Immunology and Immunogenetics Insights

ORIGINAL RESEARCH

Comparative Analysis of Genome Sequences of the Th2 Cytokine Region of Rabbit (*Oryctolagus cuniculus*) with those of Nine Different Species

E. Michael Gertz[1], Richa Agarwala[1], Rose G. Mage[2] and Alejandro A. Schäffer[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, DHHS, Bethesda, MD, 20894, USA. [2]Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, DHHS Bethesda, MD 20892, USA. Corresponding author email: gertz@ncbi.nlm.nih.gov

Abstract: The regions encoding the coordinately regulated Th2 cytokines *IL5, IL4* and *IL13* are located on chromosomes 5 of man and 11 of mouse. They have been intensively studied because these interleukins have protective roles in helminth infections, but may lead to detrimental effects such as allergy, asthma, and fibrosis in lung and liver. We added to previous studies by comparing sequences of syntenic regions on chromosome 3 of the rabbit (*Oryctolagus cuniculus*) genome OryCun 2.0 assembly from a tuberculosis-susceptible strain, with the corresponding region of ENCODE ENm002 from a normal rabbit as well as with 9 other mammalian species. We searched for rabbit transcription factor binding sites in putative promoter and other non-coding regions of *IL5, RAD50, IL13* and *IL4*. Although we identified several differences between the two donor rabbits in coding and non-coding regions of potential functional significance, confirmation awaits additional sequencing of other rabbits.

Keywords: rabbit, genomic assembly, Th2 cytokines, tuberculosis, IL4, IL13

Immunology and Immunogenetics Insights 2011:3 59–82

doi: 10.4137/III.S7236

This article is available from http://www.la-press.com.

Introduction

Rabbits (*Oryctolagus cuniculus*), a valuable resource for diagnostic and therapeutic antibodies, are becoming increasingly important for vaccine development. The unique characteristics of their immune system make them a major source of antibodies of high affinity and specificity. Rabbits have long been models for human infectious diseases and more recently for autoimmune, neurological, ophthalmological, respiratory and cardiovascular diseases. They are widely used in development of surgical techniques, testing of therapeutics, and are also valued as a source of fur and meat in many parts of the world.

Annotation and analysis of the rabbit genome is therefore of importance for both biomedicine and agriculture and is of special importance to immunologists. NCBI maintains a Rabbit Genome Resources website (http://www.ncbi.nlm.nih.gov/projects/genome/guide/rabbit/).

The Broad Institute has submitted the second whole genome assembly of the European rabbit, completed at 6.51x coverage, to GenBank. The assembly is available in MapViewer as OryCun, build 2.0. The NIH Intramural Sequencing Center (NISC) performed clone-based sequencing of regions of the rabbit genome as part of the NISC ENCyclopedia Of DNA Elements (ENCODE) comparative sequencing project[1] and deposited the sequences in GenBank.

The ENCODE project and the Broad Institute sequenced rabbits with different genealogies and phenotypes. ENCODE sequenced an outbred New Zealand White (NZW) rabbit, whereas the Broad Institute sequenced a rabbit of the partially inbred "Thorbecke" NZW strain. The Thorbecke rabbit may have had significant immunological, physiological, and developmental abnormalities. Dorman et al[2] report that the phenotype included "ruffled fur, narrow palpebral fissures and stunted facies" and furthermore "abnormal closeness of eyes, lop ears in some animals and sedentary behavior". Rabbits of the Thorbecke strain had greater susceptibility to *M. tuberculosis* infection.[2] Despite the phenotypic abnormalities, the Thorbecke strain was chosen for sequencing at Broad Institute because it was less heterozygous than outbred NZW (personal communication to RGM). Regrettably, all Thorbecke rabbits were lost in a fire in January 2005.

In both assemblies of rabbit, the cytokine genes Interleukin 4 (*IL4*), Interleukin 13 (*IL13*), and Interleukin 5 (*IL5*) were placed near each other in the "Th2 cytokine region", with synteny to corresponding regions in human and mouse. The Broad Institute assigned the region to rabbit chromosome 3. The Th2 region, and the IL4 cytokine in particular, have been linked to the progression and severity of tuberculosis.[3–5] It was of interest to learn whether any variants in the region with *IL5, IL4, IL13,* and other nearby genes (*RAD50, KIF3A*) could have contributed to immune system deficits in the Thorbecke rabbit.

The cytokines encoded in the Th2 region are characteristic of type 2 immunity. Type 2 immunity has important protective roles in responses to helminth infections, but detrimental effects include allergy-associated IL4-induced elevations in serum IgE, IL5-induced eosinophilia and airway remodeling in asthma, and IL13-induced epithelial cell damage leading to fibrosis in lung, or in liver, during helminth infections.[6] The Th2 region was selected for sequencing by ENCODE because of the important roles that cytokines play in determining the developmental fate and effector functions of T lymphocytes in the immune system.[7] The expression of *IL4*, *IL13* and *IL5* in this region is coordinately regulated, and the finding of conserved non-coding regions suggests that the mechanism of regulation is also conserved in syntenic regions of other species.[8] The conserved structure of the Th2 region is shown in Figure 1.

To identify conserved noncoding sequences in the Th2 region, we conducted comparative genome sequence analysis in 10 mammalian species including the rabbit, mouse, and human. Previous studies have used a functional approach, usually in mice, to define roles for various transcription factors in the Th2 cytokine region. Among the transcription factors

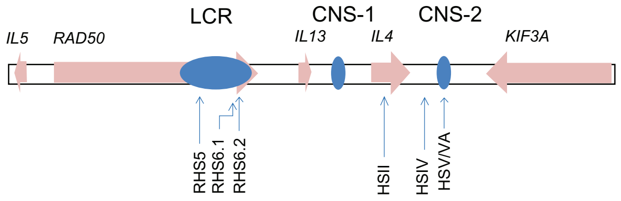

Figure 1. Conserved structure of the Th2 region. A schematic (not drawn to scale), of the structure of the Th2 region, conserved across many species, including those used in this study. The genes in the region are *IL5*, *RAD50*, *IL13*, *IL4*, and *KIF3A*, and the direction of transcription is shown using arrows. The locus control region, near the end of the *RAD50* gene, is labeled LCR.

known to bind to at least one location in the region are Ets-1,[9] GATA3,[10,11] c-Maf,[12] RBPJK,[13] Runx3,[14] IRF4,[15] JunB[16] and STAT family members.[17,18] Strempel et al[9] did multi-species bioinformatic comparisons to reach predictions of only Ets-1 and GATA binding sites, but their work included neither other transcription factors nor the rabbit.

We sought to address three general questions:

1. Do the Broad and ENCODE assemblies of the Th2 region differ in gene content, and is it possible that these differences had phenotypic consequences?
2. Are the sites predicted by Strempel et al[9] conserved in rabbit, and if so, what are the rabbit-specific binding sites?
3. Can we find transcription factor binding sites (TFBS) conserved across mammals for some transcription factors other than Ets-1 and GATA?

Results

Genomic sequences

We studied genomic sequences containing the genes *IL5*, *RAD50*, *IL13*, *IL4* and *KIF3A* from rabbit and the nine species used by Strempel et al.[9] Table 1 lists the species and the genomic sequences used in this study.

We considered the possibility of adding additional species to the study. A Th2 region syntenic to that in human exists in chicken (*Gallus gallus*).[19] We did not use the chicken genome because we found few conserved non-coding regions in chicken by a Mulan alignment (data not shown). Strempel et al[9] state the same reason for not using chicken. As of March 2011, the only other whole genome sequence in NCBI MapViewer that has a clear Th2 region belongs to Sumatran orangutan (*Pongo abelii),* which we did not add to the study since we already include two species of great ape, human and chimpanzee.

Comparison of the Broad and ENCODE sequences within predicted genes

We compared the Broad and ENCODE sequences and annotations of the genes *IL5*, *RAD50*, *IL4*, *IL13*, and *KIF3A*. We were able to confirm, by alignment, the placement of most exons in these genes (see Supplementary Data). The exceptions were that exons 4 and 5 of *IL5* could not be placed on the ENCODE sequence, that exon 6 of *RAD50* could not be placed on the Broad sequence, and that the ENCODE annotation did not include what Broad annotates as exons 10 and 11 of *KIF3A*. Further analysis suggests that *RAD50* was misassembled in Broad and that there exists insufficient evidence to support Broad's prediction of putative exons 10 and 11 in *KIF3A*.

We compared the assembled coding regions of these five genes (see Supplementary Data for details). We found a substitution of a Threonine (Thr) in Broad for a Proline (Pro) in ENCODE at amino acid 27 of IL13. The substitution is supported by traces in the NCBI trace archive. In-silico structural analysis and comparison with homologous sequences suggest that both Thr27 and Pro27 would be tolerated.

Of possible immunological interest, there is a frameshift mutation in exon 2 of *IL4* in the Broad assembly. This frameshift is supported by the trace with identifier 2047213760. A second trace, identifier 2061258363, aligns with the single nucleotide insertion, but has two gaps elsewhere in the alignment. Because the coverage of this position in *IL4* is at

Table 1. Genomic sequences used for comparative sequence analyses.

Organism	Chromosome	GenBank Id (GI)	Region start	Region stop
O. cuniculus (rabbit)	ENCODE ENm002	217273035	683500	906000
O. cuniculus (rabbit)	3	261748885	15550000	15783043
Homo sapiens (human)	5	224589817	131725000	132075000
Pan troglodytes (chimpanzee)	5	114796134	134362765	134140459
Papio anubis (baboon)	ENCODE ENm002	159461516	626932	841239
Callithrix jacchus (marmoset)	2	290467407	72733448	72922579
Otolemur garnetti (bush baby)	ENCODE ENm002	197215648	819273	1049297
Bos taurus (cow)	7	194719537	20421661	20595318
Canis familiaris (dog)	11	74030065	23810384	24049067
Rattus norvegicus (rat)	10	62750810	39029351	39200657
Mus musculus (mouse)	11	149288871	53380000	53540000

most 2x, the evidence for the insertion is weak. No traces matched the ENCODE/wild-type sequence, so there is no evidence that the sequenced rabbit was heterozygous for the *IL4* single-nucleotide insertion.

See Figure 2 for alignments of the rabbit, human and mouse protein sequences of the genes *IL5*, *IL13*, and *IL4*.

