Mol Biol Rep (2012) 39:5921–5931
DOI 10.1007/s11033-011-1404-z

Development of EST derived SSRs and SNPs as a genomic resource in Indian catfish, *Clarias batrachus*

Vindhya Mohindra · Akanksha Singh · A. S. Barman · Ratnesh Tripathi · Neeraj Sood · Kuldeep K. Lal

Received: 6 July 2011 / Accepted: 17 December 2011 / Published online: 30 December 2011

Abstract *Clarias batrachus*, an Indian catfish species, is endemic to the Indian subcontinent and potential cultivable species. The genomic resources in *C. batrachus* in the form of ESTs containing microsatellite repeats (EST-SSR) and single nucleotide polymorphisms (SNPs) that are associated with the expressed genes from spleen were mined. From a total of 1,937 ESTs generated, 1,698 unique sequences were obtained, out of which 221 EST-SSRs were identified and 54% could be functionally annotated by similarity searches. A total of 23 contigs containing 3 or more ESTs were found to contain 31 SNP loci, out of which 8 ESTs showed similarity to genes of known function and 1 for hypothetical protein. Nine ESTs with SSRs and/or SNPs identified in this study were reported to be associated with diseases in human and animals. These identified loci can be developed into markers in *C. batrachus*, which can be useful in linkage mapping, comparative genomics studies and for its genetic improvement programmes.

Keywords Indian catfish · *Clarias batrachus* · Spleen · cDNA library · Expressed sequence tags · Type I microsatellite · Single nucleotide polymorphism

Electronic supplementary material The online version of this article (doi:10.1007/s11033-011-1404-z) contains supplementary material, which is available to authorized users.

V. Mohindra (✉) · A. Singh · A. S. Barman · R. Tripathi · N. Sood · K. K. Lal
National Bureau of Fish Genetic Resources, Canal Ring Road, PO Dilkusha, Lucknow 226 002, UP, India
e-mail: vindhyamohindra@gmail.com

Introduction

Catfishes are commercially important for both the fisheries and aquaculture industry. *Clarias batrachus*, an Indian catfish species, is endemic to the Indian subcontinent and has a fairly common distribution in freshwaters of the plains throughout India [1]. It is a hardy, omnivorous, air breathing fish and well adapted to adverse ecological conditions that are fatal to most other fish. It is popular owing to its taste, medicinal and high market value. In addition, the suitability of *C. batrachus* to culture in limited space makes it a preferred and potential cultivable species [2–4]. However, its aquaculture production is limited by its slow growth rate and susceptibility to diseases. Interventions for genetic improvement in such commercially important multigenic traits depend on the availability of molecular genetic markers and use of such markers in efficient breeding programs (e.g. marker assisted selection).

The development of large genomic resources has become a prerequisite to elucidate the wide-scale evolution of genomes and the molecular basis of complex traits. High-resolution linkage maps represent a first level of integration and utilization of such resources and the primary framework for molecular analyses and mapping of quantitative trait loci (QTL) [5, 6]. Expressed sequenced tags (ESTs) represent a valuable sequence resource for research and breeding as they provide comprehensive information regarding the transcriptome, thus allowing large-scale gene expression analysis. These also facilitate breeding programs for both plants and animals by providing type I markers, simple sequence repeats (SSRs) or commonly known as microsatellites and single nucleotide polymorphisms (SNPs), as mapping type I markers directly shows the location of genes within the linkage map. The coding markers often represent genetic variations associated with economically significant

phenotypes [7]. The SSR markers, showing linkage and association with disease, can be in strong linkage disequilibrium with other functional genetic variations which truly cause the pathological phenotype. These disease-associated markers typically are represented by SNPs. They are often functionally relevant and, therefore, could be responsible for determination of the pathogenic phenotypes [8]. Although few type II microsatellite DNA markers and no SNPs have been developed, the number of markers is still insufficient for planned QTL analysis of traits such as growth or disease resistance.

To generate EST and type I markers information for *C. batrachus*, we sequenced 1,937 clones from spleen tissues of adult catfish. Spleen of fish is comparable to that of mammals and comprises the largest lymphoid tissue in teleost [9] that protects the fish against the blood borne pathogens [10]. It houses the immune cells such as T and B cells, which are able to destroy or neutralize antigens in order to prevent further manifestation of disease [11]. Therefore, a large number of immune relevant gene transcripts, involved in disease resistance, would be expected in the spleen library. Thus, the objective of the present work was to characterize type I markers from expressed sequences tags of spleen of Indian catfish species, *Clarias batrachus*, for association and linkage mapping studies for its genetic improvement programmes.

Materials and methods

Sample collection

Mature *C. batrachus* were obtained from commercial catches and acclimatized for 2 weeks before sample collection. Spleen tissues were collected from acclimatized fishes and frozen in liquid N_2 until RNA isolation.

Construction of a normalized cDNA library and generation of ESTs

A normalized cDNA library from spleens of ten individuals of *C. batrachus* was constructed directionally in plasmid vector pDNR-LIB using the Creator SMART cDNA Library Construction Kit (Clontech, Palo Alto, CA, USA). Total RNA was extracted from pooled spleen tissues using Trizol Reagent (Invitrogen, Carlsbad, CA) followed by mRNA isolation using the Oligotex mRNA Mini Kits (Qiagen, Valencia, CA, USA). First strand cDNA was prepared using the CDS-3 M adaptor, included in the TRIMMER-DIRECT kit (Evrogen, Moscow, Russia), instead of the SMART CDSIII primer. cDNA was amplified by LD-PCR according to the Creator SMART cDNA method (Clontech) using the 5′ PCR primer as the forward and reverse primer and normalized using the TRIMMER-DIRECT protocol (Evrogen). After digestion with SfiI, products smaller than 500 bp were removed using the Chroma Spin-400 column as described in the Creator SMART protocol. The resulting cDNAs were directionally cloned into the SfiI sites of pDNR-LIB (Clontech) and transformed into ElectroMAX DH10B cells (Invitrogen) by electroporation using the Gene Pulser Xcell (Bio-Rad, Hercules, CA).The clones were screened using colony PCR method and positive clones were sequenced from 5′ direction using primer pDNR.F2 [12]. The sequences obtained were cleaned using Vecscreen and separate FASTA files were generated for each EST sequence.

