major goal of the GTEx project was to identify the role of genetic variation in gene expression as a quantitative trait. Tissue enrichment in two clusters with qualitatively different I_H transcriptional pattern such as in liver (Z3 and Z6), testis (Z1 and Z2), terminal ileum (Z5 and Z9) and whole blood (Z5, Z7 and Z9) indicate CSRnc transcriptional heterogeneity in tissue donors involved in the GTEx project. This could result from

433 different tissue microenvironments (i.e., cytokine/chemokine milieu, microbiota and
434 environmental stimuli) as well as genetics, which may modify CSRnc transcription patterns
435 and possibly CSR itself. A recent study identified four SNP's in the human IgH locus
436 presumably involved in CSR that affect immunoglobulin levels [55]. Their implication in
437 modifying CSRnc transcription warrants further investigation.

438 The study of CSRnc transcription in cancer is of particular interest for several
439 reasons: **1**) The anti-tumor response is largely mediated by the presentation of tumor neo-
440 antigens to T cells and T_{reg} balance. However, antibody-mediated anti-tumor activity can be
441 achieved by antibody - dependent cytotoxicity or other mechanisms [32]. **2**) The presence
442 of ectopic (or tertiary) lymphoid structures (ELS) [32, 56] and higher densities of
443 infiltrating B cells and T follicular helper cells correlate with improved survival in lung
444 [57], breast [58] and colorectal carcinoma [59]. **3**) The tumor microenvironment, including
445 certain cytokines may modify CSR patterns in infiltrating and ELS B cells, regardless the
446 antibody effector function. **4**) AID activity is a known contributor to off-target mutagenesis
447 and genomic instability in B cell malignancies [60]. Aberrant AID and CSRnc transcription
448 in non-lymphoid tumor cells could potentially contribute to cytidine deaminases - mediated
449 kataegis [61].

450 We have found that average I_H expression in certain tumors analyzed in the TCGA
451 project resembles their non-neoplastic counterpart, however some cancer types have
452 significantly less CSRnc transcription, whereas others show the opposite. Most of the
453 evidence we have gathered so far indicate that the origin of I_H transcription is in the tumor
454 infiltrating and ELS B cells, rather than the tumor cells *per se* (**Additional file 2: Table**
455 **S2**). Thus, differences in I_H transcription in cancer may be the result of immune editing
456 [62], which may alter the amount and activation state of the infiltrating and ELS B cells.
457 Based on gene expression signature clustering, all non-hematologic cancers of the TCGA
458 project were classified into six immunologically distinct subtypes with distinctive somatic
459 aberration patterns, tumor microenvironment including the amount and cell type
460 infiltration, and clinical outcome [63, 64]. The relation between these six immunological
461 subtypes with CSRnc transcription pattern may help to understand the elusive role of
462 infiltrating B cells in the progression of different cancer types [32].

Despite the limitation of relying on public data when often the submitter researcher chooses to submit the minimal requirements of sample metadata, the SRA represents an enormous source of RNA-seq data from a highly diverse type of studies. In contrast to the standardized methodological criteria and metadata collection protocols used by the GTEx and TCGA consortiums, higher methodological variability in SRA data is expected, limiting inter-study comparisons. Nevertheless, we were able to identify a I_{μ}- I_{δ} - $I_{\alpha 2}$ signature in ileal mucosa of Crohn's disease in a treatment-naive pediatric cohort [37]. I_{μ} and $I_{\alpha 2}$ is somehow expected as the result of predominant IgM to IgA CSR on mucosal tissue (**Figures 4-6**), and its exacerbation due to increased tissue B cell infiltration in response to inflammation [37]. However, the increased I_{δ} transcription, particularly in CD associated with deep ulceration is an intriguing finding (**Figure 10**). Serum IgD levels are elevated in patients with CD [65] and other autoinflammatory syndromes [44, 66]. A high proportion of μ–δ switched IgD^{+} cells bare autoreactive and poly-reactive specificities [67, 68], a feature shared with "natural autoantibodies", which are reactive against bacterial wall components and may provide natural immunity against bacterial infection [69]. In human respiratory mucosa, μ–δ switched IgD^{+} B cells mediate the innate-adaptive immunity and inflammatory cross talk [44]. Mice incapable of undergoing classic CSR due loss of function of 53BP1 have an intestinal microbiota-dependent elevation of IgD serum titers and increased μ–δ switched IgD^{+} B cells [70]. Direct experimental testing is required to elucidate the role of I_{δ} transcription, its role in μ–δ CSR and its implications in healthy and inflamed mucosae.

Conclusions

We have performed an unbiased analysis of the transcriptional landscape of the human *IGH* locus using a vast public RNA-seq dataset. Our observations agree with previous findings regarding constitutive CSRnc transcription in naïve B cells and its upregulation upon activation. We provide a detailed analysis of CSRnc transcription in healthy tissue. As expected, CSRnc transcription correlated with the amount of associated lymphoid tissue, however, novel transcriptional signatures involving $I_{\gamma 4}$ or I_{ε} were found in testis, pituitary, thyroid and visceral adipose tissue were identified. Changes in CSRnc transcription between healthy and tumor tissue were also found, likely as a result of

493 immune editing. A novel transcribed element within the *IGHM-IGHD* intron termed I_δ was
494 discovered and highly expressed in ileal mucosae of pediatric Crohn' s disease patients.
495 Overall, this study highlights the importance of open access data for discovery and
496 generation of novel hypothesis amenable for direct testing, and is a great example of the
497 potential of the *recount2* dataset to further our understanding of transcription, including
498 regions outside the known transcriptome.

499

500 ## Materials and methods:

501 **Vaccination of human healthy volunteers:** Pre-immune (day 0), day 7, 15, 30 and 180
502 post-vaccination peripheral blood samples (18 mL) were obtained by venipuncture in 2 x 8
503 mL Vacutainer® CPT™ tubes from healthy volunteers vaccinated with Hepatitis B and/or
504 Tetanus toxoid/Diphteria (n = 16), or Trivalent Influenza Vaccine during season 2013-2014
505 (A/California/7/2009 (H1N1) pdm09; A(H3N2) A/Victoria/361/2011; B/Massachusetts/
506 2/2012) (n = 18). Written informed consent was obtained from each volunteer in each blood
507 sample draw. All procedures in human subjects were performed after Institutional Review
508 Board approval from the National Institute of Public Health (CI: 971/82-6684). Plasma and
509 PBMCs were isolated according to the manufacturer's instructions, aliquoted and stored at -
510 80ºC and liquid N_2, respectively. Total RNA was extracted from PBMCs with TRIzol and
511 stored at -80ºC until used.

512 **Quantitation of plasmablasts by FACS:** Cryopreserved PBMCs were thawed at 37º C
513 and resuspended in RPMI 10% FBS, washed with PBS 1X and fixed with 1%
514 paraformaldehyde for 20 min at room temperature. After washed with FACS solution (PBS
515 1%, sodium azide 0.05% and 2% FBS), cells were incubated for 30 min at 4 ºC with the
516 following antibody cocktail: anti-CD3 PerCP/Cy5.5 (clone SKY; Biolegend; 344808), anti-
517 CD19 FITC (clone HIB19; Biolegend; 302206), anti-CD20 PE Cy7 (clone 2H7; Biolegend;
518 302312), anti-CD27 APC (clone O323; Biolegend;356410) and anti-CD38 PE (clone HIT2;
519 Biolegend; 980302). Flow cytometry analysis was performed in a FACS Aria II (BD
520 Biosciences, San Jose, CA, USA). Doublets and CD3+ events were gated out. Plasmablasts
521 were defined as $CD3^-/CD19^+/CD20^-/CD27+/CD38^+$. 500-1000 plasmablasts were acquired
522 per sample. Analysis was performed using Flowjo software (TreeStar).

B cell *in vitro* stimulation: PBMCs were isolated by Ficoll-Paque™ density gradient from blood bank buffy coats. B cells were enriched through negative selection using B cell Isolation Kit II (MACS, Miltenyi). $1x10^6$ B cells were seeded per well on 6-well plate incubated in RPMI medium supplemented with 10% FBS, streptomycin and penicillin at 37°C with 5% CO_2. Two activation conditions were stablished at different time points (3 and 6 days post-activation: Germinal center-like activation (GC-like), 1µg/ml anti-human CD40 (G28.5), 5µg/ml CpG ODN 2006 (Invivogen) and 25ng/ml recombinant IL-21 (eBiosciences). For T-independent activation, 5µg/ml CpG ODN 2006 (Invivogen), 0.05% *S. aureus* Cowan (Pansorbin; Calbiochem), 5ng/ml Pokeweed Mitogen (Sigma).

qRT-PCR of CSRnc transcripts: Total RNA from PBMCs was extracted using TRIzol (Invitrogen). The integrity of the RNA was measured with Agilent RNA 6000 Nano. SuperScript™ III One-Step RT-PCR(Invitrogen) was used for reverse transcription and amplification. Quantitative PCR of CSRnc transcripts for *IGHM*, *IGHG*1, *IGHG*3 and AID gene was performed using specific primers and TaqMan probes(IDT). The primers and probes used to quantify the CSRnc transcripts are detailed in **Table 3**. Amplification of *HPRT* with PrimeTime® Predesigned qPCR Assays was performed as the reference gene. The fold difference was calculated using 2ΔΔCT, using resting enriched B cells as calibrator and non-B cells as negative control. *HPRT* was used as normalizer for every condition.

CSRnc transcription boundaries definition: Currently, CSRnc transcripts and switch regions (S) are not mapped as such in the current version of the human genome (GRCh38). I_H are upstream of the corresponding switch region (S_H), thus we first mapped S regions based on the frequency distribution of the AGCT motif in 500 bp bins along the whole *IGH* locus (105,583,700 – 105,863,000) [3]. *recount2* is an online resource consisting of normalized RNA-seq gene and exon counts, as well as coverage *BigWig* files (https://jhubiostatistics.shinyapps.io/recount/) that can be programmatically accessed through the R programming language [71]. To map I_H, metadata from all SRA projects contained in *recount2*, as well as through the SRA-Run selector engine were used to identify RNA-seq samples and samples performed using purified B cells. The corresponding *BigWig* files were downloaded using the *recount* Bioconductor package and

mapped read counts were visually inspected with Integrative Genomics Viewer [72]. CSRnc regions were delimited according to an expression consensus from projects described in **Additional file 2: Table S1**.

CSRnc and IGH transcription quantitation: *recount2* was used to extract read counts from each of the nine C_H constant region coding genes (*IGHM*, *IGHD*, *IGHG*3, *IGHG*1, *IGHA*1, *IGHG*2, *IGHG*4, *IGHE* and *IGHA*2), as well as from the corresponding CSRnc I exon coordinates as a *GRanges* object [73]. The $\log_2$-transformed C_H (coding) and CSRnc RPKM average per *sample* was used as an approximation of abundance of transcription. For CSRnc transcription, $\log_2$ RPKM per *sample* average adopted a quasi-normal distribution with a mean of 2.65 $\log_2$ RPKM (SD ± 3.79), which corresponds to 6.29 RPKM. As an initial exploration to which tissues and in which diseases CSRnc transcription takes place, we used the $\log_2$ RPKM average as a cut-off to define "high" expression (> 2.65 $\log_2$ RPKM) or "low" expression (< 2.65 $\log_2$ RPKM). The mean C_H transcription average $\log_2$ RPKM was 7.82 (SD± 5.16). Given the difference between coding and CSRnc transcription, and to address the relative expression between coding and CSRnc transcription for each Ig loci, coding and CSRnc $\log_2$RPKM values for each Ig gene were standardized by transformation to Z-scores.