Comparison of the broad and ENCODE promoter sequences

We aligned promoter sequences for *IL5, RAD50, IL13,* and *IL4* from the ENCODE genomic sequences to the Broad assembly; see Supplementary Data. The ENCODE *RAD50* and *IL13* promoter sequences align to the Broad assembly with full coverage and high percent identity. The Broad *IL5* and *IL4* promoters

A. Alignment of the IL5 protein for rabbit, human, and mouse.

```
XP_002710247.1   M-RMLLHWTLLALGAAYVCAMATEIRMSTVVKETLTLLSTYQSLLIGNETLMIPVPVHKNH
NP_000870.1      M-RMLLHLSLLALGAAYVYAIPTEIPTSALVKETLALLSTHRTLLIANETLRIPVPVHKNH
NP_034688.1      MRRMLLHLSVLTLSC--VWATAMEIPMSTVVKETLTQLSAHRALLTSNETMRLPVPTHKNH

XP_002710247.1   HLCIEETFRGVDTLKAQIVQGEAMDNLFQNLYLIKKYIDLQKKKCGEERRGVKHFLDYLQE
NP_000870.1      QLCTEEIFQGIGTLESQTVQGGTVERLFKNLSLIKKYIDGQKKKCGEERRRVNQFLDYLQE
NP_034688.1      QLCIGEIFQGLDILKNQTVRGGTVEMLFQNLSLIKKYIDRQKEKCGEERRRTRQFLDYLQE

XP_002710247.1   FLGVINTEWTMES
NP_000870.1      FLGVMNTEWIIES
NP_034688.1      FLGVMSTEWAMEG
```

B. Alignment of the IL13 protein for rabbit, human, and mouse.

```
XP_002710138.1   --------------MALWWAVAIAVTCLGSLVSPGPVPPPT----SLKELIEELVNITHNQ
ENCODE           --------------MALWWAVAIAVTCLGSLVSPGPVPPPP----SLKELIEELVNITHNQ
NP_002179.2      MHPLLNPLLLALGLMALLLTTVIALTCLGGFASPGPVPPST----ALRELIEELVNITQNQ
NP_032381.1      --------------MALWVTAVLALACLGGLAAPGPVPRSVSLPLTLKELIEELSNITQDQ

XP_002710138.1   KAPLCNGTMVWSVNLTGSVYCAALESLVNVSGCNAIQRTQRMLSGLCTDKAVAKQVTSVQA
ENCODE           KAPLCNGTMVWSVNLTGSVYCAALESLVNVSGCNAIQRTQRMLSGLCTDKAVAKQVTSVQA
NP_002179.2      KAPLCNGSMVWSINLTAGMYCAALESLINVSGCSAIEKTQRMLSGFCPHKVSAGQFSSLHV
NP_032381.1      -TPLCNGSMVWSVDLAAGGFCVALDSLTNISNCNAIYRTQRILHGLCNRKAPT-TVSSLP-

XP_002710138.1   RDTKIELLQFLKELRRHLQMLYRLGKFR
ENCODE           RDTKIELLQFLKELRRHLQMLYRLGKFR
NP_002179.2      RDTKIEVAQFVKDLLLHLKKLFREGQFN
NP_032381.1      -DTKIEVAHFITKLLSYTKQLFRHGPF-
```

C. Alignment of rabbit IL4 and IL4δ2, human IL4 and IL4δ2, and mouse IL4.

```
NP_001156649.1   MGLPAQLPVTLLCLLAGTAHFIQGRRGDIILPEVIKTLNILTERKTPCTKLMIADALAVPK
NP_001164577.1   MGLPAQLPVTLLCLLAGTAHFIQGRRGDIILPEVIKTLNILTERK----------------
NP_000580.1      MGLTSQLLPPLFFLLACAGNFVHGHKCDITLQEIIKTLNSLTEQKTLCTELTVTDIFAASK
NP_758858.1      MGLTSQLLPPLFFLLACAGNFVHGHKCDITLQEIIKTLNSLTEQK----------------
NP_067258.1      MGLNPQLVVILLFFLECTRSHIHGCD-KNHLREIIGILNEVTGEGTPCTEMDVPNVLTATK

NP_001156649.1   NTTEREAVCRAATALRQFYLHH-KVSWCF-----KEHGELGDLRLLRGLDRNLCSMAKLSN
NP_001164577.1   NTTEREAVCRAATALRQFYLHH-KVSWCF-----KEHGELGDLRLLRGLDRNLCSMAKLSN
NP_000580.1      NTTEKETFCRAATVLRQFYSHHEKDTRCLGATAQQFHRHKQLIRFLKRLDRNLWGLAGLNS
NP_758858.1      NTTEKETFCRAATVLRQFYSHHEKDTRCLGATAQQFHRHKQLIRFLKRLDRNLWGLAGLNS
NP_067258.1      NTTESELVCRASKVLRIFYLKHGK-TPCL-------KKNSSVLMELQRLFRAFRCLDSSIS

NP_001156649.1   CPGKEARQTTLEDFLDRLKTAMQEKYSKRQS
NP_001164577.1   CPGKEARQTTLEDFLDRLKTAMQEKYSKRQS
NP_000580.1      CPVKEANQSTLENFLERLKTIMREKYSKCSS
NP_758858.1      CPVKEANQSTLENFLERLKTIMREKYSKCSS
NP_067258.1      CTMNESKSTSLKDFLESLKSIMQMDYS----
```

Figure 2. Alignments of the rabbit, human, and mouse protein sequences for IL5, IL13, and IL4. Panel **A** shows the alignment of the IL5 proteins for rabbit (XP_002710247.1), human (NP_000870.1), and mouse (NP_034688.1). The ENCODE assembly does not encode a full length IL5 protein; it omits exons 4 and 5. Panel **B** shows the alignment of the IL13 proteins for rabbit (XP_002710138.1 and ENCODE), human (NP_002179.2), and mouse (NP_032381.1). The ENCODE IL13 protein sequence is a translation of DNA from the ENCODE assembly, and has a substitution of a Pro for a Thr at position 27 with respect to the reference IL13 sequence for rabbit (XP_002710138.1). Panel **C** shows the alignment of rabbit IL4 and IL4δ2 (NP_001156649.1 and NP_001164577.1), human IL4 and IL4δ2 (NP_000580.1 and NP_758858.1), and mouse IL4 (NP_067258.1). There is no sequence for mouse IL4δ2 in GenBank.

matched the ENCODE sequences well, but the Broad sequences had runs of the ambiguity character N that split the alignment into partial matches. Because the promoter regions in ENCODE do not contain Ns, we used the ENCODE sequences for cross-species comparison and de-novo prediction of binding sites.

Placement of Ets-1 and GATA binding sites

We placed the Ets-1 and GATA binding sites described in Strempel et al[9] on both rabbit assemblies using two methods. The first method was direct alignment by BLAST[20] of the sequences provided by Strempel et al.[9] The second method was to use the Mulan[21] and multiTF algorithms to place the binding sites. These placement methods gave similar results, but they differ from the results of Strempel et al[9] in part because Strempel et al[9] used the MatInspector program, rather than multiTF, to predict binding sites. MatInspector uses a proprietary library, and we cannot use the program due to the restrictive license on how annotations generated by MatInspector may be published.

Ets-1 and GATA binding sites placed using BLAST

Twelve of the 19 Ets-1 and GATA transcription binding sites could be unambiguously placed on both the Broad and ENCODE assemblies by alignment to the homologous sequences in the other nine species. The locations of these binding sites are shown in Table 2.

Each site in Table 2 aligns to the homologous sequence of at least eight of the species with coverage of at least 80% and E-value of at most 0.1, except HSIV and Ets-1 *IL13* Promoter. Ets-1 *IL13* Promoter cannot be confidently placed by BLAST alone, as only three homologous sequences aligned to rabbit regions with the required coverage and E-value cutoff. The multiTF program, however, predicts that the location shown is correct (see the following subsection). Only six of the nine homologs of HSIV had an alignment to rabbit with the required coverage and E-value cutoff. The three homologs of HSIV (length 21) that do not align with 80% coverage to the rabbit sequences do, however, have perfect alignments of length 16 to the putative binding site in rabbit. The alignments cover the core binding motif, and attain an E-value of 0.001.

For both assemblies, eight of the nine CNS-2(1) homologs align to the location shown in Table 2. However, six of the CNS-2(1) homologs align to a secondary location. The secondary alignment could be eliminated positionally, as it was above *IL4*, whereas CNS-2(1) should be below.

Ets-1 and GATA binding sites placed using multiTF

We used Mulan to align the ENCODE rabbit genomic sequences with the nine other species shown in Table 1. We then used an option on the Mulan website to pass the multiple alignment to multiTF. The multiTF algorithm uses the alignment and the TRANSFAC matrix library,

Table 2. Ets-1 and GATA binding sites that could be placed by BLAST.

	Binding sites[a]	ENCODE region ENm002			Broad chromosome 3		
		Start	Stop	Strand	Start	Stop	Strand
Ets-1 sites	Ets-1 *IL5* Promoter	703734	703754	–1	15575916	15575936	–1
	RHS5	787911	787931	+1	15660536	15660556	+1
	IL13 promoter*	816434	816453	+1	15689042	15689061	+1
	IL4 promoter.1	830425	830445	+1	15702565	15702585	+1
	IL4 promoter.2	830464	830482	+1	15702604	15702622	+1
	Ets-1 IL4IE	831756	831775	+1	15703899	15703918	+1
	HSIV	841204	841224	+1	15713642	15713662	+1
	CNS2	844617	844637	+1	15717059	15717079	+1
GATA sites	GATA *IL5* promoter	703763	703776	–1	15575945	15575958	–1
	RHS 6.1	795825	795838	+1	15668451	15668464	+1
	IL4P	830326	830339	+1	15702466	15702479	+1
	CNS-2(1)	844525	844538	+1	15716967	15716980	+1
	CNS-2(2,3)	844583	844607	+1	15717025	15717049	+1

Notes: [a]Binding site names follow the names in Strempel et al.[9] IL13 Promoter is marked with an asterisk to indicate that it could not be placed using BLAST alone, but that multiTF suggests that the location shown is correct.

version 10.6, to identify conserved transcription factor binding sites. TRANSFAC predicts the presence of a binding site for each species individually based on its genomic sequence; it does not use the multiple alignment. The multiTF program reports locations that are in conserved regions of the Mulan alignment and that are predicted by TRANSFAC to be binding sites in all 10 species. While only eight of the 19 binding sites reported by Strempel et al[9] were also reported by multiTF, we were able to locate 17 of 19 binding sites in a conserved region reported by Mulan; see Supplementary Table S6. The two exceptions were IL13P(2) and IL13P(3).

The reason that some positions were found in a conserved block by Mulan, but were not reported by multiTF, is that TRANSFAC did not report the binding site in all 10 species. For example, the elements *IL4* Promoter.1 and *IL4* Promoter.2 were placed by BLAST in the ENCODE sequence at the coordinates shown in Table 2. Mulan, in fact, places these coordinates within a conserved region that extends from 830030 to 830919. The multiTF program, however, does not find a conserved Ets-1 binding site within that block.

