


SNP discovery from liver transcriptome in the fish *Piaractus mesopotamicus*

Vito Antonio Mastrochirico-Filho¹ · Milene Elissa Hata¹ · Lucas Seiti Sato¹ · Paulo Henrique Jorge¹ · Fausto Foresti² · Manuel Vera Rodriguez³ · Paulino Martínez³ · Fábio Porto-Foresti⁴ · Diogo Teruo Hashimoto¹ 

Received: 1 February 2016 / Accepted: 3 February 2016 / Published online: 9 February 2016
© Springer Science+Business Media Dordrecht 2016

Abstract Pacu (*Piaractus mesopotamicus*) is a Neotropical freshwater fish threatened by overfishing, and one of the species of highest commercial value for aquaculture. Genetic variability analysis through molecular markers is an essential technique to genetic management of the wild and cultivated stocks. The main objective of this study was to identify and validate gene-associated SNPs of the liver transcriptome of pacu. Through genetic analysis in one natural population (Paraná River), 32 polymorphic SNPs were successfully genotyped and validated, some of them related to immune system genes. The observed and expected heterozygosity ranged from 0.059 to 0.706 and 0.058 to 0.507, respectively. All loci were in Hardy–Weinberg equilibrium ($P > 0.05$). Our results showed useful genomic resources for pacu, with applicability in conservation purposes and aquaculture industry by using SNPs markers.

Keywords NGS · Pacu · Aquaculture · Genetic variability

Pacu (*Piaractus mesopotamicus*) is a Neotropical freshwater fish widely distributed in floodplain areas of the La

Plata Basin. Wild populations of pacu are threatened by overfishing (Resende 2003), particularly in Brazil (in São Paulo State, according to the Decree n° 56.031, SSP, 2010), since this species is considered of high commercial value, with large-scale catches by the industrial and recreational fisheries (MPA 2011; IBGE 2014). Furthermore, this non-model fish is very important for aquaculture and represents one of the most cultivated species in Brazil, and in aquaculture from other countries in South America (Colombia, Peru, Venezuela and Argentina) and Asia (China, Myanmar, Thailand and Vietnam) (Flores Nava 2007; Honglang 2007; FAO 2010).

Genetic studies directed to this species are still insufficient and limited in few microsatellites and mitochondrial sequences (Calcagnotto and DeSalle 2009; Iervolino et al. 2010). SNPs (*Single Nucleotide Polymorphisms*) are caused by point mutations distributed throughout the genome and they have been frequently used in fish genetic studies (Vera et al. 2013; Zhang et al. 2015; Liu et al. 2016). Moreover, this marker is considered the most adaptable to automation genotyping and able to reveal hidden polymorphisms not detected in others molecular markers (Liu and Cordes 2004). Therefore, due to environmental concerns and economic importance to aquaculture production, the purpose of this study was to develop SNPs marker by liver transcriptome sequencing in pacu, which may provide a better understanding of population structure of this species and the first base information about economically relevant traits for future molecular assisted breeding.

To perform the transcriptome sequencing, samples were collected from liver of individuals (5 juveniles and 5 adults) from three different fish farms and in one wild population, and sequenced on Roche/454 pyrosequencing platform. We obtained 212,545 trimmed reads which were

✉ Diogo Teruo Hashimoto
diogo@caunesp.unesp.br

¹ Centro de Aquicultura da Unesp, Universidade Estadual Paulista, Jaboticabal, SP, Brazil

² Departamento de Morfologia, IBB, Universidade Estadual Paulista, Botucatu, SP, Brazil

³ Departamento de Genética, Facultad de Veterinaria, Universidad de Santiago de Compostela, Lugo, Spain

⁴ Departamento de Ciências Biológicas, FC, Universidade Estadual Paulista, Bauru, SP, Brazil

deposited in Short Read Archive (SRA) of NCBI, under the accession number SRA312243. De novo assembly strategy was performed using CLC Genomics Workbench (version 7.5.1; CLC bio; Aarhus; Denmark) and yielded 4110 assembled transcripts, as a result of 193,247 reads overlapped (71,581,413 bp) and N50 of 871 bp. Functional annotation was attributed to assembled transcripts by BLASTX (cutoff *e*-value 1E-3) through homology searches against NCBI nonredundant protein database, using Blast2Go software. A total of 2051 sequences was identified with the highest homology to *Astyanax mexicanus* sequences. We found 1665 annotated sequences with Interpro accession numbers and 1773 sequences classified in gene ontology (GO) terms, where cellular and metabolic process were highly represented (Fig. 1). Meanwhile, we have still identified 619 enzyme code numbers from KEGG mapping results which revealed a total of 1023 transcripts mapped to 111 different enzyme pathways, being the most frequent related to purine metabolism, thiamine metabolism and biosynthesis of antibiotics.