Clustering of ESTs into contigs and determination of gene identities

For determination of uni-genes, ESTs were clustered using the CAP3 program [13]. The linear assembly algorithm was used and the criteria for clustering were set at a minimum overlap of 30 bases (default is 20 bases). After the cluster analysis, each cluster was visually inspected to ensure fidelity of alignment to avoid pseudo-clusters caused by repetitive elements or long strings of microsatellite repeats. ESTs belonging to contigs and singletons were recorded. Redundancy number was calculated as the number of clones in the contigs divided by the number of contigs [14]. To establish the identities of ESTs, BLAST searches were conducted using BLASTN and subsequently TBLASTX searches against the non-redundant (nr) database with a BLAST cut off of 1×10^{-10} and 1×10^{-5}, respectively, to confirm gene identities. All the BLAST results were visually inspected to ensure that the matches were not due to simple amino acid stretches or repeat regions.

Identification of EST-SSRs

For estimation of the proportion of genes containing microsatellites from unigenes, identification of repeat motifs within the ESTs sequences was preformed with the program Msatfinder http://www.genomics.ceh.ac.uk/msatfinder/ [15] with all the default values. The threshold limit for di, tri, tetra and pentanucleotide repeats was 6, 5, 5 and 5, respectively.

Identification of single nucleotide polymorphism from contigs

The CAP3 assembled contigs having at least three member ESTs were manually screened for the presence of SNPs, to rule out whether the putative SNPs represent sequence errors and nature of polymorphism represented by SNPs was identified [16].

Functional annotation of ESTs

Gene ontology (GO) terms

Gene ontology (GO) annotations for consensus and singleton sequences were assigned using the program Blast2GO. Consensus and singleton sequences were submitted for GO annotation to the BLAST2GO program [17]. The BLAST2GO program uses BLAST to find homologous sequences for input sequences and extracts gene ontology (GO) terms to each hit using existing annotations. These GO terms are assigned to the query sequence to give an assessment of the biological process, the molecular function and the cellular components represented. Annotated accession numbers and GO numbers were derived with NCBI's QBLAST, with an expectation *E*-value $\leq 10^{-3}$ and an HSP length cut-off of 33. Contig sequences were then annotated according to the following parameters: a pre-*E*-value-Hit-Filter of 10^{-6}, a pro-Similarity-Hit-Filter of 15, an annotation cut-off of 55, and a GO weight of 5. Graphs were generated using a sequence filter of 5, an alpha score of 0.6 and a 0 node score filter. From these annotations, pie charts were made using 2nd level GO terms based on biological process, molecular function, and cellular component.

Kegg orthology (KO) terms

In addition, the ESTs were annotated according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) by the KEGG Automatic Annotation Server (KAAS) [18]. The query sequences are compared against the existing genes in KEGG using BLASTP for protein sequences and BLASTX and TBLASTN for nucleotide sequences. Those that were most similar to existing genes were then mapped onto the existing pathways. The sequences were analysed using the bi-directional best hit (BBH) method to obtain the KO terms for the query sequences, with a blast threshold of 40. Once genes are assigned KO identifiers or K numbers by the ortholog annotation procedure, the collective body of K numbers was mapped to BRITE functional hierarchies.

Results

Identification of ESTs containing microsatellite repeats (EST-SSRs) and distribution of microsatellite repeat types

A total of 1,937 ESTs were generated from the normalized cDNA library of *C. batrachus* and clustered into 184 contigs and 1,514 singletons (Table 1). Out of resultant 1,698 unique ESTs, a total of 221 (13.02%) unique ESTs containing microsatellites (EST-SSRs), were identified. The dinucleotide repeats were the most abundant within *C. batrachus* ESTs. These accounted for 53.8% of all microsatellite-containing ESTs, followed by 34.8%, 10.3% and 1%, respectively for tri, tetra and pentanucleotide repeats (Fig. 1). A total of 40 ESTs contained more than one type of repeats (Supplementary Table 1).

Of the dinucleotide repeats, CA/TG and GA/TC were the most abundant accounting for 88% of all dinucleotide repeats found and of these, CA/TG (55.7%) was the most abundant dinucleotide repeat type, followed by GA/TC accounting for 34.8% of all dinucleotide repeats. The AT repeats had much lower occurrence, at 9.5%, while the CG repeat was totally absent (Fig. 2).

Among the trinucleotide repeats, ATT, TTG, TAA and TGA were the most abundant accounting for 75.24% each followed by CTT, GAA, CAT and AAC (~4% each). Remaining six types of trinucleotide repeat types (CAC, GCA, CCT GTG, GCT, TAG) were found at a low level (<1.9%) (Fig. 3). Similar to the observation in dinucleotide

Table 1 Summary of analysis of ESTs containing SSRs and SNPs in *Clarias batrchus*

1.	Total number of ESTs analyzed	1,937
2.	Total number of contigs formed	184
3.	Total number of ESTs in contigs	423
4.	No of singletons	1,514
5.	Total number of Unique ESTs (2 + 4)	1,698
6.	Total number of unique ESTs containing SSRs	221 (13.02%)
7.	Contigs containing SNPs	23
8.	Total number of SNP loci identified	31

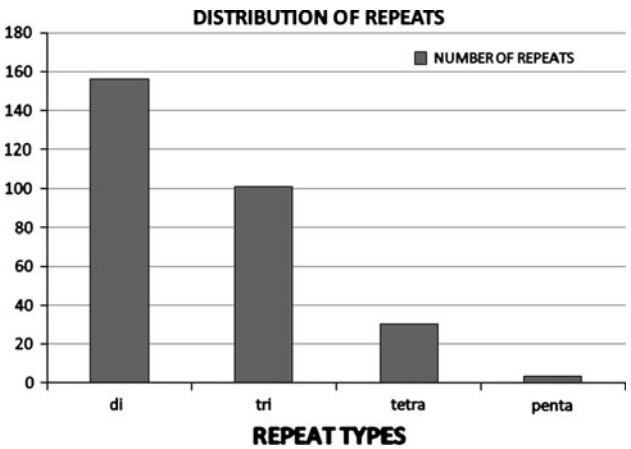

Fig. 1 Distribution of repeat types in *Clarias batrachus* microsatellite containing ESTs

