Regulation of Gene Expression in Autoimmune Disease Loci and the Genetic Basis of Proliferation in CD4+ Effector Memory T Cells

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1371/journal.pgen.1004404

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:12717376

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Regulation of Gene Expression in Autoimmune Disease Loci and the Genetic Basis of Proliferation in CD4⁺ Effector Memory T Cells

Xinli Hu¹,²,³,⁴,⁵,⁶,⁷, Hyun Kim¹,²,³,⁸, Towfique Raj²,⁴,⁷, Patrick J. Brennan¹, Gosia Trynka¹,²,³,⁴, Nikola Teslovich¹,², Kamil Slowikowski¹,²,³,⁴,⁵, Wei-Min Chen⁸, Suna Onengut⁸, Clare Baecher-Allan⁹, Philip L. De Jager⁴,⁷, Stephen S. Rich⁹, Barbara E. Stranger¹⁰,¹¹, Michael B. Brenner¹, Soumya Raychaudhuri¹,²,³,⁴,¹²

¹Division of Rheumatology, Immunology and Allergy, Department of Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, ²Division of Genetics, Department of Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, ³Partners Center for Personalized Genetic Medicine, Boston, Massachusetts, United States of America, ⁴Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, ⁵Harvard Medical School, Boston, Massachusetts, United States of America, ⁶Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts, United States of America, ⁷Program in Translational Neuropsychiatric Genomics, Institute for the Neurosciences, Department of Neurology, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, ⁸Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, United States of America, ⁹Department of Dermatology/Harvard Skin Disease Research Center, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, ¹⁰Section of Genetic Medicine, University of Chicago, Chicago, Illinois, United States of America, ¹¹Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois, United States of America, ¹²Faculty of Medical and Human Sciences, University of Manchester, Manchester, United Kingdom

Abstract

Genome-wide association studies (GWAS) and subsequent dense-genotyping of associated loci identified over a hundred single-nucleotide polymorphism (SNP) variants associated with the risk of rheumatoid arthritis (RA), type 1 diabetes (T1D), and celiac disease (CeD). Immunological and genetic studies suggest a role for CD4-positive effector memory T (CD⁴⁺ TEM) cells in the pathogenesis of these diseases. To elucidate mechanisms of autoimmune disease alleles, we investigated molecular phenotypes in CD4⁺ effector memory T cells potentially affected by these variants. In a cohort of genotype and healthy individuals, we isolated high purity CD⁴⁺ TEM cells from peripheral blood, and assayed relative abundance, proliferation upon T cell receptor (TCR) stimulation, and the transcription of 215 genes within disease loci before and after stimulation. We identified 46 genes regulated by cis-acting expression quantitative trait loci (eQTL), the majority of which we detected in stimulated cells. Eleven of the 46 genes with eQTLs were previously undetected in peripheral blood mononuclear cells. Of 96 risk alleles of RA, T1D, and/or CeD in densely genotyped loci, eleven overlapped cis-eQTLs of which five alleles completely explained the respective signals. A non-coding variant, rs3898625, increased proliferative response (p < 4.75 x 10⁻⁹). In addition, baseline expression of seventeen genes in resting cells reliably predicted proliferative response after TCR stimulation. Strikingly, however, there was no evidence that risk alleles modulated CD⁴⁺ TEM abundance or proliferation. Our study underscores the power of examining molecular phenotypes in relevant cells and conditions for understanding pathogenic mechanisms of disease variants.

Introduction

Memory T cells are an important component of the adaptive immune system. They circulate between lymphoid organs, blood, and peripheral tissues, and facilitate faster and more aggressive immune response to antigens after re-exposure. CD4-positive effector memory T (CD⁴⁺ TEM) cells are known to migrate to peripheral sites of inflammation upon activation, and rapidly produce both Th1 and Th2 cytokines [1]. Investigators have long suggested their involvement in autoimmune diseases including rheumatoid arthritis (RA), type I diabetes (T1D), and celiac disease (CeD) [2–5]. However, whether changes in cell population subsets and functions are causal or reactive to disease is uncertain. One strategy to answer this question is to examine potential intermediate molecular phenotypes, and identify those modulated by genetic variants. In order to understand the pathogenic roles of CD⁴⁺ TEM cells in autoimmunity, we aimed to characterize the variation in their phenotypic and functional markers in a healthy
population, and to identify whether these markers intersect with the genetic basis for autoimmune disease.

The majority of autoimmune disease risk variants are located in non-coding regions of the genome. It is reasonable to hypothesize that a subset of them causes disease by altering gene regulatory mechanisms as expression quantitative trait loci (eQTL) [6-9]. So far, studies of gene regulation have largely been carried out in cell lines and primary resting blood cells including undifferentiated CD4+ T cells, B cells, monocytes, and dendritic cells [8,10-12]. However, to understand the pathogenic mechanisms of risk variants, especially when studying the immune system where cells are highly diverse and functionally specialized, it is crucial to focus on relevant cell types and stimulated cellular states.

We have previously shown that genes within RA risk loci were most specifically expressed in CD4+ TEM cells, compared to more than 200 other immune cell types of various lineages and developmental stages ($p = 1.00 \times 10^{-10}$; Figure S1) [13]. Celiac disease and T1D loci were also enriched for genes specifically expressed in CD4+ TEM cells ($p = 1.43 \times 10^{-7}$ and $1.29 \times 10^{-8}$, respectively; Figure S1) [13]. Non-coding single nucleotide polymorphisms (SNPs) associated with RA significantly overlap chromatin marks of trimethylation of histone H3 at lysine 4 (H3K4me3) specifically in CD4+ regulatory and memory T cells ($p = 1.5 \times 10^{-4}$ and $7.0 \times 10^{-4}$, respectively) [14].

We hypothesized that the risk alleles of these conditions might influence CD4+ TEM quantitative molecular phenotypes: 1) the expression of immune-related genes; 2) the relative abundance of CD4+ TEM cells in peripheral blood; and 3) proliferative response to T cell receptor (TCR) stimulation. To this end, we undertook a large immunoprofiling study in a healthy population of 174 European-descent individuals, by cross-analyzing genotype, transcription abundance, and proliferative response in primary CD4+ TEM cells. Because the post-stimulation activation of CD4+ TEM cells is presumably crucial for their autoimmune response, we assayed cells not only at rest, but also after T cell receptor (TCR) stimulation with anti-CD3/CD28 beads. As such, this study is the first to our knowledge to map expression quantitative trait loci and examine immunological cellular traits in primary CD4+ TEM cells under multiple states.

Using the ImmunoChip platform, investigators recently densely genotyped 186 loci that originally arose through genome-wide association studies (GWAS) in case-control samples for RA, CeD, and inflammatory bowel disease [15-17], as well as T1D (unpublished data). Dense genotyping allowed localization of association signals within these disease loci to a set of alleles that are very likely to be causal. Within these loci, we have a greater ability to identify co-localization between alleles driving variation in molecular phenotypes (such as eQTLs) and the disease risk alleles. However, in instances where multiple variants are in perfect linkage, we cannot pinpoint the exact causal variant without functional evaluation.

