Plant Omics Journal POJ

POJ 7(6):450-460 (2014) ISSN:1836-3644

Features of transcriptome in trioecious papaya revealed by a large-scale sequencing of ESTs and comparative analysis in higher plants

Fanchang Zeng[1, 2], Qingyi Yu[3], Shaobin Hou[4], Paul H. Moore[5], Maqsudel Alam[4], Ray Ming[2, 6]

[1]State Key Laboratory of Crop Biology, Shandong Agricultural University, Tai'an, Shandong, China 271018, China
[2]Department of Plant Biology, University of Illinois at Urbana-Champaign, Illinois 61801 USA
[3]Texas A&M AgriLife Research, Department of Plant Pathology & Microbiology, Texas A&M University System, Dallas, TX 75252, USA
[4]Advanced Studies in Genomics, Proteomics and Bioinformatics, University of Hawaii, Honolulu, HI 96822, USA
[5]Hawaii Agriculture Research Center, Kunia, HI 96759, USA
[6]FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China

***Corresponding author: rming@life.illinois.edu**

Abstract

Papaya (*Carica papaya* L.) is a major fruit crop in tropical and subtropical regions worldwide. Papaya has three sex types, male, female, and hermaphrodite, and it is a model system for SunUp the evolution of plant sex chromosomes. Although the papaya genome has been sequenced, a systematic transcriptome analysis has not been performed. We report the generation of 75,847 raw sequences from Sanger sequencing of normalized cDNA libraries of the papaya cultivar SunUp prepared from 8 tissues collected from 14 developmental stages and the 3 sex types male, female and hermaphrodite. After trimming, processing and analysis, 16,362 unique sequences were obtained. Functional classification and domain analysis of the collection of unique ESTs indicates there are several protein families with functional domain, gene regulatory network and signal transduction that are involved in a variety of developmental processes including sex differentiation. Comparison of the trioecious papaya transcriptome with that of monoecious plants revealed a unique sex related transcriptome profile. In addition, 106 unigenes were assigned to X/Y chromosomes. Functional annotation revealed sex development related biological processes in which these sex unigenes are involved. Extensive simple sequence repeats (SSR) and single nucleotide polymorphhism (SNP) were identified in papaya ESTs. These sequences will be useful for developing new molecular markers for genetic mapping, detecting allelic polymorphisms associated with phenotypic variations, and facilitating gene cloning for sex determination and agronomically important traits.

Keywords: Papaya, Expressed Sequence Tags (ESTs), Comparative transcriptome, Sex determination, Molecular polymorphisms.
Abbreviations: ESTs_Expressed Sequence Tags; HSY_Hermaphrodite-Specific region of the Yh chromosome; LGs_Link Groups; SSR_Simple Sequence Repeats; SNP_Single nucleotide polymorphism.

Introduction

Papaya (*Carica papaya* L.) ranks behind mango and pineapple as a major fruit crop in tropical and subtropical regions worldwide. Papaya has become a model system for plant genomic research of its relatively small genome size of 372Mb and completion of its draft genome sequence. The fact that papaya can produce fruit in as few as 9 months makes it a potential model organism for fruit tree crops (Ming et al., 2001). An added incentive for analysis of the papaya genome is the discovery that it has a pair of primitive sex chromosomes (Liu et al., 2004), which has commercial implications, as hermaphrodites have preferred agronomic characteristics.

Large-scale sequencing and analysis of ESTs remain a fundamental part of genomic research to enable gene discovery, validation of genome annotation and functional molecular marker mining. EST sequencing projects have been completed or are under way for many plant species. These projects have provided useful tools for intragenomic comparisons (Schlueter et al., 2004) and intergenomic comparisons (Fulton et al., 2002), gene discovery (Ewing, 1999; Ronning et al., 2003; Hughes et al., 2004), molecular marker identification (Michalek et al., 2002), and microarray development (Wisman et al., 2000; Kawasaki et al., 2001; Alba et al., 2004; Arpat et al., 2004; Close et al., 2004). A thorough description of the papaya transcriptome, involving a wide array of tissues and organs, would facilitate additional gene discovery for diverse applications related to agronomic traits and sex determination.

The objective of this study was to produce extensive collections of EST sequences and cDNA clones for papaya genome annotation and to support the production of cDNA microarrays. In this paper, we report the sequencing and analysis of 75,847 sequence reads obtained through Sanger sequencing of EST libraries generated from a variety of tissues, developmental stages, and sex types. Papaya unigenes were correlated with terms derived from the Gene Ontology (Gene Ontology Consortium 2001). Analyses of intragenomic and intergenomic comparisons were conducted against specialized databases to assign a functional annotation based upon similarities to identified protein domains. A set of protein families including putative transcription factors associated with sex and fruit development was identified. The present study represents a

systematic and comprehensive characterization of the transcriptome of trioecious papaya having nascent sex chromosomes and a comparative analysis of the papaya transcriptome with that of other higher plants. This papaya collection of ESTs constitutes an important resource for functional and evolutionary genomics of papaya and related species.

Results

Sequence analysis of the cDNA library

A total of 75,847 trimmed EST papaya sequences were generated from whole-life-cycle normalized cDNA libraries by random sequencing of clones of the cDNA libraries. The cDNA libraries were constructed from the cultivar SunUp which is the same variety used for sequencing the draft genome (Ming et al., 2008) and constructing and sequencing the sex chromosome physical map (Gschwend et al., 2012; Na et al., 2012; Wang et al., 2012). The cDNA libraries were normalized to reduce redundancy by subtraction using the ZAP-cDNA Synthesis kit (Stratagene, CA). Normalization and pooling strategies were used to increase the probability of identifying unique and diverse set transcripts. Clones were randomly picked and sequenced until library sequencing appeared to be saturated so there would be only a low probability of discovering a new mRNA due to the redundancy effect. These trimmed sequences were clustered and assembled into 9,080 contigs leaving 7,282 singletons, yielding a total of 16,362 integrated unique sequences that varied in length from 101 to 2862bp, with an average length of 754bp (Supplementary Fig. 1). A total of 15,064 (92.1%) of the 16,362 unigenes derived from expressed sequence tags (ESTs) matched sequences of the papaya draft genome (Ming et al., 2008), indicating high quality of our genome assembly.