SRA RNA-seq samples Ontology mapping: Although all RNA-seq samples in TCGA and GTEx follow a homogeneous ontology categorization, metadata associated to SRA projects is widely heterogeneous and commonly insufficient. To obtain a more homogenous categorization of nearly half of our dataset, we used disease [33] annotations retrieved from MetaSRA, version 1-2 [74].

Enrichment test: To define CSRnc transcription profile variation in healthy tissue (GTEx dataset), we performed tissue sample enrichment analysis according to Z-cluster using a two sided Fisher's Exact Test. The H_0 is that there is no difference in the probability distribution between Z clusters and tissue. A 2 x 2 contingency table was built for each tissue between the number of samples belonging to a given Z cluster and the remaining samples not belonging to that cluster. A two-sided Fisher's test was performed with the R function fisher.test (c, alternative = "two.sided"). P value adjustment with the Benjamini-Hochberg method was performed to correct for multiple testing using the R function p.

adjust(p, method = "bh", n = length(p)). A False Discovery Rate (FDR) < 0.01 was considered as a significant enrichment.

CSRnc transcription profile clustering: To define CSRnc transcription profiles, Z-scores for each I-exon were subjected to *k*-means clustering using the Cluster software 3.0 [75]. Ten clusters were generated, with 100 iterative runs and Euclidean distance as distance metric. Clustered data was visualized with java Treeview 3.0 (https://sourceforge.net/projects/ jtreeview/).

Differential expression analysis: Differential expression of coding IGH and CSRnc transcripts was analyzed using the functions lmFit() and eBayes() from the *limma* R package v3.34 [76]. If technical replicates were present for a given study, the induced correlation was adjusted for using the duplicationCorrelation() function from *limma*. The resulting Bonferroni-adjusted p-values less than 0.05 were determined to be statistically significant. We used Bonferroni instead of FDR given that we tested 10 regions instead of the usual number of thousands of genes.

598 # Tables:

Table 1. Mapping I exons (I_H) and switch regions (S) in human chromosome 14.GRCh38.p10

Gene	Start	End	Length (bp)
IGHM			
Iμ	105,861,311	105,862,213	903
Sμ	105,856,501	105,860,500	4,000
IGHD			
Iδ	105,847,549	105,847,857	309
IGHG3			
$I\gamma_3$	105,775,023	105,775,702	680
$S\gamma_3$	105,773,001	105,774,000	1,000
IGHG1			
$I\gamma_1$	105,747,322	105,748,299	978
$S\gamma_1$	105,744,501	105,745,500	1,000
IGHA1			
$I\alpha_{1.2}$	105,711,304	105,711,907	604
$I\alpha_{1.1}$	105,712,639	105,712,892	254
Sα1	105,709,501	105,712,500	3,000
IGHG2			
$I\gamma_2$	105,647,803	105,648,549	747
$S\gamma_2$	105,627,501	105,628,500	1,000
IGHG4			
$I\gamma_4$	105,629,000	105,629,500	501
$S\gamma_4$	105,623,001	105,624,000	1,000
IGHE			
Iε	105,605,036	105,605,357	322
Sε	105,602,501	105,604,500	2,000
IGHA2			
$I\alpha_2$	105,589,300	105,590,300	1,001
$S\alpha_2$	105,589,001	105,591,500	2,500

599

600

Table 2. General sequencing run metrics used for CSR-ncRNA expression analysis

	log_2RPKM	**Samples**	**(%)**
Total	NA	70,603	**100.0**
Highly expressed	> 2.65	21,017	**29.8**
Low-expressed	< 2.6, > 0	23,074	**32.1**
Not expressed	0	26,512	**37.1**
Clustering set	> 0	44,091	**62.9**

601

602

603

Table 3. Oligonucleotides used for qRT-PCR of Iμ, Iγ$_3$ and Iγ$_1$

	Forward	Reverse	Probe
Iγ$_1$	CAAGCCCCTCCGTTCAC	ACGAGGAACATGACTGGATG	5-F/AGGAGGCAG/Z/CAGAGCGAGG/3IB
Iγ$_3$	CCTGTTGTGGCGAGGTACA	TTAGTGTTTGCAGCGTGGAG	5-F/CTGTCAGCT/Z/GCCACTTGTCTTCCT/3IB
Iμ	CCAGGTGTTGTTTTGCTCAGT	CACTTCTGGTTGTGAAGAGGTG	UPL_ROCHE#61
AID	TGGACACCACTATGGACAGC	GCGGACATTTTTGAATTGGT	UPL_ROCHE#69

F= FAM; Z = ZEN; IB = Iowa Black

604

605

Figure Legends:

Figure 1. CSRnc transcription boundaries definition. Selected projects using isolated B cells were used to define the limits of CSRnc transcription in human chromosome 14 (See **Additional file 2: Table S1**). The telomeric region is towards the right. *BigWig* files were downloaded using *recount2* and visualized using IGV [75] to inspect to determine the boundaries for each CSRnc transcript. **A**) Coverage graph of project SRP045500 showing the CRIP1 locus (*left panel*) and the *IGHM* locus (*right panel*) in B (*red track*) and non-B cells. The CRIP1 gene is transcribed in B and non-B cells, whereas *IGHM* and I_{μ} (within black vertical lines) is transcribed only in B cells and in peripheral blood (*Black track*). Predicted S_{μ} region is shown in the bottom track (*blue*) and annotated genes (GenCode V24) are shown in *green*. **B**) A view of the *IGHM – IGHD* intron locus displaying coverage graphs from normal purified naive tonsillar (SRP021509, blue track) and peripheral blood $CD19^{+}$ B cells (SRP060715, red track). The vertical black arrow shows I_{δ}. *IGHD* annotation GenCode V24 is shown in *green*. The I_{δ} exon overlaps with the Σμ region [22, 43, 45] and is centromeric regarding the mapped sites for $\mu - \delta$ CSR junctions (Dotted black arrows) [43]. Antisense primer (18156) used by Kluin *et al*. [43] and Arpin, *et al*. (P4) [45] are shown in purple and green, respectively. The S_{δ} Sense primer used by Chen, *et al*. [44] is shown in orange. The *blue asterisks* indicate *Hind* III sites [22, 23, 43]. **C**) Transcriptional landscape of the *IGHA1* locus displaying coverage graphs of the *IGHA1*, GenCode V24 annotation (*green track*), tonsilar naive (SRR834982, blue) and peripheral blood B cells (SRR2097501, red). Both $I_{\alpha 1.1}$ and $I_{\alpha 1.2}$ transcripts are shown. The $S_{\alpha 1}$ region (*black bar*) is shown.

Figure 2. CSRnc transcripts quantitation *in vitro* and in vaccination. A qPCR Taq-Man assay for I_{μ}, $I_{\gamma 3}$ $I_{\gamma 1}$ and AID was used to quantitate CSRnc transcription with the $2^{\Delta\Delta Ct}$ method ($\log_2$). **A-D**) Enriched B cells cultured for 3 and 6 days with T-dependent like activation (IL-21, a-CD40 and CpG) and T-independent activation (PWM, SAC and CpG). Bar plots represent the mean fold-change (non-activated enriched B cells /activated B cells) of two independent experiments. Non-B cells were used as control. **E-H**) qPCR Taq-Man assay for I_{μ}, $I_{\gamma 3}$ $I_{\gamma 1}$ and AID from total RNA obtained from donors' PBMCs taken at pre-immunization (day 0) against Hepatitis B and/or Tetanus-Diphteria and on days 7 and 14

post-immunization. (Wilcoxon test. $p<0.05$). **I**) Plasmablast ($CD3^-CD19^+CD20^-CD27^+CD38^+$) mobilization in peripheral blood expressed as a percentage of $CD19^+$ B cells (Wilcoxon test. $p = 0.005$). **J**) Positive correlation between day 0/7 plasmablast ratio and day 0/7 I_μ ratio (LTS regression method. Adjusted R^2: 0.53, p-value: 0.015). **K**) Negative correlation between day 0/7 plasmablast ratio and day 0/14 $I_{\gamma 3}$ ratio (LTS regression method. Adjusted $R^2 = 0.63$, p-value: 0.011). **L-O**) No significant changes in CSRnc transcription assessed by qPCR were observed 7 and 14 days after trivalent inactivated Influenza vaccination (Wilcoxon test. $P > 0.05$). **L**) I_μ, **M**) $I_{\gamma 3}$, **N**) $I_{\gamma 1}$ and **O**) AID. Dotted red line indicates no change in expression (Fold-change = 1.0).

Figure 3. Quantitative analysis of CSRnc transcription using *recount2*. **A**) Average $\log_2$-transformed RPKM distribution of the 10 I_H exons per sample. There were 44,091 samples with non-zero RPKM. Higher than the mean ($\log_2$RPKM > 2.65) was considered "highly" transcribed (shown in red). "Low" CSRnc transcription was defined as a mean $\log_2$ RPKM < 2.65 (green). **B)** CSRnc transcription profiling by Z-score clustering. Log_2RPKM CSRnc transcription values of GTEx + TCGA + SRA datasets were transformed to standardized Z-scores and subjected to *k*-means clustering with a predefined number of 10 clusters (*Left panel*). Clustered data is represented in a heatmap where I_H's are columns and Z clusters are in rows. Negative Z-scores (i.e. < 2.65 $\log_2$RPKM) are shown in green, positive Z-scores (i.e. > 2.65 $\log_2$RPKM) are shown in red. Absent values (RPKM = 0) are shown in grey. Z values near 0 are shown in black. The pattern expression of each cluster is represented in boxplots in **Additional file 1: Figure S6**.

Figure 4. CSRnc transcription in healthy adult tissues. The GTEx dataset was partitioned according to tissue type. Violin plot of average $\log_2$RPKM distribution of GTEx project representative tissues, ordered form left to right according to each tissue median $\log_2$RPKM. The violin area was scaled according to sample count and median and quartiles are shown. A dotted black line marks the mean average $\log_2$RPKM (2.65). For simplicity, only brain cortex was included as a representative sample for central nervous system.

Figure 5. **Regression analysis of CSRnc transcription as a function of tissue type and I-exon**. Whole blood CSRnc transcription was used as reference tissue (*rows*) for regression analysis according to I-exon (*columns*). The estimate for each comparison is

expressed as a heatmap. Higher CSRnc transcription than in blood is represented in blue tones, whereas lower transcription is shown in pink tones. Zero estimate values are shown in ivory. Missing values (NA's) are shown in white. Euclidian distance was used for hierarchical clustering by row.