Results

The experimental protocol (Figure 1) is described in detail in Methods and Text S1. Briefly, we obtained peripheral blood mononuclear cells (PBMCs) from the whole blood of healthy individuals via Ficoll-Paque centrifugation, and then used magnetic- and fluorescence-activated cell sorting to isolate CD4+ TEM cells at a high degree of purity (>90%; see Figure S2A). We acquired genome-wide genotype data of about 640,000 SNPs on Illumina Infinium Human OmniExpress Exome BeadChips [18]. For each individual we then measured three quantitative phenotypes: 1) the expression of 215 genes (see Table S1) before and after T cell receptor (TCR) stimulation with anti-CD3/CD28 antibody beads; 2) the relative abundance of CD4+ TEM cells (CD45RA-CD45RO+/CD62L-TEM) as a proportion of total CD4+ T cells; and 3) proliferation upon stimulation. Since we had low numbers of primary cells for expression profiling, we used the highly sensitive NanoString nCounter assay to avoid biases potentially induced by cDNA preparation. Out of the 215 genes assayed, 115 were within densely genotyped disease risk loci (see Tables S2 and S3). We quantified CD4+ TEM cell abundance with X-Cyt, an automated statistical method that accurately identifies cell populations in cytometry data [19].

Mapping cis-eQTLs that regulate genes in risk loci

We first aimed to identify SNP variants that regulated expression of genes in the cis. To best localize eQTL signals, we imputed 1000 Genomes variants within 250 kb from the transcription start site (TSS) of each gene (excluding five HLA genes and five long non-coding RNAs). We tested SNPs in gene-coding and non-coding regions in both resting and stimulated CD4+ TEM cells. We included gender and the top five principal components of the genotype data (calculated by EIGENSTRAT) as covariates in regression. To adjust for multiple hypothesis testing, we conducted 10,000 permutations within each gene region to calculate empirical $p$-values, and then reported associations at a false discovery rate of 5%.

In total, we observed 46 variants (22.4%) with cis-eQTL signals, including 17 in resting cells and 43 in stimulated cells (Tables 1 and 2, Figure 2A). For 14 of the 46 genes (30.4%), we detected eQTL signals in both resting (14/17, 82.4%) and stimulated (14/43, 32.6%) states. In four of these 14 genes (FHHL3, GRB10, IL18R1, and PIK3G), the lead eQTL SNPs from the two states were in perfect linkage. For the remaining five genes (C1QTNF6, PRDM1, SKAP2, DDX6, and LRMT), the lead SNPs are in tight LD ($r^2 = 0.80\sim1$; based on 1000 Genomes Release 2, European samples). For the remaining five genes (BLK, TAMR353, CD101, ORMDL5, and GSDMB), the lead SNPs from the two states were in partial LD $0.42 < r^2 < 0.56$. In these five cases, we could not be confident that the eQTLs across stimulation states were tagging the same variant.
We wanted to assess whether the eQTLs might act by altering gene regulatory elements in CD4+ TEM cells. To this end we asked whether the eQTL SNPs co-localized with marks of active promoters or enhancers. We utilized H3K4me3 marks from the NIH Roadmap Epigenomics Mapping Consortium [20] measured by ChIP-seq in primary CD4+ memory T cells. For the SNP with the strongest association to each gene, we queried the distance of the nearest H3K4me3 mark to this SNP or its LD partners (r²>0.8). We compared this distance measure between two sets of SNPs: the 46 SNPs with significant eQTL associations (FDR<5%, resting or stimulated), and the SNPs most strongly correlated with the other 159 genes but did not reach significance threshold. Indeed, the 46 significant eQTL SNPs were located at smaller distances to H3K4me3 marks (p = 1.10x10⁻⁷, one-sided Mann-Whitney test, Figure S3A). In addition, we queried the height of each H3K4me3 mark’s peak, which reflected the number of reads at a given position compared to genomic controls as defined by the MACS software package. A tall peak gives us confidence that the mark is present in a large proportion of cells. Comparing the marks nearest to the two sets of SNPs, we saw that the 46 eQTL SNPs were also located near taller peaks (p = 9.56x10⁻³, Figure S3B).

Many eQTLs are CD4+ TEM cell-specific

We compared the cis-eQTLs we discovered to those found in heterogeneous peripheral blood mononuclear cells (PBMC) in a large genome-wide eQTL meta-study (n = 5,331) conducted by Westra et al. [8]. At 5% FDR, eleven of the 46 eQTL genes we identified showed no detectable signal in PBMCs at 50% FDR. We saw significant associations in 131 genes at 50% FDR, 53 of which had no signal in PBMCs at 50% FDR (Tables 1 and 2). We hypothesized that these genes tended to be more specifically expressed in CD4+ TEM cells, thus making eQTLs readily detectable in the purified cell population. To assess this, we examined cell-specific expression of the genes the ImmGen dataset, which assayed the genome-wide expression in 247 murine immunological cell types [13,21]. We found that the genes with CD4+ TEM cell-specific eQTLs (at 50% FDR) were more specifically expressed in CD4+ TEM cells than genes with eQTLs detected in both datasets (p = 0.044, one-sided Mann-Whitney test).

Autoimmune disease alleles affect the transcription of genes in cis

We then focused on 115 genes near 96 risk alleles of RA, T1D, and/or CeD in densely genotyped loci (182 gene-SNP pairs, including two risk alleles shared by at least two diseases, see Tables S2 and S3). We discovered that eleven (11.4%) disease-associated SNPs (6 of 24 RA SNPs, 5 of 37 T1D SNPs, and 3 of 37 CeD SNPs) correlated significantly with the expression of ten genes in either resting or stimulated state (Table S3). In addition, there was substantial enrichment of nominally significant associations (p<0.05) among disease SNPs. By random chance, we expected about nine SNP-gene pairs to reach nominal association (‘state-specific’; see Tables 1 and 2). For example, rs12746918T increased the expression of PLCH2 significantly only in resting cells; and βrest was approximately twice as large as βstim (Figure 2C). We note that 1 of the 3 eQTLs in resting cells was state-specific (p<0.05), and 13 out of 29 eQTLs seen in stimulated cells were state-specific (p<0.05). Of the 14 eQTLs that were shared between resting and stimulated cells, only 4 of them, BLC (Figure 2D), CD101, PBOC, and PRDM1, had different β’s across states. The abundance of eQTLs detected exclusively in stimulated cells underscores the importance of studying cells in different cellular states.
in each stimulation state. However, we observed 26 pairs (14.2%) with nominal association in resting cells ($p=4.67 \times 10^{-7}$, one-tailed binomial test). Even more strikingly, we observed 45 pairs (24.7%) with nominal association in stimulated cells ($p=10^{-15}$, one-tailed binomial test).

To identify those instances where the disease-associated SNP could explain the entire eQTL signal in the gene region, we applied conditional analysis to identify any residual signals after controlling for the disease SNP. In five of the ten genes (BLK, C5orf30, GSDMB, IRF5, PLEK), conditioning on the disease SNP obviated any remaining eQTL signal in the region (no SNP with permutation $p$-value $<0.05$; Figure 3), suggesting that there was a single variant (the disease-associated SNP or one in very high LD to it) that drove variation in expression. Interestingly, as previously noted, the lead SNPs in resting and stimulated cells ($p<10^{-15}$, one-tailed binomial test).