Because some of the binding sites were not predicted as conserved by multiTF in the 10-species comparison, we asked whether they were at least conserved between rabbit and mouse. We performed two more multiTF queries; one with the ENCODE genomic sequence and the mouse sequence, the other with the Broad sequence and the mouse sequence. For these queries, multiTF located 18 of the 19 binding sites from Strempel et al,[9] including the IL13P(2) site that was not in a conserved block of the 10 species alignments. These queries did not identify a conserved homolog of the IL13P(3) binding site (see next subsection). Table 3 shows the locations of the binding sites. Figure 3 shows a map of the binding sites placed relative to the *IL5, RAD50, IL13*, and *IL4* genes on the Broad assembly.

Comparisons of Table 2 with Table 3 show that for the elements common to both tables, the results from multiTF confirm the results from direct alignment. Small differences in extent are not relevant; the extent from Table 2 should be used. Table 3 has five entries that are not found in Table 2: CNS-1, RHS6.2, IL13P(1), IL13P(2) and GATA IL4IE. CNS-1 is the only one of the five for which multiTF predicts a conserved binding site in the 10 species comparison. The Broad and ENCODE genetic sequences were identical at the positions listed in Tables 2 or 3.

IL13P(3) may not be a GATA binding site in rabbit

The binding site IL13P(3) seems to be lost in rabbit. In the ENCODE sequence, start of transcription for

Table 3. Binding sites predicted by multiTF.

		ENCODE start	ENCODE stop	Broad start	Broad stop	Length
Ets-1 sites	Ets-1 *IL5* promoter	703741	703752	15575923	15575934	12
	RHS5	787915	787927	15660540	15660552	13
	IL13 promoter	816436	816453	15689044	15689061	18
	Ets-1 IL4IE	831760	831769	15703903	15703912	10
	HSIV	841207	841224	15713645	15713662	18
	CNS2[a,b]	844623	844645	15717065	15717087	23
GATA sites	GATA *IL5* promoter	703764	703776	15575946	15575958	13
	RHS6.1	795825	795837	15668451	15668463	13
	RHS6.2[a]	796913	796921	15669539	15669547	9
	IL13P(1)[a]	815922	815931	15688527	15688536	10
	IL13P(2)[a,c]	815948	815961	15688553	15688566	14
	CNS-1[a]	822927	822935	15695429	15695437	9
	IL4P	830327	830336	15702467	15702476	10
	GATA IL4IE[a]	831382	831395	15703525	15703538	14
	CNS-2(1)[a]	844527	844536	15716969	15716978	10
	CNS-2(2,3)[a]	844584	844603	15717026	15717045	20

Notes: [a]Found to be conserved when mouse and rabbit were compared, but not when all 10 species were used; [b]The CNS2 site found in mouse-rabbit comparison was wider than the one found in the 10 species comparison; [c]Not only was the binding site not predicted in the 10 species alignment, but IL13P(2) is not fully contained in any of the Mulan aligned blocks.

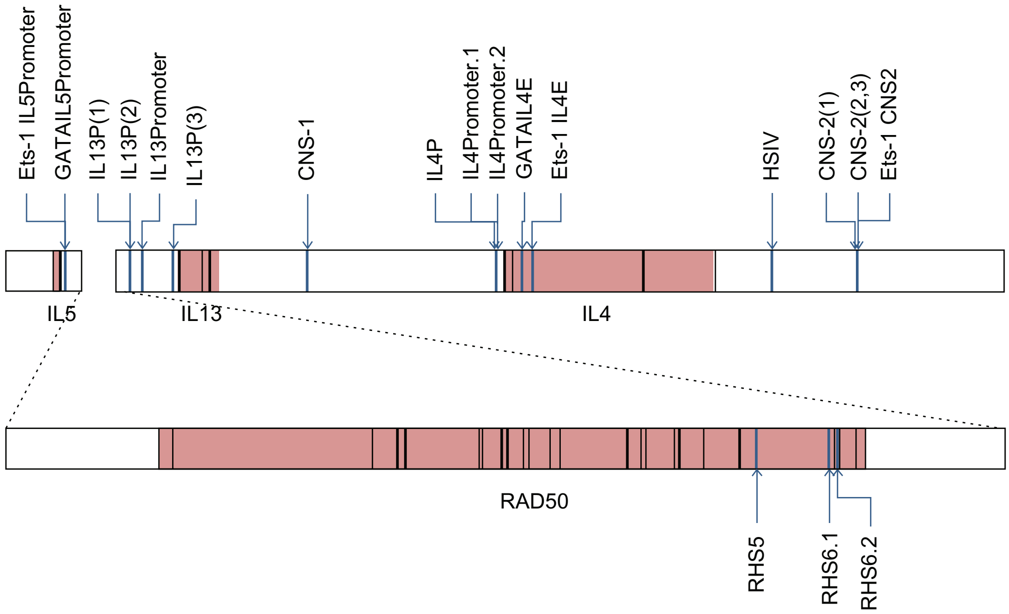

Figure 3. GATA and Ets-1 binding sites in Th2 region of rabbit. Diagram of the Th2 region in ENCODE assembly of rabbit, spanning rabbit sequence NT_165851.1, bases 701372 to 850370. Coordinates for genes and exons were obtained by aligning the rabbit reference mRNA sequences to the ENCODE assembly; note that two exons of *IL5* were not found. The coordinates for the TFBS are as computed in this document, Table 3. IL13P(3) is included in the figure, though it is not predicted to be a binding site in rabbit. Blue lines represent binding sites, pink boxes are genes and black boxes are exons within gene. Arrows point to binding sites, so the color information is redundant. Because the *RAD50* gene is large, the gene and surrounding intergenic region from bases 704372 to 850370 are drawn separately and at an approximately 3x compressed scale.

IL13 is at 817825. In mouse, IL13P(3) is 72 bases upstream, so we assume that IL13P(3), if conserved, would be located near 817825 in the ENCODE sequence. Mulan finds a conserved block that spans bases 817397 to 818100 in the 10 species alignment. The mouse and human orthologs of IL13P(3) are located in this block and are aligned with each other. The multiTF program does not predict any conserved GATA binding sites in the rabbit sequence within this block, and indeed the alignment has a gap in the rabbit sequence near the putative binding site suggested by Mulan. The gap at this location for the NZW is supported by 29 traces. The identical gap appears in the Broad sequence, supported by 11 traces, giving further evidence that IL13P(3) is missing in rabbit.

Binding sites for additional transcription factors

We used multiTF with the 10-species alignment to find putative binding sites for the transcription factors listed in Table 4. The sites predicted by multiTF are shown in Table 5. Because several transcription factors were predicted to bind to more than one site, the sites were each assigned a distinct identifier, shown in the leftmost column. The block start and block stop are the beginning and end of the ENCODE

Table 4. Transcription factor binding sites analyzed by comparative sequence analyses.

Transcription factor	Recognition matrices
IRF4	VIRF_Q6, VIRF_Q6_01
JunB	V$AP1_Q2_01, V$AP1_Q4_01, V$AP1_Q6_01
MAFG	V$CMAF_01, V$TCF11MAFG_01, V$VMAF_01
NFAT	V$NFAT_Q4_01, V$NFAT_Q6
NFκB	V$NFKB_C, V$NFKB_Q6, V$NFKB_Q6_01
PU.1	VETS_Q6, VPU1_Q6
RBPJK	V$RBPJK_01, V$RBPJK_Q4
Runx3	VAML_Q6, VPEBP_Q6
STAT5	V$STAT5A_01, V$STAT5A_02, V$STAT5A_03, V$STAT5A_04*, V$STAT5B_01, V$STAT_01*, V$STAT_Q6

Table 5. Transcription factor binding sites predicted in the Th2 region.

Binding site ID	Site start	Site stop	Block start	Block stop	Location
07_MAFG	795722	795743	795478	796187	LCR
08_JunB	795730	795738	795478	796187	LCR
09_IRF4	795803	795817	795478	796187	LCR
15_NFκB	817633	817648	817397	818100	*IL13* Promoter
16_NFAT	817636	817645	817397	818100	*IL13* Promoter
17_Runx3	817666	817680	817397	818100	*IL13* Promoter
18_STAT5	823049	823061	822978	823420	Near CNS-1
20_Runx3	823173	823187	822978	823420	Near CNS-1
21_NFκB	830358	830371	830030	830919	*IL4* Promoter
22_IRF4	830404	830414	830030	830919	*IL4* Promoter
24_NFAT	830523	830534	830030	830919	*IL4* Promoter
25_STAT5	831768	831794	830931	831858	HSII
26_Runx3	841024	841038	840789	841259	HSIV
28_RBPJK	844690	844700	844506	844845	HSV/VA

rabbit sequence in the aligned block in the Mulan alignment. Figure 4 shows the location of each site within the Th2 region.

Table 5 does not contain all the sites we expected to exist in rabbit; in particular, we expected a STAT5 site in the locus control region (LCR). We sought a larger list of putative binding sites so that we could examine the Mulan alignment to determine why some of the expected sites were not found. If a site predicted by multiTF to be conserved in human, mouse and rabbit was not found in the 10-species alignment, that site was selected for further study. The sites found in the three-species alignment, but not found in the 10-species alignment, are shown in Table 6. Supplementary Tables S7–S9 show the multiple alignments for all transcription factors we studied, when such an alignment could be generated. Among the transcription factors we considered, multiTF did not predict any overlapping binding sites.

In Tables 5 and 6, sites were assigned a putative conserved noncoding region using positional reasoning described in Supplementary Data.