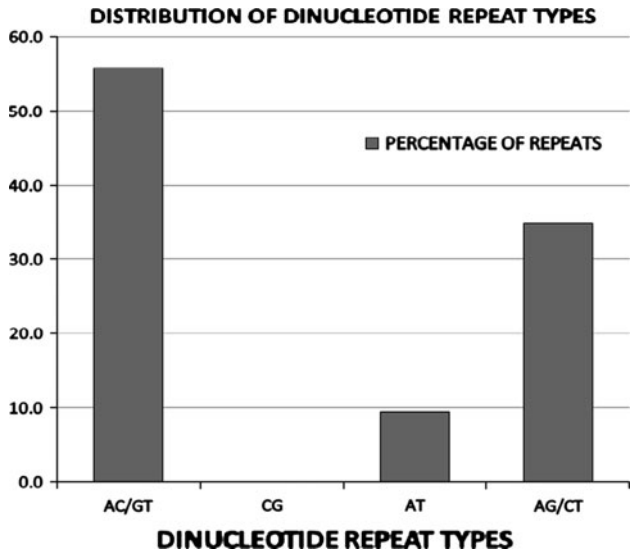

Fig. 2 Distribution of dinucleotide repeat types in *Clarias batrachus* microsatellite-containing ESTs

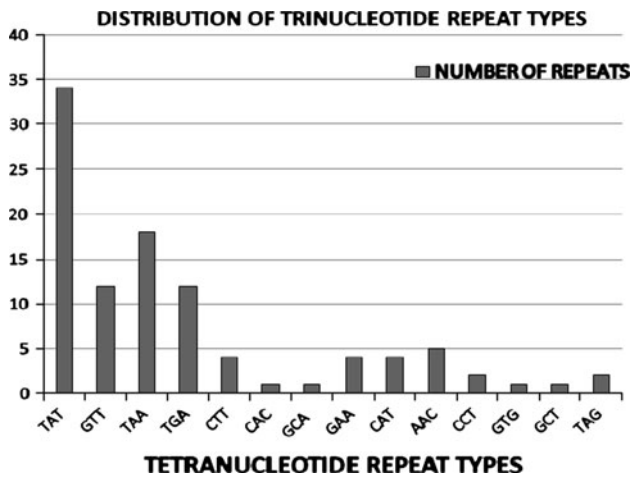

Fig. 3 Distribution of trinucleotide repeat types in *Clarias batrachus* microsatellite containing ESTs

repeats, there was an absence of GC rich repeat motifs, as CGG and CCG motif were not found.

Out of the twelve tetra-nucleotide repeats observed, GTTT, TAAA and TTCT repeat types were the most abundant (Fig. 4). All other types of tetra-nucleotide repeats (CTTC, TCTA, TATT, TTAA, TGGA, TCTG, TGTA, GTGC and AGAA) were found only once in microsatellite-containing ESTs, except AGAA that was found twice.

Functional annotation of ESTs containing microsatellites

Out of the 221 ESTs containing microsatellites, 46 had significant hits for their putative identities with BlastX (Supplementary Table 1) including 12 annotated for specific

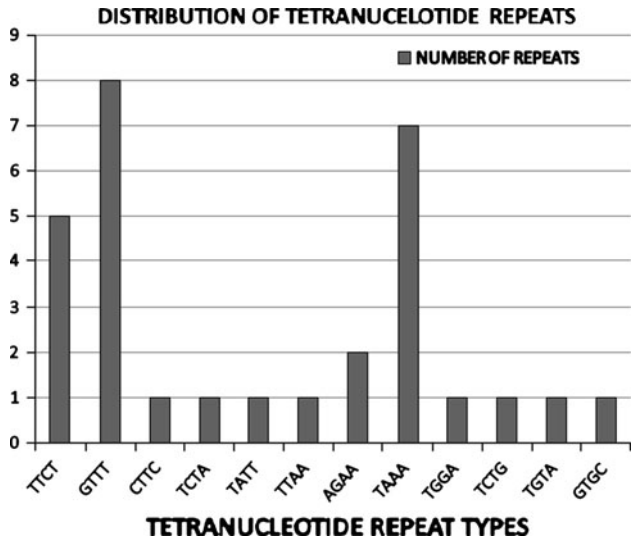

Fig. 4 Distribution of tetranucleotide repeat types in *Clarias batrachus* microsatellite-containing ESTs

functions. Additional BLASTN and TBLASTX searches against the EST database (dbEST) revealed similarity with unigene clusters for 64 ESTs including 14 ESTs annotated with functions and with the Channel catfish and blue catfish ESTs for 34 ESTs. While a significant fraction i.e. 81 ESTs (36%) could not be identified by similarity searches (Supplementary Table 1). Thus a total of 27 (12.22%) EST containing SSR sequences showed significant similarity to known protein providing product or gene names, including unassigned protein matches (hypothetical proteins) for 7 ESTs. According to the GO terms retrieved, the most abundant genes were involved in binding activity (12) under molecular function. Under biological process the most abundant were cellular process (13) followed by localization (7), biological regulation (6) and developmental processes (5). The most important genes identified were the ones under response to stimulus (3) and immune system process (3) (Fig. 5, 6, 7). KEGG pathway analysis revealed genes involved in lipid metabolism (1), genetic information processing (2) under protein export (1) and protein processing in endoplasmic reticulum (1), environmental information processing under MAPK signaling pathway (1) and cell adhesion molecules (CAMs) (1), cellular processes (1) and immune system (2) (Fig. 5). The genes mapped by GO and/or KO terms are given in Table 2 and Supplementary Table 1 and 2.

The five genes with SSRs identified under immune process from KEGG and GO analysis were WAS; Wiskott–Aldrich syndrome protein, DOCK2; dedicator of cytokinesis 2, CD163; CD163 antigen, CD62P; Endothelial P-selectin and SERP1; stress associated endoplasmic reticulum protein1. Unigenes similarity search revealed another immune relevant gene FcRI; High affinity immunoglobulin gamma Fc receptor I (Table 3).

