A family of variable immunoglobulin and lectin domain containing molecules in the snail Biomphalaria glabrata

Nolwenn M. Dheilly, David Duval, Gabriel Mouahid, Rémi Emans, Jean-François Allienne, Richard Galinier, Clémence Genthon, Emeric Dubois, Louis Du Pasquier, Coen M. Adema, et al.

Developmental and Comparative Immunology 48 (2015) 234–243

Contents lists available at ScienceDirect

Developmental and Comparative Immunology

journal homepage: www.elsevier.com/locate/dci

A family of variable immunoglobulin and lectin domain containing molecules in the snail *Biomphalaria glabrata*

Nolwenn M. Dheilly [a,b,*,1], David Duval [a,b], Gabriel Mouahid [a,b], Rémi Emans [a,b], Jean-François Allienne [a,b], Richard Galinier [a,b], Clémence Genthon [c], Emeric Dubois [c], Louis Du Pasquier [d], Coen M. Adema [e], Christoph Grunau [a,b], Guillaume Mitta [a,b], Benjamin Gourbal [a,b,**]

[a] CNRS, UMR 5244, Ecologie et Evolution des Interactions (2EI), Perpignan F-66860, France
[b] Université de Perpignan Via Domitia, Perpignan F-66860, France
[c] MGX-Montpellier GenomiX, Montpellier Genomics and Bioinformatics Facility, Montpellier F-34396, France
[d] University of Basel, Institute of Zoology and Evolutionary Biology, Basel CH-4051, Switzerland
[e] Department of Biology, Center for Evolutionary and Theoretical Immunology, University of New Mexico, Albuquerque, NM 87131, USA

ARTICLE INFO

Article history:
Received 26 August 2014
Revised 17 October 2014
Accepted 18 October 2014
Available online 28 October 2014

Keywords:
FREPs
C-type lectin
Galectin
RNAseq
Immunoglobulin superfamily
VIgL

ABSTRACT

Technical limitations have hindered comprehensive studies of highly variable immune response molecules that are thought to have evolved due to pathogen-mediated selection such as fibrinogen-related proteins (FREPs) from *Biomphalaria glabrata*. FREPs combine upstream immunoglobulin superfamily (*IgSF*) domains with a C-terminal fibrinogen-related domain (FreD) and participate in reactions against trematode parasites. From RNAseq data we assembled a *de novo* reference transcriptome of *B. glabrata* to investigate the diversity of FREP transcripts. This study increased over two fold the number of bonafide FREP subfamilies and revealed important sequence diversity within FREP12 subfamily. We also report the discovery of related molecules that feature one or two IgSF domains associated with different C-terminal lectin domains, named C-type lectin-related proteins (CREPs) and Galectin-related protein (GREP). Together, the highly similar FREPs, CREPs and GREP were designated VIgL (Variable Immunoglobulin and Lectin domain containing molecules).

1. Introduction

Pathogens and especially specialized parasites impose considerable selective pressures on their hosts. Therefore it is not surprising that animals of different phyla have independently acquired capabilities for individually diversifying parasite recognition capabilities and immune responses. In fact, diversification of genes or gene families of immune factors is now considered a common feature of animal immunity across phylogeny (Bowden et al., 2007; Ghosh et al., 2011; Hauton and Smith, 2007). Highly variable molecules assuming potential immune function may be found in any phylum of invertebrates. Among those, 185/333 proteins, Toll-like receptors and Nod-like receptors from sea urchins (Buckley and Rast, 2012; Nair et al., 2005), variable chitin binding proteins (VCBP) in cephalochordates (Cannon et al., 2002) and tunicates (Dishaw et al., 2011), Down's syndrome adhesion molecules (DSCAMs) in arthropods (Brites et al., 2008, 2013; Neves and Chess, 2004; Watson et al., 2005; Watthanasurorot et al., 2011), and fibrinogen-related proteins (FREPs) in mollusks (Adema et al., 1997b; Zhang and Loker, 2003; Zhang et al., 2004), have been studied in detail. The list of such highly variable immune factors continues to expand: additional families of potential immune receptors or effectors with an ability for specific recognition of pathogens are regularly described such as the C-type lectins (CTLD) of *Caenorhabditis elegans* (Schulenburg et al., 2008) and NACHT/NB-ARC in the coral *Acropora digitifera* (Hamada et al., 2012; Shinzato et al., 2011).

To date, most of our knowledge regarding the diversity of these molecules derives from traditional transcriptomic approaches (generation of ESTs and targeted PCR) (Brites et al., 2008; Zhang et al., 2004) and proteomics (Dheilly et al., 2009, 2012, 2013; Moné et al., 2010). However, the comprehensive study of the diversity of these highly variable immune molecules has been challenging because each individual sequence variant is expressed at low level such that

* Corresponding author. CNRS, UMR 5244, Ecologie et Evolution des Interactions (2EI), Perpignan F-66860, France. Tel.: (+33)6 25 25 35 14; fax: (+33)4 68 66 22 81.
E-mail address: nolwenn.dheilly@stonybrook.edu (N.M. Dheilly).
** Corresponding author. CNRS, UMR 5244, Ecologie et Evolution des Interactions (2EI), Perpignan F-66860, France. Tel.: (+33)4 30 19 23 12; fax: (+33)4 68 66 22 81.
E-mail address: benjamin.gourbal@univ-perp.fr (B. Gourbal).
[1] Present address: School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY, USA.

http://dx.doi.org/10.1016/j.dci.2014.10.009
0145-305X/

detection is difficult (at protein level) or expensive (at mRNA or genomic DNA level by traditional Sanger sequencing). Therefore, these traditional approaches have allowed us to glimpse, but not fully explore in depth, the extent of sequence diversity and its role in the defense response capabilities of invertebrates to counter infections. To meet such limitations, we used Next Generation Sequencing (NGS) technologies that carry low costs and provide high sequencing coverage for study of the diversity of transcripts that result from somatic diversification of a multigenic family such as FREPs of the gastropod *B. glabrata* (Dheilly et al., 2014).

Functional and genomic characteristics suggest that FREPs play a key role in the immune processes underlying immunocompatibility between *B. glabrata* and the trematode parasite *Schistosoma mansoni*. The immune role of FREPs is indicated by their differential expression over the course of an infection (Hanington et al., 2010b; Hertel et al., 2005) and their specific interaction with polymorphic mucins (SmPoMucs), antigens of *S. mansoni* (Moné et al., 2010). Furthermore, anti-trematode resistance of *B. glabrata* is reduced following FREP knockdown through RNA interference (Hanington et al., 2010a, 2012). FREPs consist of one or two immunoglobulin domains (called IgSF1 and IgSF2) and a carboxyl terminal fibrinogen (FBG) domain (Adema et al., 1997b). They belong to a multigenic family of at least 14 members of which six full-length sequences have been obtained at cDNA or DNA level (Adema et al., 1997b; Léonard et al., 2001; Zhang et al., 2001, 2008; Zhang and Loker, 2003). The combination of allelic polymorphism (Zhang et al., 2004) and somatic modification of FREP genes (Hanington et al., 2010a; Zhang et al., 2004) leads to a remarkable diversification within individual snails. Because of the considerable challenge to study this high diversity, FREPs are thus perfect candidates to test the potential of NGS in deciphering the diversity of transcript sequences derived from diverse, complex, multi-domain gene families.

In this study, we optimized the generation of a transcriptome *de novo* from Illumina sequencing of cDNA, in absence of a reference genome (see supplementary File S1 for details on transcriptome assembly). This assembly was inspected for transcripts that encoded complete and partial FREPs. This study greatly expanded the number of FREP gene subfamilies from *B. glabrata* and revealed a great diversity within the FREP12 subfamily. In addition, it leads to the discovery of new molecules that feature immunoglobulin domains similar to IgSF1 and IgSF2 domains of FREPs associated with either a C-type lectin domain or a galectin domain.

2. Materials and methods

2.1. Snail biological material and sampling

The *de novo* transcriptome assembly was conducted for the BgBRE strain of *Biomphalaria glabrata*, originally collected in Recife, Brazil and maintained in the laboratory for 30 years. This strain has a poor neutral genetic diversity. Microsatellite (neutral genetic markers) characterization indicated that the expected heterozygosity (He) was 0.387, allelic richness (AR) was 3, and Fis was 0.252. This study used two samples of 10 juvenile snails of BgBRE (shell diameter from 4 to 7 mm), two samples of 10 mature adult snails (shell diameter from 8 to 11 mm) and two samples of old adult snails (shell diameter from 12 to 16 mm). All snails were healthy, and were not intentionally immunized; the objective was to assemble the transcriptome of constitutively expressed genes.