Functional annotation and classification of the ESTs

Analysis of the 16,362 unisequences, by the tBLASTx program, revealed that 2954 (18.1%) of them had no match with non-redundant protein database (data not shown).The top five species which papaya unisequences closely match are *Vitis vinifera, Ricinus communis, Populus trichocarpa, Arabidopsis lyrata, Arabidopsis thaliana* (Supplementary Fig. 2). Sequence similarity and E-value distribution for whole ESTs collection in papaya transcriptome present general high similarity of papaya ESTs to those of other plants in public databases (Supplementary Fig. 3 and 4). This result should be expected as the ESTs in the present papaya study were from cDNA libraries constructed from a variety of tissues and organs across a range of developmental stages, including three sex type flowers both before and after meiosis. Thus, many of the novel ESTs may mainly represent tissue-specific genes or sex determination and differentiation related genes. Our set of unisequences was annotated by Blast2go tool (Götz et al., 2008), using similarity searches in nucleotide and protein databases, as well as in domain searches based on InterProScan. KEGG pathway and gene ontology terms were also assigned (Supplementary Table1,Supplementary Table 2, Supplementary Fig.5 and Supplementary Fig.6). In total, 12,526 (76.6%) of the 16,362 unisequences were classified according to the Gene Ontology (GO) database (Ashburner et al., 2000; Gene Ontology Consortium 2001) (Supplementary Table 2). Of the 12,526 unisequences, 8,061, 7,046 and 7,700 were classified by the GO terms (a) cellular components, (b) molecular functions, and (c) biological processes (Fig. 1). In category of molecular function, the largest set of genes (1,706) was assigned to the nucleotide binding category. Genes involved in protein kinase activity (596), transcription factor activity (377), and RNA binding (330) formed the second, the third, and the fourth largest groups, respectively. Plastid, mitochondrion, and plasma membrane formed the first, the second, and the third largest groups, respectively in gene category of cellular component. And the top three largest biological process groups are response to stress, protein modification process, response to abiotic stimulus, respectively (Fig. 1).

Identification and characterization of protein family

The unisequences were also analyzed for their protein domains through interPro protein domain database to assess assignment to characterized protein families, of which 1,093 protein domains were identified in 13,116 (8.3%) exemplar sequences (data not shown). Perhaps the number of identifiable domains was so small is because of incomplete gene sequences in the exemplars and most protein models were not based on plants. Even the model plant rice and *A. thaliana* were poorly represented in the exemplar sequences. The most abundant types of interPro transcriptional domains found in the collection of exemplar sequences were zinc finger protein family (ZnF), WD40-repeat protein family (WD40), helix-turn-helix transcription repressor (WHTH), and transcription factor GRAS (GRAS) (Fig. 2). These four types of domains constituted about 82% of those identified and are mostly related to gene transcription regulation functions.

Features of transcriptome of papaya and comparative analysis of higher plants

Biological annotation for the unique ESTs in papaya were inferred by the conserved homolog and GO annotation and were compared with those found in genomes of monoecious plants including Arabidopsis, cotton, poplar, pine, maize, and soybean. With the global genome-wide comparison of function annotation profile (Fig. 3 a, c, e) in papaya with these six plant genomes, we can identify specific functional elements and biological process involved in specific growth and development in different species. The wide divergence of gene distribution among different plant genomes Fig. 3(a, c, e) reflect the variety of different functional factors that these species possess. For example, papaya has higher proportion of genes involved in biological adhesion and rhythmic process than those in the other six genomes. Also, papaya has many more genes involved in growth and development and multi-organism and multicellular organismal processes compared all other genomes except that of Arabidopsis. Specifically, Arabidopsis showed considerable similarity with papaya on biological annotation profile than other plant genomes in this study (Fig. 3 a). This similarity might be expected since these two species are in the same order Brassicales and related more closely than to the other species. Particularly, exceptionally higher proportion of specific biological processes involved in sex determination and differentiation are identified in the unique trioecious papaya compared with the other species, all of which are hermaphrodites. These sex-related biological processes include meristem determinacy, meristem initiation and maintenance, pattern specification process, sexual reproduction, vegetative to reproductive phase transition, stem cell maintenance, fertilization, genetic transfer (Fig. 3 b, d, f).

Mapping of the sequenced ESTs to sex chromosomes and autosomes

A total of 15,064 (92.1%) of the 16,362 unigenes having sequence

Table 1. Distribution of the unisequences in the papaya genome.

LG	1	2	3	4	5	6	7	8&10	9&11	Unknown Location
Unique EST #	1462	1600	1348	1194	1098	1636	1181	1187	1493	2865
Percentage	8.94%	9.78%	8.24%	7.30%	6.71%	10.00%	7.22%	7.25%	9.12%	17.51
Genetic Distance (cM)	120.6	138.8	132.4	120.6	103.6	82.9	96.4	112.8	84.4	
Density (# per cM)	12.1	11.5	10.2	9.9	10.6	19.7	12.3	10.5	17.7	
Total #	12199									2865
	15064									
Total	74.6%									17.5%
Percentage	92.1%									

