genes

MDPI

Article

Variation of the 3′RR1 HS1.2 Enhancer and Its Genomic Context

Carla Jodice [1,*], Patrizia Malaspina [1], Bianca Maria Ciminelli [1], Cristina Martinez-Labarga [1], Michela Biancolella [1], Giuseppe Novelli [2] and Andrea Novelletto [1]

[1] Department of Biology, University of Rome Tor Vergata, 00133 Rome, Italy; patrizia.malaspina@uniroma2.it (P.M.); bianca.ciminelli@uniroma2.it (B.M.C.); martine@uniroma2.it (C.M.-L.); michela.biancolella@uniroma2.it (M.B.)

[2] Department of Biomedicine and Prevention, University of Rome Tor Vergata, Tor Vergata University Hospital, 00133 Rome, Italy; novelli@med.uniroma2.it

[*] Correspondence: carla.jodice@uniroma2.it; Tel.: +39-06-7259-4321

Abstract: In humans, the HS1.2 enhancer in the Ig heavy-chain locus is modular, with length polymorphism. Previous studies have shown the following features for this variation: (i) strong population structuring; (ii) association with autoimmune diseases; and (iii) association with developmental changes in Ig expression. The HS1.2 region could then be considered as a contributor to inter-individual diversity in humoral response in adaptive immunity. We experimentally determined the HS1.2-length class genotype in 72 of the 1000 Genomes CEU cell lines and assigned the HS1.2 alleles to haplotypes defined by 18 landmark SNPs. We also sequenced the variable portion and ~200 bp of the flanking DNA of 34 HS1.2 alleles. Furthermore, we computationally explored the ability of different allelic arrangements to bind transcription factors. Non-random association between HS1.2 and Gm allotypes in the European population clearly emerged. We show a wealth of variation in the modular composition of HS1.2, with five SNPs further contributing to diversity. Longer alleles offer more potential sites for binding but, for same-length alleles, SNP variation creates/destroys potential binding sites. Altogether, the arrangements of modules and SNP alleles both inside and outside HS1.2 denote an organization of diversity far from randomness. In the context of the strong divergence of human populations for this genomic region and the reported disease associations, our results suggest that selective forces shaped the pattern of its diversity.

Keywords: HS1.2 Ig enhancer; Ig allotypes; immunoglobulins; transcription factor binding

check for updates

Citation: Jodice, C.; Malaspina, P.; Ciminelli, B.M.; Martinez-Labarga, C.; Biancolella, M.; Novelli, G.; Novelletto, A. Variation of the 3′RR1 HS1.2 Enhancer and Its Genomic Context. *Genes* **2024**, *15*, 856. https://doi.org/10.3390/genes15070856

Academic Editor: David Lando

Received: 28 May 2024
Revised: 24 June 2024
Accepted: 27 June 2024
Published: 29 June 2024

1. Introduction

The human locus containing the coding segments for the immunoglobulin heavy-chain constant domains is a result of tandem duplication, in which the telomeric member harbors, among others, the active gene segments IGHG3, IGHG1, and IGHA1, whereas the other member contains IGHG2, IGHG4, IGHE, and IGHA2. All these gene segments encode for the heavy-chain constant domains of different immunoglobulin classes. They are transcribed from telomere to centromere on the minus (−) DNA strand. A regulatory region (3′RR) is present in each of the duplicated blocks, with the telomeric and centromeric paralogs named 3′RR1 and 3′RR2, respectively. In humans, each of these latter regions contains three enhancer elements, indicated as HS3, HS1.2, and HS4 (Figure 1 in refs. [1,2]). Both regions have been categorized as super-enhancers, i.e., elements that drive expression of genes that control and define cell identity. Typically, super-enhancers are formed by multiple elements, each capable of binding transcription factors synergistically [3].

Several aspects of the functioning and role of the 3′RR have been investigated in mice (reviewed in ref. [4]), in which a single paralog is present. Moreover, in mice, the entire region includes inverted repeat sequences flanking HS1.2, which diverge from humans in their internal sequence [5]. In a multi-species comparison, it seems that evolution maintained a ‘quasi-palindromic’ organization, at least in mammals, making it tempting to speculate that this unique arrangement is a structural feature of this regulatory region [6]. Evidence accumulated by studying several knock-out mouse models shows that the 3′RR portion including the quasi-palindrome dictates antigen-dependent locus remodeling (global somatic hypermutation and class switch recombination to major isotypes) in activated B cells, and antibody production in plasma cells [7].

In humans, the 3′RR1 HS1.2 is itself modular, with length polymorphism resulting from variation in the number of internal modules across gene copies. Traditionally, four size classes (named 1 to 4 from smallest to largest) can be identified through PCR and gel electrophoresis. Each class differs by approximately 60 bp, on a total length determined by the particular amplification protocol used [8,9]. Previous studies have shown the following features for this variation: (i) strong population structuring; (ii) association with autoimmune diseases; and (iii) association with developmental changes in Ig expression [10–17]. Conversely, the 3′RR2 HS1.2 displays low polymorphism, with only two alleles in the European population, the major of which accounts for 93% [9].

Due to these features, the HS1.2 region could be considered as a contributor to inter-individual diversity in humoral response in adaptive immunity. By resequencing the region immediately telomeric to HS1.2 from allele 1 and 2 in homozygous individuals, a strong non-random association (linkage disequilibrium, LD) between HS1.2 and four neighboring SNPs (rs12896746-rs12896897-rs7144089-rs7143677) was detected [18]. In these authors' data, these SNPs are arranged in virtually two haplotypes only (A-C-G-A and G-T-C-G), that become predictors of the HS1.2 allele in cis. By leveraging this information, we inferentially suggested that in diverse human populations, some degree of disequilibrium extends all the way from HS1.2 to IGHA1, IGHG1, and IGHG3, the latter lying 60 kb apart [19]. In this context, it is worth noting that variation in IGHG1 and IGHG3 includes missense variants long shown to be responsible for the immunogenic determinants of Gm allotypes [20].