2.2. RNA extraction

Total RNA was extracted from individual juvenile, adult and old snails. First, the shell was removed and whole snail body tissues were disrupted under liquid nitrogen with pestle and mortar. Powdered tissues were mixed with 500 µl Trizol reagent and stored at –80 °C. Total RNA was extracted according to the manufacturer's instructions. Briefly, 100 µl of chloroform was added and the sample was homogenized for 15 min at room temperature. Samples were centrifuged 15 min at 12,000 × *g* and the aqueous phase was transferred into a new tube. RNA was precipitated with 500 µl of isopropanol and centrifuged 10 min at 12,000 × *g*. RNA pellets were washed two times with 500 µl of cold ethanol (70%) and dissolved in water. RNA concentrations were determined using a ND-1000 spectrophotometer (Nanodrop Technologies).

2.3. cDNA library construction and Illumina SOLEXA sequencing

Equimolar amounts of RNA from juvenile, adult and old *B. glabrata* were pooled to constitute two biological replicates of 30 individuals: Bre1 and Bre2. RNA concentrations were determined spectrophotometrically (ND-1000 Nanodrop Technologies). RNA integrity was checked on a 2100 Bioanalyzer (Agilent) using RNA 6000 Nano kits (Agilent Technologies). The RNA Integrity Number (RIN) was not considered because of the hidden break in 28S rRNA from *B. glabrata* that causes 28S RNA to break into two fragments that run alongside 18S as in other invertebrate species (Dheilly et al., 2011; Ishikawa, 1977; Winnebeck et al., 2010).

Two paired-end 72bp cDNA libraries were generated using the mRNA-Seq kit for transcriptome sequencing on the Illumina Genome analyzer II platform. Three samples were multiplexed for each lane. Library construction and sequencing were performed by MGX (Montpellier Genomix, c/o Institut de Génomique Fonctionnelle, Montpellier, France). Each library was purified and quantified using a DNA 1000 Chip on a 2100 BioAnalyzer (Agilent Technologies). cDNA fragment size ranged from 220 to 500 bp with an average size of 300 bp. Adapters were ligated to cDNA before analysis on the Illumina Genome Analyzer II platform. The numbers of 72 bp reads resulting from samples Bre1 and Bre2 were 99,316,948 and 73,000,210, respectively. Only reads that passed the default quality filtering performed by the Illumina pipeline were retained. Reads were further cleaned using a workflow created in a local instance of Galaxy (Giardine et al., 2005). This workflow used the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and included the removal of sequencing artifacts, sequence trimming, and clipping of adapters. Quality control revealed an unexpected bias in nucleotide distribution associated with a reduced error per base calling for the initial 13 nucleotides at the 5′ terminus of each read (Fig. S1) possibly due to the random primers used for library generation. Additionally, the sequencing quality was reduced for the three last nucleotides at the 3′ terminus of each read (Fig. S1). Thus all reads were trimmed to remove these low quality sequence intervals (retaining bases 13–69), resulting in paired-end reads of 56 nucleotides each. Subsequently reads without a pair, named orphan reads, were removed (about 10% of total reads). The resulting two high quality libraries with 90,331,578 and 66,588,544 paired-reads for Bre1 and Bre2, respectively, were used for transcriptome assembly.

2.4. De novo *assembly of transcriptomes*

Each sequence dataset was assembled using the python script provided by Oases (version 0.2.06; http://www.ebi.ac.uk/~zerbino/oases/) that runs velvet (version 1.2.02; http://www.ebi.ac.uk/~zerbino/velvet/) for individual single-k assemblies of short reads using the de Bruijn graph algorithm. Oases exploit paired-end information to construct transcript isoforms and compute a merged assembly (Schulz et al., 2012; Zerbino and Birney, 2008). The resulting contigs were merged into unigene clusters with CD-HIT-EST (version 4.5.4 and parameters -c 0.95 -n 10 (Fu et al., 2012)). To optimize the *de novo* transcriptome assembly, the effect of k-mers usage, multiple k-mers assembly and sequencing depth was assessed (Supplementary File S1, Fig. S2). In addition, biological and

technical variabilities of the method were evaluated by comparing the identity of transcripts generated independently from Bre1 and Bre2 samples. The comparison was performed by merging the Bre1 and Bre2 transcriptomes using CD-HIT-EST, to produce a transcriptome called Bre1 + 2 (Fig. S3). Finally, a high quality reference transcriptome was produced with the 157 million paired-reads from Bre1 and Bre2 comprising 117,269 transcripts with a N50 of 983 nucleotides (Figs. S3 and S4).

2.5. Comparative analysis

The biological and technical variabilities of the method were further evaluated by comparing the expression levels of transcripts present in the Bre1 and Bre2 samples (Supplementary File S1, Fig. S4). Illumina reads were mapped against the reference transcriptome using BWA (Li and Durbin, 2009) and analyzed using SAMtools (Li and Durbin, 2009; Li et al., 2009). For each transcript, the number of reads per kilobase per millions of reads mapped (rpkm) was extracted. To compare Bre1 and Bre2 libraries (Supplementary File S1, Fig. S4), a statistical analysis was performed using the DESeq R package following the developer's instructions. Differential expression analysis (Anders and Huber, 2010) was applied using the blind method in fit-only sharing mode with an adjusted *P*-value cutoff of 0.1.

2.6. Transcriptome annotation

Transcripts were automatically annotated using Blast2GO version 2.4.2 (Conesa et al., 2005) using BlastX (e-value cutoff of 1e–3) against the NCBI non-redundant (nr) protein database, Gene Ontology functional annotation (e-value hit-filter 1e–6) (GO; http://www.geneontology.org/), InterProScan functional domain search, and enzyme annotation using the Kyoto Encyclopedia of Genes and Genomes (KEGG). Analyses of Blast2GO annotations are provided in Supplementary File S2 and Fig. 1 and Fig. S5.

2.7. Single nucleotide polymorphism discovery

Reads were initially mapped with Bowtie2 version 2.0.5 with the following parameters: -O3 -m64 -msse2 -funroll -loops -g3; size of int, long, long long, void1, size_t, off_t: 4, 8, 8, 8, 8, 8. Then a mpileup file was generated with SamTools. Single nucleotide polymorphisms (SNPs) were extracted with VarScan with the following options: minimum coverage: 8; minimum reads: 2; minimum quality: 15; minimum variant allele: 0.01; *p*-value: 99e–02.

2.8. Phylogenetic analysis

All transcripts in the reference transcriptome that showed either a high level of similarity (at the amino acid level) to FREP sequences or presenting classical constituent domain (IgSF, and FBG) of FREP sequences were aligned and compared to previous FREP entries available from GenBank (Fig. S6), NCBI (http://www.ncbi.nlm.nih.gov/). Three protein-encoding sequence region characteristics for *B. glabrata* FREPs were analyzed further and only sequences containing full-length domains were conserved: immunoglobin superfamily 1 domain (IgSF1); 26 sequences available: 7 NCBI entries and 19 BgBre sequences; immunoglobin superfamily 2 domain (IgSF2); 39 sequences available: 10 NCBI entries and 29 BgBre sequences, and the FBG domain; 25 sequences available: 13 NCBI entries and 12 BgBre sequences. Sequence accession numbers and abbreviations of sequences used for phylogenetic analyses are provided in Supplementary Table S1.

Multiple sequence alignments (MSA) of the predicted amino acid sequences were obtained using the Muscle component of the software MEGA 5.2 (Tamura et al., 2011) and refined by Gblocks 0.91b (Castresana, 2000; Dereeper et al., 2008, 2010).

Pairwise distances were calculated using MEGA 5.2, applying the number of different residues to determine percentage identity. These values were used to generate a matrix of identity for each FREP region (IgSF1, IgSF2, FBG). Each matrix was used to generate a graphical distribution of percentage of identity frequencies (Fig. S7).