Figure 6. **CSRnc transcription profile variation in healthy tissue**. The GTEX dataset was categorized according to tissue (*x* axis) and the relative proportion of samples (y-axis) belonging to each Z cluster (in *colors*). Many central nervous system tissues with highly similar pattern were removed for simplicity. Proportions were hierarchically clustered using Pearson correlation as distance metric with Cluster3.0 [75]. The red dotted line marks a correlation $R^2 = 0.7$. The black asterisk indicates a significant enrichment of a given tissue in a particular Z cluster (Exact Fisher's test, Benjamini-Hochberg adjustment for multiple correction. FDR < 0.01).

Figure 7. CSRnc I_H and C_H transcription in fetal tissues. Heatmap of Z-score average per tissue (n = 3-8) of I_H and C_H (*columns*) in 9-22 weeks of gestation fetal tissues (*rows*). Higher transcription (Z-scores) are shown in green-yellow, lower expression and no expression is shown in purple and gray, respectively. Higher than average expression of $I_{\gamma 4}$ is observed in all tested tissues. Data from this figure was obtained from study SRP055513 [31].

Figure 8. CSRnc transcription in cancer. RNA-seq data from the TCGA project was used to analyze CSRnc transcription in 33 cancer types. Violin plots of average $\log_2$RPKM, ordered from left to right according to increasing median. Violin area is scaled to each tumor sample count. A dotted black line marks the average $\log_2$RPKM = 2.65.

Figure 9. Comparison of CSRnc transcription in healthy tissue and its tumoral counterpart. Violin plots of the average $\log_2$RPKM distribution in healthy tissue (*blue*) in comparison with its cancer tissue counterpart (*purple*). Violin area are not scaled to sample count. Median (*black dot*) and quartiles are shown for each violin. Dashed black line marks the mean average $\log_2$RPKM (2.65). A reduction of CSRnc transcription in tumors was observed in **A-E**. An increase of CSRnc transcription in tumors vs. healthy tissue was observed in **G-K**. Some types of kidney cancers and melanomas showed an opposite patterns regarding the healthy tissue counterpart (**F** and **L**). The conducted statistical test

were Wilcoxon rank sum test with continuity correction for two-sample comparisons, and Kruskal-Wallis test with *post hoc* Dunn's test correction for multiple comparisons.

Figure 10. Increased I_δ transcription in ileal mucosa of Crohn's disease. Pediatric Crohn's disease patients from project SRP042228 were classified in two groups according to ileal mucosa involvement (iCD) on non-involvement (cCD) [37]. Patients were further classified according to the degree of ileal mucosa inflammation. Ileal mucosa from non-inflammatory bowel disease (not-IBD) and ulcerative colitis (UC) were used as controls [37]. **A**) Boxplot showing increased Id transcription (RPKM, *y axis*) in deep ulcerated Crohn's disease with ileal involvement (one factor ANOVA, Tukey multiple comparisons of means. $P < 0.005$). **B**) Transcriptional landscape of the *IGHD-IGHM* locus in human chromosome 14 showing increased I_δ transcription in representative samples of deep ulcerated iCD (*red track*), cCD with microscopic inflammation (*blue track*), iCD with macroscopic inflammation (*purple track*), non-IBD (*black track*) and UC (*orange track*).

Additional files

Additional file 1: Figure S1, Distribution of the number of samples per RNA-seq project analyzed. **Figure S2**, CSRnc transcription is B cells-specific. **Figure S3,** CSRnc transcription pattern in whole blood is similar to peripheral blood sorted $CD19^+$ B cells. **Figure S4**, Epigenetic marks in I_δ. **Figure S5**, CSRnc transcription according to I_H and project dataset. Figure S6, CSRnc transcriptional profiles identified by *k*-means clustering of the recount2 dataset **Figure S7**, Correlation between CSRnc I_H transcription and coding C_H transcription. **Figure S8**, CSRnc and C_H coding transcription in Bone marrow lymphoid precursors. **Figure S9**, CSRnc and C_H coding transcription in thymic lymphoid precursors. **Figure S10**, Comparison of CSRnc transcription in healthy tissue and its tumor counterpart (PDF).

Additional file 2: **Table S1**, Selected SRA projects used to map I_H boundaries. **Table S2**, CSRnc transcription analysis in cancer cell lines. **Table S3**, Differential expression analysis in I_H and C_H. (XLSX).

Ethics approval and consent to participate

The participants in this study did so voluntarily after written consent. The study was approved by the INSP Institutional Review Board (CI: 971/82-6684).

Competing interest

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank Analí Migueles and Everardo Millan for their support in data pre-processing, and Fernando Riveros Mackay for differential expression analysis; Leopoldo Santos-Argumedo for kindly donating αCD40 hybridoma G28.5), Martha Patricia Rojo for support on flow cytometry; Tomás Salmerón Enciso, Yvonne Rosenstein, Fernando Esquivel and José Moreno for academic input, and Menno Van Zelm for critically reviewing the manuscript. HKM is a PhD student of Programa de Doctorado en Ciencias Biomédicas, UNAM, and was supported by a CONACyT scholarship.

Abbreviations

AID: Activated-Induced Cytidine Deaminase; CD: Crohn's disease; CSRnc: Class Switch Recombination non-coding; CSR: Class Switch Recombination; CG: Germinal Center; C_H: Heavy chain Constant; ELS: Ectopic Lymphoid Structures; EBV: Epstein Barr Virus; GTEx: Genotype-Tissue Expression Project; I_H: I exon; MALT: Mucosal Associated lymphoid Tissue; PAMP: Pathogen-Associated Molecular Pattern; PB: Peripheral blood; SRA: Sequence Read Archive; RPKM: Reads Per Kilobase (transcript) Per Million (reads); TCGA: The Cancer Genome Atlas.

Funding:

This research was funded by the Consejo Nacional de Ciencia y Tecnología (CONACyT) /Fondo Sectorial de Investigación en Salud y Seguridad Social (FOSISSS) grant # 142120;

HKM is a PhD student of Programa de Doctorado en Ciencias Biomédicas, UNAM, and was supported by a CONACyT scholarship.

Authors' contributions

This study was conceived and designed by HKM and JMB. The experimental procedures were performed by HKM, MOM, HVT and JMTS. Flow cytometry acquisition and analysis was conducted by HKM and LBA. RNA-seq data bioinformatics and statistical analysis were done by HKM, LCT and JMB, and supervised by AJ. HKM and JMB drafted the manuscript. LBA, LCT and AJ critically reviewed the manuscript. All authors read and approved the final manuscript.

References:

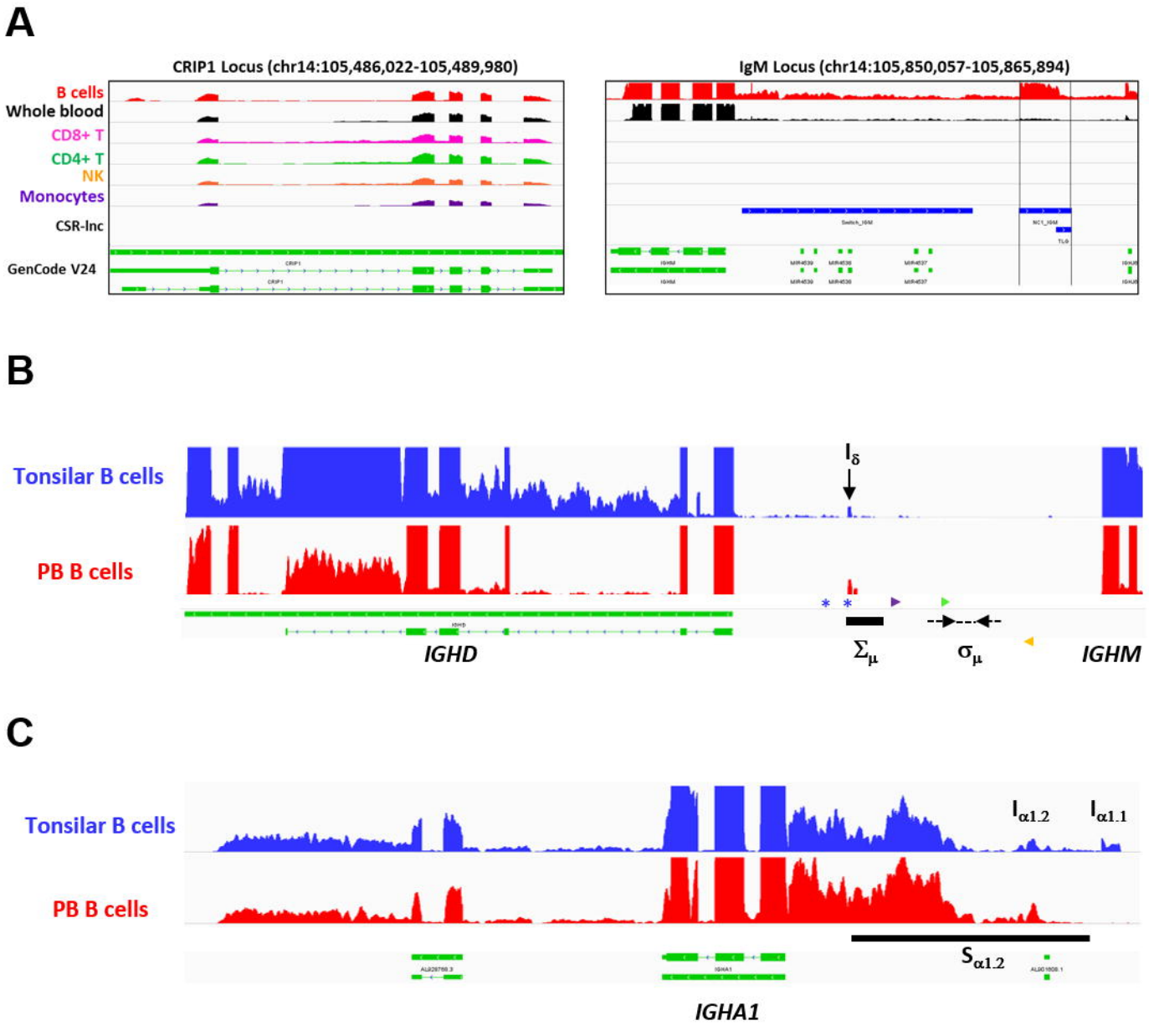
1. Victora GD, Nussenzweig MC: **Germinal centers.** *Annu Rev Immunol* 2012, **30:**429-457.
2. Taketoshi Yoshida HM, Thomas Dorner, Falk Hiepe, Andreas Radbruch, Simon Fillatreau, Bimba F. Hoyer: **Memory B and memory plasma cells.** *Immunological Reviews* 2010, **237:**117-139.
3. Xu Z, Zan H, Pone EJ, Mai T, Casali P: **Immunoglobulin class-switch DNA recombination: induction, targeting and beyond.** *Nat Rev Immunol* 2012, **12:**517-531.
4. Lorenz M, Jung S, Radbruch A: **Switch transcripts in immunoglobulin class switching.** *Science* 1995, **267:**1825-1828.
5. Chowdhury M, Forouhi O, Dayal S, McCloskey N, Gould HJ, Felsenfeld G, Fear DJ: **Analysis of intergenic transcription and histone modification across the human immunoglobulin heavy-chain locus.** *Proc Natl Acad Sci U S A* 2008, **105:**15872-15877.
6. Yewdell WT, Chaudhuri J: **A transcriptional serenAID: the role of noncoding RNAs in class switch recombination.** *Int Immunol* 2017, **29:**183-196.
7. Alt FW, Zhang Y, Meng FL, Guo C, Schwer B: **Mechanisms of programmed DNA lesions and genomic instability in the immune system.** *Cell* 2013, **152:**417-429.
8. Consortium GT: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45:**580-585.
9. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, et al: **Genetic effects on gene expression across human tissues.** *Nature* 2017, **550:**204-213.
10. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45:**1113-1120.
11. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al: **Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.** *Cell* 2018, **173:**291-304 e296.
12. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT: **Reproducible RNA-seq analysis using recount2.** *Nat Biotechnol* 2017, **35:**319-321.