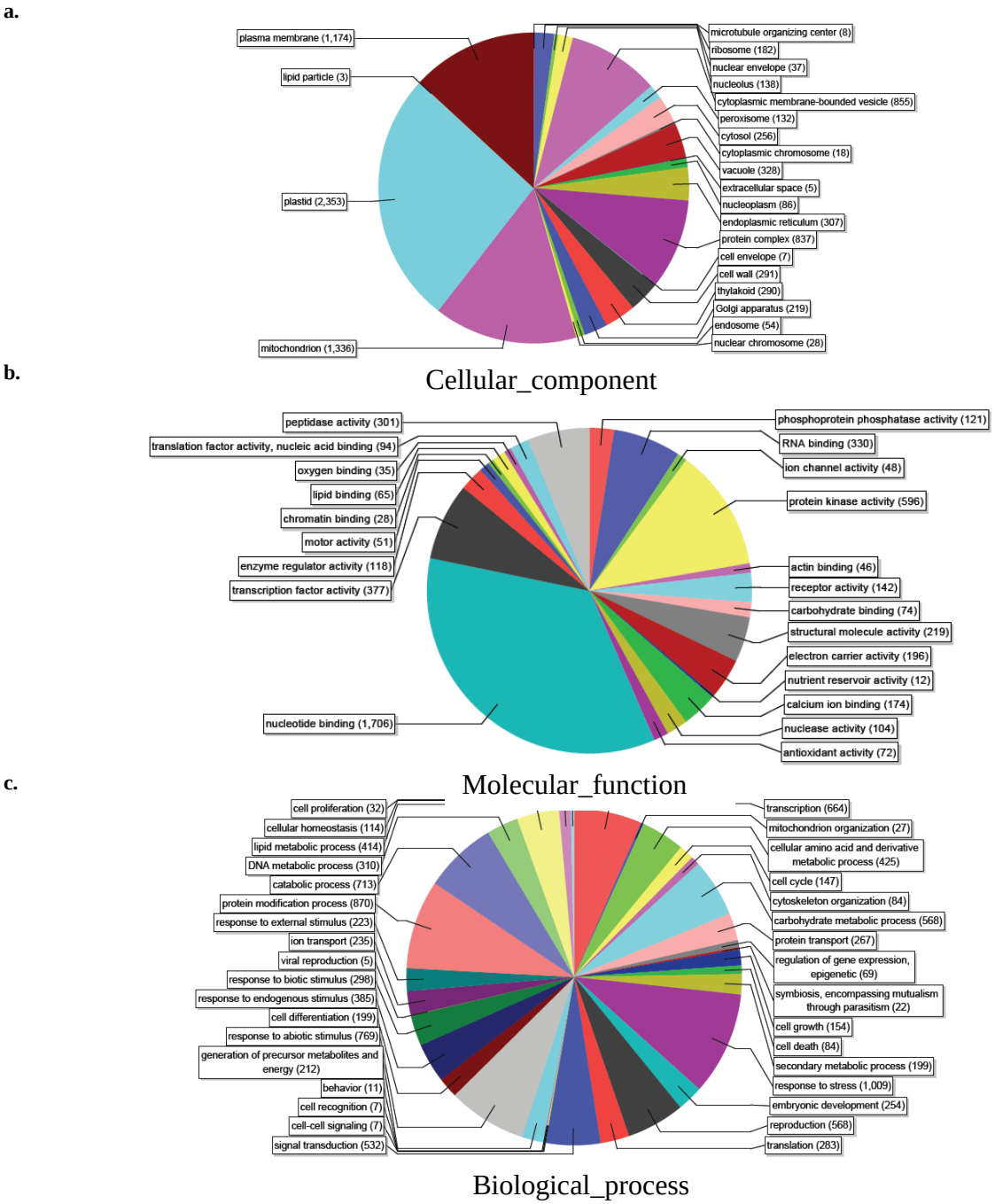


Fig 1. Gene annotation distribution for whole ESTs collection in papaya transcriptome. Cellular component; b. Molecular function; c. Biological process.

Table 2. ESTs statistics mapped to sex chromosome determination region HSY/X and autosome LG.

Sex determination region	HSY total	HSY specific	HSY-X shared	X specific	X total	Autosomes	Unmatched
Unique EST #	104	31	73	6	79		
Percentage	0.64%	0.19%	0.45%	0.04%	0.48%	10737	1194
Total #	183						
Total Percentage	1.12%					65.62%	7.30%

homology with the draft genomic sequences of papaya variety SunUp matched the female draft genome with its XX sex chromosome (Table 1). Among these matches, 12199 (74.6%) were anchored and distributed throughout all papaya LGs (Table 1). 2865 sequences (17.5%) showed homology with papaya genomic sequences but they are not anchored to the integrated genetic and physical map (Yu et al., 2009). The most recently publishedgenetic map of papaya spans a total of 1068.6 cM (Yu et al., 2009). Our present EST map covers 1027.4cM or 96.1% of the papaya genome, and has a resolution of 10 EST sites per cM. The average density of EST sites differed among the chromosomes (Table 1). LG6 had the highest density of EST sites (19.7 sites/cM), and LGs 4 had the lowest density which is approximately half of LG6 (9.9 sites/cM). This high density and annotation of the ESTs will greatly facilitate gene identification for isolation and the comparative study of gene evolution. One hundred and eighty three unique ESTs were anchored to the hermaphrodite-specific region of the Y^h chromosome (HSY) and its X counterpart. We identified 104 unigenes mapped to the sex chromosome HSY region pseudomolecule based on the physical map (Na et al., 2012) of the SunUp hermaphrodite papaya (Table 2). 10737 UniESTs matched to autosomes only and were not included on the X and Y chromosomes. Approximately 1194 (7.3%) UniESTs showed no sequence homology with any papaya genomic sequences including those of the Y/X chromosome (Table 2).

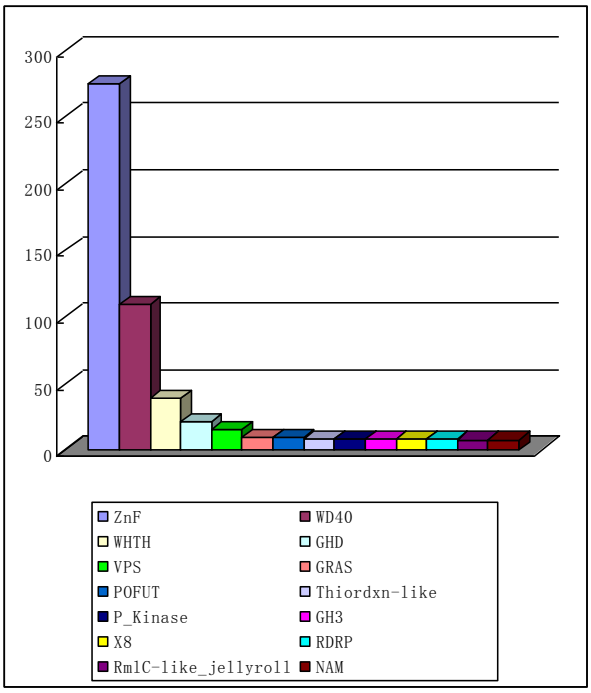


Fig 2. The top categories of protein domains as identified by InterPro analysis of the transcriptome sequences. The total bar height indicates the number of exemplar sequences containing each domain. Categories with <8 members are not shown.

