CrossMark

Transcriptome sequencing and development of novel genic SSR markers for *Dendrobium officinale*

**Meng Xu · Xin Liu · Jian-Wen Wang ·
Shi-Yuan Teng · Ji-Qing Shi · Yuan-Yuan Li ·
Min-Ren Huang**

Abstract *Dendrobium officinale* is a critically endangered and valuable medicinal herb. In this study, two transcriptome libraries were constructed from 2-year-old stems of two cultivated varieties and sequenced with the Ion Torrent PGM™ platform. A total of 51,683 all-transcripts with average length of 505 bp, and 40,405 all-unigenes with N50 of 644 bp were achieved. Totally, 8527 potential genic simple sequence repeats (SSRs) were identified from 7332 (18.15%) unigene sequences, of which 1023 (2.53%) unigenes contained more than one SSR. Finally, 68 locus-specific primer pairs were designed, and 17 of them produced polymorphic products in *D. officinale* individuals. The number of alleles per locus ranged from 2 to 15, the observed and expected heterozygosity estimates ranged from 0.194 to 0.903 and from 0.360 to 0.893, respectively. These loci were further tested for cross-species transferability to *D. huoshanense*. All the 17 primer pairs were successfully amplified to yield locus-specific PCR products in *D. huoshanense*. Due to their high level of polymorphism and transferability, these genic SSR markers are valuable tools not only for germplasm conservation of this species but also for phylogenetic studies of *Dendrobium*.

Introduction

Dendrobium, one genus of *Orchidaceae*, has approximately 1450 species, of which 74 species have been identified in China (Takamiya et al. 2011). Stems of some *Dendrobium* species contain various types of polysaccharides that exhibit anti-inflammatory, immune-enhancing, antioxidant, and anti-glycation activities (Hsieh et al. 2008; Ng et al. 2012; Pan et al. 2014), which makes them valuable as traditional medicines used for tonifying the stomach and promoting fluid, nourishing "yin" and clearing heat, improving eyesight, and relieving sore throat. *Dendrobium officinale*, a well-known traditional Chinese medicinal herb and officially listed in the Chinese Pharmacopeia (The State Pharmacopeia Commission of P. R. China 2010), has both ornamental values and a broad range of therapeutic effects. Due to over-exploitation and habitat deterioration, wild *D. officinale*, like many other Dendrobium species, has been exploited to near extinction and is now listed as an IUCN critically endangered plant (http://www.incnredlist.org/search).

M. Xu (✉) · X. Liu · J.-W. Wang · M.-R. Huang
Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, 159 Longpan Road, Nanjing, Jiangsu 210037, China
e-mail: xum@njfu.edu.cn

S.-Y. Teng · J.-Q. Shi
Suzhou Shenyuan Bio-Tech Co., Ltd, Suzhou 215200, China

Y.-Y. Li
Guizhou Academy of Sciences, Guiyang 550001, China

Microsatellites or simple sequence repeats (SSRs) are made up of tandem arrays of short nucleotide motifs and randomly distributed across prokaryotic and eukaryotic genomes. The SSR markers have distinguishing features such as good repeatability, polymorphism, and codominance. These make them a powerful high-resolution tool for the study of population genetics, molecular ecology, and marker-assisted selection (MAS) studies in plants.

The genomic resources of *D. officinale* are still limited, although there exists a limited availability of simple sequence repeat (SSR) markers for this species (Gu et al. 2007; Xu et al. 2011; Lu et al. 2013). Considering the urgency for germplasm conservation of this species and its important role in phylogenetic studies of *Dendrobium*, it is necessary to develop more genic SSR markers, which will contribute to the conservation of the species and related genetic research in *Dendrobium*. In this study, 17 polymorphic genic SSR markers were isolated and characterized using the Ion Torrent sequencing platform for *D. officinale*.