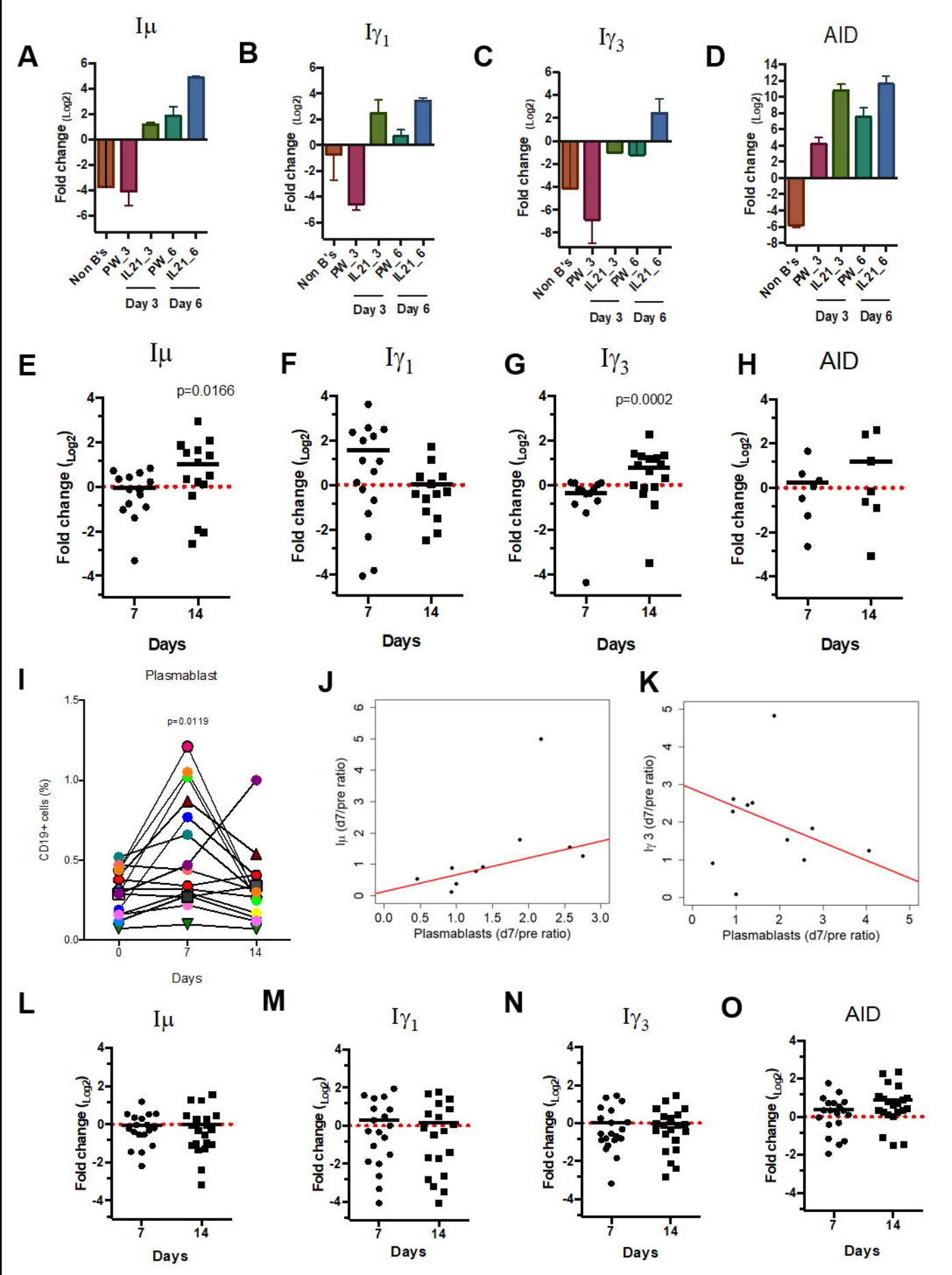
13. Litinskiy MB, Nardelli B, Hilbert DM, He B, Schaffer A, Casali P, Cerutti A: **DCs induce CD40-independent immunoglobulin class switching through BLyS and APRIL.** *Nat Immunol* 2002, **3:**822-829.
14. Fear DJ, McCloskey N, O'Connor B, Felsenfeld G, Gould HJ: **Transcription of Ig Germline Genes in Single Human B Cells and the Role of Cytokines in Isotype Determination.** *The Journal of Immunology* 2004, **173:**4529-4538.
15. Henn AD, Wu S, Qiu X, Ruda M, Stover M, Yang H, Liu Z, Welle SL, Holden-Wiltse J, Wu H, Zand MS: **High-resolution temporal response patterns to influenza vaccine reveal a distinct human plasma cell gene signature.** *Sci Rep* 2013, **3:**2327.
16. Beguelin W, Popovic R, Teater M, Jiang Y, Bunting KL, Rosen M, Shen H, Yang SN, Wang L, Ezponda T, et al: **EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation.** *Cancer Cell* 2013, **23:**677-692.
17. Pascual M, Roa S, Garcia-Sanchez A, Sanz C, Hernandez-Hernandez L, Greally JM, Lorente F, Davila I, Isidoro-Garcia M: **Genome-wide expression profiling of B lymphocytes reveals IL4R increase in allergic asthma.** *J Allergy Clin Immunol* 2014, **134:**972-975.
18. Linsley PS, Speake C, Whalen E, Chaussabel D: **Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis.** *PLoS One* 2014, **9:**e109760.
19. Koues OI, Kowalewski RA, Chang LW, Pyfrom SC, Schmidt JA, Luo H, Sandoval LE, Hughes TB, Bednarski JJ, Cashen AF, et al: **Enhancer sequence variants and transcription-factor deregulation synergize to construct pathogenic regulatory circuits in B-cell lymphoma.** *Immunity* 2015, **42:**186-198.
20. Hoek KL, Samir P, Howard LM, Niu X, Prasad N, Galassie A, Liu Q, Allos TM, Floyd KA, Guo Y, et al: **A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination.** *PLoS One* 2015, **10:**e0118528.
21. Kushwaha G, Dozmorov M, Wren JD, Qiu J, Shi H, Xu D: **Hypomethylation coordinates antagonistically with hypermethylation in cancer development: a case study of leukemia.** *Hum Genomics* 2016, **10 Suppl 2:**18.
22. Akahori Y, Handa H, Imai K, Abe M, Kameyama K, Hibiya M, Yasui H, Okamura K, Naito M, Matsuoka H, et al.: **Sigma region located between C mu and C delta genes of human immunoglobulin heavy chain: possible involvement of tRNA-like structure in RNA splicing.** *Nucleic Acids Res* 1988, **16:**9497-9511.
23. Yasui H, Akahori Y, Hirano M, Yamada K, Kurosawa Y: **Class switch from mu to delta is mediated by homologous recombination between sigma mu and sigma mu sequences in human immunoglobulin gene loci.** *Eur J Immunol* 1989, **19:**1399-1403.
24. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al: **Ensembl 2018.** *Nucleic Acids Res* 2018, **46:**D754-D761.
25. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518:**317-330.
26. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
27. Wrammert J, Smith K, Miller J, Langley WA, Kokko K, Larsen C, Zheng NY, Mays I, Garman L, Helms C, et al: **Rapid cloning of high-affinity human monoclonal antibodies against influenza virus.** *Nature* 2008, **453:**667-671.
28. Odendahl M, Mei H, Hoyer BF, Jacobi AM, Hansen A, Muehlinghaus G, Berek C, Hiepe F, Manz R, Radbruch A, Dorner T: **Generation of migratory antigen-specific plasma blasts and mobilization of resident plasma cells in a secondary immune response.** *Blood* 2005, **105:**1614-1621.