Discussion

The availability of two rabbit sequences for the Th2 cytokine region enabled us to do a variety of cross-species

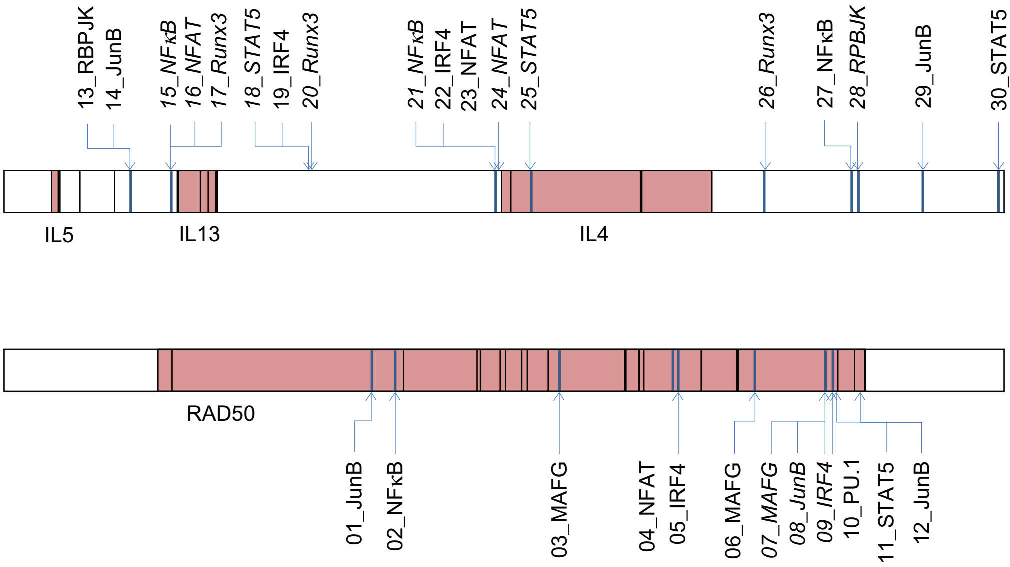

Figure 4. Other transcription factor binding sites in Th2 region of rabbit. Diagram of the Th2 region in ENCODE for rabbit. Coordinates, genes, and exons are the same as those used for Figure 3. The TFBS shown are the union of the sites listed in Tables 5 and 6, with sites from Table 5 shown in *italicized* font.

Table 6. Sites found in three-species alignment for human, mouse and rabbit not found in ten-species alignment.

Promoter	Site start	Site stop	Block start	Block stop	Location
01_JunB	745281	745289	744976	745405	*RAD50*: exon3
02_NFκB	747927	747942	747818	747970	*RAD50*: exon4
03_MAFG	766120	766138	765955	766200	*RAD50*: exon13
04_NFAT	778852	778863	778807	778950	*RAD50*: exon17
05_IRF4	779453	779467	779437	779720	*RAD50*: exon19
06_MAFG	787960	787978	787608	788040	*RAD50*: intron21
10_PU.1	796560	796561	796552	796561	LCR
10_PU.1	796562	796567	796562	796636	LCR
11_STAT5	796780	796794	796637	796967	LCR
12_JunB	799629	799630	799629	799630	LCR
12_JunB	799631	799640	799631	800321	LCR
13_RBPJK	815975	815985	815954	815995	*IL13* Promoter
14_JunB	816038	816050	815996	816081	*IL13* Promoter
19_IRF4	823076	823090	822978	823420	Near CNS-1
23_NFAT	830408	830419	830030	830919	*IL4* Promoter
27_NFκB	844406	844412	844287	844412	HSV/VA
27_NFκB	844413	844422	844413	844505	HSV/VA
29_JunB	847238	847246	847058	847351	*IL4-KIF3A*
30_STAT5	850188	850195	850109	850228	*IL4-KIF3A*

and cross-rabbit analyses. The phylogenetic study of Strempel et al[9] identified Ets-1 and GATA binding sites within major Th2 cis-regulatory elements that map to extensive (300–600 bp) regions that are highly conserved between mice and humans, but that study did not include rabbit. Because the DNA donor for the Broad 6.51x OryCun 2.0 assembly was from a partially inbred strain that had developmental defects and was more susceptible to Mycobacterial infection than outbred NZW, such as the ENCODE project's DNA donor,[2] we sought to identify differences in the exons and transcription factor binding sites that might be associated with the phenotypic differences.

As summarized in Supplementary Table S3, we found a substitution in IL13 and a frame shift in IL4 that might be relevant to the phenotypic differences. Other discrepancies in the assemblies included missing exons (IL5 in ENCODE) and extra exons (KIF3A in OryCun 2.0). That the only available full rabbit genome assembly is from an extinct strain with an abnormal phenotype poses problems for future rabbit genomic studies. The OryCun 2.0 assembly has hundreds of regions of contiguous assembled sequence that are not placed on any chromosome, many stretches with ambiguity characters (Ns) and poor coverage of some regions such as the potential frameshift in IL4.

Comparative analyses of binding sites and promoter regions

We could place 18/19 Ets-1 and GATA binding sites described in Strempel et al[9] on the Broad OryCun 2.0 assembly and the ENCODE rabbit region ENm002. The sequence that was not placed was identified by Strempel et al[9] as a GATA binding site, but was not predicted to be a GATA binding site in rabbit. The rabbit sequence does align to the orthologous binding sites, but there is a single nucleotide deletion in rabbit, causing the rabbit sequence not to be predicted as a GATA binding site. Twelve of 19 sequences could be directly placed using BLAST, and so are highly likely to be identified correctly. Six others could be placed by multiple alignment, plus a prediction of transcription factor binding sites by multiTF. There were binding sites that could only be placed by BLAST and were not predicted by multiTF.

We conducted analyses of additional transcription factor binding sites. Among the sites that were not conserved across all species, the 11_STAT5 site particularly caught our attention because this TFBS is located in the locus control region. The *Papio anubus* sequence was not predicted to have a STAT5 binding site at the homologous location, and there were 19 traces in the trace archive that support the *Papio anubus* sequence. *Callithrix jacchus* has a gap in the alignment where the STAT5 binding site would be. More generally, the "no" entries in Supplementary Table S9 define a set of (species, transcription factor, site) combinations that merit further investigation. If these sites are not present in all mammals, then this would have implications for evolution of T cell regulation.

Roles for RAD50 and KIF3A

This study includes analysis of *RAD50* and *KIF3A,* although they do not encode Th2 cytokines. Why are these genes conserved in syntenic relationships to the cytokine genes, including avian species thought to have diverged from the mammalian lineage 300 million years ago? Although *RAD50* is widely expressed, it appears to serve a secondary function in its location by harboring locus control sequences in its 3′ untranslated region.[17,22,23] Locating the LCR within *RAD50* but near the *IL4* and *IL13* cytokine genes may be advantageous because *RAD50* is accessible and transcribed as a housekeeping gene and at the same time, the LCR contributes to the regulation of the adjacent *IL4* and *IL13*. Similarly *KIF3A* may be preserved in the syntenic relationship to serve secondary functional roles. There is complex epigenetic control of the polarization steps toward characteristics of activated Th2 cells.[24–27] Chromatin remodeling brings together distant sites within the locus.[28] Th2 cell activation upregulates production of SATB1[29,30] which then binds to CNS1, CNS2 and 9 other sites extending from *IL5* past *KIF3A*. CTCF also binds between *IL5* and the neighboring *IRF1* and within the *KIF3A* gene, helping to segregate the Th2 domain from surrounding regions.[31]

Tuberculosis and the Th2 cytokine region

At the start of this project, we hypothesized that it was possible that variants in the region with *IL4, IL13*, and other nearby genes of immunological interest (*IL5, RAD50*) could have contributed to some immune system deficits in the partially inbred Thorbecke rabbit that led to this strain's decreased resistance to tuberculosis.[2] Recently, in a family-based association study of human tuberculosis, potential risk haplotypes contributing to tuberculosis susceptibility were suggested to reside on a region of human chromosome 5 encompassing Th2 cytokines within a three-marker haplotype of SNPs in *SLC22A4, SLC22A5* and *KIF3A*.[32] This haplotype may influence cytokine expression levels and influence the magnitude of T-cell responses to *Mycobacterium tuberculosis.*

The sequence differences we found between the Broad and ENCODE assemblies appear to be largely due to N's in the Broad assembly or sequencing errors, not true differences in the DNA sequences of the two donor rabbits. A splice variant equivalent to human IL4δ2[33] was reported in rabbits in 2000,[34] in several primates[35] and in mice.[36] There is a possible frameshift mutation in exon 2 of the IL4 gene in the rabbit used for the Broad assembly. If correct, this could force production of the alternatively spliced IL4δ2 variant product that lacks exon 2, at least from one allele. A pathological role of IL4 and other type 2 cytokines during responses to pulmonary infections with *Mycobacterium tuberculosis* has been suggested.[5] Although patients with increased expression of IL4 mRNA had more extensive disease, they were also observed to exhibit greater expression of IL4δ2.[3] Accurate measurements of message levels are complicated by relative instability of IL4δ2 message.[4] In addition, determinations of mRNA expression levels in cells obtained from sites of infection may be more relevant than measurements of levels produced by cells from peripheral blood.[4,5] The rabbit is an excellent model for human pulmonary tuberculosis because lung pathology in both man and rabbit includes pulmonary granulomas with caseous necrosis.[2] Increased IL4 production in tuberculosis was associated with development of pulmonary cavities.[37,38] Recently Luzina et al reported differences in pulmonary cytokines and cellular infiltrates elicited when human or murine full-length IL4 or IL4δ2 was virally expressed in mouse lungs.[39,40] Their studies demonstrate functional roles for IL-4δ2 independent and distinct from IL4. Even if the possible frameshift were a sequencing error, further studies of *Mycobacterium tuberculosis* models in rabbits should evaluate IL4 levels and screen for expression levels of both the long and IL4δ2 forms of IL4.

Our sequence analysis of the Th2 region in rabbit and other mammals suggests areas for further investigation in at least four directions. First, the transcription factor binding sites in the Th2 region appear to be variably conserved in mammalian evolution. The immunological function of binding sites present in some mammals and absent in others should be tested. Second, there are likely sequence differences between rabbits in the exons of Th2 region genes. Third, laboratories currently using the rabbit model do observe different responses to experimental infection with *M. tuberculosis*.[41,42] Potential differences in binding sites and in the coding regions of *IL4* and *IL13* reported here may be confirmed and extended in future studies using rabbits that develop differential disease presentation when infected with the same species and strain of Mycobacterium. Finally, deficiencies in the assembly and annotation of the current OryCun2.0 rabbit genome sequence emphasize the need for further sequencing of rabbits from other strains of this species and improved assembly of the many complex regions of interest to immunologists.

Methods

Ets-1 and GATA binding sites

The supplemental data found in Strempel et al[9] provided the genomic sequences of 19 evolutionarily conserved sites in the nine species listed in Table 1. One hundred seventy, rather than 171, sequences are listed because no sequence for the RHS 6.2 site was available for *Callithrix jacchus.* Ten of 19 sites have sequences length 14, eight have length 21, and one has length 25. Eight of the listed sites are Ets-1 binding sites, and 11 are GATA binding sites.

Alignments using BLAST

We used NCBI BLAST[20] to align the sequences of the 170 binding sites to the Broad OryCun 2.0 and ENCODE rabbit sequences cited in Table 1. We used version 2.2.23 of BLAST with word size 4, match reward 2, mismatch penalty -3 and no filtering (options: -r 2 -q -3 -W 4 –FF). In our usage, the purpose of using –FF was to show that even with filtering off, multiple placement was not a problem. For some results, we filtered the BLAST output further. We excluded alignments with less than 80% coverage of the query sequence. Coverage is defined as the extent of the alignment in the query, divided by the full length of the query. We also applied a filter to the BLAST results that excluded alignments with E-value $>$ 0.1. We did not use the –E option to BLAST because this option affects some of the internal heuristics.