In these instances the disease-associated SNP and remaining lead signal are in partial linkage disequilibrium ($r^2=0.36–0.73$). In these cases, we could not conclude whether the disease SNPs drove the alteration in expression, or whether the true causal SNPs were in partial linkage and caused spurious associations. It is probable that disease risk alleles were indeed causal, yet we could not confidently fine-map the effect due to experimental noise in expression assays or inadequate sampling.

We note that another 26 genes within disease loci associated contained *cis*-eQTL signals, but that these *cis*-eQTL signals did not co-localize with RA, T1D, or CeD alleles. As these loci had been fine-mapped using Immunochip, the lack of overlap strongly suggested that these *cis*-eQTLs and disease-causing variants were distinct. For example, rs798000 is an RA risk allele located in a non-coding region upstream of CD2, CD58, and PTGFRN. However, it was not associated with the expression of any of these genes ($p>0.5$). Another example was rs6911690, an RA allele located about 60 kb 5' of PRDM1, that was not associated with the expression of the gene at rest or after stimulation ($p>0.5$). The lead eQTL SNP associated to PRDM1 was rs578653 (FDR $<10^{-5}$), which was not in LD with the disease allele ($r^2<0.05$).
<table>
<thead>
<tr>
<th>Chr</th>
<th>Top SNP</th>
<th>Minor allele</th>
<th>Gene</th>
<th>Rest</th>
<th>Stim</th>
<th>Fold &amp; per Allele</th>
<th>Assoc. p-value</th>
<th>FDR</th>
<th>Fold &amp; per Allele</th>
<th>Assoc. p-value</th>
<th>FDR</th>
<th>pA*</th>
<th>CD4 TEM-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>n12746918</td>
<td>T</td>
<td>PLCH2</td>
<td>X</td>
<td></td>
<td>1.61</td>
<td>4.75E-09</td>
<td>&lt;0.001</td>
<td>0.18</td>
<td>2.35E-03</td>
<td>&gt;0.20</td>
<td>2.70E-03</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>n10789226</td>
<td>G</td>
<td>IL23R</td>
<td>X</td>
<td></td>
<td>0.78</td>
<td>1.05E-05</td>
<td>0.022</td>
<td>-0.21</td>
<td>2.34E-03</td>
<td>&gt;0.20</td>
<td>6.78E-01</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>n6704162</td>
<td>G</td>
<td>RGS1</td>
<td>X</td>
<td></td>
<td>0.77</td>
<td>2.96E-05</td>
<td>0.036</td>
<td>-0.14</td>
<td>1.71E-03</td>
<td>&gt;0.20</td>
<td>1.25E-01</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>n3746271</td>
<td>A</td>
<td>SIKPG</td>
<td>X</td>
<td>-0.24</td>
<td>1.67E-03</td>
<td>&lt;0.20</td>
<td>0.44</td>
<td>3.29E-28</td>
<td>&lt;0.001</td>
<td>1.91E-09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>n11123823</td>
<td>A</td>
<td>LONRF2</td>
<td>X</td>
<td>-0.09</td>
<td>1.52E-01</td>
<td>&gt;0.20</td>
<td>0.65</td>
<td>7.09E-17</td>
<td>&lt;0.001</td>
<td>4.26E-06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>n10058074</td>
<td>A</td>
<td>SLC22A5</td>
<td>X</td>
<td></td>
<td>0.04</td>
<td>5.35E-01</td>
<td>&gt;0.20</td>
<td>0.78</td>
<td>1.81E-09</td>
<td>&lt;0.001</td>
<td>1.11E-04</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>n942793</td>
<td>G</td>
<td>ZMIZ1</td>
<td>X</td>
<td></td>
<td>0.00</td>
<td>9.85E-01</td>
<td>&gt;0.20</td>
<td>1.23</td>
<td>3.67E-21</td>
<td>&lt;0.001</td>
<td>7.70E-07</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>n14315026</td>
<td>A</td>
<td>SULT1A1</td>
<td>X</td>
<td>-0.35</td>
<td>2.60E-03</td>
<td>&gt;0.20</td>
<td>0.59</td>
<td>4.04E-10</td>
<td>&lt;0.001</td>
<td>2.08E-01</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>n1399754</td>
<td>A</td>
<td>ABDH6</td>
<td>X</td>
<td>-0.08</td>
<td>2.61E-01</td>
<td>&gt;0.20</td>
<td>0.79</td>
<td>1.11E-12</td>
<td>&lt;0.001</td>
<td>4.07E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>n3807306</td>
<td>G</td>
<td>IRF5</td>
<td>X</td>
<td>-0.19</td>
<td>3.32E-03</td>
<td>&gt;0.20</td>
<td>0.80</td>
<td>3.07E-15</td>
<td>&lt;0.001</td>
<td>7.41E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>n39984</td>
<td>T</td>
<td>Csof30</td>
<td>X</td>
<td>-0.09</td>
<td>1.77E-01</td>
<td>&gt;0.20</td>
<td>0.87</td>
<td>1.90E-08</td>
<td>&lt;0.001</td>
<td>4.36E-01</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>n12138115</td>
<td>G</td>
<td>SF3A3</td>
<td>X</td>
<td>0.02</td>
<td>3.29E-01</td>
<td>&gt;0.20</td>
<td>1.06</td>
<td>1.63E-08</td>
<td>&lt;0.001</td>
<td>7.10E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>n6674126</td>
<td>A</td>
<td>TNFISF14</td>
<td>X</td>
<td>0.06</td>
<td>1.18E-01</td>
<td>&gt;0.20</td>
<td>1.12</td>
<td>1.28E-09</td>
<td>&lt;0.001</td>
<td>1.89E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>n7140</td>
<td>C</td>
<td>SPNS1</td>
<td>X</td>
<td>0.07</td>
<td>8.07E-03</td>
<td>&gt;0.20</td>
<td>1.16</td>
<td>2.73E-12</td>
<td>&lt;0.001</td>
<td>3.10E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>n1021469</td>
<td>A</td>
<td>METTL21B</td>
<td>X</td>
<td>0.20</td>
<td>2.72E-03</td>
<td>&gt;0.20</td>
<td>1.35</td>
<td>1.22E-09</td>
<td>&lt;0.001</td>
<td>2.09E-01</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>n10167650</td>
<td>G</td>
<td>PLEX</td>
<td>X</td>
<td>0.25</td>
<td>6.30E-03</td>
<td>&gt;0.20</td>
<td>1.49</td>
<td>4.98E-08</td>
<td>&lt;0.001</td>
<td>2.11E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>n11265501</td>
<td>A</td>
<td>CD244</td>
<td>X</td>
<td>0.23</td>
<td>1.50E-02</td>
<td>&gt;0.20</td>
<td>1.70</td>
<td>1.46E-12</td>
<td>&lt;0.001</td>
<td>1.06E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>n71383688</td>
<td>G</td>
<td>CTSH</td>
<td>X</td>
<td>0.02</td>
<td>7.63E-01</td>
<td>&gt;0.20</td>
<td>0.76</td>
<td>2.19E-07</td>
<td>0.001</td>
<td>1.55E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>n6498114</td>
<td>A</td>
<td>CIITA</td>
<td>X</td>
<td>-0.07</td>
<td>3.70E-01</td>
<td>&gt;0.20</td>
<td>1.17</td>
<td>6.38E-07</td>
<td>0.001</td>
<td>7.36E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>n149241987</td>
<td>T</td>
<td>PDHB</td>
<td>X</td>
<td>0.01</td>
<td>6.61E-01</td>
<td>&gt;0.20</td>
<td>0.93</td>
<td>1.94E-07</td>
<td>0.001</td>
<td>1.28E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>n11639897</td>
<td>C</td>
<td>TUFM</td>
<td>X</td>
<td>0.20</td>
<td>1.21E-02</td>
<td>&gt;0.20</td>
<td>1.13</td>
<td>3.05E-07</td>
<td>0.001</td>
<td>3.43E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>n4659444</td>
<td>G</td>
<td>PTGFRN</td>
<td>X</td>
<td>0.10</td>
<td>1.27E-01</td>
<td>&gt;0.20</td>
<td>1.31</td>
<td>1.18E-07</td>
<td>0.001</td>
<td>3.09E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>n14645192</td>
<td>T</td>
<td>EIF3C</td>
<td>X</td>
<td>0.07</td>
<td>4.84E-01</td>
<td>&gt;0.20</td>
<td>1.38</td>
<td>4.06E-06</td>
<td>0.002</td>
<td>2.96E-02</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>n971173</td>
<td>T</td>
<td>PHTF1</td>
<td>X</td>
<td>-0.06</td>
<td>1.24E-01</td>
<td>&gt;0.20</td>
<td>0.92</td>
<td>7.76E-06</td>
<td>0.004</td>
<td>7.24E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>n11711261</td>
<td>A</td>
<td>PXX</td>
<td>X</td>
<td>0.10</td>
<td>1.12E-01</td>
<td>&gt;0.20</td>
<td>1.11</td>
<td>5.72E-06</td>
<td>0.004</td>
<td>9.36E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>n2847224</td>
<td>A</td>
<td>ICOSLG</td>
<td>X</td>
<td>-0.11</td>
<td>5.22E-02</td>
<td>&gt;0.20</td>
<td>0.82</td>
<td>1.44E-06</td>
<td>0.005</td>
<td>1.90E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>n11066028</td>
<td>C</td>
<td>ATXN2</td>
<td>X</td>
<td>0.01</td>
<td>6.64E-01</td>
<td>&gt;0.20</td>
<td>1.07</td>
<td>1.86E-05</td>
<td>0.005</td>
<td>1.63E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>n3858706</td>
<td>C</td>
<td>PTPN11</td>
<td>X</td>
<td>0.11</td>
<td>2.81E-01</td>
<td>&gt;0.20</td>
<td>1.05</td>
<td>2.17E-05</td>
<td>0.009</td>
<td>3.65E-01</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>n2291668</td>
<td>A</td>
<td>TNFSF14</td>
<td>X</td>
<td>0.12</td>
<td>9.21E-02</td>
<td>&gt;0.20</td>
<td>1.28</td>
<td>5.16E-06</td>
<td>0.017</td>
<td>2.44E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>n8587</td>
<td>A</td>
<td>COX4ANB</td>
<td>X</td>
<td>0.07</td>
<td>1.96E-01</td>
<td>&gt;0.20</td>
<td>1.13</td>
<td>8.44E-06</td>
<td>0.018</td>
<td>4.28E-01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The genetic basis of CD4\(^+\) TEM cell proliferation