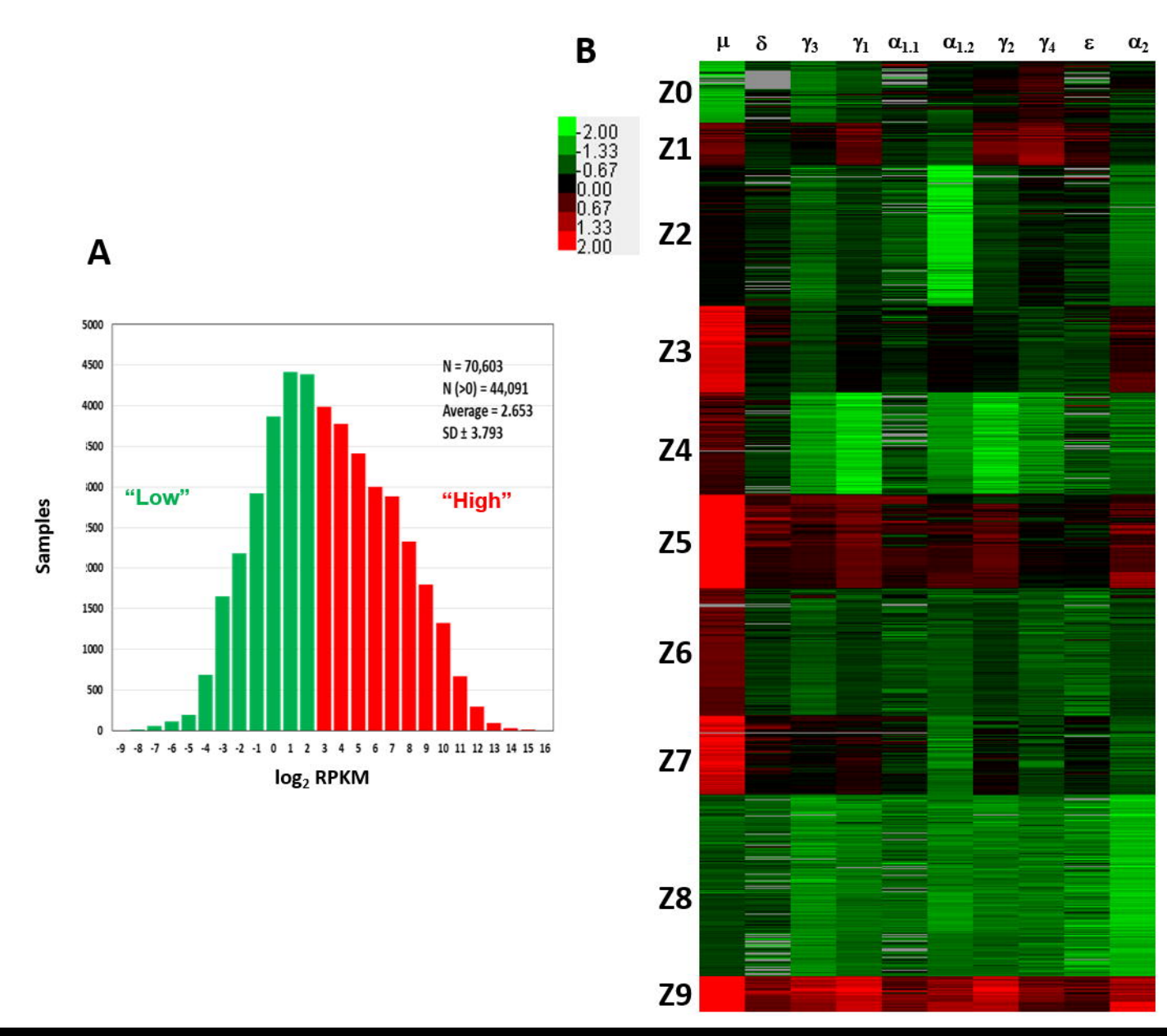
29. Frolich D, Giesecke C, Mei HE, Reiter K, Daridon C, Lipsky PE, Dorner T: **Secondary immunization generates clonally related antigen-specific plasma cells and memory B cells.** *J Immunol* 2010, **185:**3103-3110.
30. Casero D, Sandoval S, Seet CS, Scholes J, Zhu Y, Ha VL, Luong A, Parekh C, Crooks GM: **Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages.** *Nat Immunol* 2015, **16:**1282-1291.
31. Roost MS, van Iperen L, Ariyurek Y, Buermans HP, Arindrarto W, Devalla HD, Passier R, Mummery CL, Carlotti F, de Koning EJ, et al: **KeyGenes, a Tool to Probe Tissue Differentiation Using a Human Fetal Transcriptional Atlas.** *Stem Cell Reports* 2015, **4:**1112-1124.
32. Colbeck EJ, Ager A, Gallimore A, Jones GW: **Tertiary Lymphoid Structures in Cancer: Drivers of Antitumor Immunity, Immunosuppression, or Bystander Sentinels in Disease?** *Front Immunol* 2017, **8:**1830.
33. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, et al: **Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data.** *Nucleic Acids Res* 2015, **43:**D1071-1078.
34. Tran TM, Jones MB, Ongoiba A, Bijker EM, Schats R, Venepally P, Skinner J, Doumbo S, Quinten E, Visser LG, et al: **Transcriptomic evidence for modulation of host inflammatory responses during febrile Plasmodium falciparum malaria.** *Sci Rep* 2016, **6:**31291.
35. Mei B, Ding X, Xu HZ, Wang MT: **Global gene expression changes in human peripheral blood after H7N9 infection.** *Gene* 2014, **551:**255-260.
36. Hung T, Pratt GA, Sundararaman B, Townsend MJ, Chaivorapol C, Bhangale T, Graham RR, Ortmann W, Criswell LA, Yeo GW, Behrens TW: **The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression.** *Science* 2015, **350:**455-459.
37. Haberman Y, Tickle TL, Dexheimer PJ, Kim MO, Tang D, Karns R, Baldassano RN, Noe JD, Rosh J, Markowitz J, et al: **Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature.** *J Clin Invest* 2014, **124:**3617-3633.
38. Kinoshita K, Harigai M, Fagarasan S, Muramatsu M, Honjo T: **A hallmark of active class switch recombination: transcripts directed by I promoters on looped-out circular DNAs.** *Proc Natl Acad Sci U S A* 2001, **98:**12620-12623.
39. Iwasato T, Shimizu A, Honjo T, Yamagishi H: **Circular DNA is excised by immunoglobulin class switch recombination.** *Cell* 1990, **62:**143-149.
40. Matsuoka M, Yoshida K, Maeda T, Usuda S, Sakano H: **Switch circular DNA formed in cytokine-treated mouse splenocytes: evidence for intramolecular DNA deletion in immunoglobulin class switching.** *Cell* 1990, **62:**135-142.
41. von Schwedler U, Jack HM, Wabl M: **Circular DNA is a product of the immunoglobulin class switch rearrangement.** *Nature* 1990, **345:**452-456.
42. Rouaud P, Saintamand A, Saad F, Carrion C, Lecardeur S, Cogne M, Denizot Y: **Elucidation of the enigmatic IgD class-switch recombination via germline deletion of the IgH 3' regulatory region.** *J Exp Med* 2014, **211:**975-985.
43. Kluin PM, Kayano H, Zani VJ, Kluin-Nelemans HC, Tucker PW, Satterwhite E, Dyer MJ: **IgD class switching: identification of a novel recombination site in neoplastic and normal B cells.** *Eur J Immunol* 1995, **25:**3504-3508.
44. Chen K, Xu W, Wilson M, He B, Miller NW, Bengten E, Edholm ES, Santini PA, Rath P, Chiu A, et al: **Immunoglobulin D enhances immune surveillance by activating antimicrobial, proinflammatory and B cell-stimulating programs in basophils.** *Nat Immunol* 2009, **10:**889-898.

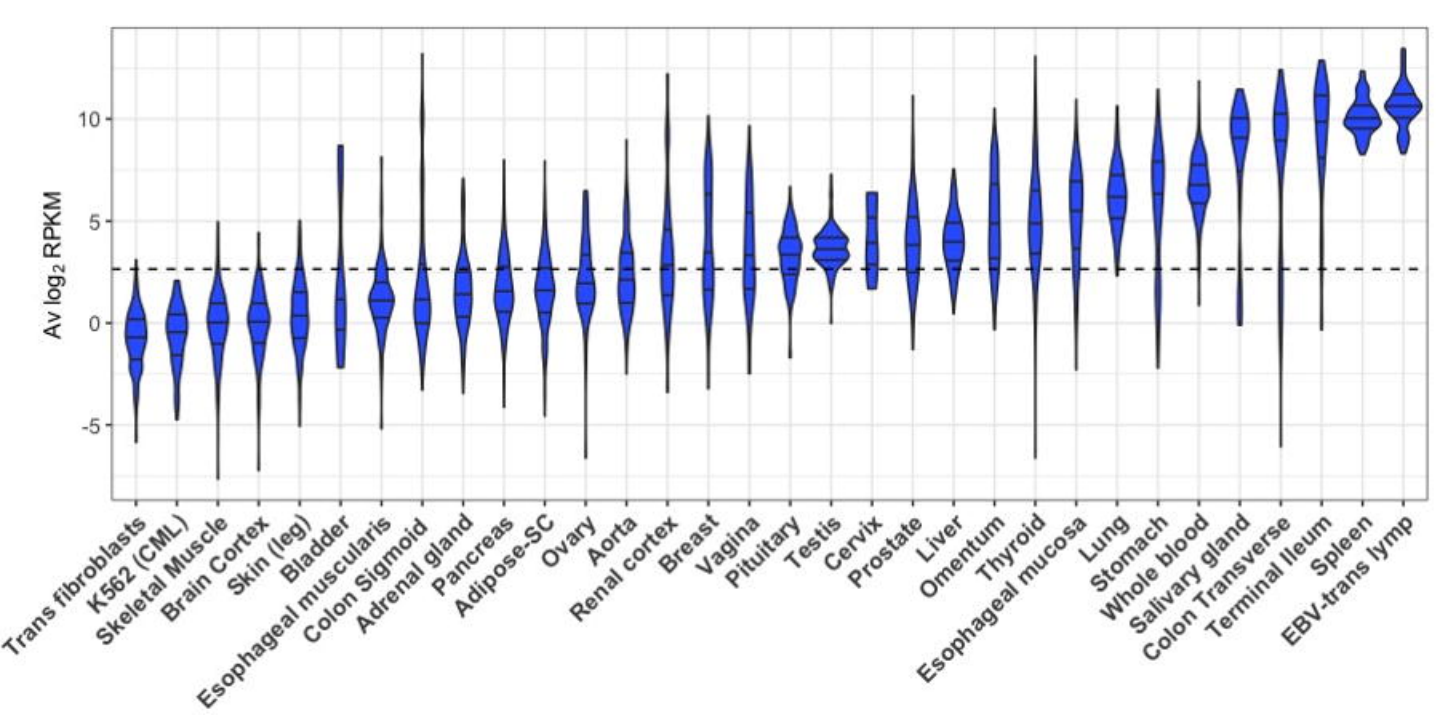
45. Arpin C, de Bouteiller O, Razanajaona D, Fugier-Vivier I, Briere F, Banchereau J, Lebecque S, Liu YJ: **The normal counterpart of IgD myeloma cells in germinal center displays extensively mutated IgVH gene, Cmu-Cdelta switch, and lambda light chain expression.** *J Exp Med* 1998, **187:**1169-1178.
46. Guthmiller JJ, Wilson PC: **Harnessing immune history to combat influenza viruses.** *Curr Opin Immunol* 2018, **53:**187-195.
47. Li SC, Rothman PB, Zhang J, Chan C, Hirsh D, Alt FW: **Expression of I mu-C gamma hybrid germline transcripts subsequent to immunoglobulin heavy chain class switching.** *Int Immunol* 1994, **6:**491-497.
48. Bolland DJ, Wood AL, Afshar R, Featherstone K, Oltz EM, Corcoran AE: **Antisense intergenic transcription precedes Igh D-to-J recombination and is controlled by the intronic enhancer Emu.** *Mol Cell Biol* 2007, **27:**5523-5533.
49. Afshar R, Pierce S, Bolland DJ, Corcoran A, Oltz EM: **Regulation of IgH gene assembly: role of the intronic enhancer and 5'DQ52 region in targeting DHJH recombination.** *J Immunol* 2006, **176:**2439-2447.
50. Perlot T, Alt FW, Bassing CH, Suh H, Pinaud E: **Elucidation of IgH intronic enhancer functions via germ-line deletion.** *Proc Natl Acad Sci U S A* 2005, **102:**14362-14367.
51. Winer DA, Winer S, Chng MH, Shen L, Engleman EG: **B Lymphocytes in obesity-related adipose tissue inflammation and insulin resistance.** *Cell Mol Life Sci* 2014, **71:**1033-1043.
52. Chen Q, Deng T, Han D: **Testicular immunoregulation and spermatogenesis.** *Semin Cell Dev Biol* 2016, **59:**157-165.
53. Kolkhir P, Metz M, Altrichter S, Maurer M: **Comorbidity of chronic spontaneous urticaria and autoimmune thyroid diseases: A systematic review.** *Allergy* 2017, **72:**1440-1460.
54. Aalberse RC, Stapel SO, Schuurman J, Rispens T: **Immunoglobulin G4: an odd antibody.** *Clin Exp Allergy* 2009, **39:**469-477.
55. Jonsson S, Sveinbjornsson G, de Lapuente Portilla AL, Swaminathan B, Plomp R, Dekkers G, Ajore R, Ali M, Bentlage AEH, Elmer E, et al: **Identification of sequence variants influencing immunoglobulin levels.** *Nat Genet* 2017, **49:**1182-1191.
56. Pitzalis C, Jones GW, Bombardieri M, Jones SA: **Ectopic lymphoid-like structures in infection, cancer and autoimmunity.** *Nat Rev Immunol* 2014, **14:**447-462.
57. Germain C, Gnjatic S, Tamzalit F, Knockaert S, Remark R, Goc J, Lepelley A, Becht E, Katsahian S, Bizouard G, et al: **Presence of B cells in tertiary lymphoid structures is associated with a protective immunity in patients with lung cancer.** *Am J Respir Crit Care Med* 2014, **189:**832-844.
58. Gu-Trantien C, Loi S, Garaud S, Equeter C, Libin M, de Wind A, Ravoet M, Le Buanec H, Sibille C, Manfouo-Foutsop G, et al: **CD4(+) follicular helper T cell infiltration predicts breast cancer survival.** *J Clin Invest* 2013, **123:**2873-2892.
59. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, Angell H, Fredriksen T, Lafontaine L, Berger A, et al: **Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer.** *Immunity* 2013, **39:**782-795.
60. Qian J, Wang Q, Dose M, Pruett N, Kieffer-Kwon KR, Resch W, Liang G, Tang Z, Mathe E, Benner C, et al: **B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity.** *Cell* 2014, **159:**1524-1537.
61. Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, Rada C, Stratton MR, Neuberger MS: **DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis.** *Elife* 2013, **2:**e00534.
62. Teng MW, Galon J, Fridman WH, Smyth MJ: **From mice to humans: developments in cancer immunoediting.** *J Clin Invest* 2015, **125:**3338-3346.