Characterization of sex determination associated ESTs in papaya

The HSY and X counterpart ESTs encode several interesting targets with molecular function related to sex determination and flower differentiation including “protein binding”, “ser/Thr protein kinase activity”, “signal transducer”, and “transcription factors”(Fig. 4 c, Fig. 5 c). These molecular functions are consistent with and confirmed by biological process sets of HSY and X ESTs such as “floral meristem determinacy”, “negative regulation of flower development”, “regulation of transcription”, “maintenance of floral meristem identity”, “regulation of anatomical structure morphogenesis”, “flower whorl development”, etc. (Fig. 4 b and Fig. 5 b). In additiion, extensive genes involved in sex determination and differentiation were identified only in the HSY compared to the X (Fig. 6 b).Such expressed genic elements involved in the flowering process above, particularly in the HSY, are therefore interesting candidates for further cloning and function verification because they likely regulate sex determination and differentiation in papaya and may be indicative of sex chromosome evolution in higher plant. The proportion of HSY or X sequences in the categories related to sex determination and flower differentiation were greater than the categories formed by the classification of total genome unigenes by ‘biological process’ (Fig. 6 b). This situation differs from the proportion of HSY or X sequences in the category ‘fruit development’ which was much smaller suggesting that the UniESTs set associated with HSY and X identified from normalized library in this study enhanced the discovery of new sex determination and differentiation related genes and reinforce the value of using the normalized library-based EST sequencing approach for mining rarely expressed new genes.

Functional molecular marker mining for application in breeding for agronomically important traits

Identification and development of genic microsatellite (SSR) markers

All 16,362 tentatively unique sequences (TUSs) were mined for the presence of SSRs using the MIcroSAtellite (MISA) tool (Thiel et al., 2003), which gave 5,424 SSRs at the frequency of 1/2.3kb in coding regions in 3,774 TUSs (23.1%) (Table 3). 1,110 ESTs contained more than one SSR and 1,039 SSRs were found as compound SSRs. In terms of distribution of different classes of SSRs, mononucleotide SSRs contributed to the largest proportion 3,749, (69.1%). Di- and tri- nucleotide SSRs was the second largest group and accounted for 991 (18.3%) while tri- nucleotide was the third largest group and accounted for SSRs 624 (11.5%). Only a limited number of SSRs of other classes were found. For instance, 33 tetrameric, 7 pentameric and 20 hexameric microsatellites were found. The most frequently occurring di-nucleotide motifs was AT (212), followed by CT (138) and AG (137). Among tri-nucleotides GAA, TCT and TTC were the top three highest SSR motifs.

Table 3. Summary of SSRs identified from papaya ESTs.

SSR mining results	No.
Total number of TUSs examined:	16362
Total size of examined sequences (bp):	12,332,907
Total number of identified SSRs:	5424
Number of SSR containing sequences:	3774
Number of sequences containing more than 1 SSR:	1110
Number of SSRs present in compound formation:	1039
Frequency of SSR	1/2.3kb
SSR Type	
Mono-nucleotide repeats	3749
Di-nucleotide repeats	991
Tri-nucleotide repeats	624
Tetra-nucleotide repeats	33
Penta-nucleotide repeats	7
Hexa-nucleotide repeats	20
Total	5424

MISA, MIcroSAtellite; SSR, simple sequence repeats, TUS, tentative unique sequence.

Single nucleotide polymorphism (SNP) mining

Putative SNP were mined by using an integrated AutoSNP pipeline for large scale SNP discovery. The putative SNP occurred in a contig ≥5 ESTs from two or more genotype was considered for large scale SNP discovery to improve the reliability of SNPs identification. From 169 contigs containing SNPs, *in silico* analysis showed a total of 308 SNPs in total length 114,512 bp, with an average frequency of 1/372bp (Table 4).

Discussion

The whole-life-cycle papaya cDNA libraries were normalized to reduce redundancy by subtraction using the ZAP-cDNA Synthesis kit. This efficient technique includes a particularly important normalization step to enrich important transcripts expressed at much lower levels. So normalization and pooling strategies used in our study has the main advantage of increasing the probability of identifying unique and diverse set transcripts involved in comprehensive biological processes as shown in Fig. 1. For functional annotation and classification of the papaya ESTs by the tBLASTx, the top three species which papaya unisequences closely match are *Vitis vinifera, Ricinus communis, Populus trichocarpa, n*ot with papaya itself or to the Arabidopsis species, duo to the most abundant protein database of these three species in public non-redundant protein database used for tBLASTx. Our discovery of such large numbers of genes for functional categories including nucleotide binding, protein kinase activity, and transcription factor activity was somewhat of a surprise and indicates that considerable regulation network and signal transduction is appropriately involved in a variety of development regulatory process. Such regulation could be achieved via cell surface receptors that perceive external stimuli and respond by relaying a signal to within the cell. Kinases have been identified with the role of transduction of the signal from the cell membrane to the action site. These protein kinases are involved in protein phosphorylation and molecule binding regulation of other successive transducers in the signal transduction pathway (reference needed?). The extensive transcription factors identified in our study may be involved in triggering downstream genes and regulating various cellular functions. Exceptionally high predicted nucleotide/RNA binding function genes were identified in our study showing potentially related to concert development interaction process in papaya. These molecules may bind and interact with specific functional RNA, and then triggering locally or long-distance downstream target through intercellular or intracellular transport. These proteins regulate various specific biological functions in papaya including fruit development and sex determination and flower differentiation process. These findings are consistent with and extend that there is the complicated and concerted communication process involving multiple cellular pathways is responsible for growth and development process in papaya as in other species (Mahalingam et al., 2003; Zeng et al., 2006). These findings suggest that papaya has a highly developed signal transduction and regulation system and these genes can support traits needed for fruit development and sex determination in this important agronomically and scientifically relevant tree plant. Identification of the *cis* elements and the cognate related partner like transcription factors that bind to them during sex determination is the first step toward characterization of higher-order nucleoprotein complexes. Identification of knockouts mutants (deletion or site mutation) or construction of antisense lines for these transcription factor genes and the use of them for RNAseq next generation sequencing or for expression profiling with microarrays created from the sex chromosome differentially expressed cDNA collection will facilitate identification of the targets for these transcription factors. Additionally, extensive genes involved in sex determination and differentiation are specifically identified only in HSY compare with X (Fig. 6 b), which indicate that there is complicated and concerted regulatory process involving more factor and cellular pathways on HSY is responsible for sex determination compare with X counterpart in papaya. These findings suggest that Y has been evolved with a highly developed sex regulation system and these sex related factors on HSY can support traits needed for thriving in fluctuating and competitive environments. This pattern is consistent with the two Y types of male Y and hermaphrodite Y in papaya and nature of higher mutation frequency on male and hermaphrodite related phenotype compare with female, such as male to female sex reversal and peduncle length variation.