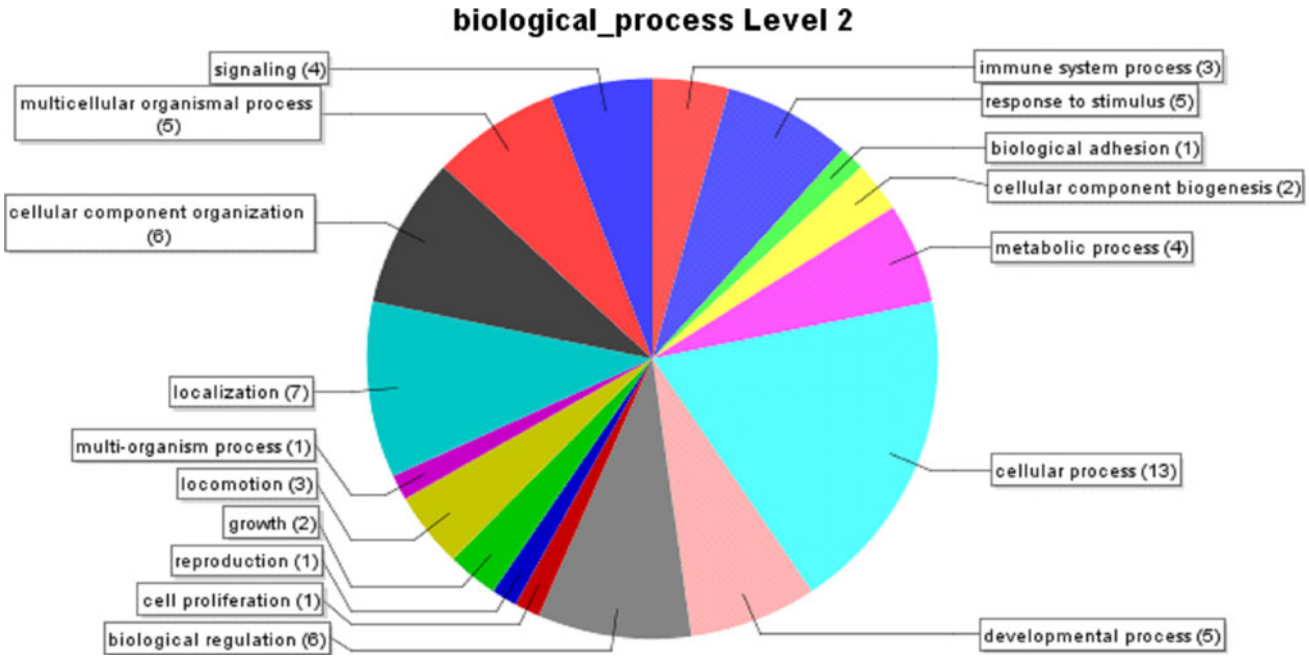

Fig. 5 Pie chart of 2nd level gene ontology (GO) terms for biological process in microsatellite containing ESTs in *Clarias batrachus*

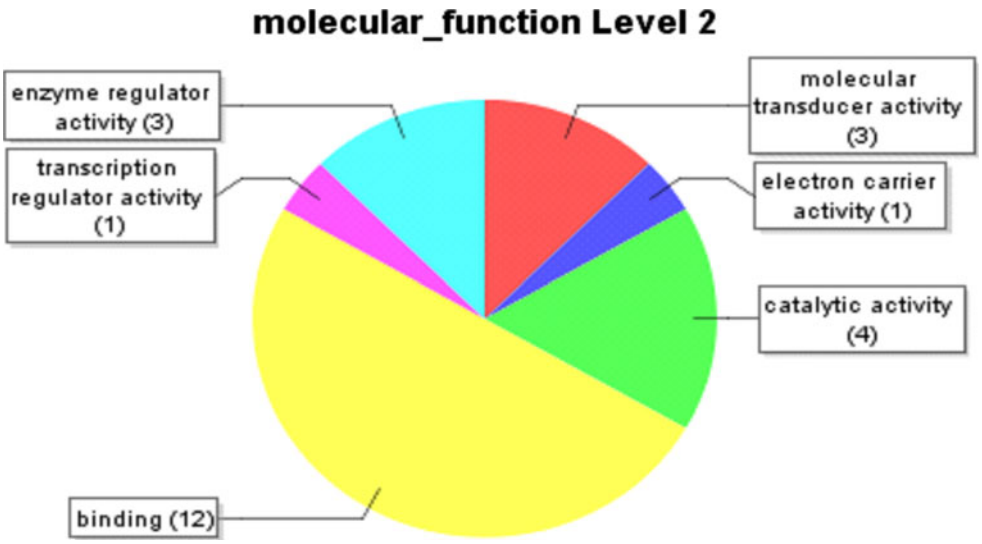

Fig. 6 Pie chart of 2nd level gene ontology (GO) terms for molecular function in microsatellite containing ESTs in *Clarias batrachus*

Single nucleotide polymorphism (SNPs)

As given above the normalized cDNA was used to construct the library, 1,937 ESTs assembled into only 184 contigs, generated by 423 ESTs. When analyzed for presence of SNPs, 23 contigs containing 3 or more ESTs were found to contain 31 SNP loci (Table 4). Transition events were observed at 24 loci and the rest 7 transversions. Search for gene identities for EST-contigs containing SNP loci revealed 8 EST-contigs showing similarity to genes of known function and 1 for hypothetical protein. Similarity to unigenes was found for 3 EST-contigs with known genes and for 7 with evidence at transcript level only, while for 4, no similarities could be established. KEGG biochemical mappings could be done for 3 EST-contigs (Supplementary Table 3).

Discussion

The present study identified large scale expressed sequence tags for development of Type I markers through identification of transcribed SSRs and SNPs in Indian catfish, *Clarias batrachus*. Type I markers are associated with genes that are conserved in a wide spectrum of species allowing gene mapping, comparative genome analysis and study of genome evolution [19]. A high proportion of microsatellite