SNP calling was performed using CLC Genomics Workbench (version 7.5.1; CLC bio; Aarhus; Denmark). In total, 802 putative SNPs were found in 229 transcripts (Table 1). Abundant and repetitive SNPs in small areas were excluded. After the filtering step, we identify 464 SNPs. For gene location of the SNPs, ORF (*Open Reading Frame*) regions were found by comparing the sequences against the NCBI protein database (cutoff *evaluate* 1E-10):

Table 1 Putative SNPs identification parameters, through de novo assembly of the pacu (*Piaractus mesopotamicus*) transcriptome

SNP features	Value
Contig with SNPs	229
Contig average length (min–max) (bp)	1461.8 (209–5150)
SNP amount	802
SNP per kilobase	2.4
Average coverage of reads (min–max)	54.8x (15–1619x)
Transitions	
A-G	301 (37.5 %)
C-T	267 (33.3 %)
Transversions	
G-T	72 (9.0 %)
A-C	61 (7.6 %)
A-T	52 (6.5 %)
C-G	49 (6.1 %)
Transition:transversion ratio	2.43

330 (71.1 %) SNPs were distributed in coding regions (cds), being 233 synonymous (50.2 %) and 97 non-synonymous (20.9 %); 80 SNPs (17.2 %) were located in 3'UTR and 46 (10.0 %) in 5'UTR.

50 SNP loci (2 Multiplex) were selected to perform validation and genotyping analysis with the Sequenom MassARRAY platform, in CeGen (Genotyping National Center, Santiago de Compostela, Spain). The technique is increasingly used in the genotyping of SNPs in fish