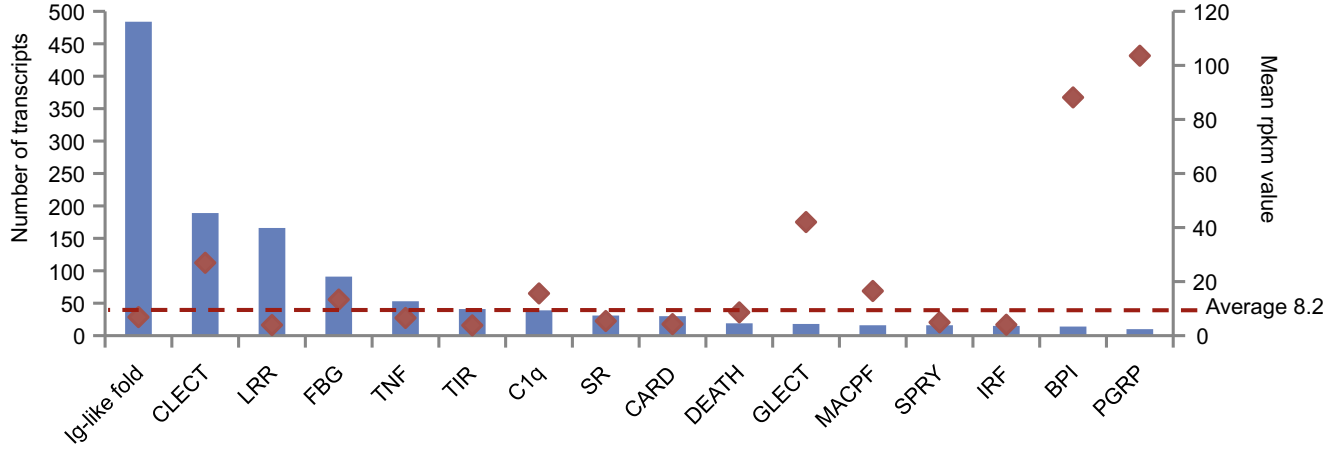

Fig. 1. A high diversity of immune relevant protein domains in *B. glabrata* transcriptome. The histogram provides the number of transcripts with immune-relevant protein domains annotated by InterProScan (blue histograms, left axis) and mean rpkm values (red squares, right axis). The dotted red line shows the mean rpkm value of all transcripts within the transcriptome (8.2). Ig-fold: Immunoglobulin-like fold, IPR013783. LRR: leucine-rich repeat, IPR001611. CLECT: C-type lectin (CTL) or carbohydrate recognition domain (CRD), IPR001304. Ig: immunoglobulin, IPR003599. FBG: fibrinogen-related domains (FReDs, or FBG), IPR002181. TNF: tumor necrosis factor family, IPR006052. TIR: toll-interleukin 1-receptor, IPR000157. C1q: complement component C1q domain, IPR001073. SR: scavenger receptor cystein-rich, IPR017448. CARD: caspase recruitment domain, IPR001315. DEATH: DEATH domain, found in proteins involved in cell death, IPR000488. GLECT: galectin, IPR001079. MACPF: membrane-attack complex/perforin, IPR020864. SPRY: domain in SPIa and the RYanodine receptor, similar to pyrin domain, IPR018355. IRF: interferon regulatory factor, IPR001346. BPI: BPI/LBP/CETP: lipopolysaccharide-binding protein, IPR001124 and IPR017942. PGRP: animal peptidoglycan recognition protein homolog, IPR006619. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Phylogenetic analyses were performed on nucleotide and amino-acid sequences using the neighbor-joining (NJ) and the maximum parsimony (MP) methods using MEGA 5.2. The maximum-likelihood (ML) method was performed with the PhyML program on the Seaview platform (Gouy et al., 2010). Reliability for internal branch was assessed using bootstrapping procedure (2000 replicates for NJ and 1000 replicates for MP and ML). For each sequence dataset, the probabilistic model of sequence evolution (Nei and Kumar, 2000) and the gamma distribution (G) to approximate rate heterogeneity among sites used for the phylogenetic reconstruction was estimated using MEGA 5.2. The phylogenetic analyses yielded congruent topologies for sequences that were represented in each of the three given datasets for IgSF1, IgSF2, FBG. The ML trees are provided in Figs. 2 and 3 (IgSF2) and Figs. S8–S13. Support values below the 50% significance level were not discussed. Some of the deep nodes are not supported.

2.9. *Validation of* de novo *transcripts assembly*

Several *de novo* assembled transcripts were validated by traditional Sanger sequencing of PCR products. Briefly, BgBRE total RNA was reverse transcribed with random primers and RevertAid premium enzyme (Thermo scientific). Two microliters of the RT reaction was then used for PCR (Advantage 2 PCR system, Invitrogen, Carlsbad, CA, USA) with primers that were designed to specially target and amplify novel predicted transcripts (Supplementary Table S2). Amplicons were cloned (pCR4-TOPO, Invitrogen), and sequenced (GATC Biotech, Konstanz, Germany). The Sequencher 4.5 program (Gene Codes, Ann Arbor, MI, USA) was used to align sequences from PCR products and for computational assembly of *de novo* transcripts (Supplementary File S3).

2.10. *Data availability*

Raw read fastq files were submitted to the Sequence Read Archive at NCBI (http://trace.ncbi.nlm.nih.gov/Traces/sra/) under the reference PRJNA213050. The reference transcriptome is publicly available from the transcriptomic database of *Biomphalaria glabrata* at the 2ei website (http://2ei.univ-perp.fr/?page_id=89). The sequences of CREP 1–4 and GREP have been submitted to NCBI database under the accession numbers KM975643, KM975644, KM975645, KM975646, KM975647, respectively.

3. Results

3.1. *An extended array of potential immune-related molecules*

RNAseq data were used to generate *de novo* a high quality reference transcriptome for *B. glabrata* (Supplementary File S1, Figs. S1–S5). The annotated transcriptome was employed to search for and calculate the constitutive expression values of (known and novel) immune-related molecules and gene families (Supplementary File S2, Fig. 1). We used a combined approach of keywords and protein domain searches similar to the one employed by Philipp et al. (2012). Immunoglobulin-like fold, C-type lectin domains (CLECT), leucine-rich repeats (LRR), and FBG domains were the most frequent domains, followed with tumor necrosis factor (TNF) and Toll-interleukin-1 receptor (TIR) domains (Fig. 1). Twenty predicted FREPs were included IgSF domains and 81 of the FBG domain containing transcripts (90% of the transcripts) were predicted to be FREPs. The most abundant transcripts had peptidoglycan recognition protein (PGRP) domains, (lipopolysaccharide-binding protein (LBP/BPI) domains, galectin (GLECT) domains and CLECT domains (expressed in rpkm values; Fig. 1). Proteins with perforin (MACPF) domains, complement component C1q (C1q) domains and FBG domains are also among the top 10% of most highly expressed transcripts. The present study focused on the analysis of the diversity presented by FREPs and related molecules.

3.2. *Diversity of fibrinogen-related proteins (FREPs)*

3.2.1. *FREPs characterization*

Further Blast searches identified 173 complete and partial transcripts that were highly similar to previously characterized *B. glabrata* FREPs. Among those, 86 transcripts incorporated characteristic combinations of diagnostic constituent domains that define FREP sequences. In order, from N- to C-terminal, these are signal peptide, IgSF1, small connecting region (SCR), IgSF2, interceding region (ICR) and FBG. Note that the structure for FREPs with a single IgSF domain is SP, IgSF, ICR, FBG, with the sequence and intron–exon structure of the IgSF domain most similar to the IgSF2 found in the longer FREPs that present two IgSF domains. For instance, 24 transcripts (27.9%) contained a complete or partial IgSF, a complete ICR and a complete or partial FBG domains, 58 transcripts (67.4%) contained complete or partial IgSF + partial ICR domain and finally 4 transcripts (4.6%) were full length and contained all FREP domains (see alignment of Locus-2616_transcript_20 with FREP12 and FREP13 in Fig. S6).

Transcript lengths ranged from a partial sequence of 179 nt (Locus_2479_Transcript_77) to a full-length FREP-encoding sequence of 2513 nt (Locus_2616_transcript_20). The IgSF and FBG domains were highly conserved among all transcripts whereas the ICR varied considerably in sequence and length (from 195 nt for Locus_2616_transcript_20 to 854 nt for Locus_3534_transcript_156). Characteristic imperfect repeats (consisting of amino acid triplets frequently composed of I, K, E residues) were often present within long ICR. The *de novo* assembly of 13 new FREPs (nine partial and four complete sequences) was validated by sequencing of PCR products obtained using transcript-specific primers (Supplementary File S3, Supplementary Table S2).