Multi-species alignment

We used the Mulan[21] alignment algorithm (http://mulan.dcode.org/), to align the genomic sequences shown in Table 1. We generated four distinct alignments using Mulan: one that aligned the ENCODE sequence with the sequences from the other nine species; one that aligned the ENCODE sequence with the syntenic region from mouse; one that aligned the Broad OryCun 2.0 sequence to the syntenic sequence from mouse; and one that aligned human, mouse, and the ENCODE sequence for rabbit.

Prediction of binding sites using multiTF

We used multiple alignments produced by Mulan as input to multiTF (http://multitf.dcode.org/), a program that uses the TRANSFAC 10.6 library to identify conserved transcription factor binding sites.

The binding factors in the TRANSFAC 10.6 library with identifiers beginning with “V$CETS” or “V$ETS”, except for “V$ETS2_B”, were considered to identify Ets-1 binding sites. The binding factors with identifiers starting with “V$GATA” were considered to identify GATA binding sites.

We sought binding sites for the additional transcription factors listed in column one of Table 4. The matrices used by multiTF to recognize the transcription factor binding sites are shown in the second column. Putative binding sites were found in the Th2 region for all the matrices listed in Table 4, except for the two matrices marked with an asterisk. Both of these would recognize binding sites for STAT5. The V$ETS_Q6 matrix, which recognizes PU.1, also recognizes the transcription factor Ets-1, so we filtered known Ets-1 binding sites from the list of predicted PU.1 binding sites.

Strempel et al[9] warn of incompleteness of the *Callithrix jacchus* sequence near the Th2 locus control region. We found no specific case in which the *Callithrix jacchus* genome alone prevented recognition of a binding site for one of the transcription factors listed in Table 4. Therefore, we did not handle *Callithrix jacchus* in any special way.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health, NLM and NIAID. Thanks to David Margulies for advice on using PyMol. Thanks to Jinfang Zhu for helpful suggestions on the design of the study and on the text of the manuscript. We also appreciate additional comments on the manuscript from Alan Sher, Michael Mage, and Laura Via.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 2004;306:636–40.
2. Dorman SE, Hatem CL, Tyagi S, et al. Susceptibility to tuberculosis: clues from studies with inbred and outbred New Zealand White rabbits. *Infect Immun*. 2004;72:1700–5.
3. Seah GT, Scott GM, Rook GAW. Type 2 cytokine gene activation and its relationship to extent of disease in patients with tuberculosis. *J Infect Dis*. 2000;18:385–9.
4. Dheda K, Chang J-S, Huggett JF, et al. The stability of mRNA encoding IL-4 is increased in pulmonary tuberculosis, while stability of mRNA encoding the antagonistic splice variant, IL-4δ2, is not. *Tuberculosis*. 2007;87:237–41.
5. Rook GA. Th2 cytokines in susceptibility to tuberculosis. *Current Molecular Medicine*. 2007;7:327–37.
6. Paul WE, Zhu J. How are T_H2-type immune responses initiated and amplified? *Nat Rev Immunol*. 2010;10:225–35.
7. Zhu J, Yamane H, Paul WE. Differentiation of effector CD4 T cell populations. *Annu Rev Immunol*. 2010;28:445–89.
8. Loots GG, Locksley RM, Blankespoor CM, et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*. 2000;288:136–40.
9. Strempel JM, Grenningloh R, Ho I-C, Vercelli D. Phylogenetic and functional analysis identifies Ets-1 as a novel regulator of the Th2 cytokine gene locus. *J Immunol*. 2010;184:1309–16.
10. Agarwal S, Avni O, Rao A. Cell-type-restricted binding of the transcription factor NFAT to a distal IL-4 enhancer in vivo. *Immunity*. 2000;12: 643–52.
11. Yamashita M, Ukai-Tadenuma M, Kimura M, et al. Identification of a conserved GATA3 response element upstream proximal from the interleukin-13 gene locus. *J Biol Chem*. 2002;277:42399–408.
12. Kim JI, Ho I-C, Grusby MJ, Glimcher LH. The transcription factor c-Maf controls the production of interleukin-4 but not other Th2 cytokines. *Immunity*. 1999;10:745–51.
13. Amsen D, Blander JM, Lee GR, Tanigaki K, Honjo T, Flavell RA. Instruction of distinct CD4 T helper cell fates by different Notch ligands on antigen-presenting cells. *Cell*. 2004;117:515–26.
14. Djuretic IM, Levanon D, Negreanu V, Groner Y, Rao A, Ansel KM. Transcription factors T-bet and Runx3 cooperate to activate Ifng and silence Il4 in T helper type 1 cells. *Nat Immunol*. 2007;8:145–53.
15. Hu C-M, Jang SY, Fanzo JC, Pernis AB. Modulation of T cell cytokine production by interferon regulatory factor-4. *J Biol Chem*. 2002;277: 49238–46.
16. Li B, Tournier C, Davis RJ, Flavell RA. Regulation of IL-4 expression by the transcription factor JunB during T helper cell differentiation. *EMBO J*. 1999;18:420–32.
17. Lee DU, Rao A. Molecular analysis of a locus control region in the T helper 2 cytokine gene cluster: a target for STAT6 but not GATA3. *Proc Natl Acad Sci U S A*. 2004;101:16010–5.
18. Zhu J, Cote-Sierra J, Guo L, Paul WE. Stat5 activation plays a critical role in Th2 differentiation. *Immunity*. 2003;19:739–48.
19. Avery S, Rothwell L, Degen WD, et al. Characterization of the first nonmammalian T2 cytokine gene cluster: the cluster contains functional single-copy genes for IL-3, IL-4, IL-13, and GM-CSF, a gene for IL-5 that appears to be a pseudogene, and a gene encoding another cytokinelike transcript, KK34. *J Interferon Cytokine Res*. 2004;24:600–10.
20. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–402.
21. Ovcharenko I, Loots GG, Giardine BM, et al. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res*. 2005;15:184–94.
22. Lee GR, Fields PE, Griffin TJ IV, Flavell RA. Regulation of the Th2 cytokine locus by a locus control region. *Immunity*. 2003;19:145–53.
23. Fields PE, Lee GR, Kim ST, Bartsevich VV, Flavell RA. Th2-specific chromatin remodeling and enhancer activity in the Th2 cytokine locus control region. *Immunity*. 2004;21:865–76.
24. Spilianakis CG, Flavell RA. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat Immunol*. 2004;5:1017–27.
25. Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA. Interchromosomal associations between alternatively expressed loci. *Nature*. 2005; 435:637–45.
26. Rowell E, Merkenschlager M, Wilson CB. Long-range regulation of cytokine gene expression. *Curr Opin Immunol*. 2008;20:272–80.
27. Wilson CB, Rowell E, Sekimata M. Epigenetic control of T-helper-cell differentiation. *Nat Rev Immunol*. 2009;9:91–105.
28. Göndör A, Ohlsson R. Transcription in the loop. *Nature Genetics* 2006;38: 1229–30.
29. Galande S, Purbey PK, Notani D, Kumar PP. The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1. *Curr Opin Genet Dev*. 2007;17:408–14.
30. Cai S, Lee CC, Kohwi-Shigematsu T. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet*. 2006;38:1278–88.
31. Ribeiro de Almeida C, Heath H, Krpic S, et al. Critical role for the transcription regulator CCCTC-binding factor in the control of Th2 cytokine expression. *J Immunol*. 2009;182:999–1010.
32. Ridruechai C, Mahasirimongkol S, Phromjai J, et al. Association analysis of susceptibility candidate region on chromosome 5q31 for tuberculosis. *Genes Immun*. 2010;11:416–22.
33. Sorg RV, Enczmann J, Sorg UR, Schneider EM, Wernet P. Identification of an alternatively spliced transcript of human interleukin-4 lacking the sequence encoded by exon 2. *Exp Hematol*. 1993;21:560–3.
34. Perkins HD, van Leeuwen BH, Hardy CM, Kerr PJ. The complete cDNA sequences of IL-2, IL-4, IL-6 and IL-10 from the European rabbit (*Oryctolagus cuniculus*). *Cytokine*. 2000;12:555–65.
35. Gautherot I, Burdin N, Seguin D, Aujame L, Sodoyer R. Cloning of interleukin-4 delta2 splice variant (IL-4δ2) in chimpanzee and cynomolgus macaque: phylogenetic analysis of δ2 splice variant appearance, and implications for the study of IL-4-driven immune processes. *Immunogenetics*. 2002;54:635–44.
36. Yatsenko OP, Filipenko ML, Voronina EN, Khrapov E, Sennikov SV, Kozlov VA. Alternative splicing of murine Interleukin-4 mRNA. *Bull Exp Biol Med*. 2004;137:179–81.
37. Mazzarella G, Bianco A, Perna F, et al. T lymphocyte phenotypic profile in lung segments affected by cavitary and non-cavitary tuberculosis. *Clin Exp Immunol*. 2003;132:283–8.
38. van Crevel R, Karyadi E, Preyers F, et al. Increased production of interleukin 4 by CD4+ and CD8+ T cells from patients with tuberculosis is related to the presence of pulmonary cavities. *J Infect Dis*. 2000;181:1194–7.

39. Luzina IG, Lockatell V, Todd NW, Keegan AD, Hasday JD, Atamas SP. Splice isoforms of human interleukin-4 are functionally active in mice in vivo. *Immunology*. 2011;132:385–93.
40. Luzina IG, Lockatell V, Todd NW, et al. Alternatively spliced variants of interleukin-4 promote inflammation differentially. *J Leukoc Biol*. 2011, in press, doi: 10.1189/jlb.0510271.
41. Via LE, Lin PL, Ray SM, et al. Tuberculous granulomas are hypoxic in guinea pigs, rabbits, and nonhuman primates. *Infect Immun*. 2008;76:2333–40.
42. Mendez S, Hatem CL, Kesavan AK, et al. Susceptibility to tuberculosis: composition of tuberculous granulomas in Thorbecke and outbred New Zealand White rabbits. *Vet Immunol Immunopathol*. 2008;122:167–74.