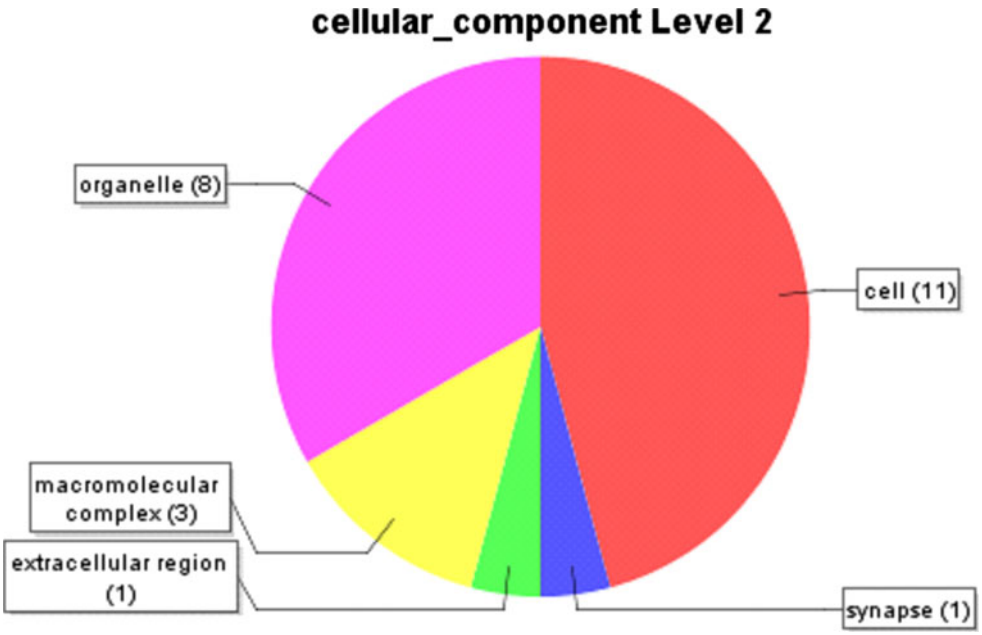

Fig. 7 Pie chart of 2nd level gene ontology (GO) terms for cellular component in microsatellite containing ESTs in *Clarias batrachus*

Table 2 Spleen ESTs containing SSRs mapped by GO and/or KO terms in *Clarias batrachus*

S. no.	GO/KO	Clone no.	Accession no.	Annotation
1	GO annotated	CbSpn1295	GW397085	tpa_inf: alsin
2		CbSpn1362	GW787373	Stress-associated endoplasmic reticulum protein 1
3		CbSpn2737	GW397126	Transposable element tcb1 transposase
4		CbSpn 2532	GR955292	Sorting nexin-25
5	KO Annotated	Contig2 (CbSpn0745, CbSpn3261)	GW397115 GW397149	RASA2, GAP1M; Ras GTPase-activating protein 2
6		CbSpn0827	GW707107	CD62E; SELE; selectin, endothelial cell
7		CbSpn0736	GW397130	orf 73
8		CbSpn3527	GW397163	Nuclear receptor coactivator partial
9		CbSpn1164	GR955287	Uncharacterized protein
10	Both	Contig3 (CbSpn2759, CbSpn2906)	GW397127 GR955348	TYK3, FER; fer (fps/fes related) tyrosine kinase
11		Contig8 (CbSpn4184, CbSpn 837)	GW397191 GW397105	WAS; Wiskott–Aldrich syndrome protein
12		CbSpn 2574	GW707077	SRP19; signal recognition particle subunit SRP19
13		CbSpn0820	GW397131	TRAM1; translocating chain- associated membrane protein 1
14		CbSpn1335	GW397132	Serine palmitoyltransferase
15		CbSpn3499	GW840489	CD163; CD163 antigen
16		CbSpn1472	GR955288	DOCK2; dedicator of cytokinesis 2

containing *C. batrachus* genes (11.3%) was observed in this study. A high level of redundant sequencing of highly expressed microsatellite-containing transcripts has been suggested to be one of the reasons for this observed trend [20]. However, the redundancy number (2.17) in *C. batrachus* (in present study) which is comparable to that of other catfishes excludes this possibility. The proportion of microsatellite containing genes observed in the study was found to be similar to that reported in Channel catfish (11.2%) [20, 21] and Japanese pufferfish (~12%) [22]; however higher than that found in Chinese shrimp (2.2%), [23]; zebrafish (6.2%) [24] and pacific Oyster [25]; and lower than black tiger shrimp (13.7%) [26].

In *C. batrachus*, the dinucleotide repeats were the dominant repeat type (53.8%) in microsatellite containing ESTs. Dominance of AT rich repeats has been observed in

Table 3 Genes with SSRs identified under immune process from KEGG and GO analysis

SNo.	Genes	Clone name	Accession no.	Repeat	Accession of closest homology
1	WAS; Wiskott–Aldrich syndrome protein,	CbSpn4184	GW397191	$(AC)_6$	NP_956232.2
2	DOCK2; dedicator of cytokinesis 2,	CbSpn1472	GR955288	$(AG)_9$	XP_002664309.1
3	CD163; CD163 antigen	CbSpn3499	GW397160	$(TTG)_{11}$	XP_688939.3
4	CD62P; Endothelial P-selectin	CbSpn0827	GW707107	$(TTTC)_{10}$ $(TCTT)_9$	XP_001336824.2
5	SERP1; stress associated endoplasmic reticulum protein1	CbSpn1362	GW787373	$(ATC)_6$	BAE38513.1
6	FcRI; High affinity immunoglobulin gamma Fc receptor I	CbSpn4566	GW840553	$(AATA)_6$	CK411091.1

di- and trinucleotide repeats similar to that in Channel catfish genome and most aquaculture species [20, 25, 26]. AC/GT repeat types were most common in microsatellite containing ESTs of Indian catfish in accordance to that found in animal genomes [27, 28], thus pointing to the AT rich nature of *C. batrachus* genome. And total absence of CG repeat motifs from the microsatellite containing ESTs further emphasizes the Indian catfish *C. batrachus* genome to be AT rich. It may be due to the fact that these CG repeats may contain highly mutable CpG dinucleotides within and additionally, long CCG repeats may interfere with the efficiency and accuracy of splicing [28]. Moreover, there was also no occurrence of CGG and CCG motifs, further emphasizing the AT rich nature of the Indian catfish genome.

In the present study, only 12.22% ESTs having SSRs could be annotated to known protein providing product or gene names. This could be related to nucleotide sequences corresponding to the 5′ or 3′ untranslated (UTR) region [29]. Repeats have been reported to be more commonly present in the 5′ and 3′ UTR regions as repeat variation within the coding segment would affect the normal gene activity and can cause phenotypic changes [30] as the evolutionary constraint rates, within gene-coding sequences, are lower than those in non-coding genomic sequences. However, the unidentified transcripts are still valuable sources of microsatellite markers, and can be further sequenced if determined to be important in QTL analysis or expression profiling with microarray. Additionally, many of these currently unknown transcripts will likely be identified when they cluster with additional transcripts produced in the future. In the present study, polymorphism could not correlated to the number of microsatellite repeats, since repeats of lower dinucleotides repeat units were polymorphic while longer ones were not [23].