Utilization of ESTs for large-scale gene discovery and marker development has been documented in many plant species. EST based markers have been developed because they can be used to assay the functional genetic variation compared to other classes of genetic markers (Varshney et al., 2005). The set tentative unique sequences (TUSs), based on ESTs generated in this study, were used for identification and development of molecular markers. This study mines a large set of genome wide new marker including SSR and single SNP markers for papaya. As these markers are derived directly from coding parts of the genome, they provide good opportunities to identify the 'perfect functional marker' for traits of interest.

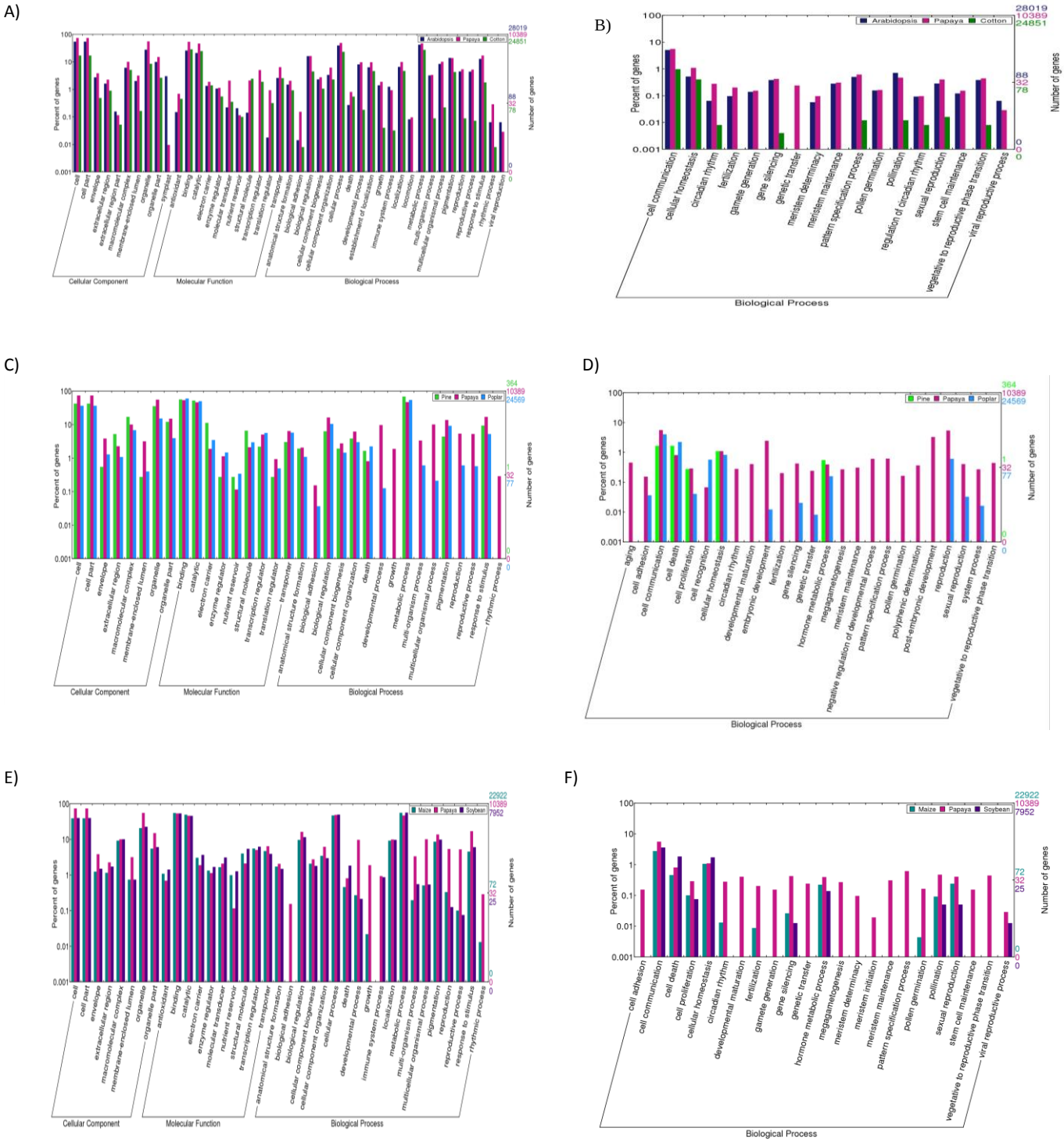
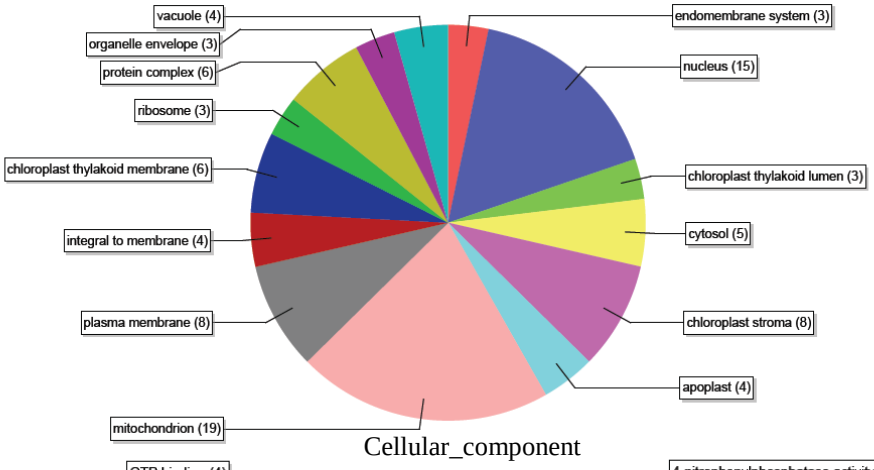