Supplemental Material

Alignment of reference sequences to the rabbit genomic regions

We used the Splign[1] program to align the reference mRNA sequences shown in Table S1 to the full length of the Broad and ENCODE genomic sequences summarized in Table 1. For both assemblies, the placement derived from this alignment is shown in Table S2. The alignment almost entirely confirmed the exon placement recorded in the Entrez Nucleotide record of the Broad sequence, except for exon 6 of *RAD50*, which is explained below. This confirmation is not surprising, as the placement recorded in the database was derived from a similar alignment. *IL5* could not be placed by Splign in the ENCODE sequence, but exons 1 and 2 could be placed by BLAST, and the locations of these exons are shown in the Table S2.

To provide cross-species support for the placement of coding exons, we used TBLASTN[2] with low-complexity filtering disabled (option –FF) to align the human protein sequences listed in Table S1 to both rabbit assemblies. The alignments mostly confirm the placement of the exons found by rabbit mRNA alignment, with occasional differences in length as expected for cross-species comparison (data not shown). The alignments did not confirm the placement of exon 6 of *RAD50* on the Broad assembly, and did confirm the absence of IL5 exons 3 and 4 in the ENCODE genomic sequences. Furthermore, there are no exons of human *KIF3A* that correspond to exons 10 and 11 of rabbit reference mRNA.

Exon 6 of *RAD50* may be missassembled in the broad assembly

Alignment of the rabbit *RAD50* transcript to the ENCODE sequence suggests exon 6 has a length of 129 bases, rather than the 12 bases assigned to it in the Broad assembly. Furthermore, alignment of the rabbit mRNA *RAD50* transcript to the Broad assembly does not assign a genomic location to the 117 bases putatively assigned a location in exon 6 by the ENCODE assembly.

Since the discrepancy in exon 6 could be explained by an error in the Broad assembly, or by a deletion in the animal, we performed some additional tests. First, we aligned the human RAD50 protein to the rabbit genomic sequences. Exon 6 of the human protein aligned to the same 129 bases putatively assigned to *RAD50* exon 6 by rabbit mRNA alignment. No placement for exon 6 of the human protein was found in the Broad assembly. We then searched the NCBI trace archive, and were able to find a complete sequence for *RAD50* exon 6 among the traces submitted as part of the sequencing efforts for Broad OryCun 2.0 (NCBI trace identifier 2052675903). These data taken together suggest *RAD50* exon 6 was not correctly assembled in the Broad sequence.

Table S1. Rabbit and human sequences used to confirm the location of exons on both assemblies.

Gene	Rabbit reference mRNA	Human reference protein
IL5	XM_002710201.1	NP_000870.1
RAD50	NM_001171348.1	NP_005723.2
IL13	XM_002710092.1	NP_002179.2
IL4	NM_001163177.1	NP_000580.1
KIF3A	XM_002710093.1	NP_008985.3

The existence of exons 10 and 11 in rabbit *KIF3A* is poorly supported

There are two plausible rabbit KIF3A proteins, the reference sequence (RefSeq identifier XP_002710139.1), and a sequence (GenBank identifier 217273045) submitted by the ENCODE sequencing project. The protein 217273045 was formerly the reference protein for rabbit, but the RefSeq record was suppressed with the stated reason “currently there is not sufficient data to support this transcript.” Ignoring a difference in the putative start of translation, the difference between XP_002710139.1 and 217273045 is exactly an alternate splice in the mRNA that skips exon 10 and 11. The alternate splice does not result in a frame shift. Though exon nine contains a base after the final full in-frame codon, a splice to either exon 10 or exon 12 results in a complete codon encoding Gly, and a continuation in the same reading frame.

It is difficult to resolve whether exons 10 and 11 are part of an in vivo transcript of rabbit KIF3A, or whether the prediction of exons 10 and 11 is an artifact of the gene prediction algorithm used. The RefSeq sequence XP_002710139.1 was predicted using the Broad assembly and the gene prediction algorithm GNOMON. The sequence 217273045 was predicted using the ENCODE assembly and the algorithm JIGSAW.[3]

Table S2. Location of the exons of genes *IL5*, *RAD50, IL13*, *IL4*, and *KIF3A* on the Broad and ENCODE genomic sequences. As explained in the text, *RAD50* exon 6 spans 129 bases in ENCODE and only 12 bases in Broad (the Entrez Nucleotide record for the Broad sequence makes it 13 bases rather than 12).

Gene	Exon	Strand	ENCODE postion		Broad position	
			Start	Stop	Start	Stop
IL5	Exon04	–1	–	–	15573697	15573795
IL5	Exon03	–1	–	–	15573897	15574025
IL5	Exon02	–1	703282	703316	15575464	15575496
IL5	Exon01	–1	703512	703655	15575694	15575837
RAD50	Exon01	+1	721523	721651	15593649	15593777
RAD50	Exon02	+1	723173	723256	15595298	15595381
RAD50	Exon03	+1	745150	745301	15617269	15617420
RAD50	Exon04	+1	747862	748047	15619981	15620166
RAD50	Exon05	+1	748742	748946	15620861	15621065
RAD50	Exon06	+1	757025	757153	15623038	15623049
RAD50	Exon07	+1	757405	757570	15630029	15630196
RAD50	Exon08	+1	759507	759700	15632134	15632327
RAD50	Exon09	+1	760091	760297	15632716	15632922
RAD50	Exon10	+1	762028	762210	15634650	15634832
RAD50	Exon11	+1	762563	762720	15635180	15635337
RAD50	Exon12	+1	764905	765080	15637522	15637697
RAD50	Exon13	+1	765958	766195	15638574	15638811
RAD50	Exon14	+1	773474	773663	15646088	15646277
RAD50	Exon15	+1	775011	775137	15647625	15647751
RAD50	Exon16	+1	775549	775742	15648163	15648356
RAD50	Exon17	+1	778818	778928	15651432	15651542
RAD50	Exon18	+1	779283	779375	15651897	15651989
RAD50	Exon19	+1	779453	779566	15652067	15652180
RAD50	Exon20	+1	782061	782188	15654674	15654801
RAD50	Exon21	+1	785914	786138	15658542	15658766
RAD50	Exon22	+1	796574	796659	15669200	15669285
RAD50	Exon23	+1	797124	797266	15669750	15669892
RAD50	Exon24	+1	798970	799103	15671596	15671729
RAD50	Exon25	+1	800157	800343	15672783	15672969
IL13	Exon01	+1	817825	817956	15690430	15690561
IL13	Exon02	+1	818784	818837	15691387	15691440
IL13	Exon03	+1	819039	819143	15691642	15691746
IL13	Exon04	+1	819408	819515	15692011	15692118
IL4	Exon01	+1	830654	830788	15702794	15702928
IL4	Exon02	+1	830964	831011	15703106	15703154
IL4	Exon03	+1	836049	836207	15708470	15708628
IL4	Exon04	+1	838902	839003	15711332	15711433
KIF3A	Exon19	–1	850323	850371	15722771	15722819
KIF3A	Exon18	–1	852498	852622	15726037	15726161
KIF3A	Exon17	–1	853800	853868	15727338	15727406
KIF3A	Exon16	–1	854177	854230	15727717	15727770
KIF3A	Exon15	–1	855715	855840	15729255	15729380
KIF3A	Exon14	–1	857855	857965	15731387	15731497
KIF3A	Exon13	–1	858176	858356	15731707	15731887
KIF3A	Exon12	–1	859084	859240	15732615	15732771
KIF3A	Exon11	–1	862232	862240	15735757	15735765
KIF3A	Exon10	–1	864320	864391	15737834	15737905
KIF3A	Exon09	–1	866247	866345	15739761	15739859
KIF3A	Exon08	–1	875538	875712	15749407	15749581

(*Continued*)

Table S2. (*Continued*)

Gene	Exon	Strand	ENCODE postion		Broad position	
			Start	Stop	Start	Stop
KIF3A	Exon07	–1	875992	876189	15749861	15750058
KIF3A	Exon06	–1	876558	876697	15750427	15750566
KIF3A	Exon05	–1	880392	880497	15753922	15754027
KIF3A	Exon04	–1	886291	886375	15759815	15759899
KIF3A	Exon03	–1	886503	886647	15760027	15760171
KIF3A	Exon02	–1	905793	906066	15779859	15780132
KIF3A	Exon01	–1	908713	908763	15782993	15783043

The experimentally validated transcript AJ920325 that Evidence Viewer supplies to support the *KIF3A* RefSeq mRNA is not long enough to include or exclude exons 10 and 11. Though exons 10 and 11 were predicted using the Broad assembly, identical sequence for these exons is present in the ENCODE assembly. The Broad and ENCODE sequences are 99% identical from 633 bases upstream of exon nine to at least 1000 bases downstream of exon 12, and are 100% identical within these four exons and for 21 bases upstream or downstream of these four exons. Thus, there are no obvious gross changes in sequence that explain the different prediction of exons 10 and 11.

The evidence against a transcript including exons 10 and 11 is that none of the species in the NCBI HomoloGene database record (identifier 38266) for *KIF3A* have a transcript listed that includes exons 10 and 11. The HomoloGene proteins include human and mouse KIF3A, which are presumably the best studied. There are, however, *KIF3A* sequences in NCBI's non-redundant protein collection, often sequences produced by gene prediction algorithms, that do include exons 10 and 11. The species with a *KIF3A* entry in the HomoloGene database were *Homo sapiens*, *Pan troglodytes*, *Canis familiaris*, *Bos taurus*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Drosophila melanogaster*, *Anopheles gambiae*, and *Caenorhabditis elegans*.

Comparision of coding regions

Table S3 shows the difference between the coding regions of five genes of interest in the Broad assembly and the corresponding coding regions in the ENCODE assembly. A 21 nucleotide margin was added to both the start and end of each putative coding exon before the comparison was done.

Table S3. Sequence differences between the ENCODE and Broad assembly of the coding exons of *RAD50*, *IL13*, *IL4*, and *KIF3A*, for those exons that could be located in both assemblies. Each of the two coding exons of *IL5* that was present in the ENCODE assembly was identical to its counterpart in the Broad assembly. A boundary of 21 bases was added to both ends of each exon. Changes in these boundary bases are labeled "Intronic" in column 8. The column labeled Traces shows the number of reads that could be found in the Broad WGS traces to support Broad's version of the sequence.

Gene	Exon	ENCODE position	Codon/ triplet	Broad position	Codon/ priplet	Strand	Traces	Subst.
RAD50	11	762554	TAC	15635171	AAC	+	1[a]	Intronic
RAD50	25	800307	GTG	15672933	GTA	+	6	V->V
IL13	1	817903	CCA	15690508	ACA	+	7	P->T
IL13	4	819534	GGT	15692137	CAG	+	5	Intronic
IL4	2	831003	GTC	15703145	GTCC	+	1	Frame shift
KIF3A	2	905778	TTC	15779844	TTG	–	3	Intronic
KIF3A	2	906051	AAA	15780117	AAG	–	4	K->K
KIF3A	3	886647	GAA	15760171	GGA	–	4	E->G
KIF3A	4	886393	TAA	15759917	AAA	–	6	Intronic
KIF3A	13	858356	GCA	15731887	ACA	–	9	A->T

Note: [a]For *RAD50* exon 11, the six highest scoring matching traces contradicted the change to AAC, while the seventh highest scoring trace has AAC.