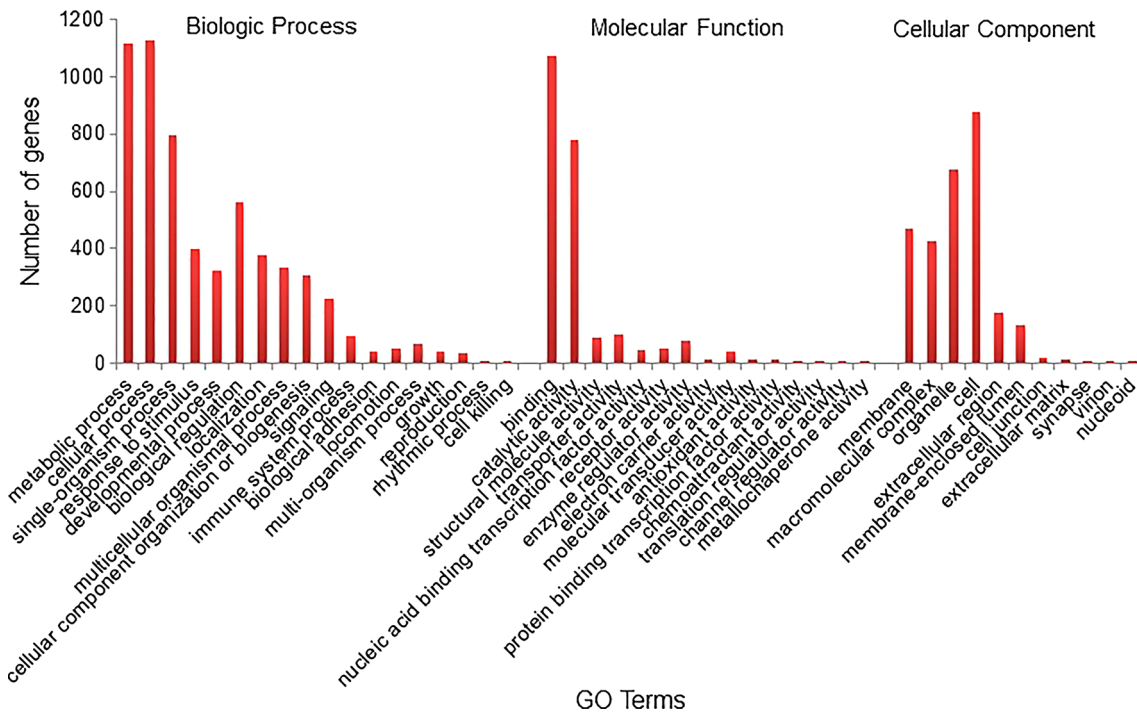


Fig. 1 Gene ontology categories of *Piaractus mesopotamicus* sequences

Table 2 Variants and diversity values of the 32 feasible SNPs of pacu (*Piaractus mesopotamicus*) from Paraná River population, which corresponds to 34 individuals

SNP_ID	iPLEX	PCR primers (F and R) and extension primer (E)	Allele	MAF	HW	H _{obs}	H _{exp}	Fis	Gene Description
C4_231	M1	F:GCAGCCAGATGCCAAAGTTC R:AGGCTTCTGGAGAGATTGTG E:cctTGGCCTGGTTCATGAGAAAGTCTCC	T/C	0.324	1.000	0.471	0.444	-0.0602	Unnamed protein product
C1459_108	M1	F:ACCAITGAACATCAGGCCAC R:CAGCTGTGTCTATGAGAACC E:caCCTTCAGATTGAAAACCTTTCGAG	T/G	0.344	0.0045	0.688	0.458	-0.512	Unnamed protein product
C30_132	M1	F:TCAGCACCCAGCACCTG R:ACATCGACTTCTCCCTGCG E:CCGTGTGAAGAGGAAAGAA	G/A	0.176	0.249	0.235	0.295	0.205	Ribosomal protein S9
C1013_445	M1	F:ATACAACAACAGCCGCTCCC R:TTTAGCTGCAGGAAAGCATC E:GGAAAAGCATCAAAACGAG	C/T	0.426	0.289	0.382	0.496	0.233	Vitronectin-like
C857_201	M1	F:AAGAGTGGACTGTGGCTTG R:GAGACACAGTGAAGTCAGAG E:tCCCCCTTCAGGACAAAAC	A/G	0.279	0.233	0.324	0.409	0.211	Inter-alpha-trypsin inhibitor heavy chain H3-like
C83_761	M1	F:GACACAGAACAGGATTAGTC R:CATCCGTCTGATCAGTCAAC E:cttaCAAAACACTAAATCGACCCCA	G/A	0.191	0.570	0.265	0.314	0.159	Glutathione peroxidase 3 precursor
C627_936	M1	F:TCTGAGGGTGAGAACTATG R:GTAATAATAACTACAGATAC E:AATAACTACAGATACAATAAGAATTAG	A/T	0.397	0.285	0.382	0.486	0.216	Glutathione S-transferase A-like
C579_153	M1	F:CAAGAAGATCGAAGCCAGAC R:TCTTGGCCCTTCATGACCTTG E:ggatAAGTCGTTCAITTTGGC	C/T	0.235	0.152	0.471	0.365	-0.294	60S ribosomal protein L14-like
C147_351	M1	F:TTCCTACTCCAGACGTTACC R:GAATTAGTCTGCACCTGTGTC E:cgggaAACACACGGCGCAGGCAAGTT	G/A	0.162	0.562	0.324	0.275	-0.179	Vitellogenesis membrane outer layer protein 1 homolog
C178_243	M1	F:TTGGATTCAAGGAAGGCGAG R:GAGAACATGTTCTGCACACAGG E:cccCCCAGCAAATTTCCAGGA	T/C	0.162	0.562	0.324	0.275	-0.179	Haptoglobin
C458_2209	M1	F:TCTCGGTTTTTCCGCTCG R:ACGTACAGGGAGGCCATC E:cctctCCGGGATTCAATCTCCGAATT	C/T	0.044	1.000	0.088	0.086	-0.0313	Polyadenylate-binding protein 1-like