The relative peripheral abundance of CD4\(^+\) TEM cells varied between individuals (mean = 9.57%; SD = 4.85%), and was reproducible 33 individuals with two separate blood draws more than one month apart (Pearson’s r = 0.87, p = 1.77 \times 10^{-11}, see also Figure S2B). Consistent with other studies, we observed that the relative proportion of CD4\(^+\) TEM cells increased with age by 0.11% per year (β_age = 1.92 \times 10^{-7}) [22]. We also observed that on average men had 2.22% more CD4\(^+\) TEM cells than women (β_gender = 3.80 \times 10^{-2}; see Figure S4). Upon anti-CD3/CD28 stimulation, there was a substantial inter-individual variation in proliferation measured by both division index (DI, average number of divisions undergone by all cells; mean = 1.46, SD = 0.35), and proliferation index (PI, average number of divisions undergone only by dividing cells; mean = 2.16, SD = 0.21). Proliferation metrics were also reproducible in the 35 individuals (Pearson’s r_DI = 0.57; Pearson’s r_PI = 0.62, Figures S2G and S2D). Interestingly, proliferation was negatively correlated to the proportion of CD4\(^+\) TEM cells (p_DI = 1.28 \times 10^{-5}, p_PI = 1.95 \times 10^{-5}), but was not associated to age or gender (p > 0.3). This negative correlation needs to be replicated in an independent dataset. Effector functions of TEM cells with higher proliferative capacities need to be examined to understand whether they represent a hyperactive subset whose abundance is controlled to maintain immune homeostasis. Possibly individuals with a lower proportion of TEM cells are relatively enriched for these subsets.

We tested genome-wide SNPs for association to relative abundance, division index, and proliferation index, considering \(p<5 \times 10^{-8}\) as the threshold for significance. For abundance, we included gender, age, and the top five principal components of genotypes as covariates. Given the correlation with proliferation, we also included the measured CD4\(^+\) TEM relative abundance as an additional covariate. We observed associations to division index in several loci, including 13q34 led by rs3898622 (p = 4.75 \times 10^{-8}, Figure 4A). This SNP is a non-coding variant located 30 kb upstream of RASL5, and 70 kb upstream from CDC16. Both genes have known roles in regulating cell proliferation or differentiation [23,24]. This SNP was also strongly associated with proliferation index (p = 2.75 \times 10^{-7}). Additionally, there was a strongly suggestive association to rs3775500 on chromosome 4, located in the intron of DAPPI, which encodes the Bam32 protein (p = 5.40 \times 10^{-2}; Figure 4B), which is an adaptor protein expressed solely in antigen presenting B cells. Interestingly, mutations in this gene have been shown by several groups to affect T cell activation [25,26], suggesting the possibility that B cells may indirectly regulate T cell function in autoimmunity. We did not observe any significant association with the relative abundance of CD4\(^+\) TEM cells.

When we extracted the association statistics of 118 densely genotyped risk alleles of CeD, RA, and/or T1D, they showed no inflation in association p-values for relative abundance of CD4\(^+\) TEM cells (Figure 5A, Table S2). This suggested that risk variants did not modify risk via modulation of CD4\(^+\) TEM peripheral abundance. We recognized that the power to detect significant associations might have been limited in our study by the sample size. However, this negative finding was corroborated by results from a recently published study with data from \(\sim 2800\) individuals, in which the same set of risk alleles also showed no significant association to CD4\(^+\) TEM (see Figure S5) [27]. Similarly, the same set of risk alleles did not show significant association to proliferative response (Figure 5B, Table S2). Based on these data, it was unlikely that SNP variants associated to