A 12.22% EST sequences showed significant similarity to known protein providing product or gene names and the remaining 7.8% showed no similarities to any proteins or unclassified with GO identifier. This could be due to the sequences being too short, incomplete or are novel proteins of the known database. In addition in present studies, 2.71% of genes with SSRs were identified under immune process.

Genes with repeat motifs (SSRs) and SNPs identified associated with diseases

Fishes (zebra fish, a model species) form a relevant vertebrate system for modeling human cancer, displaying many similarities in tumorigenic pathways, as genes relevant to cancers, homologous to those found in humans and other mammals, have been reported in fish [31]. The functional genes with the type I markers, EST-SSRs and SNPs, identified in this study, have been reported earlier including in human, as markers for many important diseases and disorders. Microsatellite markers, found in FcRi gene in the chromosome 11q13 region might prove to be a molecular marker for loci regulating physiological traits closely associated with asthma in human [32]. In EST-SSRs located at the 3′ UTR region of the *Slc11a1* bovine solute carrier family 11 a1 (*Slc11a1*) gene along with ARO28 situated about 0.6 cM upstream of the same gene was used for typing the 34 European, 18 Asian, 20 Creole and 23 hybrid bovines for polymorphisms, as allelic variants of several genes have been implicated in the genetic susceptibility to tuberculosis in some human populations. High level of diversity and heterozygosity was found in most of the cattle surveyed except the Europeans bovines and especially Holsteins in relation to the 3′ UTR microsatellite locus [33].

Nucleotide changes within the NACHT domain of other NLRP proteins have been associated with hereditary fever syndromes and chronic inflammatory diseases. A single nucleotide polymorphism within the NACHT domain of NLRP2 has been reported to contribute to the amplification of inflammatory responses due to a reduction of inhibitory signals on the NF-kappa B pathway [34]. Casabonne et al. [35] reported a strong correlation between high risk of chronic lymphocytic leukemia in patients with SNP in MMP9 gene and were at highest risk of this disease. The

Table 4 SNP loci observed in *Clarias batrachus spleen* ESTs

S. no.	Contig no.	Contig members	Accession no.	No of member ESTs	Annotations	Position; nature of SNP
1	5	CbSpn0091	GW774926	3	Novel NACHT domain containing protein (Danio rerio)	384; C ↔ T
		CbSpn2822	GW672553			677; G ↔ A
		CbSpn2493	GW707075			
2	13	CbSpn0295	GW774932	3	Homo sapiens INO80 complex subunit D (INO80D)	733; G ↔ A
		CbSpn0576	GW672506			742; C ↔ T
		CbSpn1672	GW707011			
3	14	CbSpn0308	GW492626	3	Unknown	58; C ↔ T
		CbSpn0206	GW492724			
		CbSpn1888	GW672559			
4	16	CbSpn0367	GW397111	3	Yippee-like 5	166; C ↔ T
		CbSpn0072	GW787317			340; G ↔ A
		CbSpn550	GW397112			
5	30	CbSpn0813	GW787346	3	Expressed sequence Ipu.18865	843; G ↔ A
		CbSpn4490	GW840404			
		CbSpn1122	GW672452			
6	38	CbSpn1040	GW774963	4	Expressed sequence Ipu.2829	510; C ↔ T
		CbSpn0229	GW492680			
		CbSpn1754	GW707032			
		CbSpn0389	GW492738			
7	41	CbSpn1090	GW774966	5	Expressed sequence Ipu.28970	469; C ↔ T
		CbSpn4481	GW840401			
		CbSpn3025	GW775064			
		CbSpn3721	GW836363			
		CbSpn0429	GW774939			
8	49	CbSpn1246	GW787365	3	Unknown	512; G ↔ T
		CbSpn2335	GW492686			
		CbSpn3362	GW836258			
9	57	CbSpn1351	GW774973	3	Transcribed locus, moderately similar to NP_997824.1 DnaJ (Hsp40) homolog, subfamily B, member 12 (Danio rerio)	598; G ↔ A
		CbSpn3989	GW836441			
		CbSpn3538	GW836306			
10	60	CbSpn1483	GT157710	3	Gamma-aminobutyric acid receptor-associated protein like 1	292; C ↔ T
		CbSpn2868	GT271593			431; C ↔ T
		CbSpn3614	GW836328			
11	85	CbSpn2261	GW707062	3	Cytoplasmic dynein 1 heavy chain 1 [Danio rerio]	591; T ↔ A
		CbSpn1722	GT271624			
		CbSpn2465	GT157731			
12	91	CbSpn2405	GW672543	3	Unknown	346; G ↔ T
		CbSpn2437	GW672544			
		CbSpn2152	GW672566			
13	100	CbSpn2575	GW707078	3	Danio rerio zgc: 153976 Hypothetical protein LOC777625	309; C ↔ T
		CbSpn2089	GW775045			553; C ↔ T
		CbSpn429	GW774939			
14	110	CbSpn2791	GW707088	3	Expressed sequence Ipu.2031 transcribed locus, moderately similar to NP_571677.1 death effector domain-containing 1 (Danio rerio)	495; C ↔ T
		CbSpn2585	GW397102			
		CbSpn4628	GW840581			