In this study, we used a more direct experimental approach to investigate linkage disequilibrium in the entire genomic region surrounding HS1.2, spanning from 100 kb centromerically to 60 kb telomerically of it. We experimentally determined the HS1.2-length class genotype in 72 of the cell lines (Table S1) that were fully sequenced for CEPH European Utah residents with Northern and Western European ancestry (CEU) subset of the 1000 Genomes project [21] and assigned the HS1.2-length alleles to haplotypes defined by 18 landmark SNPs. Among these, 11 correspond to missense variants, i.e., 2, 2, 4, and 3 in IGHA2, IGHA1, IGHG1, and IGHG3, respectively, 7 of which determine allotypic diversity in IGHG1 and IGHG3 and the ensuing changes in immunogenicity [22]. Furthermore, a relevant role of this set of variants on the biology of immunoglobulin genes derives from the differential association of their genotypes with expression level or splicing of IGHC gene segments, thus qualifying them as eQTL and/or sQTL [23]. These effects are often tissue-specific and disease-related. For example, a signal of association with the intrathecal synthesis of IgG was found for SNPs around IGHG3 in multiple sclerosis [24].

In this study, we aimed at providing a clearer picture of the non-random association between HS1.2, the adjacent non-coding SNPs, and the more distant Gm allotypes in a cohort of subjects of European ancestry.

During the development of this work, we also noticed subtle variation within HS1.2 length classes, that was detectable upon agarose gel electrophoresis but not clearly sizeable. As the diversity of HS1.2 cannot be safely captured by high throughput methodologies with short reads, due to its modular nature, we Sanger sequenced amplicons including the variable portion of 34 HS1.2 alleles and ~200 bp of flanking DNA from homozygous individuals or after cloning and recombinant selection.

We show a wealth of variation that could not be detected with the separation methods used so far and predicts a much larger repertoire with larger screenings.

Multiple lines of evidence denote that the 3′RR region harbors variation contributing to the qualitative and quantitative transcriptional output of IGH gene segments. First, flag SNPs showing association with serum Ig class ratios have been identified in the region [25]. Second, variable mRNA expression in lymphoblastoid cell lines, depending on the HS1.2-inferred genotype, was reported [19]. Third, an estrogen-dependent sex-biased Chip-seq enrichment was observed in the mouse [26,27] and was also paralleled by an excess of estrogen-binding sites in humans [28]. Interestingly, a sex-biased response was also reported for clinical implications in COVID-19 [16].

Experimental data support the ability of super-enhancers to bind an array of transcription factors more than the sum of their internal elements [3]. Binding sites for transcription factors have repeatedly been reported for HS1.2 [6,8,9,29]. The new sequence diversity here reported prompted us to explore computationally the allele-specific content in putative transcription factor binding sites. In doing this, we wanted to explore the role of changes in the modular composition of HS1.2-length alleles and internal SNPs by analyzing the length-invariant CORE, END, and length-variable (VAR) sub-portions separately.

2. Materials and Methods

2.1. HS1.2-Length Typing

The list of the 72 cell line DNAs used is reported in Table S1. For both lymphoblastoid cell and donor blood DNA, fragments containing the VNTR region of HS1.2 were PCR-amplified according to [8], with minor modifications. In particular, the reactions were carried out in 30 µL of a mix made up of 200 ng genomic DNA, 0.2 mM of each dNTP, 0.1 mM of each primer, 1× Taq Platinum plus buffer and 0.5 U Taq Polymerase (Platinum plus, Invitrogen^{TM}, Thermo Fisher Scientific Inc., Waltham, MA, USA). PCR was performed on a Step One (Applied Biosystem$^{®}$, Thermo Fisher Scientific Inc., Waltham, MA, USA) thermal cycler. Denaturation was at 94 °C/3 min, followed by 10 cycles consisting of 94 °C/45 s, 65 °C/55 s, and 72 °C/45 s, and 20 additional cycles with annealing at 63 °C/105 s. A final elongation step was carried out at 72 °C for 10 min. The PCR products were then analyzed on 1% agarose gel stained with ethidium bromide, and the alleles were classified according to the length of the PCR fragments (allele 1, ca. 300 bp; allele 2, ca. 350 bp; allele 3, ca. 400 bp; allele 4, ca. 450 bp).

2.2. Handling of the 1000 Genomes Data

Individual genotypes at each of the 18 SNPs listed in Table S1 were obtained with the data slicer available at http://www.ensembl.org/Homo_sapiens/Tools/DataSlicer (accessed on 2 February 2024).

Genotypes at rs61984162 and rs12433324 in cell lines NA10839, NA12752, NA12003, and NA12145 were experimentally validated with TaqMan assays (Thermo Fisher Scientific^{TM}, Waltham, MA, USA) n. C__90305355_10 and C__26668621_10, according to the manufacturer's instructions.

The whole 72 × 18 dataset was handled in a spreadsheet to insert the experimental results of HS1.2 typing (Table S1) and to produce the Phase2 [30] input file. Missing genotypes at rs1045853 (3) and rs77307099 (10) were replaced with ??, as per the instruction manual. The program was run with an elongated chain (1000, 10, and 1000 for the number of iterations, thinning interval, and burn-in, respectively).

In order to project the population coverage of the haplotypes identified here, we inferred the occurrence of the haplotypes defined by the same 18 SNPs in the full series of 503 subjects of European descent (EUR) of the 1000 Genomes project, retaining the original phasing [21].

Measures of LD and the visualization of haplotype blocks were obtained with Haploview [31]. As the implemented methods only allow for strictly biallelic markers, only homozygotes and heterozygotes for HS1.2-length alleles 1 and 2 were included (as reported in Table S1).

2.3. DNA Sequencing

We selected cell line/individual DNAs to undergo HS1.2 sequencing according to the following criteria: (1) putative homozygotes showing a single sharp band on agarose gel; (2) apparent homozygotes for length alleles departing from the canonical sizes reported above (single puffy band); and (3) heterozygotes for large (3,4) alleles uncommon among CEUs. Subsequently, 2 ng/100 bp of PCR products from homozygous individuals were dried for 40 min, at 65 °C and cleaned up with exonuclease I and shrimp alkaline phosphatase, according to standard protocols. Sequencing was performed on both strands with each of the same primers used in the original PCR.