3.2.2. *Diversity of FREP subfamilies*

None of the 173 transcripts was 100% identical to a FREP sequence entry in GenBank. The discovery of numerous new potential FREPs expands the sequence data available for comparison and facilitates a sharpening of the criteria used to assign FREPs as members of different sub-families. Zhang et al. (Zhang and Loker, 2004a; Zhang et al., 2004) previously suggested to assign two sequences to the same subfamily if their nucleotide identity was equal to or greater than 85%. For three FREP regions, IgSF1, IgSF2 (combined with the IgSF sequence of FREPs with a single IgSF domain), and FBG, we generated matrices of amino acid identity and nucleotide identity with all BgBre sequences and all previously reported FREPs from GenBank. Phylogenetic approaches were used to cluster sequences and allowed subfamily assignments at the transcript and protein level. The branches of the resulting trees were strongly supported and distinguished different gene subfamilies of FREPs. At nucleotide level, 90% or greater sequence identity of constituent domains correctly captures all FREP sequences that cluster together in “gene trees”. This level of sequence difference groups FREP proteins that differ in amino acid composition by no more than 15% (≥85% identity at AA level) and also provides a sensitive criterion for distinction of products from different FREP gene subfamilies (Fig. S7). Congruent subfamilies assignments were obtained from analysis of the different FREP domains.

According to the analysis of the IgSF2 domain, six BgBre FREP transcripts were identified as new members of FREP families 2, 3 and 12 (see the ML trees provided in Fig. 2 and Figs. S8–S13): L50943T1 is highly similar to NCBI FREP2, L3534T147 and L3534T148 were highly similar to NCBI FREP3 and L2479T70, L2479T84 and L2479T128 were highly similar to NCBI FREP12. The remaining FREP

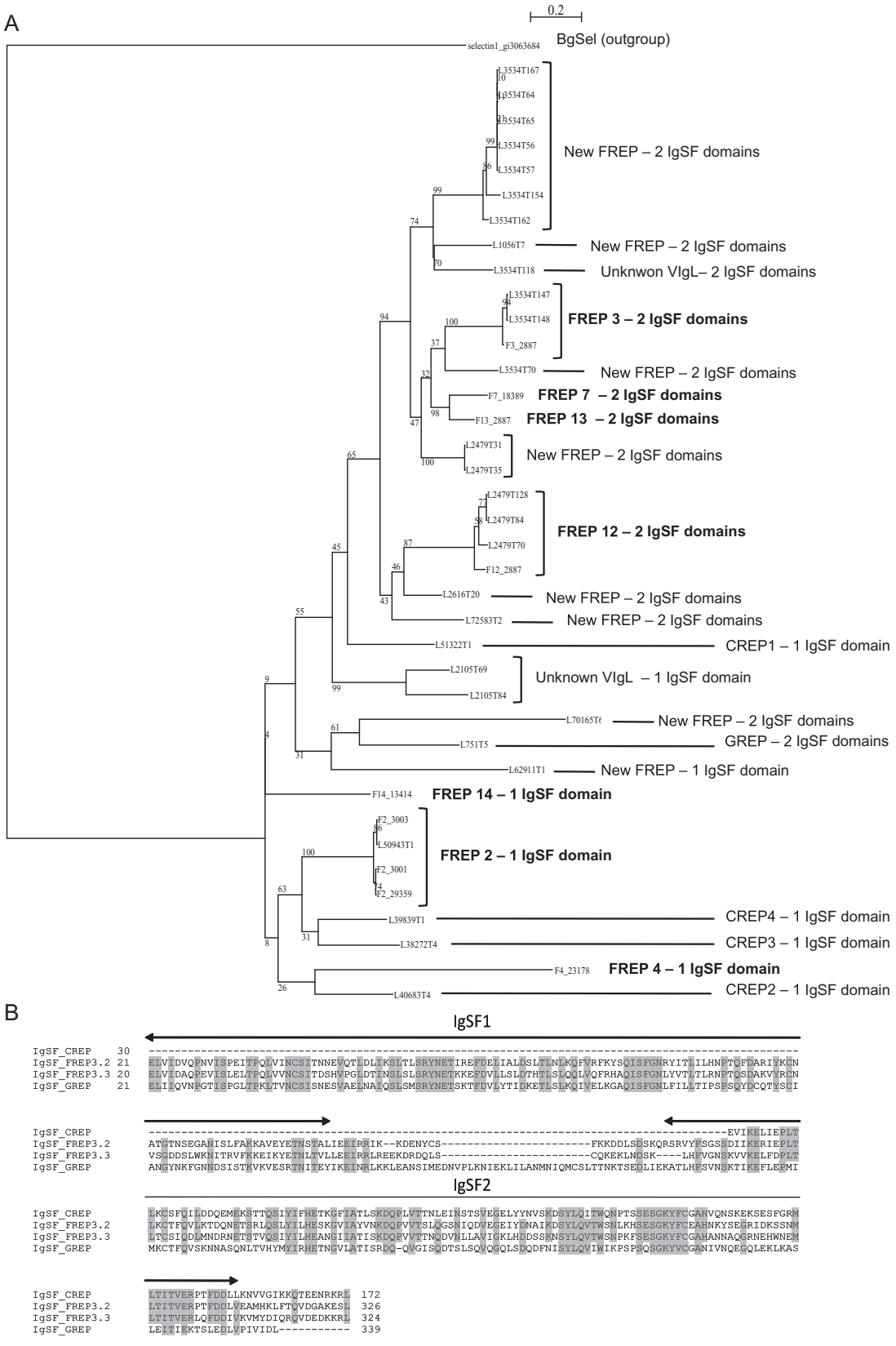

Fig. 2. Subfamily grouping of immunoglobulin (IgSF2) domains of variable immunoglobulin and lectin domain containing molecules (VIgLs). A. Maximum-likelihood tree showing the relationships between FREPs, CREPs and GREP based on alignment of 38 IgSF2 domains. The clustering pattern demonstrates the similarities of IgSF2 sequences among VIgL molecules, yet show that CREPs and GREP remain distinct from all FREP subfamilies identified up to date. BgSel was used as outgroup to confirm that CREPs are more similar to FREPs than to BgSel. See Supplementary Table S1 for the gene sequence abbreviations and accession numbers. Combined with the analysis of the percentage of identity between sequences, this analysis identified new members of the FREP2, FREP3 and FREP12 families and new FREP subfamilies (eight new subfamilies featuring both IgSF domains and FBG domains) and unknown VIgL subfamilies (two new subfamilies with unknown lectin domain). B. Alignment of IgSF2 domains of FREP3.3, FREP3.2, CREP and GREP. Identical amino acid residues between CREP, GREP and FREP3 are boxed in gray.