Table 2 continued

SNP_ID	iPLEX	PCR primers (F and R) and extension primer (E)	Allele	MAF	HW	H _{obs}	H _{exp}	F _{is}	Gene Description
C213_629	M1	F:CATGTCCACAAACTGTCCTG R:AGTTCGTGATTCACCTCAG E:gcAGAGTTCTGCAAGCTGGACG	C/T	0.471	0.165	0.647	0.506	-0.285	Prostaglandin-H2 D-isomerase-like
C238_1041	M1	F:AAACTGAACCACTCTGGAG R:TCATTTTGGACACCTCAC E:TGGACACCCCTCACTTGTAATG	A/T	0.044	1.000	0.088	0.086	-0.0313	Coagulation factor V-like
C239_1594	M1	F:TCCCAAAGAACATAAAAGC R:GTTTGGCAGCATGGGATTTG E:atgCTTGATGCTAACTGCAGTG	C/A	0.339	0.706	0.500	0.454	-0.102	Basigin isoform X1
C240_1549	M1	F:TGCGAAGAGACACAATTCCC R:GTATCTTATTGCTATGGCT E:cGCTATGGCTTATGAACAG	T/C	0.206	0.599	0.294	0.332	0.115	Heme oxygenase-like
C260_818	M1	F:CACTTTGAACAGAGGGCAC R:AGACGAGTTCTACTGTGTGG E:TGTGGGCTTCCGAAAT	A/G	0.471	0.0363	0.706	0.506	-0.404	Complement C5-like
C271_399	M1	F:ATATTAGGCAAGCGGCTAAG R:CATGGCCGAATACCTGTTTG E:AGCCCTTTAAACCC	T/A	0.029	1.000	0.059	0.058	-0.0154	Ferritin, heavy subunit-like
C348_245	M1	F:ATCATCTTTGTGCCTGTCC R:TGCCGCTGAATTTCTCTCC E:cccaatTCGGGCACAAGCCGCAC	A/G	0.485	1.000	0.500	0.507	0.0141	Ribosomal protein S7
C379_275	M1	F:AGCTGGTTTATGTGGGCTGTG R:TGGTACACTTGTCCACCATC E:CCACCATCATGACTATGC	T/A	0.265	1.000	0.412	0.395	-0.0429	UPF0762 protein C6orf58 homolog
C455_315	M1	F:CATGAAAATGTTTACAGG R:ATTGAATCCCATGGCTGTTG E:AGCTTATAATGAATACTGTTAAGA	C/T	0.368	1.000	0.500	0.472	-0.0605	15-hydroxyprostaglandin dehydrogenase [NAD(+)]isoform X1
C417_302	M1	F:AGAGTATCCTTTCATGGGC R:CATATTGCTTGGCTGTGTGGG E:ggcGTTGGATGCCCTCTATGC	G/A	0.191	1.000	0.324	0.314	-0.0312	Alpha-1-antitrypsin homolog isoform X6
C437_455	M2	F:TCA GCACACTAAGACCACATG R:AGCAGGGAGGGCGATTACTT E:gaAGGGGGAITACTTCTT	G/A	0.132	1.000	0.265	0.233	-0.138	40S ribosomal protein S18
C391_875	M2	F:CGAAAACGGTCAGATGATG R:ATGTGGCTCGCATTGAAC	G/A	0.426	0.177	0.618	0.496	-0.249	Fibronectin