Amino acid substitution in IL13

There is a substitution of a Threonine (Thr) in Broad for a Proline (Pro) in ENCODE at amino acid 27 of IL13. The codon for Thr27 is supported by seven traces from Broad with no support for Pro27 in the sequences from Broad. The Pro27 allele in the NZW rabbit, however, is supported by multiple traces from two different BACs. Multiple alignment of IL13 homologs places position 27 in rabbit in a column predominantly filled with Thr, including a Thr in the corresponding position in human (data not shown). Thus sequence conservation supports that a Thr in that position will be tolerated. Moreover, the only available structure for IL13 (Protein Data Bank identifier 3BPO[4]) is from human, and this structure has a Thr in the corresponding position.

On the other hand, the NZW rabbit appears healthy so the Pro at position 27 does not have any known deleterious phenotype. Pairwise alignments of the IL13 protein from dog and camel place a Pro in the place corresponding to position 27 in rabbit, but the multiple alignment algorithm COBALT[5] places mismatches and gaps rather than a Pro in the column corresponding to position 27 in rabbit. Analysis of the human structure 3PBO using PyMol[6] suggests that the position corresponding to 27 in rabbit is not part of an alpha helix that would be disturbed by a Pro in that location. Thus, we have no reason to believe the Pro at position 27 is not tolerated.

Frameshift in *IL4* in the Broad assembly

Of possible immunological interest, there is a frameshift mutation in exon 2 of *IL4* in the Broad assembly. See the section Comparison of the Broad and ENCODE within Predicted Genes in the main manuscript.

Alignment of promoter regions

We extracted putative promoter regions for the *IL5, RAD50, IL13*, and *IL4* genes from the ENCODE genomic sequence. Putative promoter regions were defined to start 1500 bases upstream of the putative start of transcription and extend to include the transcription start site and 50 additional bases, a total length of 1550 bases. We then aligned these promoter sequences to the Broad rabbit genomic sequences, the human syntenic region, and the mouse syntenic region. The alignments of the Broad and ENCODE sequences are shown in Table S4.

The ENCODE *IL13* promoter aligned to the Broad sequence with full coverage and 98% identity, but with mismatches and a short gap. The Broad *RAD50* promoter aligns with the ENCODE *RAD50* promoter with no mismatches and a single gap of length one. The Broad *IL5* and *IL4* promoters matched the ENCODE sequences well, but the Broad sequences had runs of the ambiguity character N that split the alignment into partial hits. The full range of the 1550 base *IL5* promoter in the ENCODE sequence is from 703605 to 705154; the full range of the *IL4* promoter is from 829140 to 830689. For both these promoters, a long region near each end aligns to the Broad sequence with high percent identity (see Table S4).

Comparison between Broad and ENCODE binding sites

The Broad and ENCODE genetic sequences were identical at the positions listed in Tables 3 or 6; when a site was listed in both tables, the coordinates of Table 3 were used. We compared the genetic sequences for 100 bases above (in genomic coordinates) and 100 bases

Table S4. Comparison of promoter sequences for *IL5, RAD50, IL13*, and *IL4* between the two assemblies.

Promoter	% identity	Length	Gaps	ENCODE		Broad		Strand[a]
				Start	Stop	Start	Stop	
IL5 Promoter	99.89	952	0	703605	704556	15575787	15576738	–1
	100.00	157	0	704998	705154	15577124	15577280	–1
RAD50 Promoter	99.94	1550	1	720023	721572	15592150	15593698	+1
IL13 Promoter	98.32	1550	1	816325	817874	15688933	15690479	+1
IL4 Promoter	98.35	424	2	829140	829562	15701332	15701753	+1
	99.81	535	0	830155	830689	15702295	15702829	+1

Note: [a]The strand relative to start of transcription is indicated (it is the same for both Broad and ENCODE), but coordinates are shown with respect to the forward strand.

Table S5. Sequence differences between ENCODE and Broad near binding sites.[a]

	Binding site	ENCODE		Broad	
		Position	Sequence	Position	Sequence
Ets-1 sites	RHS5	787823	C	15660448	T
	IL4 Promoter.1	830485	G	15702625	T
	IL4 Promoter.2	830484	G	15702624	T
	Ets-1 IL4IE	831843	AAAA	15703986	GGCT
	HSIV	841287	AA	15713725	TT
GATA sites[b]	IL13P(1)	815879	C	15688484	–
	IL13P(2)	815879	C	15688484	–
	IL13P(2)	816056	A	15688661	G
	CNS-1	822964	T	15695466	A

Notes: [a]None of these changes are within the sites themselves; [b]IL13P(3) omitted because it may not be a GATA binding site in Rabbit.

below the binding site. Table S5 shows the differences that were found. For gaps, the position before the gap is given. It is correct that the same gap is shown for IL13P(1) and IL13P(2); the sites are close together.

Multiple alignments of binding sites

We took the wider of the ranges of the predicted binding sites in Tables 3 and 6. We then used the ENCODE coordinates to locate the binding site in the alignment generated by Mulan. Furthermore, because we consistently found that the aligned region in mouse intersected the region predicted by Strempel et al,[7] we extended the alignment to include the full extent of mouse. In the case of IL13P(1), this also involved deleting some positions from the end; (compare Table 3 and Table S6). The alignments shown below Table S4 are oriented to match the forward strand in rabbit, regardless of the orientation shown in Strempel et al.[7]

Placement of trancription factor binding sites in conserved noncoding regions

Trancription factor binding sites were assigned a putative conserved noncoding region in the column labeled "Location" of Tables 5 and 6. Those sites within *RAD50* between the Ets-1 binding site known to be in the locus control region (LCR) and the 3′

Table S6. The placement of the rabbit transcription factor binding sites within the 10-species Mulan alignment. The third and fourth columns show the start and end of the site itself, the fifth and sixth column show the start and end of the block in which it is aligned. Coordinates are with respect to the ENCODE genetic sequence. The block identifier is the identify number of the block within the Mulan alignment; the full Mulan alignment is too large to be shown.

	Promoter	Site start	Site stop	Block start	Block stop
Ets-1 sites	Ets-1 *IL5* promoter	703734	703754	703484	703915
	RHS5	787911	787931	787608	788040
	IL13 promoter	816434	816453	816191	816838
	IL4 promoter.1	830425	830445	830030	830919
	IL4 promoter.2	830464	830484	830030	830919
	Ets-1 IL4IE	831755	831775	830931	831858
	HSIV	841204	841224	840789	841259
	CNS2	844617	844637	844506	844845
GATA sites	GATA *IL5* promoter	703763	703776	703484	703915
	RHS6.1	795825	795838	795478	796187
	RHS6.2	796911	796924	796637	796967
	IL13P(1)	815913	815925	815913	815953
	CNS-1	822925	822938	822890	822977
	IL4P	830326	830339	830030	830919
	GATA IL4IE	831382	831395	830931	831858
	CNS-2(1)	844525	844538	844506	844845
	CNS-2(2,3)	844583	844607	844506	844845

Table S7. Alignments.

Ets-1 *IL5* promoter (reverse complement of that in Strempel et al[7]

Hum	TGTCTTTGAGGAAATGAATAA
Pan	....
Papio	....
Calli	....
Oto	.C...................
Bos	.CC..................
Canis	.A...................
Rat	.C...................
Mus	.C..--...............
Rabbit	.C...................

RHS5

Hum	GGTAACACAGGAAGTCAGCAG
Pan	....
Papio	.............A.......
Calli	.............A.T.....
Oto	.............A.T.....
Bos	.A...........A.T.A...
Canis	.............A.T.....
Rat	.............A...A...
Mus	...............T.A...
Rabbit	....G........A.T.....

IL13 promoter (alignment differs from Strempel et al[7] in that multiTF prefers gaps to mismatches)

Hum	GTTC-GGGGAGGAAGTGGGTA
Pan	....-................
Papio	.C..-................
Calli	.C.G-...A............
Oto	.C.TA................
Bos	.C.TA..........A....G
Canis	...TA.............C.G
Rat	.CCTAA..............G
Mus	.CCTGA..............G
Rabbit	.GCT-................

***IL4* promoter.1**

Hum	GATTTCACAGGAACATTTTAC
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	.............A....-..
Mus	.............A....-..
Rabbit	....

***IL4* promoter.2**

Hum	TTTTCTCCTGGAAGAGAGGTG
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	..................A..
Mus	....
Rabbit	...................C.

(*Continued*)

Table S7. (*Continued*)

Ets-1 IL4IE

Hum	CATTTCAGTTCCTGTTTTCAT
Pan	....
Papio	....
Calli	..C..................
Oto	..C..................
Bos	..CA.................
Canis	..CA.................
Rat	..C..................
Mus	....
Rabbit	T.C..................

HSIV

Hum	TCTGCCACAGGATATGGGTAG
Pan	....
Papio	....
Calli	................A..G.
Oto	................A..T.
Bos	................AC.T.
Canis	................A..T.
Rat	................A..T.
Mus	................A..T.
Rabbit	................AC.T.

CNS2 (alignment does not include extra bases in Calli)

Hum	TGGGTCACAGGAAGCCCAAGA
Pan	....
Papio	....
Calli	.............-------.
Oto	....
Bos	....
Canis	.................-...
Rat	......G..............
Mus	......G..............
Rabbit	....

GATA *IL5* promoter (reverse complement of the one in Strempel et al[1]

Hum	AATCAGATAGAGAA
Pan	....
Papio	....
Calli	.............G
Oto	....
Bos	....
Canis	....
Rat	............G.
Mus	............G.
Rabbit	....

RHS6.1

Hum	ATCAGATAAGAGGC
Pan	....
Papio	....
Calli	....
Oto	....
Bos	............A.
Canis	............A.
Rat	..........GA..
Mus	....
Rabbit	....