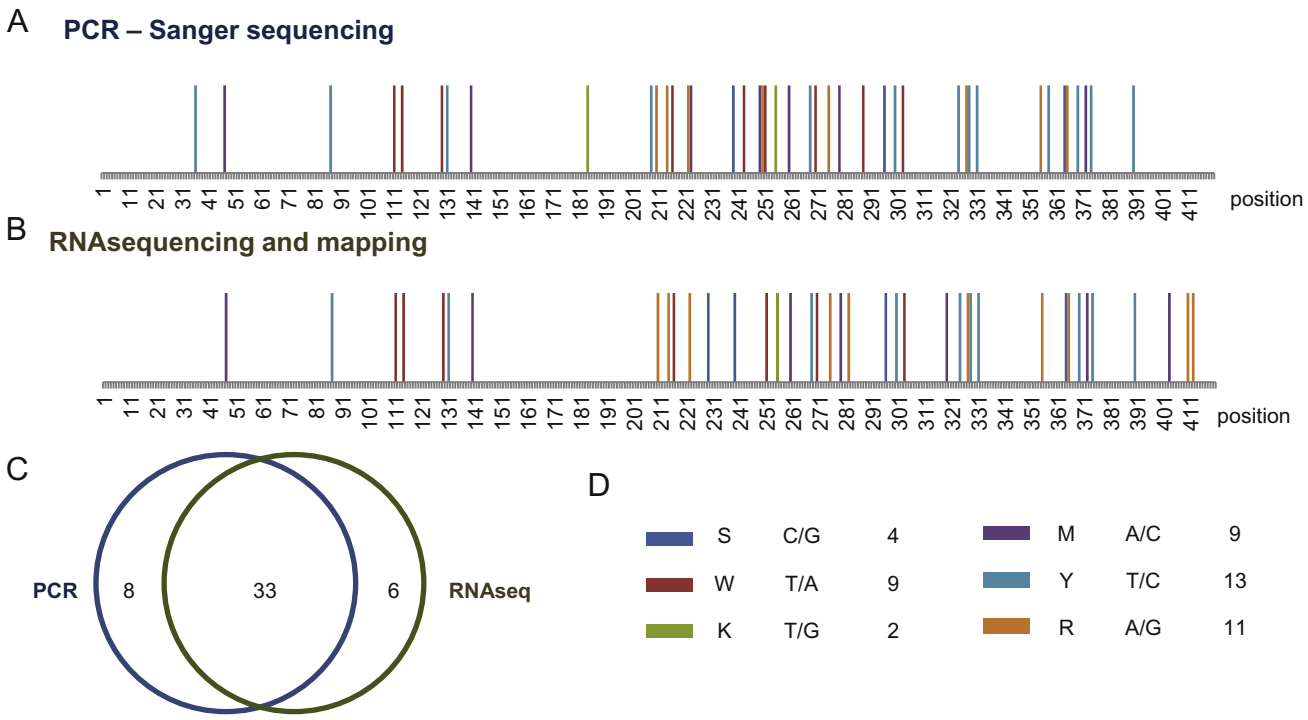

Fig. 3. High diversity within FREP12 subfamily A. Positions of SNPs as obtained from sequencing of PCR products. B. Mapping of RNA-seq reads against reference sequences. C. Venn diagram representing the number of SNPs found with both techniques. D. Legend of color-coded substitutions in A and B.

transcripts grouped into eight new FREP subfamilies and two novels, unknown subfamilies containing IgSF2. The analyses of IgSF1 and FBG showed that all the potential FREP transcripts were novel FREPs, unknown IgSF1 containing molecules or unknown FBG containing molecules – these sequences did not cluster on the same branches with previously reported FREPs (MLTrees) and did not have ≥90% nucleotide identity with any FREP in GenBank (Figs. S8–S13).

3.2.3. Diversity within a FREP subfamily

Based on >90% nucleotide identity, several transcripts were evidently members of the FREP12 subfamily. The mapping of reads from the Bre1 and Bre2 libraries against three divergent reference transcripts of a 420 nt long fragment of FREP12 (Locus_2479_Transcript_128, Locus_2479_transcript_147 and Locus_2479_Transcript_85) revealed 39 variable (SNP) nucleotide positions. From a pool of the original Bre1 and Bre2 RNA samples, 59 cloned amplicons of the same 420 bp long fragment were generated and sequenced by traditional Sanger sequencing (Supplementary File S3). Alignment of these sequences revealed 41 variable sites. The position and the nature of 33 SNPs (80%) were identical with both approaches (Fig. 3), this confirmed that 7.9% of the residues in this 420 nt long fragment represent polymorphic sites. Clearly, FREP12 is highly diverse within this strain of *B. glabrata*.

3.3. Discovery of new Variable Immunoglobulin and Lectin domain containing molecules (VIgLs)

The search for particular IgSF domains as component of FREP sequences led to identification of novel molecules that resemble FREPs in domain structure and sequence. These included transcripts that consisted of one or two IgSF and an ICR highly similar to FREP sequences followed by either a C-type lectin domain or a galectin domain (Supplementary File S4, Fig. 4). These were named C-type lectin-related protein (CREP) and galectin-related protein (GREP), respectively. The new *de novo* assembled transcripts were validated by Sanger sequencing of PCR products (Supplementary File S3, supplementary Table S2).

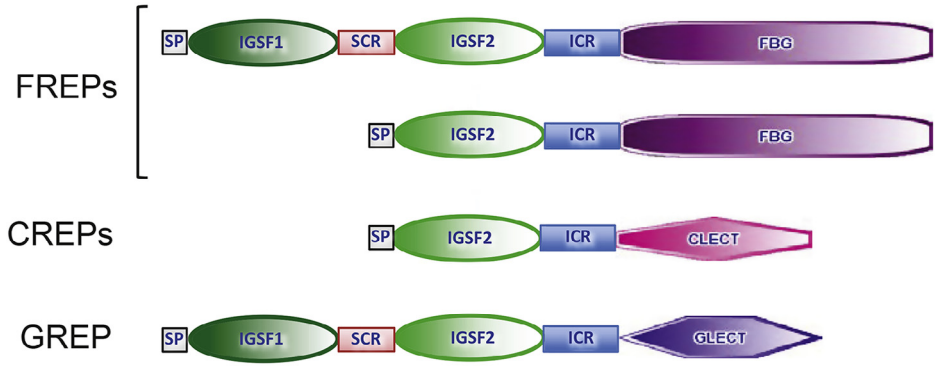

Fig. 4. Similar domain structure of VIgLs. SP: signal peptide; IgSF1: immunoglobulin superfamily domain 1; IgSF2: immunoglobulin superfamily domain 2; SCR: small connecting region; ICR: interceding region; FBG: fibrinogen domain; CLECT: C-type lectin domain; GLECT: galectin domain.

A total of four distinct CREPs composed of an IgSF domain in association with a C-terminal C-type lectin domain were identified in the reference transcriptome (Supplementary File S3). This architecture is identical to that of *Biomphalaria* selectin (BgSel, (Guillou et al., 2004), and the lectins of ~35 kDa recovered from albumen gland and egg masses of *B. glabrata* (Hathaway et al., 2010) but the IgSF domain of CREP is more similar to the second IgSF domain sequence from FREPs (60% nt identity with IgSF2 from FREP3.3; gi: 346721864). The CLECT domain of CREPs had all the 11 canonical residues that compose the ligand binding interface of C-type lectins (Fig. S14). BlastX analysis showed that the CLECT domain of CREP1 (Locus_51322_Transcript_1) was more similar to molluskan CLECT (incilarin, perlucin, C-type lectin proteins and selectins) whereas CREP2, CREP3 and CREP4 (Locus_40683_Transcript_4 Locus_38272_Transcript_4 and Locus_39839_Transcript_1, respectively) have CLECT domains that were more similar to vertebrate CLECT (CLEC4 and CLEC17).

Only one full length GREP transcript was found (Locus_751 _Transcript_1). It consisted of two tandemly arranged IgSF domains upstream of a C-terminal galectin domain. The N-terminal IgSF domain is highly similar to (the N-terminal) IgSF1 of FREPs (52% nt identity with IgSF1 from FREP3.3; gi: 28875402), and the second IgSF domain has high similarity with (the second) IgSF2 of FREPs (43% nt identity with IgSF2 from FREP14; gi: 346721861). The GLECT domain of GREP was highly divergent from other previously reported galectins. Moreover, this is the first report of a GLECT domain that associates with another type of domain. Still, the GLECT domain of GREP did contain all the eight canonical amino acids that constitute the conserved sugar binding pocket of GLECT (Fig. S14). In addition, 8 of the 11 residues that compose the dimerization interface were conserved and all 11 residues that compose the putative alternate dimerization interface were present (Fig. S14).

Due to the high percentage of identity between IgSF domains of CREPs, GREP and FREPs, IgSF2 domains of CREPs and GREP could be included in the phylogenetic analysis of IgSF2 domains of FREPs (BgSel was used as an outgroup to confirm that CREPs and GREP are more similar to FREPs; Figs. S10 and S11, Fig. 2) and IgSF1 domain of GREP could be included in the phylogenetic analysis of IgSF1 domains of FREPs (Figs. S8 and S9). Alignment demonstrates the high proportion of identical nucleotide and amino acid residues among IgSF domains of CREP, GREP and FREP3.2 and FREP3.3 (Fig. 2). Moreover, the CREP and GREP-derived IgSF sequences did not segregate out as separate clades but clustered within different branches of FREP IgSF domains (Figs. S8–S11, Fig. 3). Therefore, CREP and GREP IgSF sequences are as similar to IgSF of FREP genes as different IgSF sequences of FREP IgSF are to each other (Fig. 2). Interestingly, and in accordance with the analysis of CLECT domains, the IgSF domain of CREP1 clusters apart from the IgSF domains of CREP2, CREP3 and CREP4. The latter appear more similar to the IgSF domains of FREP2 and FREP14. In consideration of the evident high similarity of IgSF1 and IgSF2 domains of FREPs, CREPs and GREP, we grouped all these molecules within a single family distinct from other molecules with the same domain architecture (IgSF and lectin domains). FREPs, CREPs and GREP were named VIgL for variable immunoglobulin and lectin domain containing molecules (Fig. 4).