In case (1), the electropherogram was clearly readable, whereas in cases (2) and (3), the electropherogram was compatible with the overlap of two sequences differing in base composition and/or length. In these latter cases, to determine the HS1.2 sequence, PCR amplicons were cloned into pGEM®-T Easy Vector (Promega, Madison, WI, USA) to transform DH5α *Escherichia coli* competent cells. Plasmid DNA from recombinant clones was purified by standard techniques and sequenced on both strands using flanking plasmid-specific primers.

2.4. Search for Putative Transcription Factor Binding Sites and Other Elements

The sequence of each HS1.2 allele was converted to fasta format and analyzed with TFBIND (https://tfbind.hgc.jp, accessed on 2 February 2024) [32] to find putative transcription factor binding sites. This tool uses a weight matrix in the transcription factor database TRANSFAC [33], and originally estimated cut-offs. The analysis was performed with the complete sequence of each allele, and the results partitioned for the CORE (first 94 positions), VAR (variable), and END (last 83 positions) portions of each allele. This was obtained by handling the results of each run in a spreadsheet to compile a direct comparison between outputs. Notably, the assignment of a putative transcription factor binding site to each of the regions is based on the position of the first nucleotide of the binding motif, and some motifs may overlap the boundaries between two regions.

3. Results

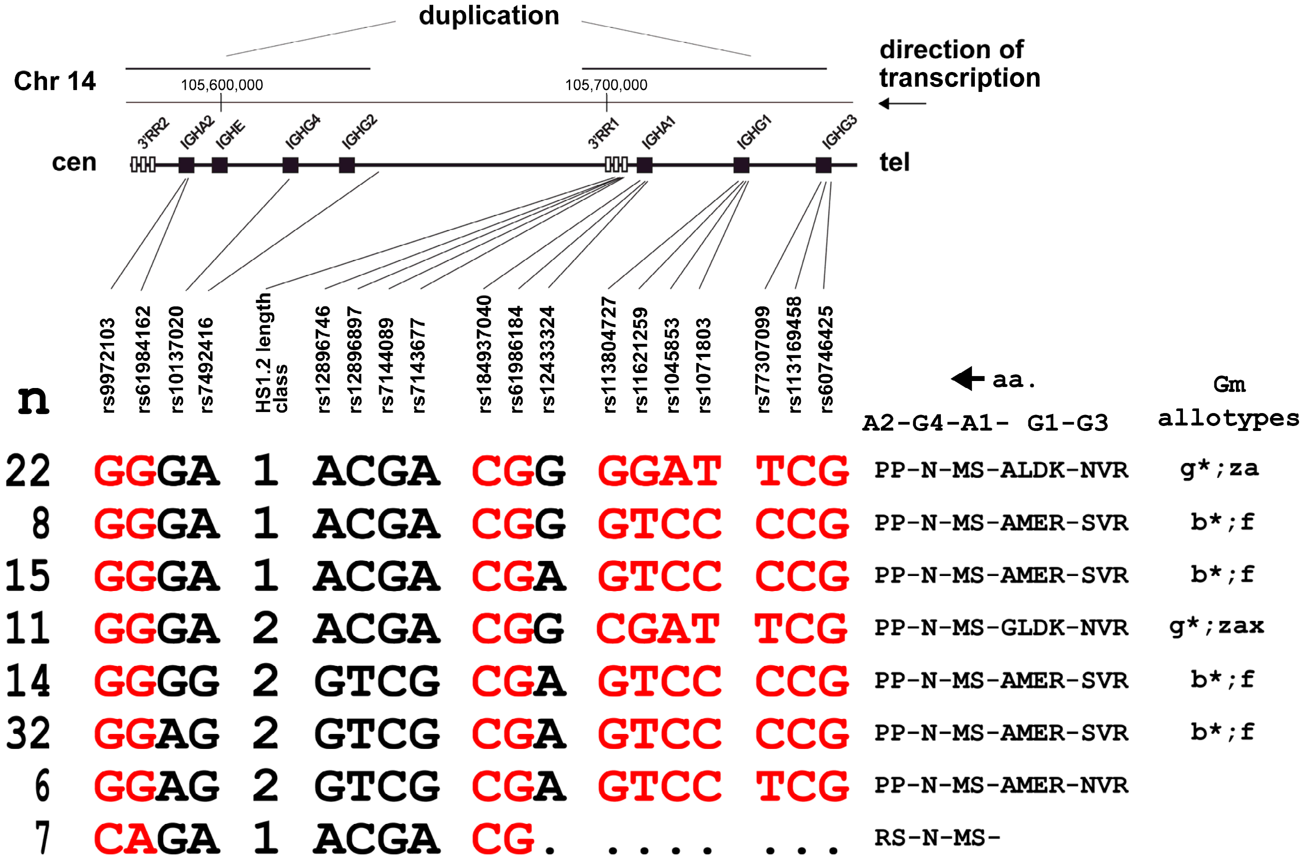
3.1. HS1.2-Length Allele Contributions to Haplotype Blocks

Interpreting the HS1.2 and 18 SNP allele arrangements in the 144 chromosomes of the 72 cell lines resulted in 28 different haplotypes (Table S2). Of these, 12 were singletons, whereas 16 had frequencies ranging from 2 to 32. The 7 most common haplotypes accounted for 75% of the 144 total haplotypes (Figure 1, rows 1–7), denoting strong non-independent arrangements of marker alleles.

HS1.2 allele 1 is clearly preferentially associated with the G-G-G-A and A-C-G-A sub-haplotypes located centromerically and telomerically to it, respectively. On the other hand, HS1.2 allele 2 is preferentially associated with the G-G-A-G and G-T-C-G sub-haplotypes at the same SNP loci.

The uncommon rs61984162 allele A in IGHA2 (Figure 1, last row) was found on four different haplotypes, three of which (Id 25, 26, and 27 in Table S2) carried the HS1.2 allele 1 (seven allele copies).