Table 4 continued

S. no.	Contig no.	Contig members	Accession no.	No of member ESTs	Annotations	Position; nature of SNP
15	111	CbSpn2793	GW787452	3	Expressed sequence Ipu.22308	627; G ↔ A
		CbSpn2183	GW707058			
		CbSpn1494	GR955330			
16	116	CbSpn2843	GW787455	3	Homo sapiens at rich interactive domain 4a (rbp1-like) transcript variant mrna	646; A ↔ T
		CbSpn1914	GW774992			
		CbSpn2935	GW836205			
17	121	CbSpn2958	GW706954	3	Unknown	377; G ↔ A
		CbSpn4782	GW840632			365; C ↔ T
		CbSpn2516	GW672578			
18	136	CbSpn3268	GW836231	3	Expressed sequence Xl.83249 transcribed locus, strongly similar to NP_001135634.1 forkhead box K2 (Xenopus (Silurana) tropicalis)	660; G ↔ T
		CbSpn2044	GW840437			
		CbSpn2041	GW787414			
19	160	CbSpn4127	GW836477	3	Matrix metalloproteinase-9 [Ictalurus punctatus]	409; C ↔ T
		CbSpn1516	GW840427			
		CbSpn0376	GW840408			
20	163	CbSpn4184	GW397191	3	Wiskott–Aldrich syndrome (eczema-thrombocytopenia)	419; C ↔ T
		CbSpn0837	GW397105			
		CbSpn1987	GW672526			
21	165	CbSpn4235	GW840339	3	Protein tyrosine phosphatase type IVA, member 1 (ptp4a1)	470; C ↔ T
		CbSpn5109	GW840712			
		CbSpn0800	GT157699			
22	169	CbSpn4433	GT157699	3	Expressed sequence Ipu.8352	292; C ↔ G
		CbSpn2163	GW775081			582; C ↔ T
		CbSpn2167	GW707056			
23	181	CbSpn4985	GW840684	9	Eukaryotic initiation factor 4a-iii	212; A ↔ T
		CbSpn1236	GW840422			305; G ↔ A
		CbSpn1961	GW840435			
		CbSpn4005	GW840509			
		CbSpn2425	GW840448			
		CbSpn5085	GW840703			
		CbSpn1428	GW836115			
		CbSpn1131	GW836146			
		CbSpn4022	GW840510			

human homolog of yippee-like 2 (YPEL2) along DEAH (Asp-Glu-Ala-His) box polypeptide 40 (DDX40) were found to be detected as two flanking positional candidate genes with an intragenic SNP in rs2572886 region of locus HSA8q24.3 and this SNP was found to be positively correlated with cellular susceptibility to HIV-1 [36]. Pharmacological evidence suggests the involvement of polymorphisms (SNPs) in gamma-aminobutyric acid (GABA) displayed significant associations with mood disorders [37] as well as with schizophrenia [38]. Mutations in an intermediate chain dynein (*DNAI1*; *IC78*) have been described in primary ciliary dyskinesia patients, with outer dynein arm (ODA) defects. Primary ciliary dyskinesia (PCD) is a genetically heterogeneous, autosomal recessive disorder caused by abnormal ciliary ultrastructure and function, characterized clinically by oto-sino-pulmonary disease [39]. The SNPs found in the genes of death domain (DD) superfamily, composed of the DD, death effector domain (DED) and caspase recruitment domain (CARD) families of proteins, plays a pivotal role in signaling events that regulate apoptosis and in turn of innate immunity and inflammation [40]. Wiskott–Aldrich syndrome, WAS, an X-linked recessive immunodeficiency disorder is caused by mutations in the WAS gene. The mutation includes base deletion that

produces a frame shift and premature termination of translation and/or point mutation that change the same arginine residue to either a histidine or a leucine in WAS patients [41]. Thus, type I markers based screening strategies can be used in the fields of veterinary and medical parasitology and for molecular studies of infectious diseases. This includes mapping and further identification of genes responsible for resistance to parasites and pathogens and the identification of genes controlling drug resistance in pathogenic organisms [7].

Conclusions

In summary, cDNA library was constructed from the *C. batrachus* spleen for development of expressed sequence tags and type I markers, EST-SSRs and SNPs were characterized, which represent functional genes. Thus the identified EST-SSR and SNP loci that hit with disease-related and other important genes in GenBank, are believed to be beneficial in the development of EST-SSR and SNP markers, which can be useful in linkage mapping, comparative genomics studies and for its genetic improvement programmes.

Acknowledgments Financial support received from the Department of Biotechnology, Government of India, for this study is thankfully acknowledged. Excellent technical assistance provided by Mr. R. S. Sah, Mr. Rajesh Kumar and Mr. Sree Ram is duly acknowledged.

References

1. Pouyaud L, Sudarto Paradis E (2009) The phylogenetic structure of habitat shift and morphological convergence in Asian *Clarias* (Teleostei, Siluriformes: Clariidae). J Zool Sys Evol Res 47: 344–356
2. Chonder SL (1999) Biology of fin fishes and shellfishes. SCSC Publishers, Howrah
3. Sahoo SK, Giri SS, Sahoo AK (2004) Effect of stocking size of *Clarias batrachus* fry on growth and survival during fingerling hatchery production. Asian Fish Sci 17:229–233
4. Hossain Q, Hossain LA, Parween S (2006) Artificial breeding of *Clarias batrachus* (Linnaeus, 1758). Scientific World 4:32–37
5. Liu ZJ, Cordes JF (2004) DNA marker technology and their applications in aquaculture genetics. Aquaculture 238:1–37
6. Teh SL, Chan WS, Abdullah JO, Namasivayam P (2011) Development of expressed sequence tag resources for Vanda Mimi Palmer and data mining for EST-SSR. Mol Biol Rep 38:3903–3909
7. Chistiakov DA, Hellemans B, Volckaert FAM (2006) Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. Aquaculture 255:1–29
8. Schork NJ, Fallin D, Lanchbury JS (2000) Single nucleotide polymorphisms and the future of genetic epidemiology. Clin Genet 58:250–264
9. Press CM, Evensen O (1999) The morphology of the immune system in teleost fishes. Fish Shellfish Immunol 9:309–318
10. Zapata AG, Chibá A, Varas A (1996) Cells and tissues of the immune system of fish. In: Iwama G, Nakanishi T (eds) The fish immune system: organism, pathogen and environment. Academic Press, New York, pp 1–62
11. Kaattari SL, Piganelli JD (1996) The specific immune system: humoral defense. In: Iwama G, Nakanishi T (eds) The fish immune system: organism, pathogen and environment. Academic Press, New York, pp 207–254
12. Douglas SE, Knickle LC, Kimball J, Reith ME (2007) Comprehensive EST analysis of Atlantic halibut (*Hippoglossus hippoglossus*), a commercially relevant aquaculture species. BMC Genom 8:144–155
13. Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877
14. Kocabas AM, Li P, Cao D, Karsi A, He C, Patterson A, Ju Z, Dunham RA, Liu Z (2002) Expression profile of the channel catfish spleen: analysis of genes involved in immune functions. Mar Biotechnol (NY) 4:526–536
15. Thurston MI, Field D (2005) Msatfinder: detection and characterisation of microsatellites. Distributed by the authors at http://www.genomics.ceh.ac.uk/msatfinder/. CEH Oxford, Mansfield Road, Oxford OX1 3SR
16. Hubert S, Bussey JT, Higgins B, Curtis BA, Bowman S (2009) Development of single nucleotide polymorphism markers for Atlantic cod (*Gadus morhua*) using expressed sequences. Aquaculture 296:7–14
17. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435
18. Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–W185. http://www.genome.jp/tools/kaas/
19. Liu ZJ, Tan G, Kucuktas H, Li P, Karsi A, Yant DR, Dunham RA (1999) High levels of conservation at microsatellite loci among Ictalurid catfishes. J Hered 90:307–312
20. Serapion J, Kucuktas H, Feng JN, Liu ZJ (2004) Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol (NY) 6:364–377
21. Liu ZJ, Tan G, Li P, Dunham RA (1999) Transcribed dinucleotide microsatellites and their associated genes from channel catfish, *Ictalurus punctatus*. Biochem Biophys Res Commun 259: 190–194
22. Edwards YJ, Elgar G, Clark MS, Bishop MJ (1998) The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses. J Mol Biol 278: 843–854
23. Wang HX, Li FH, Xiang JH (2005) Polymorphic EST-SSR markers and their mode of inheritance in *Fenneropenaeus chinensis*. Aquaculture 249:107–114
24. Maneeruttanarungroj C, Pongsomboon S, Wuthisuthimethavee S, Klinbunga S, Wilson KJ, Swan J, Li Y, Whan V, Chu KH, Li CP, Tong J, Glenn K, Rothschild M, Jerry D, Tassanakajon A (2006) Development of polymorphic expressed sequence tag derived microsatellites for the extension of the genetic linkage map of the black tiger shrimp (*Penaeus monodon*). Anim Genet 37:363–368
25. Yu H, Li Q (2008) Exploiting EST databases for the development and characterization of EST-SSRs in the Pacific oyster (*Crassostrea gigas*). J Hered 99:208–214
26. Tassanakajon A, Klinbunga S, Paunglarp N, Rimphanitchayakit V, Udomkit A, Jitrapakdee S, Sritunyalucksana K, Phongdara A, Pongsomboon S, Supungul P, Tang S, Kuphanumart K, Pichyangkura R, Lursinsap C (2006) *Penaeus monodon* gene discovery