4. Discussion

Next generation sequencing techniques are lifting limitations that restricted access to invertebrate genomes in ways that now facilitate comparative immunologists to study structure and function of the immune system of any organism, including non-model species (Dheilly et al., 2014). In the present study, we tested the hypothesis that NGS allows the study of the diversity of highly variable molecules (Dheilly et al., 2013, 2014) by investigating *B. glabrata* FREPs diversity from a *de novo* generated transcriptome. FREPs constitute the most highly diversified immune recognition gene family described from *B. glabrata* and are actively involved in anti-schistosome immune response (Hanington et al., 2010a; Mitta et al., 2012; Moné et al., 2010). This makes FREPs good candidates to evaluate the ability of state of the art *de novo* transcriptome assembly to provide a comprehensive representation of a large diversity of highly similar sequences. To date, 14 subfamilies of FREPs have been described, and these differ not just in nucleotide content but also by having one or two immunoglobulin domains and the length of the interceding region. Moreover, FREPs within subfamilies diversify even further through different mechanisms including alternative splicing and somatic diversification through gene conversions and point mutations (Hanington et al., 2012; Zhang et al., 2004).

The initial search for FREP sequences in the reference transcriptome led to the discovery of transcripts that encode novel types of lectins with upstream IgSF domains highly similar to FREP IgSF domains associated with different types of lectin domains, C-type lectins for CREPs and galectin for GREP (Fig. 4). Sequencing of specific PCR products provided experimental confirmation for the existence of these new molecules. The recovery of four full-length CREP sequences suggests that CREPs constitute a new category of lectins in *B. glabrata*. The single full-length GREP has a novel, unique structure that reveals the existence of a new category of galectins. Indeed, it is the first report of a galectin domain associated with another domain. Galectins are a family of β-galactoside-binding lectins and are among the most conserved and ubiquitous family of lectins (Kilpatrick, 2002). They are composed of a one to four galectin domains (Vasta et al., 2012). Galectins are involved in host–pathogen interactions by recognition of exogenous ligands like glycans on the surface of viruses, bacteria, fungi and protozoa (Vasta, 2009, 2012). Moreover, *B. glabrata* galectin (BgGal) has hemagglutinating activity. BgGal is absent from cell-free plasma and immunolocalizes in the plasma membrane of some sub-populations of snail hemocytes. It has been suggested that BgGal may serve as a pattern recognition receptor that selectively recognizes and binds hemocytes to pathogens with appropriate sugar ligands (Yoshino et al., 2008). All these observations provide further support for the idea that GREP could have a role in *B. glabrata* immune response. Similarly to other galectins, GREPs may interact with each other and form dimers or multimers (Song et al., 2011; Tasumi and Vasta, 2007; Vasta, 2012; Zhang et al., 2011). The discovery of a GREP protein and of four CREPs reveals the existence in *B. glabrata* of a broader category of lectins that like FREPs, Bgselectin (Guillou et al., 2004), and IgSF-CLECT (Hathaway et al., 2010) sequences is composed of one or two closely related IgSF domains with a downstream lectin domain that may be either a fibrinogen (FBG), C-type lectin (CLECT) or galectin (GLECT). The high similarity at nucleotide and amino acid levels of the IgSF domains and ICR within FREPs, CREPs and GREP may indicate that these sequences originated from a common ancestor gene and/or that these molecules participate in the same or related biological pathways. Lectins with related N-terminal sequences are likely to be processed by similar receptors or cell types and/or interact with each other in order to provide diversity in recognition molecules of various polysaccharides. Together, FREPs, CREPs and GREP are designated VIgL, which stand for variable immunoglobulin and lectin domain containing molecules family.

The discovery of an extended family of VIgL, with the discovery of GREP and CREPs as related yet different lectins that also contain IgSF sequences similar to the IgSF domains recorded from FREPs, emphasizes the need for a precise definition of *B. glabrata* FREPs. Correct identification of a FREP requires demonstration of the association of an upstream IgSF domain followed by an FBG domain sequence. Partial transcripts that cover only IgSF domains

or IgSF and ICR cannot be named FREPs because they may derive from CREPs or GREP. Similarly, a partial sequence that contains only a FBG domain may originate from *B. glabrata* fibrinogen-related molecule (FREM), that combines an FBG domain with upstream collagen-like repeats (Zhang et al., 2008). Consequently, of the 14 previously reported FREP gene subfamilies, 8 can be unambiguously considered as FREPs because they possess both IgSF and FBG domains (FREPs 2, 3, 4, 5, 7, 12, 13 and 14) (Léonard et al., 2001; Zhang et al., 2001; Zhang and Loker, 2003, 2004b).

The current study yielded 28 transcripts that covered at least partially the IgSF domain and the FBG domain and these new FREPs clustered in at least eight new FREP subfamilies. The remaining 58 transcripts are referred to as "unknown VIgL" until a more complete sequence covering the IgSF domain and lectin domain is obtained. However, it seems that the more diversified a family is, the more the *de novo* assembly generates a proportion of partial sequences. For instance, of the 28 transcripts confirmed to be FREPs, only 14% were full length whereas the four CREP transcripts and the single GREP transcript recovered were almost completely full length. Although it indicates that at this time, for highly variable multigenic sequence families, it is difficult to use only Illumina RNA sequencing in efforts to assemble and reconstruct the complete family, this observation strongly suggests that most "unknown VIgL" subfamilies may be partial FREP sequences. The high discovery rate of novel FREP subfamilies is afforded by the unprecedented sequence coverage that is provided by the RNAseq. Additionally, the absence of some subfamilies previously entered into GenBank and the discovery of new FREP and "unknown VIgL" subfamilies may result directly from the use of a Brazilian strain of *B. glabrata*, different from those studied before (Porto Rico strains or crosses between Brazilian and Porto Rico strains). The FREP gene family may increase in size each time the transcriptome of a different strain of *B glabrata* is inspected through revealing new alleles or loci. Regardless, this study increased over twofold the number of known FREP subfamilies and suggests that it represents only a fraction of the existing FREP diversity in *B. glabrata*. The present study has provided a basic necessary knowledge of the diversity of FREPs that can now be further investigated using a targeted sequencing approach to sequence the different FREP subfamilies (Babik et al., 2009; Hughes et al., 2013).

In a previous analyses of FREPs diversity, Zhang et al. (2004) and Hanington et al. (2012) showed that gene conversions may generate additional variants within the same FREP3 gene subfamily. The phylogenetic component of the present study did not detect the occurrence of recombinatorial diversification between different FREP subfamilies. Such recombination would have been revealed by an incongruent clustering of the same sequence in the phylogenetic analysis of the different domains (IgSF1, IgSF2 and FBG). While it may be that the assembly process did not reveal all sequence variants, or that healthy animals did not produce recombinant (variant or diversified) molecules, it may perhaps be that recombinatorial diversification remains restricted to within subfamilies of FREPs. FREP gene sequences are also (somatically) diversified through random point mutations (Hanington et al., 2012; Moné et al., 2010; Zhang et al., 2004). In the present study of the Brazilian *B. glabrata* strain, FREP 12 subfamily appeared to be the most highly diversified subfamily. Different FREP12 source variants were assembled *de novo*. Then, read mapping against reference FREP12 transcripts and traditional Sanger sequencing of clones confirmed a further diversification by point mutation. Previous studies revealed a high diversity of FREP3 in *S. mansoni*-exposed resistant BS90 strain of *B. glabrata* and a high diversity of FREP2 in infected Brazilian strain of *B. glabrata*. The present study investigated the diversity of immune recognition molecules in non-stimulated snails. Indeed, *B. glabrata* is able to activate an efficient innate cellular immune response immediately after an encounter with a pathogen (Deleury et al., 2012; Mitta et al., 2012). Here, FREP12 variability is found in healthy non-challenged individuals, which suggests a constitutive expression of diversified FREPs that participate in constitutive or anticipatory immunity. Now, further studies on VIgLs diversity are necessary to determine if these molecules are more diversified in response to pathogen exposures and if there are differences between individuals and/or populations.