Materials and methods

Plant materials

Two common cultivated varieties, *D. officinale* "Brown Stem" (BS) and *D. officinale* "Green Stem" (GS), were used to develop transcriptome sequences. "Brown Stem" (BS) is a low-growing and disease-resistant, brown stem cultivar, while "Green Stem" (GS) is a fast-growing, susceptible to disease, green stem cultivar. Young leaves of 31 *D. officinale* individuals were collected and were individually ground into powder in liquid nitrogen, and DNA extraction was performed using the DNeasy Plant Mini Kit (Qiagen).

RNA extraction, library preparation, and sequencing

For transcriptome sequencing, total RNA was extracted from 2-year-old stems of two cultivated varieties with the RNeasy Plant Mini Kit (Qiagen), and two rounds of oligo (dT) selection of the poly (A) RNA were performed using the MicroPoly(A) Purist™ Kit (Ambion). Two transcriptome libraries of sequencing were generated using Ion Total RNA-Seq Kit v2 (Thermo Fisher Scientific) and sequenced in the Ion Torrent PGM™ platform (ION PGM™ sequencer with

318 chip, Thermo Fisher Scientific) following the manufacturer's recommended procedures.

Data processing, assembly, and annotation

The raw reads were cleaned by removing reads containing ploy-N and low-quality reads. All the downstream analyses were based on clean data with high quality. Trinity (http://trinityrnaseq.github.io) was used for assembly (Grabherr et al. 2011), by which, clean reads of various isoforms from one gene were assembled into distinct transcripts but the same subcomponent, which can be regarded as a gene, and the longest transcript of each subcomponent was defined as the "unigene" for annotation.

All the assembled unigenes of the two tissues were searched against the NCBI non-redundant protein sequences (Nr) database to identify the putative mRNA functions using the BLAST algorithm with an E-value cut-off of 10^{-6} (Korf et al. 2003). Gene ontology (GO) terms were extracted from the best hits obtained from the BLASTx against the Nr database using the Blast2GO program (Conesa and Götz 2008). The BLAST algorithm was also used to align unique sequences to the NCBI non-redundant nucleotide sequences (Nt), Swiss-Prot, KOG, and KO to predict possible functional classifications and signaling pathways.

Differential expression analysis

Gene expression levels were estimated by RSEM (Li and Dewey 2011) for each sample (two chips per sequenced library). Differential expression analysis of the two samples was performed using the DEGseq (Wang et al. 2010) R package. Genes with an adjusted P value of <0.05 found by DESeq were assigned as differentially expressed genes (DEGs). Volcano plot of the DEGs was drawn using the value of $\log_{10}$ (padj) and $\log_2$ (foldchange). GO and KEGG analysis of DEGs was implemented by the GOseq R packages (Young et al. 2010) and KOBAS (Mao et al. 2005).

Isolation of microsatellite markers

Perl scripts were developed to search for SSRs using the MIcroSAtellite (MISA) search module. The parameters were set to detect perfect ten-, di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of 10, 6, 5, 5, 5,

Table 1 Length distribution of assembled transcripts and unigenes

Nucleotides length	Transcripts	Unigenes
200–500 bp	35,095	28,952
500–1 kb	11,156	7719
1 k–2 kb	4835	3309
>2 kb	597	425
Total	51,683	40,405
Min length (bp)	201	201
Mean length(bp)	505	477
Median length (bp)	342	318
Max length (bp)	10,688	10,688
N50 (bp)	644	594

and 5 repeats, respectively. Primer pairs were designed using online software (http://www.genscript.com/), and their forward primers were labeled with a 6-FAM fluorescent dye.

PCR reactions were carried out under the following conditions: an initial denaturation at 94 °C for 3 min, and then followed by 30 cycles of 30 s at 94 °C, 30 s at 56 °C, and 30 s at 72 °C, followed by a final extension 3 min at 72 °C. A typical 20-μl reaction volume contained the following: 1 × PCR buffer, 1.6-mM $MgCl_2$, 0.2 mM of each dNTPs, 0.5 μM of each primer, 0.25 U of Taq DNA polymerase (Takara), and 10–25-ng

genomic DNA. Polymorphic PCR products were detected by capillary electrophoresis using the ABI 3730 DNA Analyzer (Applied Biosystems).