By inspecting the array of haplotypes in more detail, we identified six broader families (Table S2 and Figure S1), which accounted for 129 out of 144 haplotypes (89.6%). Two such families include the vast majority of HS1.2 allele 1 linked to the G-G-G-A sub-haplotype centromerically and extending to IGHA1 telomerically (orange and light blue). One family included HS1.2 allele 2 linked to the G-G-G-A sub-haplotype centromerically and extended up to IGHG1 telomerically (pink). A medium-frequency family included HS1.2 allele 2, linked only to a centromeric G-G-G-G block (green). A high-frequency family included HS1.2 allele 2 linked to the G-G-A-G sub-haplotype centromerically and extending to IGHA1 telomerically (light green). Finally, one family included HS1.2 allele 1 linked to the C-A-G-A sub-haplotype centromerically, as mentioned above, and extended to IGHA1 telomerically (ochre). Four of the five HS1.2 alleles 3 and 4 were found on two haplotypes, sharing the centromeric G-G-G-A sub-haplotype (no color).



n	rs9972103 rs61984162 rs10137020 rs7492416	HS1.2 length class	rs12896746 rs12896897 rs7144089 rs7143677	rs184937040 rs61986184 rs12433324	rs113804727 rs11621259 rs1045853 rs1071803	rs77307099 rs113169458 rs60746425	A2-G4-A1- G1-G3	Gm allotypes
22	GGGA	1	ACGA	CGG	GGAT	TCG	PP-N-MS-ALDK-NVR	g*;za
8	GGGA	1	ACGA	CGG	GTCC	CCG	PP-N-MS-AMER-SVR	b*;f
15	GGGA	1	ACGA	CGA	GTCC	CCG	PP-N-MS-AMER-SVR	b*;f
11	GGGA	2	ACGA	CGG	CGAT	TCG	PP-N-MS-GLDK-NVR	g*;zax
14	GGGG	2	GTCG	CGA	GTCC	CCG	PP-N-MS-AMER-SVR	b*;f
32	GGAG	2	GTCG	CGA	GTCC	CCG	PP-N-MS-AMER-SVR	b*;f
6	GGAG	2	GTCG	CGA	GTCC	TCG	PP-N-MS-AMER-NVR	
7	CAGA	1	ACGA	CG.	....	...	RS-N-MS-	

Figure 1. The 8 most common 19-loci haplotypes reconstructed from 72 cell lines. A simplified genomic map is reported at the top, showing the genomic position of each marker. Below, from left to right: absolute number of inferred haplotypes, haplotype composition with missense SNP alleles in red (amino acid substitutions listed in Table S1), the corresponding amino acid in the heavy-chain constant region (ordered from right to left), the corresponding Gm allotype with the classical nomenclature reconstructed from https://www.imgt.org/IMGTrepertoire/Proteins/allotypes/human/IGH/IGHC/G1m_allotypes.html (accessed on 2 February 2024). The 29 remaining haplotypes are listed in Table S2, top, and Figure S1.

The frequencies in our limited series were similar to the larger series of 503 Europeans (Table S2, bottom), with the seven most common haplotypes accounting for 63.5%. It is worth noting that relevant discrepancies were observed among the five European subpopulations, in line with the strong structuring of the IGH locus [21]. These patterns denote a fairly strong degree of LD across the entire IGH locus, which apparently strongly links HS1.2 and markers at a distance up to 100 kb on the centromeric side and at least 30 kb on the telomeric side.

We used Haploview to quantify LD and visualize haplotype blocks (Figure S2).

On the centromeric side, strong disequilibrium was confirmed between HS1.2 and two SNPs in IGHG4 and IGHG2 at approximately90 kb. On the telomeric side, four markers displayed almost complete disequilibrium, confirming the initial findings [18]. Further telomerically, the disequilibrium of HS1.2 decays abruptly in the interval rs7143677–rs61986184 and decreases with distance. Conversely, disequilibrium between SNPs remains strong all the way to rs1071803 in IGHG1.

3.2. Internal Sequence Heterogeneity of HS1.2 Alleles

Separation of HS1.2 alleles on agarose gel electrophoresis revealed subtle variation within length classes, which was not clearly sizeable. A review of the images reported in previous papers confirmed this observation, which has already prompted detailed investigations [9]. We then engaged in the characterization of this level of variation by directly sequencing a number of allele copies, obtained from apparently homozygous 1/1 or 2/2 DNA samples or from carriers of alleles 3 and 4, unusual in CEUs, upon cloning. Sequencing covered positions homologous to 14:105696578-105696260 (GRCh38/hg38), as numbered in the RefSeq NT_026437.11.

We found 10 different allele structures, graphically represented in Figure 2 and further detailed in Tables S3–S5. Over the surveyed DNA, three main regions were observed, i.e., a "CORE", invariant in length, spanning pos. 14:105696578-105696485; a portion variable in length spanning positions equivalent to 14:105696484-105696343 in RefSeq; and an "END" portion, invariant in length, spanning pos. 14:105696342-105696260. The overall allele lengths fell into the following classes: 247 to 264 bp, grouped into allele 1; 299 to 319 bp, grouped into allele 2; 352 and 353, for allele 3; and 405 to 425, for allele 4.