FREPs actively participate in *B. glabrata* resistance to *S. mansoni* (Hanington et al., 2010a, 2012). They bind to *S. mansoni* polymorphic mucins (SmPoMuc) on the parasite surface (Moné et al., 2010) and in ways thought to lead to parasite elimination (Mitta et al., 2012). The polymorphism and high diversity of interacting FREPs and SmPoMuc supports the compatibility polymorphism and represent one part of the molecular processes involved in the matching phenotype hypothesis (Mitta et al., 2012). Indeed, in natural populations, some snail/schistosome interactions are compatible (meaning that the host is successfully infected) and others are not (the host is resistant to the parasite strain). The RNAi-effected knockdown of FREP3 led to susceptibility to trematode infection in about 30% of normally resistant *B. glabrata.* This demonstrated the contribution of FREP3 in gastropod immunity and at the same time confirmed the involvement of additional determinants of anti-schistosoma immune responses (Hanington et al., 2010a, 2012). Perhaps FREPs, CREPs and GREP serve as complementary or collaborative recognition factors that would be processed through the same pathway. In addition, different VIgLs could interact with each other to form complexes. Adema et al. (1997a) showed that parasite-reactive lectins, including FREPs, occur as high molecular weight complexes under native conditions. It was not resolved whether these large multimers were of homologous or heterologous composition. Protein polymerization has previously been proposed for other highly variable molecules such as 185/333 proteins of sea urchins (Dheilly et al., 2009). In *Anopheles gambiae*, functional studies of fibrinogen-related domain (FreD) containing protein and C-type lectins have revealed complementary and synergistic functions that are mediated by inter- and intra-molecular associations. Formation of multimers provides a means of increasing the mosquito's pathogen recognition receptor repertoire and mediating anti-pathogen responses (Cirimotich et al., 2010). As stated by Schulenburg et al. (2007) regarding lectins "multimerization increases binding valency and avidity, and as such, the potential for specific recognition of parasite molecules". Interestingly, *B. glabrata* is able to mount a highly specific inducible protection against different strains or species of *Schistosoma* as demonstrated in immune priming experiments (Portela et al., 2013). Future studies may clarify whether FREPs, CREPs and GREP are able to interact by multimerization. It is tempting to hypothesize the high combinatorial diversity of different types of VIgLs, that derive from different loci with multiple alleles, including FREPs that are yet further diversified by alternative splicing and continuous somatic mutations, in service of non-self recognition in *B. glabrata*.

Acknowledgements

NMD was supported by the Agence Nationale de la Recherche (ANR) Blanc, SVSE7, project Bodyguard to FT. BG acknowledges support from ANR JCJC INVIMORY number ANR-13-JSV7-0009. CMA acknowledges support from NIH grant number P20GM103452 from the National Institute of General Medical Sciences (NIGMS).

Appendix: Supplementary Material

Supplementary data to this article can be found online at doi:10.1016/j.dci.2014.10.009.

References

Adema, C.M., Hertel, L.A., Locker, E.S., 1997a. Infection with *Echinostoma paraensei* (Digenea) induces parasite-reactive polypeptides in the hemolymph of the gastropod host *Biomphalaria glabrata*. In: Beckage, N.E. (Ed.), Parasites and Pathogens – Effects on Host Hormones and Behavior. Chapman Press, New York, pp. 77–99.

Adema, C.M., Hertel, L.A., Miller, R.D., Loker, E.S., 1997b. A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. Proc. Natl. Acad. Sci. U. S. A. 94, 8691–8696.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106.

Babik, W., Taberlet, P., Ejsmond, M., Radwan, J., 2009. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. Mol. Ecol. Resour. 9, 713–719.

Bowden, L., Dheilly, N.M., Raftos, D.A., Nair, S.V., 2007. New immune systems: pathogen-specific host defence, life history strategies and hypervariable immune-response genes of invertebrates. Invert. Surv. J. 4, 127–136.

Brites, D., McTaggart, S., Morris, K., Anderson, J., Thomas, K., Colson, I., et al., 2008. The Dscam homologue of the crustacean *Daphnia* is diversified by alternative splicing like in insects. Mol. Biol. Evol. 25, 1429–1439.

Brites, D., Brena, C., Ebert, D., Du Pasquier, L., 2013. More than one way to produce protein diversity: duplication and limited alternative splicing of an adhesion molecule gene in Basal arthropods. Evolution 67, 2999–3011.

Buckley, K.M., Rast, J.P., 2012. Dynamic evolution of toll-like receptor multigene families in echinoderms. Front. Immunol. 3, 136.

Cannon, J.P., Haire, R.N., Litman, G.W., 2002. Identification of diversified genes that contain immunoglobulin-like variable regions in a protochordate. Nat. Immunol. 3, 1200–1207.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.

Cirimotich, C.M., Dong, Y., Garver, L.S., Sim, S., Dimopoulos, G., 2010. Mosquito immune defenses against *Plasmodium* infection. Dev. Comp. Immunol. 34, 387–395.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Deleury, E., Dubreuil, G., Elangovan, N., Wajnberg, E., Reichhart, J.M., Gourbal, B., et al., 2012. Specific versus non-specific immune responses in an invertebrate species evidenced by a comparative de novo sequencing study. PLoS ONE 7, e32512.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., et al., 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36, W465–W469.

Dereeper, A., Audic, S., Claverie, J.M., Blanc, G., 2010. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol. Biol. 10, 8.

Dheilly, N., Lelong, C., Huvet, A., Favrel, P., 2011. Development of a Pacific oyster (*Crassostrea gigas*) 31,918-feature microarray: identification of reference genes and tissue-enriched expression patterns. BMC Genomics 12, 468.

Dheilly, N.M., Nair, S.V., Smith, L.C., Raftos, D.A., 2009. Highly variable immune-response proteins (185/333) from the sea urchin, *Strongylocentrotus purpuratus*: proteomic analysis identifies diversity within and between individuals. J. Immunol. 182, 2203–2212.

Dheilly, N.M., Haynes, P.A., Raftos, D.A., Nair, S.V., 2012. Time course proteomic profiling of cellular responses to immunological challenge in the sea urchin, *Heliocidaris erythrogramma*. Dev. Comp. Immunol. 37, 243–256.

Dheilly, N.M., Raftos, D.A., Haynes, P.A., Smith, L.C., Nair, S.V., 2013. Shotgun proteomics of coelomic fluid from the purple sea urchin, *Strongylocentrotus purpuratus*. Dev. Comp. Immunol. 40, 35–50.

Dheilly, N.M., Adema, C., Raftos, D.A., Gourbal, B., Grunau, C., Du Pasquier, L., 2014. No more non-model species: the promise of next generation sequencing for comparative immunology. Dev. Comp. Immunol. 45, 56–66.

Dishaw, L.J., Giacomelli, S., Melillo, D., Zucchetti, I., Haire, R.N., Natale, L., et al., 2011. A role for variable region-containing chitin-binding proteins (VCBPs) in host gut-bacteria interactions. Proc. Natl. Acad. Sci. 108, 16747–16752.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Ghosh, J., Lun, C.M., Majeske, A.J., Sacchi, S., Schrankel, C.S., Smith, L.C., 2011. Invertebrate immune diversity. Dev. Comp. Immunol. 35, 959–974.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., et al., 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 15, 1451–1455.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224.

Guillou, F., Mitta, G., Dissous, C., Pierce, R., Coustau, C., 2004. Use of individual polymorphism to validate potential functional markers: case of a candidate lectin (BgSel) differentially expressed in susceptible and resistant strains of *Biomphalaria glabrata*. Comp. Biochem. Physiol. 138, 175–181.

Hamada, M., Shoguchi, E., Shinzato, C., Kawashima, T., Miller, D.J., Satoh, N., 2012. The complex NOD-like receptor repertoire of the coral *Acropora digitifera* includes novel domain combinations. Mol. Biol. Evol. 30, 167–176.