SSR data analysis

The genetic diversity indexes including number of alleles at each locus (Na), the number of effective alleles per locus (Ne), observed heterozygosity (Ho), and expected heterozygosity (He) were estimated using GenAIEx version 6.5 (Peakall and Smouse 2012).

Results and discussion

Sequencing and assembly

Two transcriptome libraries were constructed from 2-year-old stems of two cultivated varieties and sequenced with the Ion Torrent PGM™ platform. After removing adaptor sequences, ambiguous nucleotides, and low-quality sequences, there were 5,721,513 clean reads for BS and 5,107,465 for GS, respectively. These clean reads were used for de novo assembly. The Trinity software generated 51,683 all-transcripts with average length of 505 bp, and 40,405 all-unigenes with N50 of 644 bp were achieved in Supplementary File S1. Of the unigenes, 28,952 (28.38%) were 200–
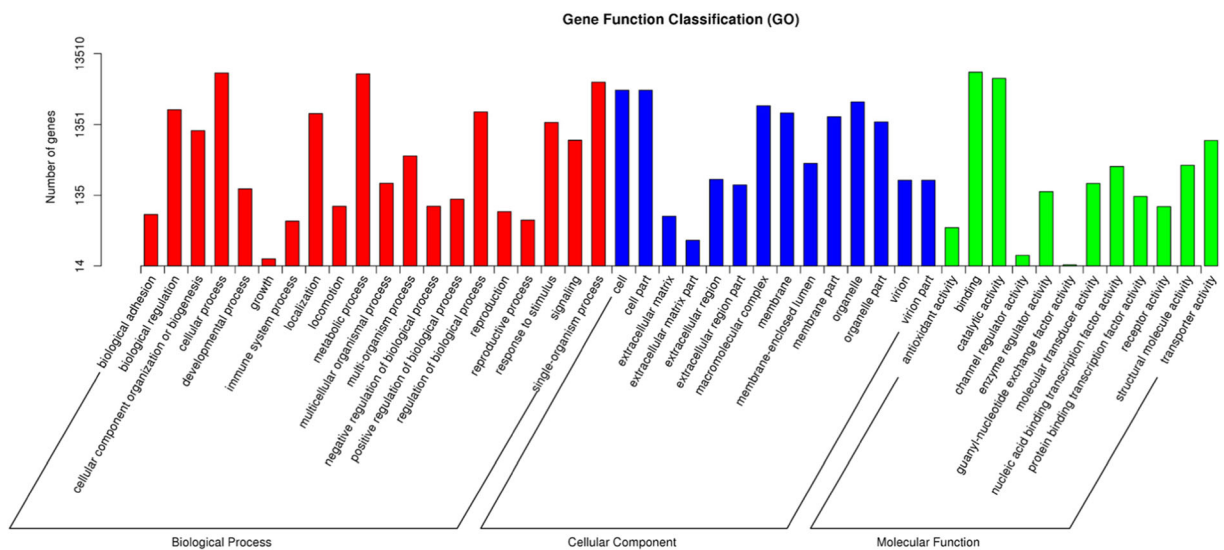
Fig. 1 Gene ontology classification of assembled unigenes. The 13,510 matched unigenes were classified into three functional categories: biological process, cellular component, and molecular function
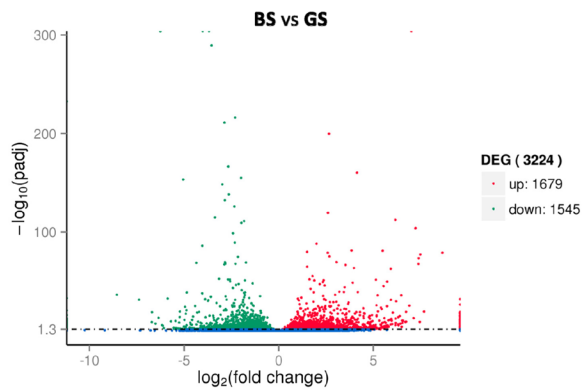
Fig. 2 Volcano plot of the differentially expressed genes (DEGs) between BS and GS. The *red and blue dots* indicated significantly upregulated or downregulated expression of genes in BS compared to GS

500 bp, 7719 (19.10%) were 500–1 kb, 3309 (8.19%) were 1–2 kb, and the remaining 425 (1.1%) were >2 kb (Table 1).