<table>
<thead>
<tr>
<th>Chr</th>
<th>Top SNP</th>
<th>Minor allele</th>
<th>Gene</th>
<th>eQTL Effect</th>
<th>Rest Fold per Allele</th>
<th>Stim Fold per Allele</th>
<th>p-</th>
<th>pFDR</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>rs7498329</td>
<td>A</td>
<td>FAM120B</td>
<td>Rest</td>
<td>0.20</td>
<td>1.22</td>
<td>0.20</td>
<td>0.042</td>
<td>2.41E-02</td>
</tr>
<tr>
<td>6</td>
<td>rs7453655</td>
<td>A</td>
<td>FAM120B</td>
<td>Stim</td>
<td>0.20</td>
<td>7.02E-01</td>
<td>0.04</td>
<td>7.02E-01</td>
<td></td>
</tr>
</tbody>
</table>

* For each row we list a SNP and the gene transcript for which it is a cis-eQTL. We indicate whether the effect is observed in resting or stimulated CD4\(^+\) effector memory T cells. For each SNP we indicate the fold change in expression + conferred per allele, assuming an additive model and the false discovery rate estimate. Finally, we indicate whether the effect is specific for CD4\(^+\) effector memory T cells; we define specificity as the absence of any cis-eQTL effect + in PBMCs at FDR 50% [8].
Table 2. Genes with resting and stimulated eQTLs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>SNP_rest</th>
<th>minor allele</th>
<th>Fold Δ per Allele</th>
<th>assoc. P-value</th>
<th>FDR_rest</th>
<th>SNP_stim</th>
<th>minor allele</th>
<th>Fold Δ per Allele</th>
<th>assoc. P-value</th>
<th>FDR_stim</th>
<th>R2</th>
<th>pΔ**</th>
<th>CD4 TEM-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLK</td>
<td>8</td>
<td>rs4840565</td>
<td>C</td>
<td>0.69</td>
<td>2.75E-05</td>
<td>0.012</td>
<td>rs4840565</td>
<td>C</td>
<td>0.45</td>
<td>1.35E-22</td>
<td>&lt;0.001</td>
<td>1*</td>
<td>7.38E-05</td>
<td></td>
</tr>
<tr>
<td>C1QTNF6</td>
<td>22</td>
<td>rs229515</td>
<td>T</td>
<td>0.63</td>
<td>6.10E-10</td>
<td>&lt;0.001</td>
<td>rs229522</td>
<td>A</td>
<td>0.61</td>
<td>8.75E-15</td>
<td>&lt;0.001</td>
<td>1</td>
<td>5.41E-01</td>
<td></td>
</tr>
<tr>
<td>CD101</td>
<td>1</td>
<td>rs620527</td>
<td>G</td>
<td>1.80</td>
<td>3.06E-09</td>
<td>&lt;0.001</td>
<td>rs9323416</td>
<td>A</td>
<td>1.96</td>
<td>1.76E-19</td>
<td>&lt;0.001</td>
<td>0.547</td>
<td>3.23E-02</td>
<td>yes</td>
</tr>
<tr>
<td>DDX6</td>
<td>11</td>
<td>rs500254</td>
<td>G</td>
<td>1.92</td>
<td>2.33E-13</td>
<td>&lt;0.001</td>
<td>rs4938544</td>
<td>A</td>
<td>1.75</td>
<td>1.48E-16</td>
<td>&lt;0.001</td>
<td>0.834</td>
<td>3.40E-01</td>
<td>yes</td>
</tr>
<tr>
<td>FHL3</td>
<td>1</td>
<td>rs7631072</td>
<td>T</td>
<td>0.63</td>
<td>6.43E-08</td>
<td>&lt;0.001</td>
<td>rs7631072</td>
<td>T</td>
<td>0.61</td>
<td>9.53E-26</td>
<td>&lt;0.001</td>
<td>1*</td>
<td>7.05E-01</td>
<td></td>
</tr>
<tr>
<td>GRB10</td>
<td>7</td>
<td>rs12536500</td>
<td>T</td>
<td>0.68</td>
<td>4.10E-06</td>
<td>0.016</td>
<td>rs12536500</td>
<td>T</td>
<td>0.61</td>
<td>4.92E-11</td>
<td>&lt;0.001</td>
<td>1*</td>
<td>3.11E-01</td>
<td></td>
</tr>
<tr>
<td>GSDMB</td>
<td>17</td>
<td>rs6038753/rs12936409</td>
<td>T</td>
<td>0.86</td>
<td>2.20E-07</td>
<td>&lt;0.001</td>
<td>rs6038753/rs12936409</td>
<td>T</td>
<td>0.83</td>
<td>2.64E-12</td>
<td>&lt;0.001</td>
<td>1*</td>
<td>3.65E-01</td>
<td></td>
</tr>
<tr>
<td>IL18R1</td>
<td>2</td>
<td>rs3771164</td>
<td>T</td>
<td>0.52</td>
<td>6.35E-16</td>
<td>&lt;0.001</td>
<td>rs3771164</td>
<td>T</td>
<td>0.48</td>
<td>3.79E-31</td>
<td>&lt;0.001</td>
<td>1*</td>
<td>4.35E-01</td>
<td></td>
</tr>
<tr>
<td>LVRM7</td>
<td>5</td>
<td>rs12522164</td>
<td>A</td>
<td>0.59</td>
<td>1.05E-08</td>
<td>&lt;0.001</td>
<td>rs12517633</td>
<td>T</td>
<td>0.51</td>
<td>1.38E-29</td>
<td>&lt;0.001</td>
<td>0.792</td>
<td>1.36E-01</td>
<td></td>
</tr>
<tr>
<td>ORMIL3</td>
<td>17</td>
<td>rs35222145</td>
<td>G</td>
<td>0.75</td>
<td>1.16E-19</td>
<td>&lt;0.001</td>
<td>rs2290400</td>
<td>C</td>
<td>0.82</td>
<td>4.44E-28</td>
<td>&lt;0.001</td>
<td>0.482</td>
<td>2.75E-01</td>
<td></td>
</tr>
<tr>
<td>PIGC</td>
<td>1</td>
<td>rs1063412</td>
<td>A</td>
<td>1.91</td>
<td>1.36E-14</td>
<td>&lt;0.001</td>
<td>rs1063412</td>
<td>A</td>
<td>3.04</td>
<td>2.65E-46</td>
<td>&lt;0.001</td>
<td>1*</td>
<td>6.79E-07</td>
<td></td>
</tr>
<tr>
<td>PRDM1</td>
<td>6</td>
<td>rs811925</td>
<td>G</td>
<td>0.72</td>
<td>2.61E-13</td>
<td>&lt;0.001</td>
<td>rs578653</td>
<td>G</td>
<td>0.81</td>
<td>3.30E-13</td>
<td>&lt;0.001</td>
<td>0.944</td>
<td>2.78E-02</td>
<td></td>
</tr>
<tr>
<td>SKAP2</td>
<td>7</td>
<td>rs4719882</td>
<td>G</td>
<td>1.37</td>
<td>1.85E-12</td>
<td>&lt;0.001</td>
<td>rs3801813</td>
<td>C</td>
<td>1.46</td>
<td>6.08E-39</td>
<td>&lt;0.001</td>
<td>0.862</td>
<td>1.17E-01</td>
<td></td>
</tr>
<tr>
<td>TMPRSS5</td>
<td>21</td>
<td>rs7283281</td>
<td>G</td>
<td>0.76</td>
<td>1.08E-06</td>
<td>0.005</td>
<td>rs277798</td>
<td>A</td>
<td>0.77</td>
<td>7.74E-10</td>
<td>&lt;0.001</td>
<td>0.55</td>
<td>5.42E-01</td>
<td></td>
</tr>
</tbody>
</table>

As in Table 1, for each row we list a SNP and the gene transcript for which it is a cis-eQTL. For each SNP we indicate the fold change in expression conferred per allele, assuming an additive model and the false discovery rate estimate. Finally, we indicate whether the effect is specific for CD4+ effector memory T cells [8].