Hanington, P.C., Forys, M.A., Dragoo, J.W., Zhang, S.M., Adema, C.M., Loker, E.S., 2010a. Role for a somatically diversified lectin in resistance of an invertebrate to parasite infection. Proc. Natl. Acad. Sci. 107, 21087–21092.

Hanington, P.C., Lun, C.M., Adema, C.M., Loker, E.S., 2010b. Time series analysis of the transcriptional responses of *Biomphalaria glabrata* throughout the course of intramolluscan development of *Schistosoma mansoni* and *Echinostoma paraensei*. Int. J. Parasitol. 40, 819–831.

Hanington, P.C., Forys, M.A., Loker, E.S., 2012. A somatically diversified defense factor, FREP3, is a determinant of snail resistance to schistosome infection. PLoS Negl. Trop. Dis. 6, e1591.

Hathaway, J.J.M., Adema, C.M., Stout, B.A., Mobarak, C.D., Loker, E.S., 2010. Identification of protein components of egg masses indicates parental investment in immunoprotection of offspring by *Biomphalaria glabrata* (Gastropoda, Mollusca). Dev. Comp. Immunol. 34, 425–435.

Hauton, C., Smith, V.J., 2007. Adaptive immunity in invertebrates: a straw house without a mechanistic foundation. Bioessays 29, 1138–1146.

Hertel, L.A., Adema, C.M., Loker, E.S., 2005. Differential expression of FREP genes in two strains of *Biomphalaria glabrata* following exposure to the digenetic trematodes *Schistosoma mansoni* and *Echinostoma paraensei*. Dev. Comp. Immunol. 29, 295–303.

Hughes, G.M., Gang, L., Murphy, W.J., Higgins, D.G., Teeling, E.C., 2013. Using Illumina next generation sequencing technologies to sequence multigene families in *de novo* species. Mol. Ecol. Resour. 13, 510–521.

Ishikawa, H., 1977. Evolution of ribosomal RNA. Comp. Biochem. Physiol. B 58, 1–7.

Kilpatrick, D.C., 2002. Animal lectins: a historical introduction and overview. Biochim. Biophys. Acta 1572, 187–197.

Léonard, P.M., Adema, C.M., Zhang, S.-M., Loker, E.S., 2001. Structure of two FREP genes that combine IgSF and fibrinogen domains, with comments on diversity of the FREP gene family in the snail *Biomphalaria glabrata*. Gene 269, 155–165.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Mitta, G., Adema, C.M., Gourbal, B., Loker, E.S., Theron, A., 2012. Compatibility polymorphism in snail/schistosome interactions: from field to theory to molecular mechanisms. Dev. Comp. Immunol. 37, 1–8.

Moné, Y., Gourbal, B., Duval, D., Du Pasquier, L., Kieffer-Jaquinod, S., Mitta, G., 2010. A large repertoire of parasite epitopes matched by a large repertoire of host Immune receptors in an invertebrate host/parasite model. PLoS Negl. Trop. Dis. 4, e813.

Nair, S.V., Del Valle, H., Gross, P.S., Terwilliger, D.P., Smith, L.C., 2005. Macroarray analysis of coelomocyte gene expression in response to LPS in the sea urchin. Identification of unexpected immune diversity in an invertebrate. Physiol. Genomics 22, 33–47.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press. 333p.

Neves, G., Chess, A., 2004. Dscam-mediated self- versus non-self-recognition by individual neurons. Cold Spring Harb. Symp. Quant. Biol. 69, 485–488.

Philipp, E.E.R., Kraemer, L., Melzner, F., Poustka, A.J., Thieme, S., Findeisen, U., et al., 2012. Massively parallel RNA sequencing identifies a complex immune gene repertoire in the lophotrochozoan *Mytilus edulis*. PLoS ONE 7, e33091.

Portela, J., Duval, D., Rognon, A., Galinier, R., Boissier, J., Coustau, C., et al., 2013. Evidence for specific genotype-dependent immune priming in the Lophotrochozoan *Biomphalaria glabrata* snail. J. Innate Immun. 5, 261–276.

Schulenburg, H., Boehnisch, C., Michiels, N.K., 2007. How do invertebrates generate a highly specific innate immune response? Mol. Immunol. 44, 3338–3344.

Schulenburg, H., Hoeppner, M.P., Weiner, J., 3rd, Bornberg-Bauer, E., 2008. Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode *Caenorhabditis elegans*. Immunobiol 213, 237–250.

Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092.

Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., et al., 2011. Using the *Acropora digitifera* genome to understand coral responses to environmental change. Nature 476, 320–323.

Song, X., Zhang, H., Wang, L., Zhao, J., Mu, C., Song, L., et al., 2011. A galectin with quadruple-domain from bay scallop *Argopecten irradians* is involved in innate immune response. Dev. Comp. Immunol. 35, 592–602.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

Tasumi, S., Vasta, G.R., 2007. A galectin of unique domain organization from hemocytes of the eastern oyster (*Crassostrea virginica*) Is a receptor for the protistan parasite *Perkinsus marinus*. J. Immunol. 179, 3086–3098.

Vasta, G.R., 2009. Roles of galectins in infection. Nat. Rev. Micro. 7, 424–438.

Vasta, G.R., 2012. Galectins as pattern recognition receptors: structure, function, and evolution. Adv. Exp. Med. Biol. 946, 21–36.

Vasta, G.R., Ahmed, H., Nita-Lazar, M., Banerjee, A., Pasek, M., Shridhar, S., et al., 2012. Galectins as self/non-self recognition receptors in innate and adaptive immunity: an unresolved paradox. Front. Immunol. 3.

Watson, F.L., Puttmann-Holgado, R., Thomas, F., Lamar, D.L., Hughes, M., Kondo, M., et al., 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science 309, 1874–1878.

Watthanasurorot, A., Jiravanichpaisal, P., Liu, H., Soderhall, I., Soderhall, K., 2011. Bacteria-induced Dscam isoforms of the crustacean, *Pacifastacus leniusculus*. PLoS Pathog. 7, e1002062.

Winnebeck, E.C., Millar, C.D., Warman, G.R., 2010. Why does insect RNA look degraded? J. Insect Sci. 10, 159.
Yoshino, T.P., Dinguirard, N., Kunert, J., Hokke, C.H., 2008. Molecular and functional characterization of a tandem-repeat galectin from the freshwater snail *Biomphalaria glabrata*, intermediate host of the human blood fluke *Schistosoma mansoni*. Gene 411, 46–58.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.
Zhang, D., Jiang, S., Hu, Y., Cui, S., Guo, H., Wu, K., et al., 2011. A multidomain galectin involved in innate immune response of pearl oyster *Pinctada fucata*. Dev. Comp. Immunol. 35, 1–6.
Zhang, S.-M., Loker, E.S., 2004a. Representation of an immune responsive gene family encoding fibrinogen-related proteins in the freshwater mollusc *Biomphalaria glabrata*, an intermediate host for *Schistosoma mansoni*. Gene 341, 255–266.
Zhang, S.-M., Léonard, P.M., Adema, C.M., Loker, E.S., 2001. Parasite-responsive IgSF members in the snail *Biomphalaria glabrata*: characterization of novel genes with tandemly arranged IgSF domains and a fibrinogen domain. Immunogenetics 53, 684–694.
Zhang, S.-M., Nian, H., Zeng, Y., Dejong, R.J., 2008. Fibrinogen-bearing protein genes in the snail *Biomphalaria glabrata*: characterization of two novel genes and expression studies during ontogenesis and trematode infection. Dev. Comp. Immunol. 32, 1119–1130.
Zhang, S.M., Loker, E.S., 2003. The FREP gene family in the snail *Biomphalaria glabrata*: additional members, and evidence consistent with alternative splicing and FREP retrosequences. Fibrinogen-related proteins. Dev. Comp. Immunol. 27, 175–187.
Zhang, S.M., Loker, E.S., 2004b. Representation of an immune responsive gene family encoding fibrinogen-related proteins in the freshwater mollusc *Biomphalaria glabrata*, an intermediate host for *Schistosoma mansoni*. Gene 341, 255–266.
Zhang, S.M., Adema, C.M., Kepler, T.B., Loker, E.S., 2004. Diversification of Ig superfamily genes in an invertebrate. Science 305, 251–254.