Functional annotation and classification

All the 40,405 assembled unigenes were annotated by sequence aligning in diverse public databases (Supplementary Table S1). Analyses showed that 23,773 unigenes (58.83%) had significant hits in the Nr database, 9076 (22.46%) in the Nt database and 16,319 (40.38%) in the Swiss-Prot database. In total, there were 24,955 unigenes (61.76%) successfully annotated in at least one of the seven public databases

Table 2 Occurrence and number of repeats of SSR motifs in *Dendrobium officinale*

Repeats	5	6	7	8	9	10	11	12	13	14	Total
A/T						2212	4666				6878
C/G						3					3
AC/GT		21	4	1	3	2	4				35
AG/CT		224	136	94	44	80	55	4			637
AT/AT		142	49	36	25	20	13	1			286
CG/CG		1									1
AAC/GTT	43	15	7	5						1	71
AAG/CTT	106	57	28	5		1					197
AAT/ATT	47	21	20	3						1	92
ACC/GGT	13	5									18
ACG/CGT	7	2	3								12
ACT/AGT	1	1									2
AGC/CTG	41	25	9								75
AGG/CCT	27	22	9								58
ATC/ATG	49	30	19	3	1						102
CCG/CGG	20	6	1								27
AAAC/GTTT	1										1
AAAG/CTTT	8		1								9
AAAT/ATTT	10										10
AATG/ATTC	1										1
AATT/AATT		1									1
AGAT/ATCT	1										1
AGCG/CGCT	1										1
AAAAG/CTTTT	2										2
AAAAT/ATTTT	1										1
AAGCT/AGCTT	1										1
AATAG/ATTCT	1										1
AGGGC/CCCTG	1										1
AAGATG/ATCTTC				1							1
ACCATC/ATGGTG			1								1
ACGGCC/CCGTGG	1										1

(Supplementary Table S1), and 2945 unigenes (7.28%) were annotated in all seven databases.

For GO analysis, the 13,510 matched unigenes were divided into three functional categories: biological process, cellular component, and molecular function (Fig. 1). For biological process category, unigenes involved in "cellular process" (7150), "metabolic process" (6960), and "single-organism process" (5317) were highly represented. The cellular component category mainly comprised proteins involved in "cell" (4099), "cell part" (4097) and "organelle" (2482). Within the molecular function category, "binding" (7355), "catalytic activity" (6018), and "transporter" (803) were highly represented.

All the unigenes were aligned to KOG database for functional prediction and classification. In total, there were 8241 (20.40%) unigenes assigned to KOG classification and divided into 25 specific categories (Supplementary Fig. S1). The "general function prediction only" (1274) was the largest group, followed by "posttranslational modification, protein turnover, chaperones" (1136), "signal transduction mechanisms" (766), "translation, ribosomal structure and biogenesis"

(673), "RNA processing and modification" (629), and "intracellular trafficking, secretion, and vesicular transport" (569).

The metabolic pathway analysis was also conducted by sequence aligning in KEGG ORTHOLOG (KO) Database for all unigenes (Supplementary Fig. S2). This process predicted a total of 264 KEGG pathways, representing a total of 9643 unigenes that were assigned to five KEGG biochemical pathways (Hierarchy1): Cellular Processes (a), Environmental Information Processing (b), Genetic Information Processing (c), Metabolism (d), and Organismal Systems (e). The pathways (Hierarchy2) involving the highest number of unique transcripts were "translation" (943), followed by "signal transduction" (851), "carbohydrate metabolism" (766) and "folding, sorting and degradation" (739).