* denotes identical SNPs.

** In cases where the lead resting and stimulated SNPs differ, pΔ compares the cross-state effect sizes of only one SNP. For each gene, the SNP with the stronger association across the two states was considered.

doi:10.1371/journal.pgen.1004404.t002
RA, T1D, or CeD conferred risk through modulation of CD4+ TEM cell abundance or proliferation.

Gene expression in resting cells predicted post-TCR stimulation proliferation

After stimulation we observed that 122 genes showed significant changes in expression in response to stimulation, including 78 whose expression at least doubled or decreased by 50% (Table S1). The gene with the greatest post-stimulation induction was GZMB (average fold change = 93.48), which encodes granzyme B, a protein involved in the apoptosis of target cells during cell-mediated immune response in cytotoxic and memory lymphocytes. The most significantly down-regulated gene was GRB10 (average fold change = 0.18), which is near rs6944602 associated with T1D and encodes growth factor receptor-bound protein 10, whose function in the immune system is unclear.

We observed that relative gene expression at rest predicted proliferative response. In 182 individuals with both proliferation and gene expression data, 17 of the 215 genes were associated with proliferation index ($p < 0.01$, two-tailed test by permuting).

Figure 3. Five disease risk alleles explained the eQTL associations with five genes. The left-sided panels show unconditional SNP-expression association results. Green dashed lines mark the TSS of the eQTL gene. The red dots indicate the risk alleles associated with the expression of respective genes shown as red arrows. The right-sided panels show adjusted association results after conditioning for the respective risk alleles. In each of the five loci, conditioning on the disease SNP obviated signals in the entire region, such that no association more significant than $p = 0.05$ remains.

doi:10.1371/journal.pgen.1004404.g003
proliferation data. Figure 6, Table S1. Increased expression of 15 of the 17 genes including CCR5, IL2RB, PRR5L, and TBX21, were correlated with reduced proliferative response, while CCR9 and the lncRNA XLOC_003479 showed significant correlation with increased proliferation. This number of correlated genes was far in excess of random chance based on a null distribution consisting of 1000 permutations \( p_{10^{-23}} \), median 2, maximum 15. The weighted sum of the 17 genes served as a “proliferation potential signature”, where we weighted the positively- and negatively-correlated genes as +1 and \(-1\), respectively. This signature strongly predicted proliferation index \( r = 0.55 \). We show the correlation between each of the 17 genes as well as the aggregate signature to proliferation as a heatmap (Figure 6A).

To assess if we were overfitting the data, we applied a two way cross-validation, where we defined the proliferation signature based on genes from half of the individuals and tested correlation to proliferation in the remaining half of the individuals. In both instances we again observed significant prediction of proliferation \( r = 0.41 \), one tailed \( p = 10^{-23} \) by permutation; \( r = 0.39 \), \( p = 10^{-23} \).

To search for biological pathways underlying genes correlated to proliferation, we applied gene set enrichment analysis (GSEA) to test for enrichment for 1,008 functional gene sets based on Gene Ontology codes [28] (Figure 6B). Genes correlated to reduced proliferation were most significantly enriched for GO:0012502 (induction of programmed cell death; one tailed \( p = 1.8 \times 10^{-4} \)); those correlated with increased proliferation were most significantly enriched for GO:0002285 (lymphocyte activation involved in immune response; one tailed \( p = 3.9 \times 10^{-4} \)).

Figure 6. Relationship between baseline expression and post-stimulatory response. A) Baseline expression of 17 genes correlated with post-stimulation proliferation. Rows in the heatmap are ordered from top to bottom by ascending proliferation index. Genes/columns are ordered from the most negatively correlated (IL23RB) to the most positively correlated (CCR9). The 17-gene signature was calculated as the weighted sum of the 17 genes, where we weighted the positively- and negatively-correlated genes as +1 and \(-1\), respectively. This signature strongly predicted proliferation index \( r = 0.55 \). We show the correlation between each of the 17 genes as well as the aggregate signature to proliferation as a heatmap (Figure 6A). To assess if we were overfitting the data, we applied a two way cross-validation, where we defined the proliferation signature based on genes from half of the individuals and tested correlation to proliferation in the remaining half of the individuals. In both instances we again observed significant prediction of proliferation \( r = 0.41 \), one tailed \( p = 10^{-23} \) by permutation; \( r = 0.39 \), \( p = 10^{-23} \).

To search for biological pathways underlying genes correlated to proliferation, we applied gene set enrichment analysis (GSEA) to test for enrichment for 1,008 functional gene sets based on Gene Ontology codes [28] (Figure 6B). Genes correlated to reduced proliferation were most significantly enriched for GO:0012502 (induction of programmed cell death; one tailed \( p = 1.8 \times 10^{-4} \)); those correlated with increased proliferation were most significantly enriched for GO:0002285 (lymphocyte activation involved in immune response; one tailed \( p = 3.9 \times 10^{-4} \)).
Using data from 29 individuals each with two samples collected at least one month apart, we replicated the observed correlation. In these samples we performed a cross-visit analysis, and observed that the same 17-gene signature from the first visit significantly predicted proliferation indices on the second visit ($r = 0.65$, $p = 0.0006$, 1-tailed permutation), and vice versa ($r = 0.53$, $p = 0.0019$).

**Discussion**

To fine-map and link risk loci to their pathogenic mechanisms, we investigated molecular and immune phenotypes potentially leading to disease end-points. The immune system is particularly complex, and different cells under various activation states have specialized functions that may not be adequately captured by examining PBMCs. Therefore, we focused on one purified cell population that had been shown to be important for the pathogenesis of several autoimmune diseases. We quantified population variation in several traits, including peripheral abundance, proliferative response to TCR stimulation, and expression of genes within autoimmune disease loci at rest and after stimulation. In **Tables 1, 2, and S2**, we provide significant cis-eQTLs and genome-wide association results.

To our knowledge this study was a first cross-examination of genetic-, transcriptional-, and cellular-level quantitative traits in CD4$^+$ TEM cells. It demonstrated the importance of focusing functional studies in a purified cell population under relevant developmental and stimulation states. By examining the proliferative response upon TCR stimulation, we identified a subset of genes whose baseline expression predicted proliferative potential. Intriguingly, these genes were involved in programmed cell death and lymphocyte activation. Whether variation in proliferative abilities correlated with cytokine production and other signaling functions, thus affecting susceptibility to autoimmunity, remains a question to be addressed by future studies.