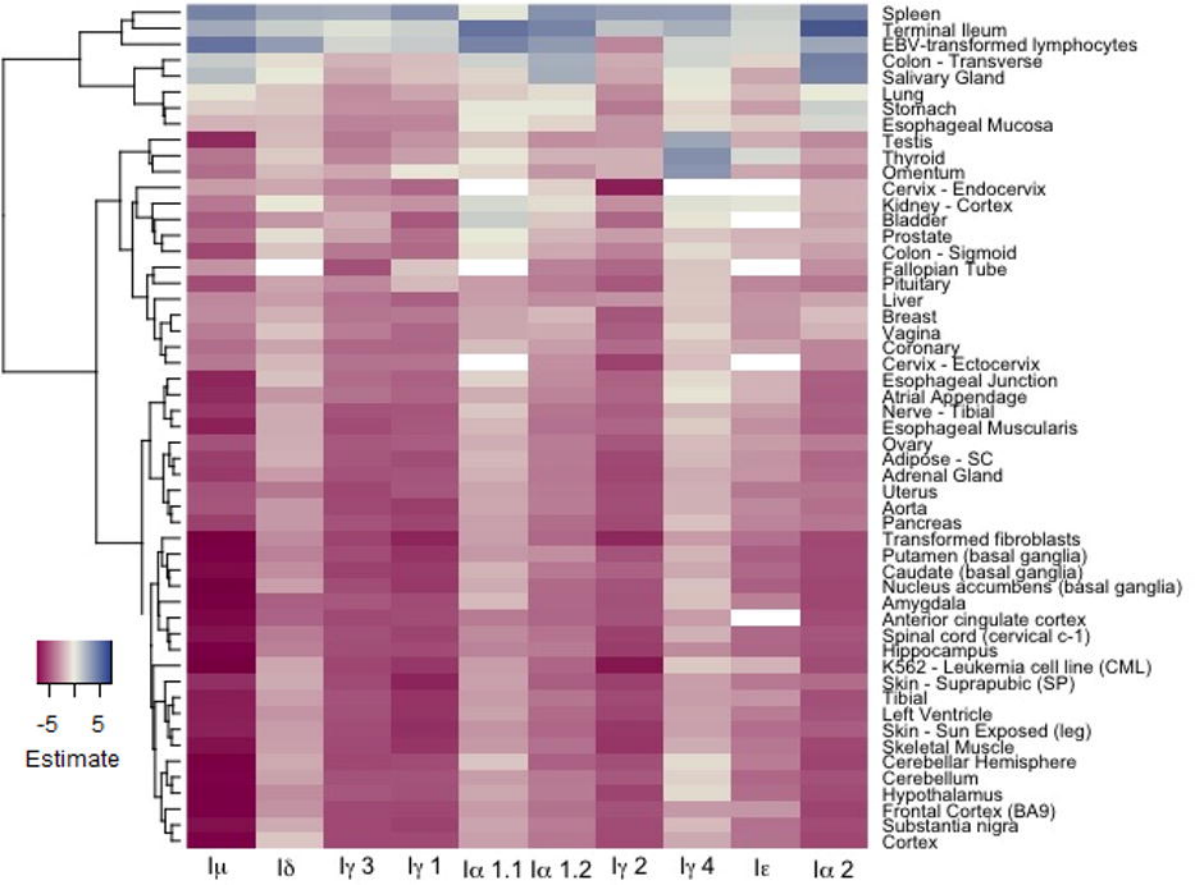
63. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, et al: **Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images.** *Cell Rep* 2018, **23:**181-193 e187.
64. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al: **The Immune Landscape of Cancer.** *Immunity* 2018, **48:**812-830 e814.
65. Cianci R, Lolli S, Pagliari D, Gambassi G, Frosali S, Marmo R, Melioli G, Orlando A, Newton EE, Serone E, et al: **The involvement of IgH enhancer HS1.2 in the pathogenesis of Crohn's disease: how the immune system can influence a multifactorial disease.** *Eur Rev Med Pharmacol Sci* 2016, **20:**3618-3627.
66. Masters SL, Simon A, Aksentijevich I, Kastner DL: **Horror autoinflammaticus: the molecular pathophysiology of autoinflammatory disease (*).** *Annu Rev Immunol* 2009, **27:**621-668.
67. Koelsch K, Zheng NY, Zhang Q, Duty A, Helms C, Mathias MD, Jared M, Smith K, Capra JD, Wilson PC: **Mature B cells class switched to IgD are autoreactive in healthy individuals.** *J Clin Invest* 2007, **117:**1558-1565.
68. Zheng NY, Wilson K, Wang X, Boston A, Kolar G, Jackson SM, Liu YJ, Pascual V, Capra JD, Wilson PC: **Human immunoglobulin selection associated with class switch and possible tolerogenic origins for C delta class-switched B cells.** *J Clin Invest* 2004, **113:**1188-1201.
69. Mannoor K, Xu Y, Chen C: **Natural autoantibodies and associated B cells in immunity and autoimmunity.** *Autoimmunity* 2013, **46:**138-147.
70. Choi JH, Wang KW, Zhang D, Zhan X, Wang T, Bu CH, Behrendt CL, Zeng M, Wang Y, Misawa T, et al: **IgD class switching is initiated by microbiota and limited to mucosa-associated lymphoid tissue in mice.** *Proc Natl Acad Sci U S A* 2017, **114:**E1196-E1204.
71. Team; RC: **R: A Language and Environment for Statistical Computing.** (Computing RFfS ed.; 2018.
72. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform* 2013, **14:**178-192.
73. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol* 2013, **9:**e1003118.
74. Bernstein MN, Doan A, Dewey CN: **MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive.** *Bioinformatics* 2017, **33:**2914-2923.
75. de Hoon MJ, Imoto S, Nolan J, Miyano S: **Open source clustering software.** *Bioinformatics* 2004, **20:**1453-1454.
76. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015, **43:**e47.

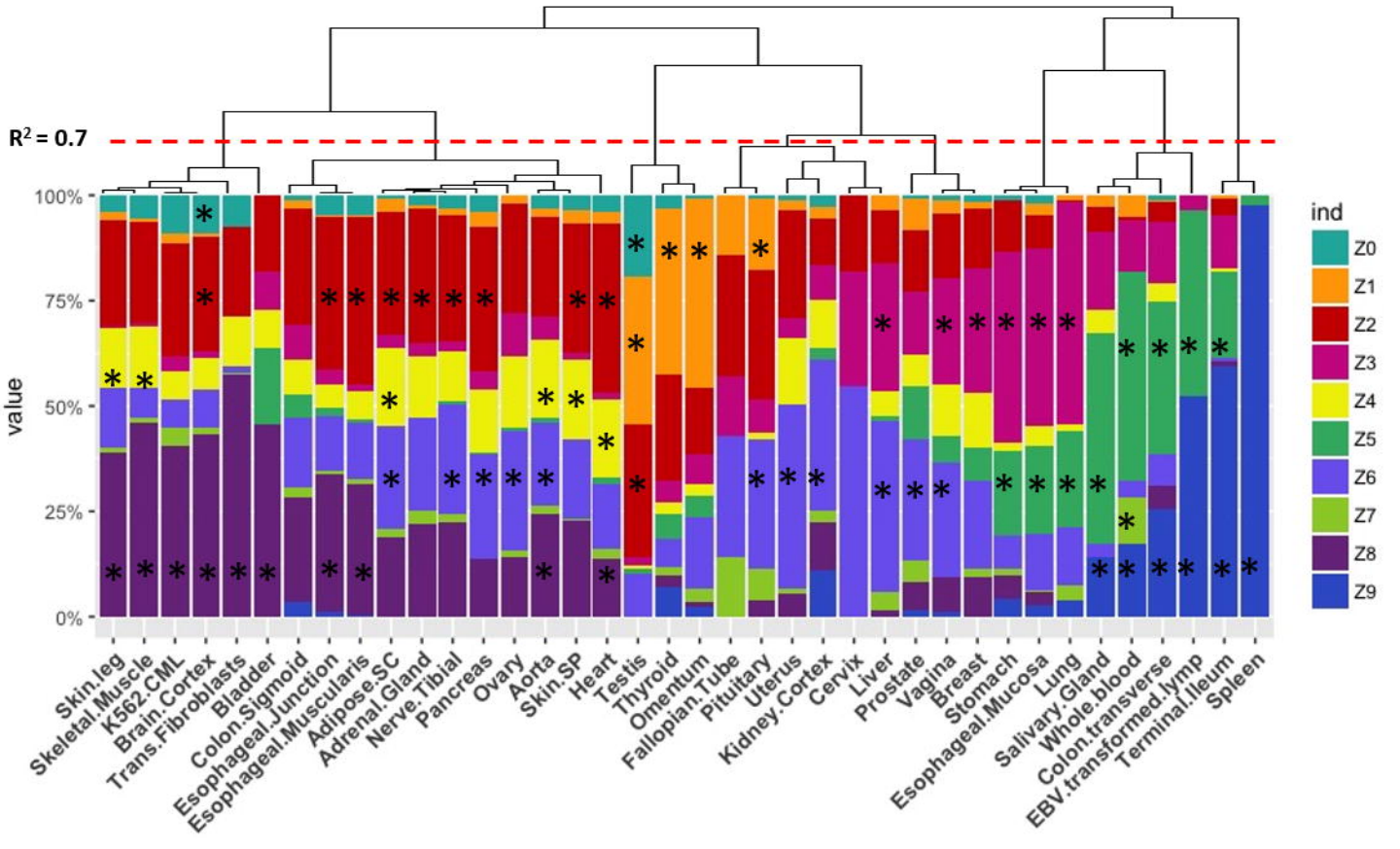

A
CRIP1 Locus (chr14:105,486,022-105,489,980)
IgM Locus (chr14:105,850,057-105,865,894)
B cells
Whole blood
CD8+ T
CD4+ T
NK
Monocytes
CSR-Inc
GenCode V24
B
Tonsilar B cells
PB B cells
Iδ
IGHD
Σμ
σμ
IGHM
C
Tonsilar B cells
PB B cells
Iα1.2
Iα1.1
Sα1.2
IGHA1