Differentially expressed genes

To detect differentially expressed genes (DEGs), the expression level of these unigenes in the two cultivars was estimated. Volcano plots showed the expression

Table 3 Characteristics of 17 polymorphic genic SSR markers in *Dendrobium officinale*, including loci name, repeat motif, primer sequences, expected size of alleles (S), the number of alleles (Na), the number of effective alleles (Ne), the observed heterozygosity (Ho), and the expected heterozygosity (He)

Locus	Repeat motif	S (bp)	Forward primer (5' to 3')	Reverse primer (5' to 3')	Na	Ne	Ho	He
Do_eSSR02	(TC)$_{13}$	229	TTTTAACTTGCCAACAATGG	GCATGGGTTCTCTGTTCATA	12	6.744	0.742	0.852
Do_eSSR20	(AAG)$_9$	173	ACACACAAGAAGAG CGAGAG	CCAGCAACCTTCCATAGTTT	7	4.124	0.548	0.758
Do_eSSR21	(GA)$_9$	117	AATGGAAGCTTTTGGGTAAG	GCTAATTATGCCCAATTCCT	15	8.580	0.613	0.883
Do_eSSR23	(AG)$_9$	147	GAAGAGGAGGCTATTGCTGT	TACTTCCAATGCCATATCCA	8	4.918	0.517	0.797
Do_eSSR30	(TCT)$_8$	142	TTCCCCTTCTTCACTCTCTC	GATTGGTTTCTGCTTCCTCT	7	2.765	0.452	0.638
Do_eSSR35	(TC)$_8$	125	CTTAATTTGTTGGCGCACT	AATGTCATGTAGCCCTGCT	13	5.539	0.839	0.819
Do_eSSR36	(TC)$_8$	204	CAATTCAGTTGCCTGAAAGA	GGCTCAACTAAAGAGTGCAG	15	9.326	0.633	0.893
Do_eSSR44	(AGA)$_8$	341	TACTTATTCCCCTCCTCCTG	AGCCTAAGCTTCGTCTTCAT	10	5.670	0.903	0.824
Do_eSSR45	(CAA)$_8$	193	CTAGGGAGCAACAACAACAG	TTTGCAGAGCCATTTAGAAG	6	3.939	0.774	0.746
Do_eSSR47	(AT)$_8$	226	GCTTGGTTTGAGGACTTGTA	AATCAATCAAGTAAGCCTTCG	10	5.508	0.742	0.802
Do_eSSR49	(TC)$_8$	108	GGAAGCTGTTGTGTATTGCT	ATCACCAAGTGCATCTTAAATC	12	7.421	0.806	0.865
Do_eSSR50	(CAA)$_7$	137	TCTGAACTCTGCCACAACTT	AGTTTTGCTTACCGAACACA	4	3.241	0.323	0.691
Do_eSSR53	(CAA)$_7$	263	TGGTGTGATAAATGGTGATG	GCATATGAATAGTCGGGTGA	7	2.056	0.516	0.514
Do_eSSR58	(GAG)$_7$	264	GAGCAGAGAGTTGTTGGTTG	GATCAACCAGCTGTACTTCG	4	2.128	0.548	0.530
Do_eSSR65	(AAG)$_6$	161	CCCCTTCTCAAATCTACCAC	TGGTACCCTTTCTTTCTTCC	3	2.038	0.194	0.509
Do_eSSR67	(AGAA)$_6$	345	ACAATGAATGGTGGATTCTG	ACAAAGGAGACAATCAGAGA TG	2	1.949	0.323	0.487
Do_eSSR68	(CTACAT)$_6$	338	CCTCATCCTCCATCAGTGTA	GATTGACGGTGGATACAGAA	6	1.563	0.355	0.360

differences of gene among these two cultivars. A total of 3224 genes (7.98% of all 40,405 genes) were identified as significant DEGs between these two cultivars, which comprised 1679 upregulated genes (accounting for 52.08% of all significant DEGs) and 1545 downregulated genes (accounting for 47.92%) in BS compared to GS (Fig. 2, Supplementary Table S2). Statistics of GO and KEGG enrichment was archived in Supplementary Tables S3 and S4, respectively.