Of the 205 genes in disease loci that we examined, 46 had cis-eQTLs. Notably, eleven of these were specific to stimulated CD4$^+$ TEM cells, and not previously found in PBMCs. We noted that approximately 10% of genes within risk loci of diseases had cis-eQTLs. However in many instances the lead eQTL SNPs were unrelated to the disease-associated SNPs. One example of a disease allele that functioned as cis-eQTL was rs39984, which was associated to lower risk of RA, and regulated the expression of C5orf30 encoding an UNC119-binding protein. This SNP variant is located in the first intron of C5orf30, and indeed explained the entire cis-eQTL signal in this gene (see **Figure 2B**). This eQTL effect was previously undetected in PBMCs, and the protein’s functional role in the immune system is largely unknown. However, a recent study showed that rs26232 (the lead GWAS SNP prior to fine-mapping, $r^2 = 0.988$ to rs39984) was correlated with lower severity of radiologic damage in RA, independent of previously established biomarkers [29]. Another gene in the locus, GIN1, is located 140 kb from rs39984; however its expression showed no correlation with the SNP ($p > 0.5$).

Another CD4$^+$ TEM cell-specific eQTL gene was DDX6, which encodes DEAD-box RNA helicase 6. However, in this case, the lead eQTL SNP (rs8193544) associated to increased expression of DDX6 in stimulated cells was not in LD with the known CeD risk allele (rs10892253, $r^2 < 0.1$) or the RA risk allele (rs4939573, $r^2 < 0.1$). Neither risk allele showed significant association to DDX6 expression ($p > 0.19$ and 0.26, respectively). Both risk alleles are also located near CXCR5, BCL9L, and TREF; none of these genes had reported cis-eQTLs in PBMCs [8]. However, we did not assay these three genes in this study, therefore could not confirm the role of disease alleles in regulating their expression in CD4$^+$ TEM cells.

Although we did not assay all genes or test for trans-acting eQTLs, based on the level of co-localization between eQTL SNPs and risk alleles observed in the study, we found it unlikely that all non-coding risk variants caused disease by altering gene expression within resting or stimulated CD4$^+$ TEM cells. In addition, while changes in proportions of lymphocyte subsets had been observed in patients of autoimmune disorders [30–35], we did not find evidence to support disease alleles’ roles in directly modulating CD4$^+$ TEM cell abundance or proliferative response. Ultimately, other cell states and cell types will need to be investigated.

We recognize several limitations to the current study. In order to conduct a focused study on a small amount of purified primary cells we used the NanoString nCounter assay system. This avoided potential biases and artifacts arising from cDNA synthesis required for microarray or RNA-seq studies, but restricted our analysis to a subset of candidate genes within risk loci of CeD, RA, and T1D, rather than a genome-wide expression analysis. Consequently we could not identify trans-eQTLs, splice variants, or epistatic effects on expression regulation. Additionally, anti-CD3/CD28 stimulation for memory T cells is not antigenic, especially while in isolation from a “natural” multi-cellular environment, thus it was only partially physiological.

This and other cell-specific studies on population variation in molecular phenotypes are only a beginning of examining potential intermediate phenotypes. Post-activation cytokine production by CD4$^+$ TEM cells are likely crucial in driving autoimmunity. Therefore, it is critical that future studies of molecular phenotypes include proteomic assays to quantify functional markers of immune response. Finally, functional experiments will need to be conducted in the future to determine whether these molecular phenotypes are indeed intermediary to disease.

**Materials and Methods**

**Ethics statement**

All research was approved by our Institutional Review Board, and informed consent was obtained from each volunteer.

**Study sample**

We enrolled 225 healthy volunteers (134 females, 91 males) of non-Hispanic Caucasian descent that proved informed consent through the Phenogenetics Project at Brigham and Women’s Hospital. Subjects’ ages ranged from 19 to 57 years with average female and male ages of 28.8 years and 34.9 years, respectively. Thirty-five subjects (18 females, 17 males) returned for a second study visit one to nine months after their initial visits.

**Genotyping**

We genotyped each subject using the Illumina Infinium Human OmniExpress Exome BeadChip. In total, we genotyped 951,117 SNPs, of which 704,808 SNPs are common variants (minor allele frequency [MAF]>0.01) and 246,229 are part of the exome. After quality control, 638,347 common SNPs remained. Of all subjects, 174 subjects had abundance, proliferation, gene expression, and quality controlled genotype data. Detailed quality control criteria are described in **Text S1**.

**SNP imputation**

For each gene, we selected a 500 kb region (250 kb each in the 3’ and 5’ directions) around the transcription start site and imputed 1000 Genomes SNPs into the genome-wide SNP data using BEAGLE Version 3.3.2. We used the European samples from 1,000 Genomes as the reference panel. We excluded markers that had MAF<0.05 in the reference panel as well as all insertion/
deletions. After imputation, we excluded markers with a BEAGLE R²<0.4 or MAF<0.01 in the imputed samples.

**CD4⁺ TEM cell isolation and stimulation**

We isolated peripheral blood mononuclear cells (PBMC) from whole blood using a Ficoll density gradient (GE Healthcare). We then isolated CD4⁺ effector memory T cells from PBMCs first by magnetic-activated cell sorting to enrich for CD4⁺ T cells, followed by fluorescent-activated cell sorting using labeled antibodies against CD45RA, CD45RO, and CD62L.

We stimulated CD4⁺ TEM cells by incubation with commercial anti-CD3/CD28 beads for 72 hours. For proliferation studies, we labeled cells with carboxyfluorescein diacetate succinimidyl ester (CFSE; eBioscience), and measured proliferation by dye dilution. Detailed isolation and purification methods are described in [Text S1](#).

**Gene expression**

We designed the NanoString codeset based on GWAS SNPs associated with CeD, RA, and T1D as of April 2012. This list of SNPs can be found in [Supplementary Table S4](#). As the numbers of associated loci with autoimmune diseases continuously expand, we refer the reader to Immunobase [https://www.immunobase.org](https://www.immunobase.org) for up-to-date disease regions. For each locus, we defined a region of interest implicated by the GWAS lead SNP [36]. We identified the furthest SNPs in LD in the 3' and 5' directions (r²>0.5). We then extended outward in each direction to the nearest recombination hotspot. If no genes were found in this region, we extended an additional 250 kb in each direction. All genes overlapping this region were considered implicated by the locus. The final NanoString codeset (prior to expression data generation) was the region of interest identified by the GWAS lead SNP and the lambda GC for CD4 TEM proportion association was 1.008; that of division index was 1.001.

**Genotype principal component analysis**

To control for any potential population stratification, we adjusted all association tests using the top five principal components of our genome-wide SNP data. Principal components were generated via EIGENSTRAT using unsupervised analysis (no reference populations were used). The top five PCs explained 6.88% (2.08%, 1.27%, 1.20%, 1.17%, and 1.16%, respectively) of the total variance. After controlling for these five PCs, the lambda GC for CD4 TEM proportion association was 1.008; that of division index was 1.001.

**Cis-eQTL analysis**

For each gene-SNP pair, we applied linear regression using the first five principal components of the genotypic data and gender as covariates. As such, normalized expression = β₀ + β₁*allelic dosage + β₂*PC₁ + β₃*PC₂ + β₄*PC₃ + β₅*PC₄ + β₆*PC₅ + β₇*factor/gender. To adjust for multiple hypothesis testing while taking into consideration the correlation among SNPs within each locus, we calculated a permutation-based P-value for each SNP. We performed 10,000 permutations of the residual expression values. We reported each SNP’s P-value the proportion of permutation P value smaller than the analytical P-value. For conditional analysis, the vector of allelic dosages of the disease-associated SNP was included as an additional covariate.