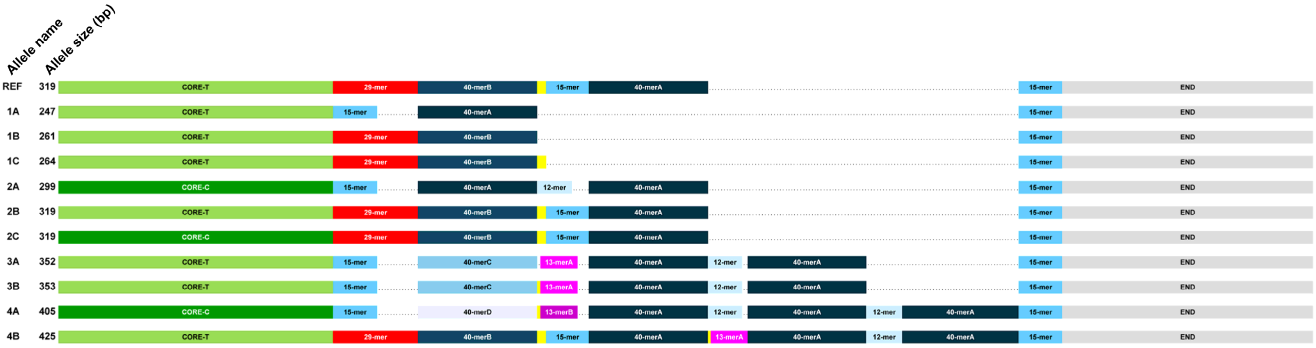

Allele name
Allele size (bp)
REF 319 CORE-T 29-mer 40-merB 15-mer 40-merA 15-mer END
1A 247 CORE-T 15-mer 40-merA 15-mer END
1B 261 CORE-T 29-mer 40-merB 15-mer END
1C 264 CORE-T 29-mer 40-merB 15-mer END
2A 299 CORE-C 15-mer 40-merA 12-mer 40-merA 15-mer END
2B 319 CORE-T 29-mer 40-merB 15-mer 40-merA 15-mer END
2C 319 CORE-C 29-mer 40-merB 15-mer 40-merA 15-mer END
3A 352 CORE-T 15-mer 40-merC 13-merA 40-merA 12-mer 40-merA 15-mer END
3B 353 CORE-T 15-mer 40-merC 13-merA 40-merA 12-mer 40-merA 15-mer END
4A 405 CORE-C 15-mer 40-merD 13-merB 40-merA 12-mer 40-merA 12-mer 40-merA 15-mer END
4B 425 CORE-T 29-mer 40-merB 15-mer 40-merA 13-merA 40-merA 12-mer 40-merA 15-mer END

Figure 2. Internal structure of HS1.2 alleles characterized in this study.

The structure of the variable portion was clearly modular, with different module compositions within each length class (Table S3). The modules are listed in Table S4. In particular, as observed previously [8], the main contributor to length differences was a 40 bp module (40-mer) which was found to be repeated one to four times in alleles 1 to 4, respectively. In between the CORE and the first 40-mer (in the order shown in Figure 2 and Table S3), either a 29-mer or a 15-mer could be found in alleles 1, 2, and 4, whereas among the two copies of allele 3, only the 15-mer was found. These two alternative modules strongly differ in their base composition, with the 15-mer having a G-C content of 93% as opposed to 76% in the 29-mer. In between the first and second 40-mer copy (alleles 2–4), a 12-, 13-, or 15-mer can be found, all of which are very G-C-rich. These can be accompanied by a $C_{1–3}$ stretch. A similar organization is present between the second and third 40-mer (alleles 3,4). Finally, in the two copies of allele 4, only the 12-mer was found between the third and fourth 40-mer. In all sequenced alleles, a copy of the 15-mer is invariably present at the boundary with the END segment.

3.3. Single Nucleotide Variation in the CORE and the Modules

A variable position was found in the CORE, identified as rs373084296. We distinguished copies carrying each of the two SNP alleles as CORE-C and CORE-T, respectively (Figure 2 and Table S3).

Within the 40-mer copies, we found four variable positions, with four arrangements in cis (Table S4 top). The four different 40-mers were thus distinguished in 40-mer A to D. It is worth noting that only the 40-mer closest to the CORE was found to vary. This copy of 40-merA was found adjacent to the 15-mer in alleles 1 and 2 only. The 40-merB was found adjacent only to the 29-mer in alleles 1, 2, and 4. The single 40-mers C and D were adjacent to the 15-mer. The second to fourth additional copies of the 40-mer in alleles 2–4 were invariably of the 40-merA type. No variation was found in the END segment.

In summary, among 34 sequenced alleles, we found three, three, two, and two different types for class 1, 2, 3, and 4, respectively (named in Tables S3 and S5, leftmost column). Overall, we sequenced 14 allele 1s (Table S5), 12 of which turned out to be of the 1A type, and 16 allele 2s, 8 of type 2A and 7 of type 2B. Allele 2B is identical to RefSeq in the DNA region surveyed. We note here that the relative proportions of the different types do not represent population frequencies, due to the strong selection of candidates to sequencing.

3.4. Putative Impact of HS1.2 Diversity on Transcription Factor Binding Sites and Other Elements

We computationally explored the ability of HS1.2 to bind transcription factors, by analyzing the CORE, the length-variable portion (VAR) and the END, separately (Table S6). We focus here on a subset of transcription factors motifs repeatedly reported in the literature.

The transcription factor motif for Oct1 was present in the CORE equally in all alleles described here, as an octanucleotide invariant across our sequences, beginning at 14:105696565, confirming previous data [9].

Motifs for SP1 were found in multiple positions, i.e., the CCCGCCCCCT(C) motif of the 15-mer, e.g., at 14:105696412 [6,34], and the GGGACACCC motif at the boundary between the 40-mer and the 12-, 13-, or 15-mer, e.g., 14:105696363 [8]. As these motifs are increasingly repeated along the allele 1–4 series, the number of predicted binding sites increases accordingly (Table S6). Interestingly, motifs for SP1 are not present in the 29-mer, reducing the number of sites in allele 1A as compared to 1B and 1C.

Conversely, the number motifs for NFkB, which is predicted to bind the boundary between the 40-mer and the 15-mer, increase regularly with allele size.

A binding motif for AP1 (11 bp) starts at 14:105696444 in the 40-mer and covers a variable position. The 40-merB has a C at rs78955324, which reduces the match to the consensus to 10/11.

AP4 and MyoD have similar binding motifs, which are found in the 29-mer (14:105696482) and at the boundary between the 15-mer and the 40-mer (14:105696405). This situation somewhat compensates between alleles with and without the 29-mer, with the opportunity to bind AP4 (Table S6).

Stat3 was reported to bind super-enhancers more than typical enhancers [3]. A site compatible with the consensus motif is invariably present in the CORE at 14:105696515. A second site is found at the boundary of 40-mer (14:105696431) when followed by C_5. Thus, alleles 1A and 1B lack this second site.

In addition to transcription factor binding motifs, other regulatory elements have been identified in HS1.2. Moreover, these elements may be affected by both sequence and structural variation. A DRE (Dioxin Response Enhancer) GCGTG motif [34] was found to start at 14:105696454. However, we predict that this motif is disrupted in the 40-merB, which carries the reference allele T at rs28624614. Therefore, alleles 1B and 1C lack this element completely, whereas other alleles have one or more copies in their 40-merA specimens.