Table 2 continued

SNP_ID	iPLEX	PCR primers (F and R) and extension primer (E)	Allele	MAF	HW	H _{obs}	H _{exp}	F _{is}	Gene Description
C191_480	M2	E:TCCCTTGCCATTGCC	C/T	0.353	1.000	0.471	0.464	−0.0154	40S ribosomal protein S2
		F:CAAGCTGCCATCATTCCTG							
		R:ACCAGTCACCTTGCAGGGTA							
C470_159	M2	E:ccGGGTTTGCCGATCTTGT	A/G	0.353	1.000	0.471	0.464	−0.0154	Peptidyl-prolyl cis-trans isomerase-like isoform X2
		F:CCCCAAAACCTGCGTCTTC							
		R:ATGGGCTAGGAGTAAAACCG							
C564_1273	M2	E:tttCTGTTTGAACCAGGATTTTC	C/T	0.273	0.383	0.485	0.403	−0.208	S-adenosylmethioninesynthase isoform type-1-like
		F:ACTGCTTCAGATCGTCAAC							
		R:TTGGCCTCTTCAGCTTCAG							
C128_1801	M2	E:AATGACACCAGGCCGGAG	T/C	0.118	0.0015	0.059	0.211	0.724	Proteoglycan 4-like isoform X2
		F:ATGGACACTGGATTCCCAAG							
		R:GTACTGAGGCATGGACAAAAG							
C585_507	M2	E:cTTTGACCACTCAGTCC	C/T	0.485	0.0432	0.324	0.507	0.365	Novel protein similar to <i>H. sapiens</i> HPN, hepsin
		F:TCTGTGTAAGGAGAGCGAG							
		R:AGCACTTCAACAATCTGTCC							
C87_1726	M2	E:GTGCTGATCTTCTTGCC	C/T	0.177	1.000	0.294	0.295	0.0030	Heparin cofactor 2-like
		F:AAGATGGCGCCAAAAGCTG							
		R:GTCCAGAGGGATAGAGTC							
C43_831	M2	E:ACGGCCAGGTGCTCG	G/A	0.309	0.686	0.382	0.433	0.119	Protein AMBP precursor
		F:TATAACTCCTCCCTCATGGC							
		R:AGACACTCCTTCTGTGTAC							
C41_428	M2	E:TGTTCTGGTTGCCCA	G/A	0.235	0.647	0.412	0.365	−0.130	Secreted phosphoprotein 24
		F:AAATGCAATCAGGCCACAGAG							
		R:GACGATTTGGATCATAGTGC							
C5_660	M2	E:tATCATAGTCCATGCCAT	C/T	0.338	0.254	0.559	0.454	−0.234	Uncharacterized protein LOC103025530
		F:TTCGACAACCTGCCATGATGC							
		R:AAAGCCTGTAGTTCAGTGTG							
		E:GTTCAGTGTGAAGCTCT							

M multiplex (1 and 2), *Allele* wild/rare, *MAF* minimum allele frequency, *HW* Hardy–Weinberg *P* value, *H_{obs}* observed heterozygosity, *H_{exp}* expected heterozygosity, *F_{is}* inbreeding coefficient

(Willians et al. 2010; Salem et al. 2012) and consists of an extension reaction in which one primer is annealed immediately above the existing polymorphism, that by mass spectrometry, the mass of extended primer is determined (Gabriel and Ziaugra 2004). Some genes of the genotyped SNPs were related to immune system, such vitronectin-like (C1013_445), coagulation factor V (C238_1041) and glutathione S-transferase (C627_936) genes (Table 2).

Statistical genetic parameters to describe the Paraná River population (São Paulo State, Brazil) were obtained through the validation process of 32 feasible polymorphic SNPs (64 %) in 34 individuals (Table 2). The observed (H_{obs}) and expected heterozygosity (H_{exp}) were calculated using Cervus 3.0.7 (Marshall et al. 1998). Inbreeding coefficient (Fis), minimum allele frequency (MAF) parameters, conformance to Hardy–Weinberg equilibrium (HW) and Linkage disequilibrium were performed using Genepop 4.0.11 (Raymond and Rousset 1995). Fis parameters were estimated regards Weir and Cockerhsam (1984). The observed and expected heterozygosity ranged from 0.059 to 0.706 and 0.058 to 0.507, respectively. MAF ranged from 0.029 to 0.485. Deviation of Hardy–Weinberg equilibrium was not significant after Bonferroni correction ($P > 0.0015$). Significant linkage disequilibrium (LD) was observed between the loci: C178_243/C191_480, C417_302/C41_428, C1013_445/C585_507, C1013_445/C87_1726 and C239_1594/C87_1726. Negative values of Fis values were predominant (heterozygote excess). The present study increases significantly the genetic resources for pacu (*Piaractus mesopotamicus*), a non-model warmwater species used in aquaculture of several countries, which has high market value, but threatened by overfishing. Furthermore, our SNPs set are useful for pre-breeding genetic programs, particularly to delineate the formation of the best families in terms of genetic variability, as well as to detect genetic structure of wild and farmed stocks of pacu.