project: the generation of an EST collection and establishment of a database. Gene 384:104–112
27. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200
28. Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967–981
29. Cerdà J, Mercadé J, Lozano JJ, Manchado M, Tingaud-Sequeira A, Astola A, Infante C, Halm S, Viñas J, Castellana B, Asensio E, Cañavate P, Martínez-Rodríguez G, Piferrer F, Planas JV, Prat F, Yúfera M, Durany O, Subirada F, Rosell E, Maes T (2008) Genomic resources for a commercial flatfish, the Senegalese sole (*Solea senegalensis*): EST sequencing, oligo microarray design, and development of the Soleamold bioinformatic platform. BMC Genom 9:508–521
30. Liu ZJ, Li P, Dunham R (1998) Characterization of an A/T-rich family of sequences from the channel catfish (*Ictalurus punctatus*). Mol Mar Biol Biotechnol 7:232–239
31. Rubinstein AL (2003) Zebrafish: from disease modeling to drug discovery. Curr Opin Drug Discov Devel 6:218–223
32. Palmer LJ, Daniels SE, Rye PJ, Gibson NA, Tay GK, Cookson WO, Goldblatt J, Burton PR, LeSöuef PN (1998) Linkage of chromosome 5q and 11q gene markers to asthma-associated quantitative traits in Australian children. Am J Respir Crit Care Med 158:1825–1830
33. Vázquez-Flores F, Alonso R, Villegas-Sepúlveda N, Arriaga C, Pereira-Suárez AL, Mancilla R, Estrada-Chávez C (2006) A microsatellite study of bovine solute carrier family 11 a1 (*Slc11a1*) gene diversity in Mexico in relation to bovine tuberculosis. Genet Mol Biol 29:503–507
34. Fontalba A, Gutierrez O, Fernandez-Luna JL (2007) NLRP2, an inhibitor of the NF-kappaB pathway, is transcriptionally activated by NF-kappaB and exhibits a nonfunctional allelic variant. J Immunol 179:8519–8524
35. Casabonne D, Reina O, Benavente Y, Becker N, Maynadié M, Foretová L, Cocco P, González-Neira A, Nieters A, Boffetta P, Middeldorp JM, de Sanjose S (2011) Single nucleotide polymorphisms of matrix metalloproteinase 9 (MMP9) and tumor protein 73 (TP73) interact with Epstein-Barr virus in chronic lymphocytic leukemia: results from the European case-control study EpiLymph. Haematologica 96:323–327
36. Loeuillet C, Deutsch S, Ciuffi A, Robyr D, Taffé P, Muñoz M, Beckmann JS, Antonarakis SE, Telenti A (2008) In vitro whole-genome analysis identifies a susceptibility locus for HIV-1. PLoS Biol 6:319–327
37. Yamada K, Watanabe A, Iwayama-Shigeno Y, Yoshikawa T (2003) Evidence of association between gamma-aminobutyric acid type A receptor genes located on 5q34 and female patients with mood disorders. Neurosci Lett 349:9–12
38. Chen J, Tsang SY, Zhao CY, Pun FW, Yu Z, Mei L, Lo WS, Fang S, Liu H, Stöber G, Xue H (2009) GABRB2 in schizophrenia and bipolar disorder: disease association, gene expression and clinical correlations. Biochem Soc Trans 37:1415–1418
39. Zariwala M, Noone PG, Sannuti A, Minnix S, Zhou Z, Leigh MW, Hazucha M, Carson JL, Knowles MR (2001) Germline mutations in an intermediate chain dynein cause primary ciliary dyskinesia. Am J Respir Cell Mol Biol 25:577–583
40. Kersse K, Vanden Berghe T, Lamkanfi M, Vandenabeele P (2007) A phylogenetic and functional overview of inflammatory caspases and caspase-1-related CARD-only proteins. Biochem Soc Trans 35:1508–1511
41. Derry JM, Ochs HD, Francke U (1994) Isolation of a novel gene mutated in Wiskott–Aldrich syndrome. Cell 26:635–644