The mouse hs1.2 homologue showed a peak of estrogen receptor binding [26,27]. A 13 bp stretch beginning at 14:105696450 in the 40-merB bears a 12/13 match with the Estrogen receptor Response Elements (ERE) consensus [16,28]. This sequence overlaps two of the variable positions detected here. For the first (14:105696448), both alternative alleles match the ERE consensus, but for the second (rs78955324), the alternative allele T does not. Thus, in this case, the cumulative ability of HS1.2 to bind may also depend on the number of 40-mer repeats and on their particular types.

4. Discussion

This study was inspired by our previous observation [19] that largely divergent 3′RR1 HS1.2 allele frequencies across global populations parallel diverse IgH allotype frequencies [35]. We thus worked out the non-random arrangement of HS1.2 and neighboring SNP alleles in the Ig heavy-chain coding segments on chromosome 14, by directly testing cell lines from subjects of European descent. In phasing our experimental results with genotype calls obtained by whole genome resequencing, we considered only a subset of SNPs enriched in missense and regulatory variants, which impact the functional and immunogenic properties of Ig heavy chains. While some of these did not display polymorphism among the 72 CEUs examined here, they are nevertheless variable in other continental backgrounds (African and East Asia), contributing to the global differentiation of this genomic region.

Our results confirm the strong disequilibrium between HS1.2-length alleles 1 and 2 and the four flanking SNPs of rs12896746, rs12896897, rs7144089, and rs7143677 [18]. However, we show that non-independent arrangements also involve the more distant IGHG4 and IGHA2 gene segments on the centromeric side and, to some extent, the IGHA1, IGHG1, and IGHG3 s well on the telomeric side. The particular haplotypic combination reported here replicates those inferentially predicted [19]. In a recent report, a larger number of haplotypic combinations at the same four SNPs was reported [16]. Our reconstructions did not confirm three of them. Among haplotypes confirmed in both studies, the A-C-G-A and the G-T-C-G arrangements in cis were associated preferentially with HS1.2 alleles 1 and 2, respectively.

Overall, the disequilibria between SNPs appear to be stronger than between SNPs and HS1.2, an indication that the latter accrues diversity at a faster pace, but not enough to reach randomization. In fact, it has been hypothesized that in HS1.2 of both rodents and humans, the generation of multiple alleles has likely been favored by its central position within a large palindromic region [29].

We stress that the above results were obtained on the CEUs of the 1000 Genomes Project [21]. In the context of the strong divergence of human populations in this genomic region [18,19], they call for validation in populations with continental ancestry other than Europeans.

Our sequencing results reveal a repertoire of HS1.2 alleles more variegated than previously described [1,6] and documented in deposited accessions. Both structural and single nucleotide variation contribute to this heterogeneity. Allelic variants of the length-variable region differ in the number and type of modules. In some cases, this diversity implies remarkable variation in the G-C content. Single nucleotide variation at several positions further adds to diversity. The finding of 10 haplotypic arrangements among 34 sequenced alleles predicts that an array of additional variants may be found upon larger screenings. Additionally, it is to be expected that populations in which alleles 3 and 4 are prevalent may display even larger repertoires, due to the higher propensity of repeated segments to undergo unequal exchanges.

Our computational search showed that the occurrence of motifs for binding transcription factors or other regulatory effectors is not simply proportional to allele length, due to two factors: (i) the particular modules contributing to an allele of a given length class; and (ii) the particular module type, especially for the 40-mer.

Overall, the results reported here add support to the concept that the investigated genomic region contributes to inter-individual variation in IgH output through at least three mechanisms: (i) the regulatory activity of HS1.2 during B-cell maturation and the isotype preferentially produced; (ii) the quantitative IgH output; and (iii) specific properties of the particular allotypes. While the joint action of these three factors might be the basis of the observed disease associations of HS1.2, it is indeed possible that some subtypes within a length class contribute disproportionately. The notion that LD spans the entire IGH locus [22] on one side complicates the interpretation of functional consequences of individual variants, but allows for potential synergy between variants in conferring peculiar properties to the final product(s) of specific haplotypic arrangements. This may call for a new generation of studies in which targeted typing could eventually reveal association to a specific allele sub-type.

In conclusion, the current results require a new wave of studies on the specific properties of each variant, both in vivo and in vitro. In this context, previously reported disease associations [11,13,14,16,17] and the degrees of variation identified here align with the observation that sequence variation occurs in transcription regulatory regions of the genome, primarily enhancers, and variation has a disproportionate impact on super-enhancer domains [3]. Finally, the identification of several haplotypes potentially associated with human diseases can serve as a basis to generate studies of molecular genetics with associated immunological factors in the susceptibility to human infectious diseases.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes15070856/s1; six supplementary tables cited in the text are in a single Excel spreadsheet. Table S1: List of SNP loci flanking HS1.2 and considered in the analysis of linkage disequilibrium, experimentally typed cell lines of the 1000 Genomes project and subset included in Haploview display; Table S2, top: Phased 18-SNP haplotypes encompassing HS1.2 in 72 CEU cell lines experimentally typed; bottom: Frequencies of haplotypes defined by the same 18 SNP in the 503 EUR subjects of the 1000 Genomes project, and 5 subpopulations; Table S3: Internal structure of HS1.2 alleles characterized in this study. The sequence on the (−) strand is shown. Variable positions are shown in red. Positions are numbered according to GRCh38/hg38. Landmark positions discussed in the text are reported in the top row; Table S4: Internal sequences of HS1.2 sub-modules. The sequence on the (−) strand is shown. Variable positions are shown in red; Table S5: Distribution of the 34 resequenced alleles by different internal structure and length class; Table S6: Summary results of the search of putative transcription binding sites with TFBIND. Each entry reports the number of identified motifs with the starting base falling in the reported portion of HS1.2; Figure S1: Same data as Table S2 (top), with color shading of similar haplotypes as discussed in the text; Figure S2: Plot of linkage disequilibrium colored according to D′ in phased chromosomes carrying HS1.2 alleles 1 or 2. The distance between markers is reported to scale. LD values for the two rightmost markers are not reported as these are monomorphic in this smaller series.