Acknowledgments This work was supported by grants from Pró-Reitoria de Pesquisa da UNESP (Prope 07/2015—DTH), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 446779/2014-8 and 305916/2015-7—DTH; and 130262/2014-5 VAMF), and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2014/03772-7—DTH and 2014/12412-4—VAMF).

References

- Calcagnotto D, DeSalle R (2009) Population genetic structuring in pacu (*Piaractus mesopotamicus*) across the Paraná-Paraguay basin: evidence from microsatellites. *Neotrop Ichthyol* 7(4):607–616. doi:10.1590/S1679-62252009000400008
- FAO (2010) The state of world fisheries and aquaculture—2010 (SOFIA). FAO, Rome
- Flores Nava A (2007) Aquaculture seed resources in Latin America: a regional synthesis. In: Bondad-Reantaso MG (ed) Assessment of freshwater fish seed resources for sustainable aquaculture. FAO Fisheries Technical Paper No 501. FAO, Rome
- Gabriel S, Ziaugra L (2004) SNP genotyping using Sequenom MassARRAY 7K Platform. *Curr Protoc Hum Genet* 42:2.12: 2.12.1–2.12.16. doi:10.1002/0471142905.hg0212s42
- Honglang H (2007) Freshwater fish seed resources in China. In: Bondad-Reantaso MG (ed) Assessment of freshwater fish seed resources for sustainable aquaculture. FAO Fisheries Technical Paper No 501. FAO, Rome
- IBGE (2014) Instituto Brasileiro de Geografia e Estatística. Produção da Pecuária Municipal 2013, vol 41. IBGE, Rio de Janeiro. <http://www.ibge.gov.br/home/estatistica/economia/ppm/2013/>
- Iervolino F, Resende EK, Hilsdorf AWS (2010) The lack of genetic differentiation of pacu (*Piaractus mesopotamicus*) populations in the Upper-Paraguay Basin revealed by the mitochondrial DNA D-loop region: implications for fishery management. *Fish Res* 101:27–31. doi:10.1016/j.fishres.2009.09.003
- Liu ZJ, Cordes JF (2004) DNA marker technologies and their applications in aquaculture genetics. *Aquaculture* 238:1–37. doi:10.1016/j.aquaculture.2004.05.027
- Liu S, Palti Y, Gao G, Rexroad CE III (2016) Development and validation of a SNP panel for parentage assignment in rainbow trout. *Aquaculture* 452:178–182. doi:10.1016/j.aquaculture.2015.11.001
- Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Mol Ecol* 7:639–655. doi:10.1046/j.1365-294x.1998.00374.x
- MPA. Ministério da Pesca e Aquicultura (2011) Boletim estatístico da pesca e aquicultura. MPA, Brasília
- Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenism. *J Hered* 83(6):248–249
- Resende EK (2003) Migratory fishes of the Paraguay-Paraná basin excluding the Upper Paraná River. In: Carolsfeld J, Harvey B, Ross C, Baers A (eds) Migratory fishes of South America: biology, fisheries and conservation states. World Bank, Victoria, pp 99–156
- Salem M, Vallejo RL, Leeds TD, Palti Y, Liu S et al (2012) RNA-Seq identifies SNP markers for growth traits in rainbow trout. *PLoS One* 7(5):e36264. doi:10.1371/journal.pone.0036264
- Vera M, Alvarez-Dios J, Fernandez C, Bouza C, Vilas R et al (2013) Development and validation of Single Nucleotide Polymorphisms (SNPs) markers from two transcriptome 454-runs of turbot (*Scophthalmus maximus*) using high-throughput genotyping. *Int J Mol Sci* 14:5694–5711. doi:10.3390/ijms14035694
- Weir BS, Cockerhsam CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370
- Willians LM, Ma X, Boyko A, Bustamante CD, Oleksiak MF (2010) SNP identification, verification, and utility for population genetics in a non-model genus. *BMC Genet* 11:32. doi:10.1186/1471-2156-11-32
- Zhang HW, Yin SW, Zhang LJ, Hou XY, Wang YY (2015) Development and validation of single nucleotide polymorphism markers in *Odontobutis potamophila* from transcriptomic sequencing. *Genet Mol Res* 14(1):2080–2085. doi:10.4238/2015