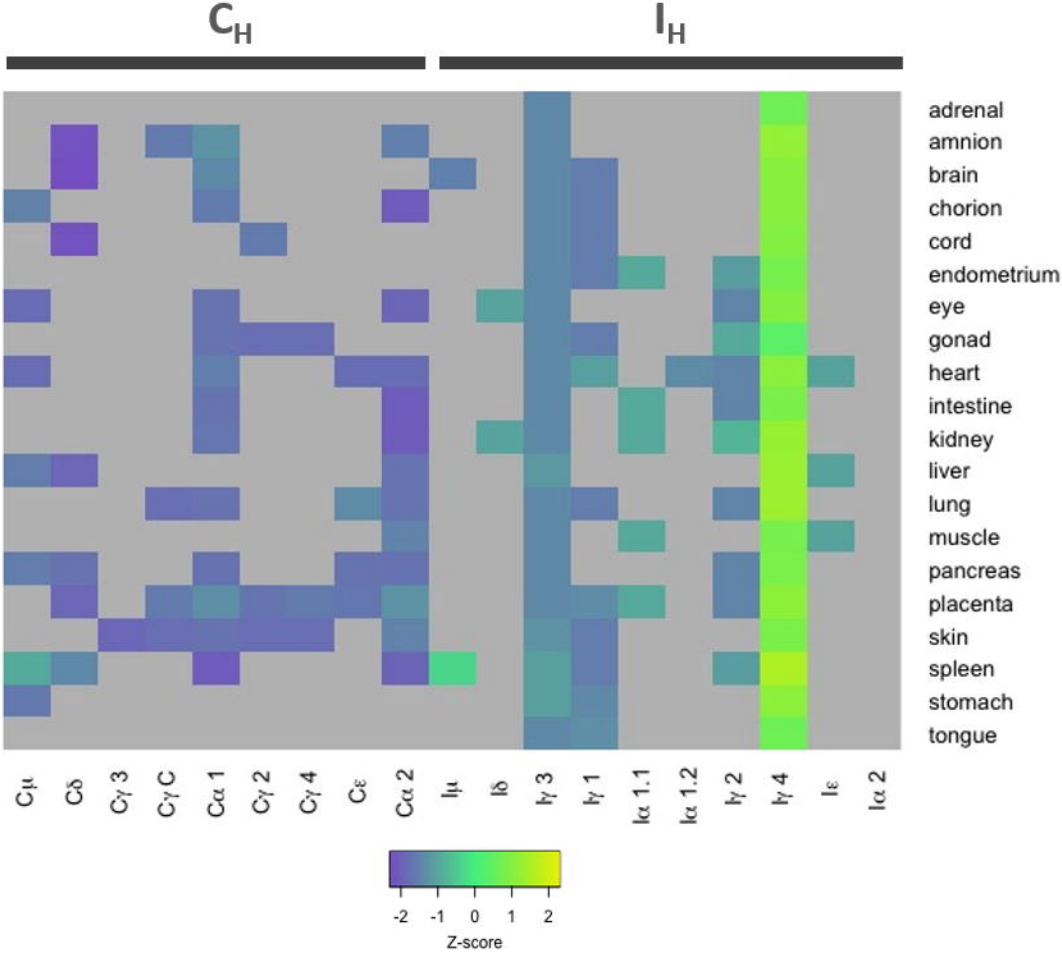

A
Iμ
Fold change (Log2)
Non B's
PW_3
IL21_3
PW_6
IL21_6
Day 3
Day 6
B
Iγ1
C
Iγ3
D
AID
E
Iμ
p=0.0166
Fold change (Log2)
Days
F
Iγ1
G
Iγ3
p=0.0002
H
AID
I
Plasmablast
p=0.0119
CD19+ cells (%)
J
Iμ (d7/pre ratio)
Plasmablasts (d7/pre ratio)
K
Iγ 3 (d7/pre ratio)
L
Iμ
M
Iγ1
N
Iγ3
O
AID

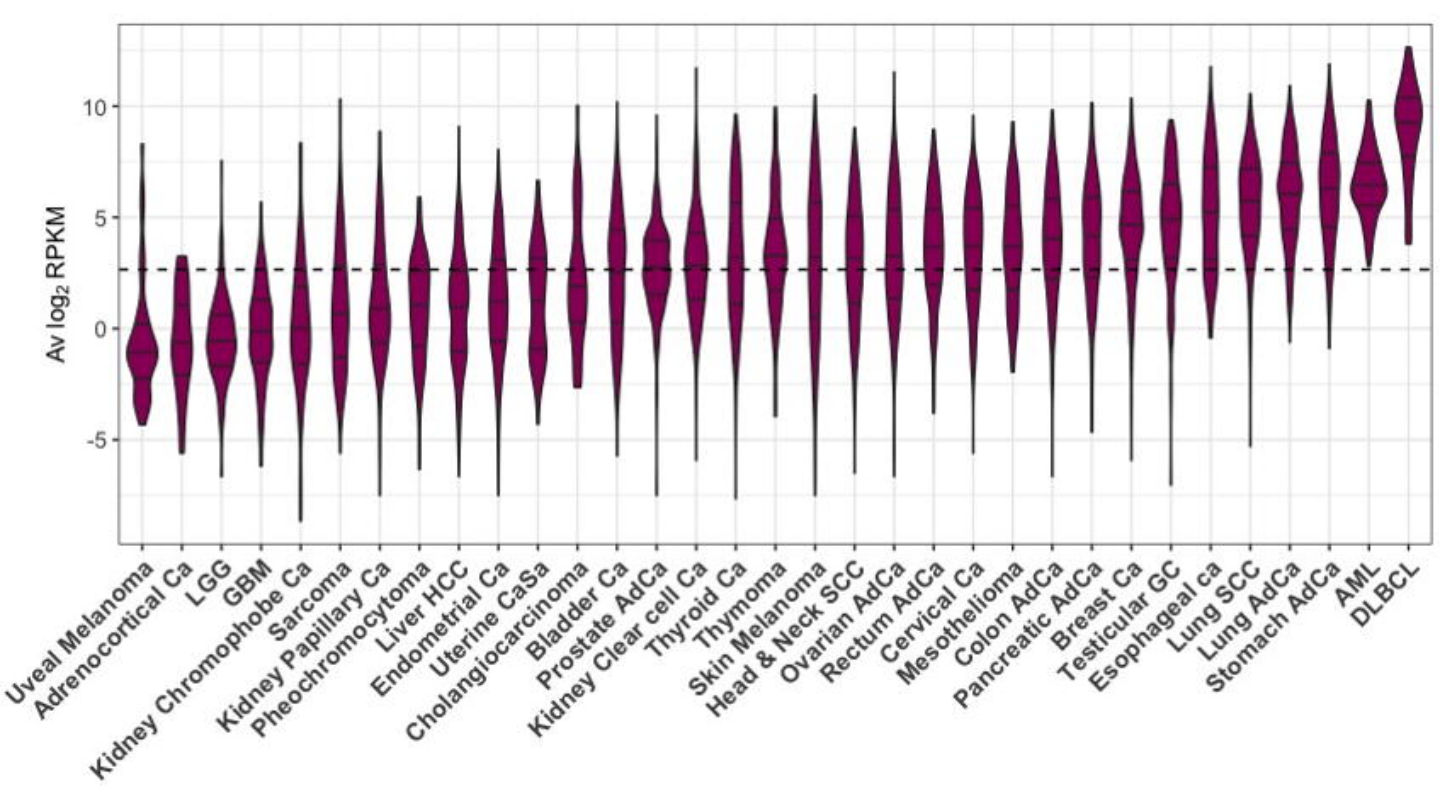

A
B
Samples
log2 RPKM
N = 70,603
N (>0) = 44,091
Average = 2.653
SD ± 3.793
"Low"
"High"
μ δ γ3 γ1 α1.1 α1.2 γ2 γ4 ε α2
Z0
Z1
Z2
Z3
Z4
Z5
Z6
Z7
Z8
Z9
2.00
1.33
0.67
0.00
0.67
1.33
2.00