**Quantification of CD4⁺ TEM cells and proliferative response**

We defined CD4⁺ TEM cells as CD45RA⁻, CD45RO⁻, and CD62L</sup>low<sup>-</sup>. In all samples CD4⁺ TEM cells were quantified using X-Gly, a mixture-modeling based clustering program for automated cell population identification (see Figure S6) [19]. We fit proliferation division peaks with one-dimensional Gaussian mixture models (see Figure S7). Detailed protocol and algorithms are described in [Text S1](#).

**Statistical analysis**

All linkage disequilibrium calculations (r²) were based on 1000 Genomes Release 3 European samples. All association tests were performed using Plink v1.07. We considered p<5×10⁻⁸ to be genome-wide significant; p<5×10⁻⁵ was considered as suggestive. CD4⁺ TEM abundance and proliferation correlations with age and gender were calculated by multivariate linear model implemented in R-3.0. We calculated two-sample comparisons (CD4⁺ TEM cell-specific expression between genes, and H3K4me3 h/d scores between SNPs) with the Mann-Whitney test. Details of statistical analyses are described in [Text S1](#).

**Data access**

We make all phenotypic data (expression, peripheral abundance, and proliferation) along with eQTL results publicly available online (http://immunogenomics.hms.harvard.edu/CD4eqtl.html). Genome-wide genotype data will become available through dbGaP and through the ImmVar project. These data are potentially useful to investigators wishing to assess the potential of genetic variants in altering these molecular phenotypes.

**Supporting Information**

**Figure S1** Enrichment of cell-specific expression of genes within risk loci. As described in Hu et al. *AJHG* 2011, A. Genes within risk loci of RA were the most specifically expressed in CD4⁺ TEM cells (p = 1.00×10⁻⁸) followed by signal in regulatory T cells (p = 5.00×10⁻⁸). B. Genes within CeD were also the most strongly enriched in CD4 TEM cells (p = 1.43×10⁻⁸) followed by regulatory T cells (p = 3.78×10⁻⁷). C. In T1D, CD8 memory T cells showed the strongest enrichment (p = 2.26×10⁻⁷), followed by regulatory T cells (p = 5.13×10⁻⁷) and CD4⁺ TEM cells (p = 1.29×10⁻⁴).

**Figure S2** A. Using a combination of magnetic and fluorescence-activated cell sorting (MACS and FACS), CD4⁺ T cells were isolated to a high degree of purity. The isolated population contained ~97% CD3⁺ cells, ~90% CD4⁺ cells, ~0.4% CD8⁺ cells, and ~0.03% CD19⁺ cells. B. The relative abundance (as a percentage of all sorted lymphocytes), C. division index (average division of all cells), and D) proliferation index (average division of all cells that went into division), were reproducible in 35 individuals with two blood draws at least one month apart. Pearson’s r = 0.87, 0.57, and 0.62, respectively.

**Figure S3** The 46 eQTL SNPs show more overlap with H3K4me3 marks. A. Cis-eQTL SNPs were located nearer H3K4me3 peaks in CD4 TEM cells than the 159 top SNPs that did not reach statistical significance at 5% FDR (p = 1.10×10⁻⁷, one-sided Mann-Whitney test). B. The 46 cis-eQTL SNPs were
near larger H3K4me4 peaks (peak height) and located at smaller distances to the summit of the peaks ($p = 9.56 \times 10^{-9}$, one-sided Mann-Whitney test).

(TIFF)

Figure S4 The relative abundance of CD4 TEM cells as the percentage of CD4 T cells. A) increased with age, at 0.11% per year; and B) was correlated with gender, where men on average as 2.2% more CD4 TEM cells than women. The associations remained significant in a multivariate linear regression.

(TIFF)

Figure S5 SNPs associated to CeD, RA, and T1D, showed no significant association to CD4 TEM cell abundance. The 119 risk alleles within densely genotyped loci showed no significant association to CD4 TEM abundance as a percentage of CD4 T cells in the study by Orru et al. The shaded area shows the 95% confidence interval.

(TIFF)

Figure S6 Quantification of CD4 TEM cells using X-Cyt. A) In each sample of enriched CD4 T lymphocytes, X-Cyt clustered all flow events based on fluorescence intensities in CD45RA, CD45RO, and CDF62L simultaneously, using a seven-component multivariate Gaussian mixture-modeling. The TEM cell population is shown in red, defined as CD45RA$^-$, CD45RA$, and CDF62Llow/-$. B) X-Cyt clustered and quantified CD4 TEM cells in all samples (four random samples are shown here) in the study following the template in A). The TEM cell population is shown as the red cluster in each sample. In Sample 3, the subset of the black population residing in the lower left quadrant is the light green population identified as “debris” in Panel A; they are CDF62L$^-$, CD45RA$, and CD45RO$^-$.

(TIF)

Figure S7 The CFSE intensity peak present in the pooled resting wells for each subject (data underlying the green fitted curve) was modeled as a single Gaussian distribution. Its mean and variance were then used to initialize the location of the first component (undivided cells) and the variance of all components in each of the stimulated wells (data underlying the red fitted curve). The CFSE dilution peaks from stimulated wells were fitted using a one-dimensional mixture model of multiple Gaussian components of equal peak-to-peak distance and equal variance via a gradient descent optimization algorithm. A maximum of six components (five divisions) was fitted to each stimulated well; the weight of each component was allowed to be 0.

(TIF)

Table S1 Table S1 lists all genes included in the NanoString codset, including those that did not pass quality control for downstream analyses. We list each gene’s relationship to GWAS SNPs, resting and stimulated expression levels, as well as the correlation between baseline expression and post-stimulation proliferation.

(XLSX)

Table S2 Table S2 lists all densely-genotyped disease-associated SNPs for CeD, RA, and T1D included in this study. We list each SNP’s association to disease(s), nearby genes included in the study, as well as association $p$-values to CD4$^+$ TEM relative abundance and proliferation.

(XLSX)

Table S3 Table S3 lists the 182 (densely-genotyped disease-associated) SNP-gene pairs included in the eQTL study. We list each pair’s effect size and $p$-value in resting and stimulated states.

(XLSX)

Table S4 Table S4 contains all GWAS associated SNPs to CeD, RA, and/or T1D, used for designing the NanoString codset in April 2012. For each gene, we list its lead SNP in LD, disease association (to the lead SNP), GO code, and functional description.

(XLSX)

Text S1 Text S1 includes detailed descriptions of materials, experimental methods, and statistical analyses used in our study. We provide protocols and analytical methods for 1) cell population collection, isolation, staining, stimulation, flow cytometry, and NanoString assays; 2) cell abundance and proliferation quantification; and 3) gene selection, expression analysis, and eQTL analysis.

(OCX)

Acknowledgments

We thank Drs. Yukinori Okada, Buhm Han, and Deepak Rao for helpful discussions.

Author Contributions

Conceived and designed the experiments: XH HK PJB MBB SR BES CBA. Performed the experiments: HK. Analyzed the data: XH GT KS SR BES NT. Contributed reagents/materials/analysis tools: TR WMC SO SSR PLDJ. Wrote the paper: XH HK GT NT PJB CBA SSR BES SR.