Fig 3. Comparative biological annotation of trioecious papaya transcriptome with monoecious plants. a, c, e. GO annotation comparison for papaya with Arabidopsis and cotton, popalar and pine as well as maize and soybean. b, d, f. Sex associated biological process in papaya compare with monoecious plants. Note that the y-axis in all the comparison histograms is displayed as log-scale.

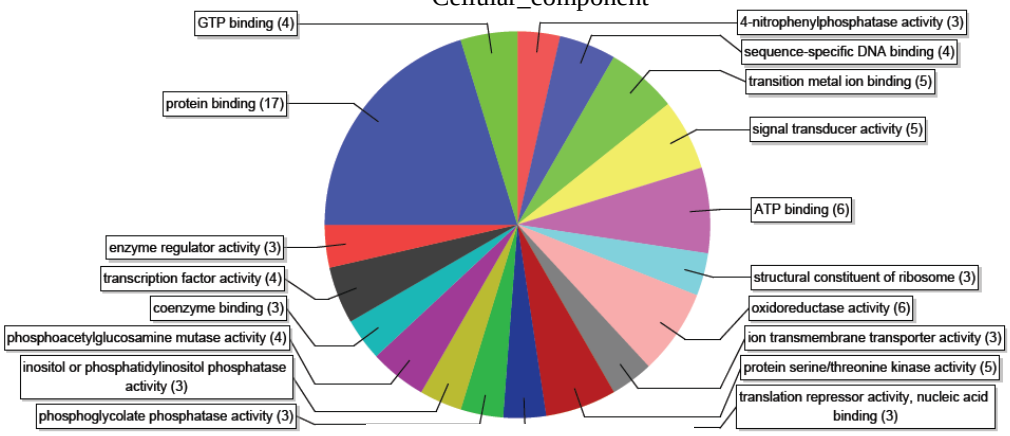
Table 4. Summary of SNPs identified from papaya ESTs.

Total number of contigs examined	9,080
Number of contigs containing SNPs	169
Total length of 169 contigs (bp)	114,512
Total number of identified SNPs in 169 contigs	308
Average SNP frequency	1/372bp

a.



b.



c.

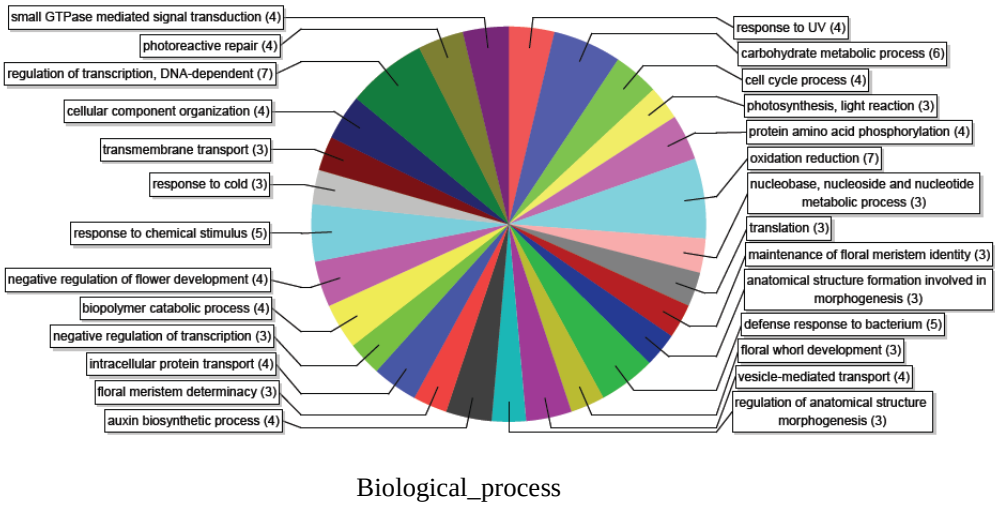


Fig 4. Gene annotation distribution for HSY related ESTs collection on sex chromosome in papaya. a: Cellular component; b.Molecular function; c. Biological process.