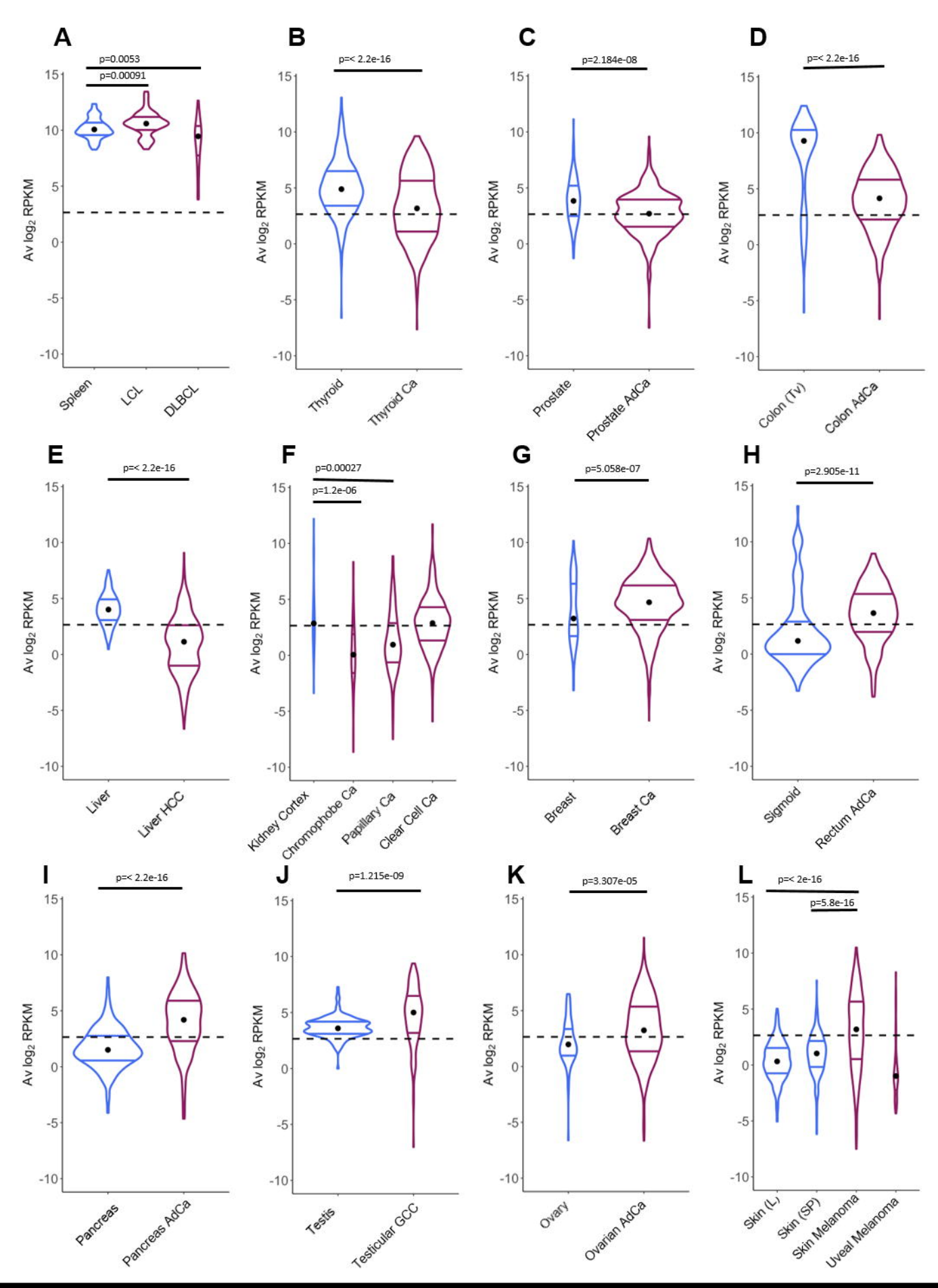

Av log2 RPKM
10
5
0
-5
Trans fibroblasts
K562 (CML)
Skeletal Muscle
Brain Cortex
Skin (leg)
Bladder
Esophageal muscularis
Colon Sigmoid
Adrenal gland
Pancreas
Adipose-SC
Ovary
Aorta
Renal cortex
Breast
Vagina
Pituitary
Testis
Cervix
Prostate
Liver
Omentum
Thyroid
Esophageal mucosa
Lung
Stomach
Whole blood
Salivary gland
Colon Transverse
Terminal Ileum
Spleen
EBV-trans lymp


Spleen
Terminal Ileum
EBV-transformed lymphocytes
Colon - Transverse
Salivary Gland
Lung
Stomach
Esophageal Mucosa
Testis
Thyroid
Omentum
Cervix - Endocervix
Kidney - Cortex
Bladder
Prostate
Colon - Sigmoid
Fallopian Tube
Pituitary
Liver
Breast
Vagina
Coronary
Cervix - Ectocervix
Esophageal Junction
Atrial Appendage
Nerve - Tibial
Esophageal Muscularis
Ovary
Adipose - SC
Adrenal Gland
Uterus
Aorta
Pancreas
Transformed fibroblasts
Putamen (basal ganglia)
Caudate (basal ganglia)
Nucleus accumbens (basal ganglia)
Amygdala
Anterior cingulate cortex
Spinal cord (cervical c-1)
Hippocampus
K562 - Leukemia cell line (CML)
Skin - Suprapubic (SP)
Tibial
Left Ventricle
Skin - Sun Exposed (leg)
Skeletal Muscle
Cerebellar Hemisphere
Cerebellum
Hypothalamus
Frontal Cortex (BA9)
Substantia nigra
Cortex
-5
5
Estimate
Iμ
Iδ
Iγ 3
Iγ 1
Iα 1.1
Iα 1.2
Iγ 2
Iγ 4
Iε
Iα 2

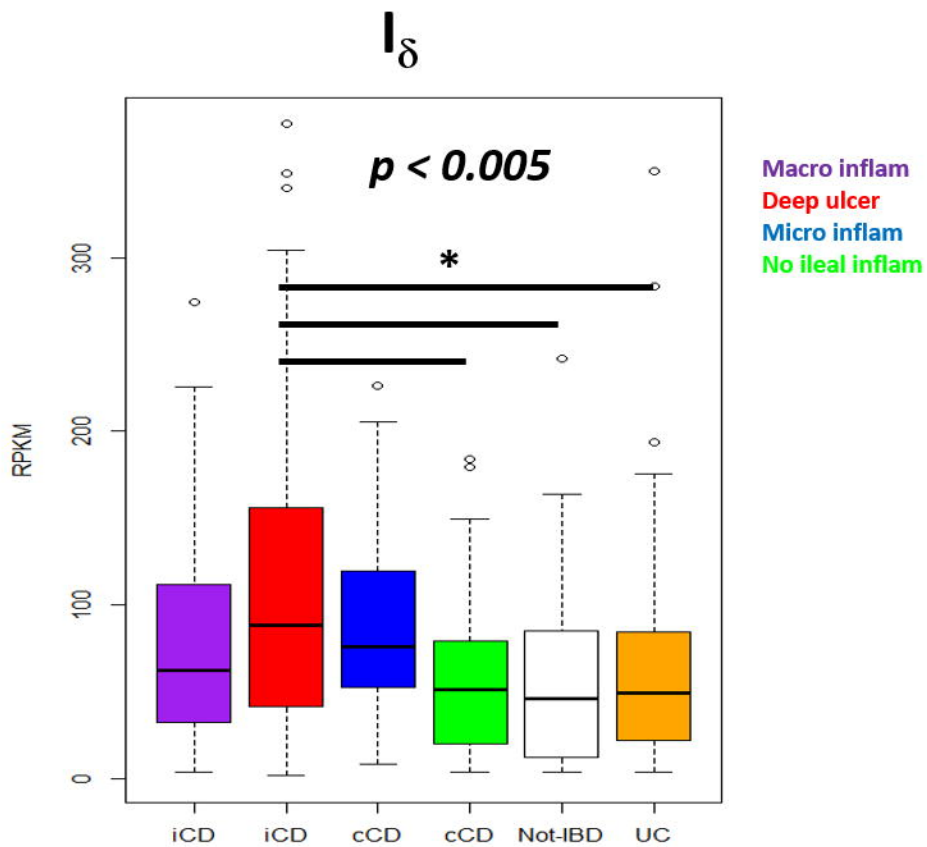

R2 = 0.7
100%
75%
50%
25%
0%
value
ind
Z0
Z1
Z2
Z3
Z4
Z5
Z6
Z7
Z8
Z9
Skin.leg
Skeletal.Muscle
K562.CML
Brain.Cortex
Trans.Fibroblasts
Bladder
Colon.Sigmoid
Esophageal.Junction
Esophageal.Muscularis
Adipose.SC
Adrenal.Gland
Nerve.Tibial
Pancreas
Ovary
Aorta
Skin.SP
Heart
Testis
Thyroid
Omentum
Fallopian.Tube
Pituitary
Uterus
Kidney.Cortex
Cervix
Liver
Prostate
Vagina
Breast
Stomach
Esophageal.Mucosa
Lung
Salivary.Gland
Whole.blood
Colon.transverse
EBV.transformed.lymp
Terminal.Ileum
Spleen


C_H
I_H
adrenal
amnion
brain
chorion
cord
endometrium
eye
gonad
heart
intestine
kidney
liver
lung
muscle
pancreas
placenta
skin
spleen
stomach
tongue
Cμ
Cδ
Cγ 3
Cγ C
Cα 1
Cγ 2
Cγ 4
Cε
Cα 2
Iμ
Iδ
Iγ 3
Iγ 1
Iα 1.1
Iα 1.2
Iγ 2
Iγ 4
Iε
Iα 2
-2 -1 0 1 2
Z-score

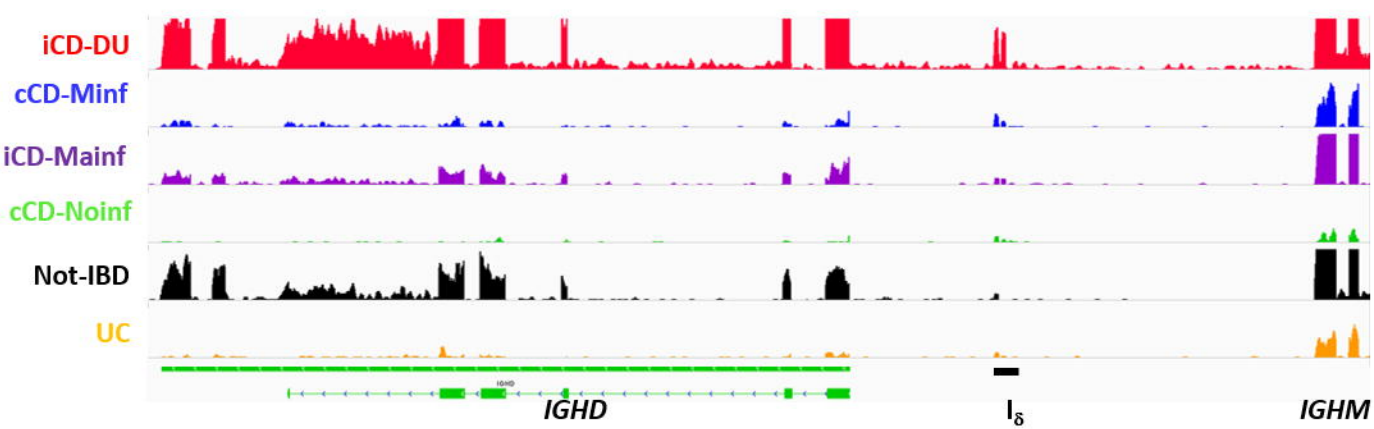

Av log2 RPKM
10
5
0
-5
Uveal Melanoma
Adrenocortical Ca
LGG
GBM
Kidney Chromophobe Ca
Sarcoma
Kidney Papillary Ca
Pheochromocytoma
Liver HCC
Endometrial Ca
Uterine CaSa
Cholangiocarcinoma
Bladder Ca
Prostate AdCa
Kidney Clear cell Ca
Thyroid Ca
Thymoma
Skin Melanoma
Head & Neck SCC
Ovarian AdCa
Rectum AdCa
Cervical Ca
Mesothelioma
Colon AdCa
Pancreatic AdCa
Breast Ca
Testicular GC
Esophageal ca
Lung SCC
Lung AdCa
Stomach AdCa
AML
DLBCL


A
p=0.0053
p=0.00091
Av log2 RPKM
Spleen
LCL
DLBCL
B
p=< 2.2e-16
Thyroid
Thyroid Ca
C
p=2.184e-08
Prostate
Prostate AdCa
D
p=< 2.2e-16
Colon (Tv)
Colon AdCa
E
p=< 2.2e-16
Liver
Liver HCC
F
p=0.00027
p=1.2e-06
Kidney Cortex
Chromophobe Ca
Papillary Ca
Clear Cell Ca
G
p=5.058e-07
Breast
Breast Ca
H
p=2.905e-11
Sigmoid
Rectum AdCa
I
p=< 2.2e-16
Pancreas
Pancreas AdCa
J
p=1.215e-09
Testis
Testicular GCC
K
p=3.307e-05
Ovary
Ovarian AdCa
L
p=< 2e-16
p=5.8e-16
Skin (L)
Skin (SP)
Skin Melanoma
Uveal Melanoma

A


Iδ
p < 0.005
*
RPKM
0
100
200
300
iCD
iCD
cCD
cCD
Not-IBD
UC
Macro inflam
Deep ulcer
Micro inflam
No ileal inflam

B


iCD-DU
cCD-Minf
iCD-Mainf
cCD-Noinf
Not-IBD
UC
IGHD
Iδ
IGHM