(*Continued*)

Table S7. (*Continued*)

RHS6.2	
Hum	TGTAGATAGGGATA
Pan	....
Papio	.A............
Calli	--------------
Oto	..-...........
Bos	..C...........
Canis	CA......T.....
Rat	C.C...........
Mus	C.C...........
Rabbit	CAG....G...TA.
IL13P(1)	
Hum	CGCTTATCGGGCCC
Pan	....
Papio	.............T
Calli	....
Oto	........CA...-
Bos	..T.....A.C...
Canis	........A.C...
Rat	.T.......AC...
Mus	.T.......AC...
Rabbit	........T.-...
CNS-1	
Hum	CCCATTATCTTCAT
Pan	....
Papio	....
Calli	....
Oto	....
Bos	.T...........C
Canis	.T............
Rat	.T............
Mus	.T............
Rabbit	.T.CC.........
IL4P	
Hum	AGCTGATAAGATTA
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	C.............
Mus	C.............
Rabbit	....
GATA IL4IE	
Hum	AAACAGATATTGAG
Pan	....
Papio	....
Calli	....
Oto	G.............
Bos	....
Canis	...T......G...
Rat	..........GA..
Mus	.T........GA..
Rabbit	..GT..........
CNS-2(1)	
Hum	TATCTGATCTGTCA
Pan	....
Papio	.G............
Calli	.G............
Oto	.G............
Bos	.G............
Canis	.G.G..........
Rat	CT.........C.C
Mus	CG...........C
Rabbit	.G............
CNS-2(2,3)	
Hum	CTTCTGATAACGTTGATAAAAGTCA
Pan	G........................
Papio	G........................
Calli	G........................
Oto	G.........T............A.
Bos	G..........A...........A.
Canis	G......................A.
Rat	G..........AC..........A.
Mus	G..........AC..........G.
Rabbit	G...........C..........A.

(*Continued*)

Table S8. Alignments of additional the sites shown in Table 5.

07_MAFG	
Hum	TATTTATGTTGAGTCATTTCTTTTCTC
Pan	....
Papio	....
Calli	.......................T..
Oto	.........................A
Bos	....
Canis	....
Rat	....
Mus	....
Rabbit	.....................----.
08_JunB	
Hum	TTGAGTCAT
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	....
Mus	....
Rabbit	....
9_IRF4	
Hum	TTCAGTTTCTTTTTT
Pan	....
Papio	....
Calli	....
Oto	..-............
Bos	....
Canis	....

(*Continued*)

Table S8. (*Continued*)

Rat	..T...........C
Mus	..T............
Rabbit	..TG...........
15_NFκB	
Hum	CTGGATTTTCCACAAA
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	............A...
Mus	............A...
Rabbit	....
16_NFAT	
Hum	GATTTTCCAC
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	.........A
Mus	.........A
Rabbit	....
17_Runx3	
Hum	AAAGATGTGGTTTCT
Pan	....
Papio	....
Calli	....
Oto	....
Bos	.............G.
Canis	....
Rat	.............TC
Mus	.............TC
Rabbit	.............A.
18_STAT5	
Hum	TCCCAGAAGCAAT
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	..........G..
Rat	....
Mus	....
Rabbit	.........TG..
20_Runx3	
Hum	TGCTGTGTGGTCAGA
Pan	....
Papio	....
Calli	..T............
Oto	....A..........
Bos	CC
Canis	.A...........CT
Rat	....
Mus	....
Rabbit	.A.............
21_NFκB	
Hum	GGTGTAATTTCCTA
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	....
Mus	....
Rabbit	....
22_IRF4	
Hum	GTTTCATTTTC
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	....
Mus	....
Rabbit	....
24_NFAT	
Hum	AAATTTCCAA
Pan	....
Papio	....
Calli	....
Oto	....
Bos	....
Canis	....
Rat	....
Mus	..T.......
Rabbit	....
25_STAT5	
Hum	GTTTTCATGGAAACACACGGCTGAGAA
Pan	....
Papio	....
Calli	....
Oto	...................A.......
Bos	..................AA.......
Canis	.................A.A.......
Rat	..................AA.......
Mus	..................CA.......
Rabbit	...................A.......
26_Runx3	
Hum	CCTGACCACAGCCAG
Pan	....

(*Continued*)

Table S8. (*Continued*)

Papio	............T..
Calli	....
Oto	....
Bos	....
Canis	....
Rat	...........T...
Mus	...........G...
Rabbit	....
28_RBPJK	
Hum	TTTCCCACAC------------A
Pan	..........------------.
Papio	..........------------.
Calli	..........------------.
Oto	..........------------.
Bos	..........------------.
Canis	..........------------.
Rat	..........------------G
Mus	..........AGGGGAGGGAGGG
Rabbit	..........------------.

Table S9. Alignments for sites shown in Table 6.

01_JunB		
Hum	yes	AGGAGTCAT
Pan	yes	....
Papio	yes	....
Calli	yes	....
Oto	yes	....
Canis	yes	....
Rat	yes	....
Mus	yes	....
Rabbit	yes	....
02_NFκB		
Hum	yes	TTGGGGTTTCCAAGGC
Pan	yes	....
Papio	yes	....
Calli	yes	....
Oto	yes	....A........AT.
Canis	no	....T.........T.
Rat	yes	..............T.
Mus	yes	..............T.
Rabbit	yes	..............T.
03_MAFG		
Hum	yes	AACTCAAGTCAACAGAATC
Pan	yes	....
Papio	yes	....
Oto	yes	....
Canis	yes	....
Rat	yes	....
Mus	yes	....
Rabbit	yes	....

(*Continued*)

Table S9. (*Continued*)

04_NFAT		
Hum	yes	CATTGGAAAAGT
Pan	yes	....
Papio	no	.G..........
Oto	yes	....
Canis	no	T......G....
Rat	no	..C.........
Mus	yes	....
Rabbit	yes	....
05_IRF4		
Hum	yes	CAAAAAGAAACTGAA
Pan	yes	....
Papio	yes	....
Oto	yes	....
Canis	yes	..........A....
Rat	yes	....
Mus	yes	....
Rabbit	yes	....
06_MAFG		
Hum	yes	CTGTTATTAGTAATCATCT
Pan	yes	....
Papio	yes	....
Calli	yes	....
Oto	no	..A.A..............
Bos	yes	....
Canis	yes	....
Rat	no	..............G.C..
Mus	yes	....C..............
Rabbit	yes	....
11_STAT5		
Hum	yes	ATTTTCTAAGAATTC
Pan	yes	....
Papio	no	.........A.....
Oto	yes	..C............
Bos	yes	....
Canis	yes	....
Rat	yes	G..............
Mus	yes	G..............
Rabbit	yes	....
13_RBPJK		
Hum	yes	TCTCCCACGCG
Pan	yes	....
Papio	yes	....
Calli	no	.......T...
Oto	yes	G..........
Bos	yes	....
Canis	no	G.CA..G.C..
Rat	yes	....
Mus	yes	....
Rabbit	yes	G..........
14_JunB		
Hum	yes	TTGACTCACCCGG
Pan	yes	....

(*Continued*)

Table S9. (*Continued*)

Calli	yes	....
Oto	yes	........TA...
Bos	no	.CC.....T..T.
Rat	yes	...........AA
Mus	yes	...........AT
Rabbit	yes	C.......T.GCC
19_IRF4		
Hum	yes	CTCACTTTCTGTTGC
Pan	yes	....
Papio	yes	....
Calli	no	....T..........
Oto	yes	....
Bos	yes	....
Canis	yes	....
Rat	yes	..............T
Mus	yes	....
Rabbit	yes	....
23_NFAT		
Hum	yes	CATTTTCCTATT
Pan	yes	....
Papio	yes	....
Calli	yes	....
Oto	yes	....
Bos	no	..........C.
Canis	yes	....
Rat	yes	........A...
Mus	yes	........A...
Rabbit	yes	....
29_JunB		
Hum	yes	TTAAATTAGTCAG
Pan	yes	....
Papio	yes	....
Calli	yes	-----........
Oto	no	........A...-
Bos	no	..C....----..
Canis	no	........A...A
Rat	yes	.CT..........
Mus	yes	.CT..........
Rabbit	yes	...----......
30_STAT5		
Hum	yes	TATTTCCA
Pan	yes	....
Papio	yes	....
Oto	no	C..A....
Bos	no	..G.....
Canis	no	..G.....
Rat	no	..-.....
Mus	yes	....
Rabbit	yes	....

Notes: The word "yes" or the word "no" follows the name of each species and indicates whether the site was predicted by TRANSFAC to exist in that species at the given location. Sites listed in Table 9. at more than one location are not shown; such sites are split across multiple blocks of the Mulan alignment. Not all species are represented in each alignment.

end of the gene were assigned to LCR. The LCR contains several DNAse 1 hypersensitive sites clustered within the 3′ end of the DNA repair gene *RAD50* and was shown to be important for the regulation of the cytokine genes.[8,9] Binding sites between the Ets-1 promoter binding site and the start of transcription of either *IL4* or *IL13* were assigned to the promoter region of the respective gene. The sites 18_STAT5 and 20_Runx3 are less than 300 bases from the GATA CNS-1 site, but are in a different block of the Mulan alignment. The sites 25_STAT5 and 28_RBPJK were putatively assigned to HSII and HSV/VA respectively based on published studies.[10,11] The 26_Runx3 site is in the same block of the Mulan alignment as Ets-1 HSIV. In Table 6, the 27_NFκB site was placed by proximity to 28_RBPJK. Other sites in Table 6 were assigned a putative location by reasoning analogous to that used for the rightmost column of Table 5.

References

1. Kapustin Y, Souvourov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct*. 2008;3:20.
2. Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology*. 2006;4:41.
3. Allen JE, Salzberg SL. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*. 2005;21:3596–603.
4. LaPorte SL, Juo ZS, Vaclavikova J, et al. Molecular and structural basis of cytokine receptor pleiotropy in the interleukin-4/13 system. *Cell*. 2008; 132:259–72.
5. Papadopoulos JS, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*. 2007;23:1073–9.
6. The PyMol Molecular Graphics System, Schrödinger, LLC.
7. Strempel JM, Grenningloh R, Ho I-C, Vercelli D. Phylogenetic and functional analysis identifies Ets-1 as a novel regulator of the Th2 cytokine gene locus. *J Immunol*. 2010;184:1309–16.
8. Li B, Tournier C, Davis RJ, Flavell RA. Regulation of IL-4 expression by the transcription factor JunB during T helper cell differentiation. *EMBO J*. 1999;18:420–32.
9. Lee GR, Fields PE, Griffin TJ IV, Flavell RA. Regulation of the Th2 cytokine locus by a locus control region. *Immunity*. 2003;19:145–53.
10. Amsen D, Blander JM, Lee GR, Tanigaki K, Honjo T, Flavell RA. Instruction of distinct CD4 T helper cell fates by different Notch ligands on antigen-presenting cells. *Cell*. 2004;117:515–26.
11. Lee DU, Rao A. Molecular analysis of a locus control region in the T helper 2 cytokine gene cluster: a target for STAT6 but not GATA3. *Proc Natl Acad Sci U S A*. 2004;101:16010–15.