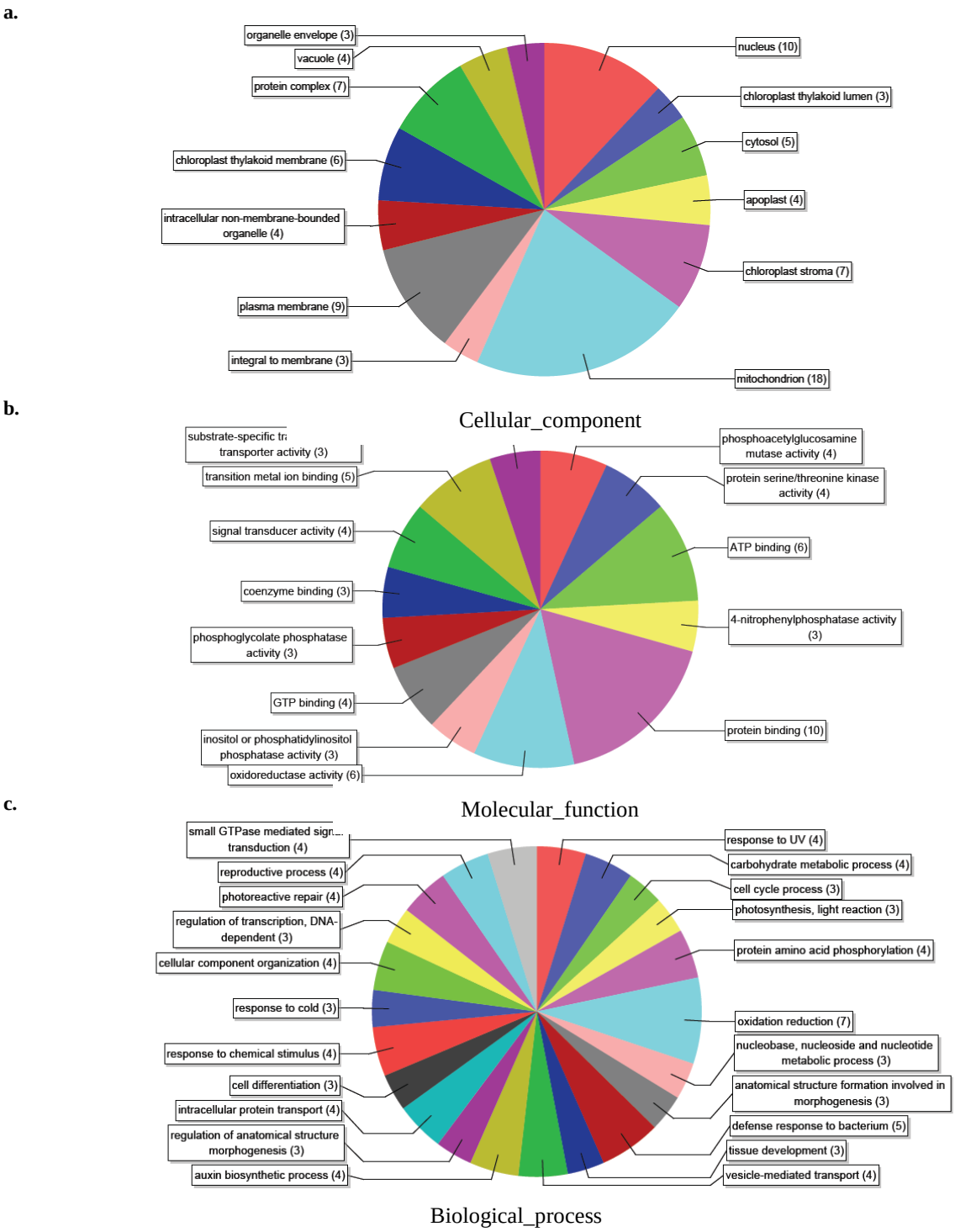


Fig 5. Gene annotation distribution for ESTs collection related to X counrterpart of HSY in papaya.a: Cellular component; b.Molecular function; c. Biological process.

EST-derived SSRs have been widely used in constructing high-density linkage maps, marker-trait association, diversity analysis, etc. in several crop species (Varshney et al., 2002). As transcripts are more highly conserved than nongenic sequence, they are useful in detecting the signature of divergent selection (Li et al., 2002). SNPs offer several advantages like high-throughput and cost effective genotyping (Varshney et al., 2009a; Raju et al., 2010) and identification of functional markers for complex trait through linkage map development or association analysis (Rafalski et al., 2002; Moreno-Vazquez et al., 2003; Varshney et al., 2006), especially with the advances in next generation sequencing technologies which have made SNP discovery cheaper and faster (Varshney et al., 2009b). So SNP marker system is becoming very popular in recent times. The putative SNP occurred in a contig ≥5 ESTs from more than one genotype was considered for large scale SNP discovery to improve the reliability of SNPs identification by using pipeline AutoSNP (Barker et al., 2003). Extended the repertoire of genic markers in our report will provide a high-throughput marker genotyping system for accelerating their use in genetics and breeding programs in papaya. Extensive genes involved in papaya transcriptome, as well as the unique transcriptome

profile of the trioecious papaya compare with other hermaphroditic plants in our results suggest that papaya sex determination is the concerted process involving multiple cellular pathways controlled by a complicated gene regulatory network. Unfolding of the mechanism and the relationships of molecular events at system-level will be necessary for understanding this important concerted process. Our results have provided a rich framework and unified platform on which future genetic and functional studies can be based for papaya functional genomics research like fruit development and sex chromosome research. Combined with the information generated through various genome projects, it provides a basis and a prerequisite for understanding the precise process of the sex determination and evolution in higher plants and, ultimately, the detailed steps by which these genes direct specific process of sex determination. The next steps towards understanding sex biology are reconstructing sex association network using system approach as previously applied in other biological process (Zeng et al., 2007). When coupled with large-scale gene functional analysis studies, this will allow rapid functional definition of these key factors like sex determination genes which have the greatest potential in elucidating the network of sex determination.

Materials and Methods

Plant Materials

Papaya cultivar SunUp (*Carica papaya* L.) plants were grown under natural field condition. 6 tissues (leaf, root, stem, flower, fruit, and seed) of SunUp papaya, including leaves from young seedlings and mature trees, roots from young seedlings, stem from young seedlings, flowers from female, male, and hermaphrodite plants before meiosis and after meiosis stage respectively, and fruits and seeds of 10 different stages from 7 days after pollination to 100% ripe fruits (Supplementary Table 3).

Library construction

The clones from a normalized whole-life-cycle cDNA library were used for obtaining EST sequences. This library was constructed by using all the tissues mentioned above (Supplementary Table 3). The cDNA libraries were constructed and normalized to reduce redundancy by subtraction using ZAP-cDNA Synthesis kit (Stratagene, CA). The normalization strategies were used to increase probability of identifying unique and diverse sets of transcripts especially for the rarely expressed genes. Clones were randomly picked and sequenced until library sequencing is saturated due to the redundancy effect.