Author Contributions: A.N., C.J. and G.N. conceptualized the study; C.J., P.M. and B.M.C. performed the experiments; C.J., A.N., G.N. and M.B. discussed the results; C.M.-L. provided reagents (samples); A.N. and C.J. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the HORIZON-HLTH-2021-DISEASE-04 program under grant agreement 101057100 (UNDINE).

Institutional Review Board Statement: Since the present study did not involve any issue relevant for the donors' health, only a subset of the WMA Declaration of Helsinki and COE Oviedo Convention prescriptions were applicable and obeyed. The use of the samples for investigations on genetic diversity was approved by the local Ethics Committee (Comitato Etico Policlinico Tor Vergata protocol numbers 88/11, 90/11, 92/11).

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in GenBank (Acc. n. PP841115-PP841124).

Conflicts of Interest: The authors declare no competing interests.

References

1. Cianci, R.; Mancino, G.; Galli, E.; Serone, E.; Massoud, R.; D'Addabbo, P.; Poscia, A.; Borghetti, A.; Porzio, O.; Marmo, R.; et al. New insight of human-IgH 3′regulatory regions in immunoglobulins switch. *Gene* **2023**, *862*, 147254. [CrossRef]
2. Sepulveda, M.A.; Garrett, F.E.; Price-Whelan, A.; Birshtein, B.K. Comparative analysis of human and mouse 3′ Igh regulatory regions identifies distinctive structural features. *Mol. Immunol.* **2005**, *42*, 605–615. [CrossRef]
3. Hnisz, D.; Abraham, B.J.; Lee, T.I.; Lau, A.; Saint-André, V.; Sigova, A.A.; Hoke, H.A.; Young, R.A. Super-Enhancers in the control of cell identity and disease. *Cell* **2013**, *155*, 934–947. [CrossRef] [PubMed]
4. Pinaud, E.; Marquet, M.; Fiancette, R.; Péron, S.; Vincent-Fabert, C.; Denizot, Y.; Cogné, M. The IgH Locus 3′ Regulatory Region: Pulling the Strings from Behind. In *Advances in Immunology*; Alt, F.W., Austen, K.F., Honj, T., Melchers, F., Uhr, J.W., Unanue, E.R., Eds.; Academic Press: Cambridge, MA, USA, 2011; Volume 110, pp. 27–70.
5. Pinaud, E.; Aupetit, C.; Chauveau, C.; Cogné, M. Identification of a homolog of the Cα3′/hs3 enhancer and of an allelic variant of the 3′IgH/hs1,2 enhancer downstream of the human immunoglobulin α1 gene. *Eur. J. Immunol.* **1997**, *27*, 2981–2985. [CrossRef]
6. D'Addabbo, P.; Scascitelli, M.; Giambra, V.; Rocchi, M.; Frezza, D. Position and sequence conservation in Amniota of polymorphic enhancer HS1.2 within the palindrome of IgH 3′Regulatory Region. *BMC Evol. Biol.* **2011**, *11*, 71. [CrossRef] [PubMed]
7. Garot, A.; Marquet, M.; Saintamand, A.; Bender, S.; Le Noir, S.; Rouaud, P.; Carrion, C.; Oruc, Z.; Bébin, A.-G.; Moreau, J.; et al. Sequential activation and distinct functions for distal and proximal modules within the IgH 3′ regulatory region. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 1618–1623. [CrossRef]