Frequency and distribution of SSR motifs

The MISA search module was used to search for SSRs with the 40,405 unigenes. In this study, only SSR loci were considered to contain mon-, di-, tri-, tetra-, penta-, and hexa-nucleotides with minimum repeat numbers of 10, 6, 5, 5, 5, and 5, respectively. Totally, 8527 potential genic SSRs were identified from 7332 (18.15%) unigene sequences, of which 1023 (2.53%) unigenes contained more than one SSR (Table 2).

Development, validation, and transferability of genic SSR markers

After adaptor removal and assemblage, more than 4000 unigenes were generated. A total of 68 locus-specific primer pairs were designed, and 17 of them produced polymorphic products in eight *D. officinale* individuals. These 17 loci were subsequently tested on 31 individuals for allele polymorphism. The number of alleles (Na), the number of effective alleles (Ne), the observed heterozygosity (Ho), and the expected heterozygosity (He) per polymorphic locus were estimated using GenAIEx version 6.5 (Peakall and Smouse 2012). As showed in Table 3, 2 to 15 alleles were counted in per polymorphic locus, the observed and expected heterozygosity estimates ranged from 0.194 to 0.903 and from 0.360 to 0.893, respectively.

These loci were further examined in *D. huoshanense* for cross-species transferability. All the 17 primer pairs were successfully amplified to yield locus-specific PCR products in *D. huoshanense*. These genic SSR markers have high levels of polymorphism and transferability, which are potentially valuable tools not only for germplasm conservation of this species but also for phylogenetic studies of *Dendrobium*.

Authors' contributions M X designed the experiments and conducted transcriptome sequencing, drafted the manuscript. XL and J-WW performed the SSR experiments. MX, S-YT, J-QS, and Y-Y Li collected plant materials. MX and J-WW analyzed the data. M-RH help to coordinate the experiments. All authors read and approved the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Conesa A, Götz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:1–13

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652

Gu S, Ding XY, Wang Y, Zhou Q, Ding G, Li X, Qian L (2007) Isolation and characterization of microsatellite markers in *Dendrobium officinale*, an endangered herb endemic to China. Mol Ecol Resour 7(6):1166–1168

Hsieh YS, Chien C, Liao SK, Liao SF, Hung WT, Yang WB, Lin CC, Cheng TJ, Chang CC, Fang JM (2008) Structure and bioactivity of the polysaccharides in medicinal plant Dendrobium huoshanense. Bioorg Med Chem 16(11): 6054–6068

Korf I, Yandell M, Bledell J (2003) BLAST-an essential guide to the basic local alignment search tool. O'Reilly Associates, Sebastopol, CA

Li B, Dewey C (2011) RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 12:323

Lu JJ, Gao L, Kang JY, Feng SG, He RF, Wang HZ (2013) Thirteen novel polymorphic microsatellite markers for endangered Chinese endemic herb Dendrobium Officinale. Conser Genet Resour 5:359–361

Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21:3787–3793

Ng TB, Liu J, Wong JH, Ye X, Sze SCW, Tong Y, Zhang KY (2012) Review of research on Dendrobium, a prized folk medicine. Appl Microbiol Biotechnol 93:1795–1803

Pan LH, Li XF, Wang MN, Zha XQ, Yang XF, Liu ZJ, Luo YB, Luo JP et al (2014) Comparison of hypoglycemic and anti-oxidative effects of polysaccharides from different Dendrobium species. Int J Biol Macromol 64:420–427

Peakall R, Smouse PE (2012) GenAIEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28:2537–2539

Takamiya T, Wongsawad P, Tajima N, Shioda N, Lu JF, Wen CL, Wu JB, Handa T, Iijima H, Kitanaka S et al (2011) Identification of Dendrobium species used for herbal medicines based on ribosomal DNA internal transcribed spacer sequence. Biol Pharm Bull 34:779–782

Wang L, Feng Z, Wang X, Wang X, Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26:136–138

Xu W, Zhang F, Lu B, Cai X, Hou B, Feng Z, Ding X (2011) Development of novel chloroplast microsatellite markers for *Dendrobium officinale*, and cross-amplification in other *Dendrobium* species (Orchidaceae). Sci Hortic 128(4):485–489

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14