8. Denizot, Y.; Pinaud, E.; Aupetit, C.; Le Morvan, C.; Magnoux, E.; Aldigier, J.C.; Cogné, M. Polymorphism of the human α1 immunoglobulin gene 3′ enhancer hs1,2 and its relation to gene expression. *Immunology* **2001**, *103*, 35–40. [CrossRef] [PubMed]
9. Giambra, V.; Fruscalzo, A.; Giufre′, M.; Martinez-Labarga, C.; Favaro, M.; Rocchi, M.; Frezza, D. Evolution of human IgH3′EC duplicated structures: Both enhancers HS1,2 are polymorphic with variation of transcription factor′s consensus sites. *Gene* **2005**, *346*, 105–114. [CrossRef]
10. Giambra, V.; Martìnez-Labarga, C.; Giufre′, M.; Modiano, D.; Simpore′, J.; Gisladottir, B.K.; Francavilla, R.; Zhelezova, G.; Kilic, S.S.; Crawford, M.; et al. Immunoglobulin enhancer HS1,2 polymorphism: A new powerful anthropogenetic marker. *Ann. Hum. Genet.* **2006**, *70*, 946–950. [CrossRef]
11. Giambra, V.; Cianci, R.; Lolli, S.; Mattioli, C.; Tampella, G.; Cattalini, M.; Kilic, S.S.; Pandolfi, F.; Plebani, A.; Frezza, D. Allele *1 of HS1.2 enhancer associates with selective IgA deficiency and IgM concentration. *J. Immunol.* **2009**, *183*, 8280–8285. [CrossRef]
12. Martínez-Labarga, C.; Lelli, R.; Tarsi, T.; Babalini, C.; De Angelis, F.; Ottoni, C.; Giambra, V.; Pepe, G.; Azzebi, E.; Frezza, D.; et al. Polymorphisms of the COL1A2, CYP1A1 and HS1,2 Ig enhancer genes in the Tuaregs from Libya. *Ann. Hum. Biol.* **2007**, *34*, 425–436. [CrossRef] [PubMed]
13. Frezza, D.; Giambra, V.; Cianci, R.; Fruscalzo, A.; Giufrè, M.; Cammarota, G.; Martìnez-Labarga, C.; Rickards, O.; Scibilia, G.; Sferlazzas, C.; et al. Increased frequency of the immunoglobulin enhancer HS1,2 allele 2 in coeliac disease. *Scand. J. Gastroenterol.* **2004**, *39*, 1083–1087. [CrossRef] [PubMed]
14. Frezza, D.; Giambra, V.; Mattioli, C.; Piccoli, K.; Massoud, R.; Siracusano, A.; Di Giannantonio, M.; Birshtein, B.K.; Rubino, I.A. Allelic frequencies of 3′ Ig heavy chain locus enhancer HS1,2-A associated with Ig levels in patients with schizophrenia. *Int. J. Immunopathol. Pharmacol.* **2009**, *22*, 115–123. [CrossRef] [PubMed]
15. Serone, E.; Daleno, C.; Principi, N.; Porretti, L.; Iacoacci, V.; Gargioli, C.; Magrini, A.; Massoud, R.; D′Addabbo, P.; Cattalini, M.; et al. The change in Ig regulation from children to adults disconnects the correlation with the 3′RR hs1.2 polymorphism. *BMC Immunol.* **2014**, *15*, 45. [CrossRef] [PubMed]
16. Colucci, M.; Frezza, D.; Gambassi, G.; De Vito, F.; Iaquinta, A.; Massaro, M.G.; Di Giambenedetto, S.; Borghetti, A.; Lombardi, F.; Panzironi, N.; et al. Functional associations between polymorphic regions of the human 3′IgH locus and COVID-19 disease. *Gene* **2022**, *838*, 146698. [CrossRef] [PubMed]
17. Cianci, R.; D′Addabbo, P.; Gambassi, G.; Lolli, S.; Serone, E.; Rizzi, A.; Pitocco, D.; Pandolfi, F.; Frezza, D. Association between IgH enhancer hs1.2 and type 1 diabetes. *Acta Diabetol.* **2018**, *55*, 443–448. [CrossRef] [PubMed]
18. D′Addabbo, P.; Serone, E.; Esposito, M.; Vaccari, G.; Gargioli, C.; Frezza, D.; Bianchi, L. Association between Psoriasis and haplotypes of the IgH 3′ Regulatory Region 1. *Gene* **2018**, *669*, 47–51. [CrossRef] [PubMed]
19. Frezza, D.; Martinez-Labarga, C.; Giambra, V.; Serone, E.; Scano, G.; Rickards, O.; D′Addabbo, P.; Novelletto, A. Concerted variation of the 3′ regulatory region of Ig heavy chain and Gm haplotypes across human continental populations. *Am. J. Phys. Anthrop.* **2020**, *171*, 671–682. [CrossRef]
20. Lefranc, M.P.; Lefranc, G. (Eds.) *Human Gm, Km, and Am Allotypes and Their Molecular Characterization: A Remarkable Demonstration of Polymorphism*; Humana Press: Totowa, NJ, USA, 2012; Volume 882.
21. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef]
22. Jefferis, R.; Lefranc, M.-P. Human immunoglobulin allotypes. *mAbs* **2009**, *1*, 332–338. [CrossRef]
23. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **2017**, *550*, 204–213. [CrossRef]
24. Buck, D.; Albrecht, E.; Aslam, M.; Goris, A.; Hauenstein, N.; Jochim, A.; International Multiple Sclerosis Genetics Consortium; Wellcome Trust Case Control Consortium; Cepok, S.; Grummel, V.; et al. Genetic variants in the immunoglobulin heavy chain locus are associated with the IgG index in multiple sclerosis. *Ann. Neurol.* **2012**, *73*, 86–94. [CrossRef] [PubMed]
25. Jonsson, S.; Sveinbjornsson, G.; de Lapuente Portilla, A.L.; Swaminathan, B.; Plomp, R.; Dekkers, G.; Ajore, R.; Ali, M.; Bentlage, A.E.H.; Elmér, E.; et al. Identification of sequence variants influencing immunoglobulin levels. *Nat. Genet.* **2017**, *49*, 1182–1191. [CrossRef] [PubMed]
26. Jones, B.G.; Penkert, R.R.; Xu, B.; Fan, Y.; Neale, G.; Gearhart, P.J.; Hurwitz, J.L. Binding of estrogen receptors to switch sites and regulatory elements in the immunoglobulin heavy chain locus of activated B cells suggests a direct influence of estrogen on antibody expression. *Mol. Immunol.* **2016**, *77*, 97–102. [CrossRef]
27. Jones, B.G.; Sealy, R.E.; Penkert, R.R.; Surman, S.L.; Maul, R.W.; Neale, G.; Xu, B.; Gearhart, P.J.; Hurwitz, J.L. Complex sex-biased antibody responses: Estrogen receptors bind estrogen response elements centered within immunoglobulin heavy chain gene enhancers. *Int. Immunol.* **2019**, *31*, 141–156. [CrossRef]
28. Hurwitz, J.L.; Penkert, R.R.; Xu, B.; Fan, Y.; Partridge, J.F.; Maul, R.W.; Gearhart, P.J. Hotspots for vitamin-steroid-thyroid hormone response elements within switch regions of immunoglobulin Heavy Chain loci predict a direct influence of vitamins and hormones on B Cell class switch recombination. *Viral Immunol.* **2016**, *29*, 132–136. [CrossRef] [PubMed]
29. Guglielmi, L.; Truffinet, V.; Magnoux, E.; Cogné, M.; Denizot, Y. The polymorphism of the locus control region lying downstream the human IgH locus is restricted to hs1,2 but not to hs3 and hs4 enhancers. *Immunol. Lett.* **2004**, *94*, 77–81. [CrossRef]
30. Stephens, M.; Smith, N.J.; Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **2001**, *68*, 978–989. [CrossRef]
31. Barrett, J.C.; Fry, B.; Maller, J.; Daly, M.J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **2005**, *21*, 263–265. [CrossRef]
32. Tsunoda, T.; Takagi, T. Estimating transcription factor bindability on DNA. *Bioinformatics* **1999**, *15*, 622–630. [CrossRef]

33. Wingender, E.; Dietze, P.; Karas, H.; Knüppel, R. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **1996**, *24*, 238–241. [CrossRef]
34. Snyder, A.D.; Ochs, S.D.; Johnson, B.E.; Sulentic, C.E.W. Aryl hydrocarbon receptor-induced activation of the human IGH hs1.2 enhancer: Mutational analysis of putative regulatory binding motifs. *Mol. Immunol.* **2020**, *120*, 164–178. [CrossRef]
35. Cavalli-Sforza, L.L.; Menozzi, P.; Piazza, A. *The History and Geography of Human Genes*; Princeton University Press: Princeton, NJ, USA, 1994.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.