Sequencing and assembly of cDNA sequences

The cDNA clones were sequenced from the 5' ends. The sequences were trimmed by removing those that showed homology to sequences of cloning vectors, *Escherichia coli*, mitochondria and chloroplast contaminating sequences and those that were <100 bp by LUCY. Repeat sequences were masked by RepBase databases, TIGR Plant Repeat Databases and papaya repeat database. A total of 75,847 sequences trimmed were then clustered and assembled by PTA (paracel transcript assembler) package of Paracel Inc. (Huang et al., 1996). The EST clusters were assembled into contigs by multiple-sequence alignment that generates a consensus sequence for each of the cluster. Clusters containing only one sequence were grouped as singletons.

Annotation and genome-wide comparison of the ESTs

Gene Ontology (GO) assignation and high-level functional category for UniESTs were performed based on Blast2Go function annotation system filtered by #seqs with cutoff=3.0 (E-value < $1e^{-6}$) (Conesa et al., 2005; Götz et al., 2008), according to Gene Ontology (GO) (Gene Ontology Consortium 2001), eukaryotic orthologous groups (KOGs), and KEGG metabolic pathways. For annotated papaya UniESTs, where possible, assigned predicted protein functions using a combination of sequence comparison with BlastP and domains identification with interProScan by Blast2Go tool (Conesa et al., 2005; Götz et al., 2008). Large-scale genomics GO functional comparison for all of these related plant species were explored and plotted by program WEGO (Ye et al., 2006). Comparison histograms were displayed with GO items with significant relationship for the dataset compared base on Pearson Chi-Square test (Significance level is below the 0.05) at appropriate different levels. This process allowed assignment of unigenes to the GO functional categories of biological process, cellular component and molecular function. Distribution of unigenes was further investigated in terms of their assignment to sub-categories of the main GO categories.

Protein family identification and analysis

The exemplar sequences were analyzed for their protein domains to assess assignment to characterized protein families. Protein domains were predicted using InterProScan (Götz et al., 2008) against integrated protein databases (PRINTS, Pfam, ProDom, PROSITE, SMART) with default set. The best InterProScan search hit and corresponding protein domains present in the exemplar sequences were determined for papaya uniESTs database protein family annotation.

Mapping of ESTs

The ESTs were mapped to the papaya HSY, X and autosomes using the BLASTN program by searching for homologous genomic sequences with known chromosomal locations (Ming et al., 2008; Yu et al., 2009) using a stringent criterion (sequence identity greater than 95% and E-value less than10^{-5}) for sequence alignment.

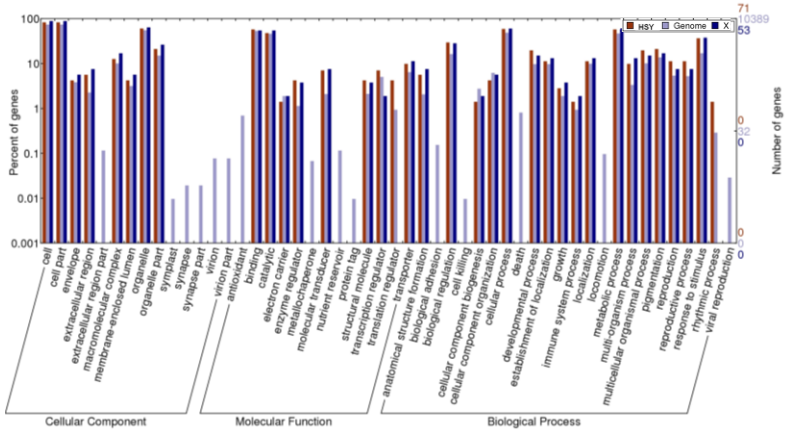
Genic microsatellite (SSR) identification and analysis

A total of 16,362 unigenes were mined with a Perl script program, MISA (MIcroSAtellite) (Thiel et al., 2003) for identification of SSRs. MISA search provides information about the type and localization of each individual microsatellite. The SSR motifs, with repeat units more than five times in di-, tri-, tetra-, penta- and hexa- nucleotides were considered as SSR mining criteria in MISA script.

Single nucleotide polymorphism (SNP) detection

Putative SNP/indels were mined by using an integrated AutoSNP pipeline for large scale SNP discovery (Barker et al., 2003). The pipeline utilized the CAP3 output files as input to detect SNPs/indels based on the nucleotide redundancy in the multiple sequence alignments. The AutoSNP pipeline generated text file includes contig ID, number of sequences in the contig ID, consensus length, number of SNPs and SNP frequency. The threshold for identification of SNPs was based on the number of sequences (≥5) in each consensus sequence and two or more sequences from different genotypes.

a.



b.

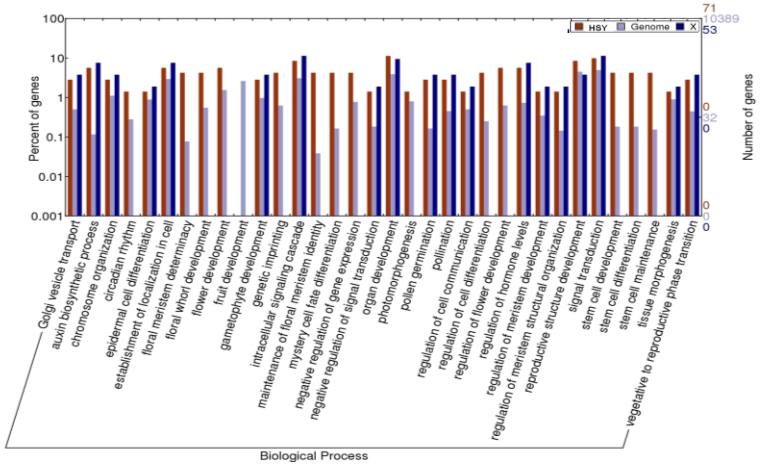


Fig 6. Biological annotation comparison of unique ESTs collection related to HSY, X and whole transcriptome in papaya. a. General GO annotation comparison for HSY, X and whole genome in papaya. b. Detailed biological process associated with sex determination condensed on HSY and X compare with genome in papaya.

Accession numbers

The GenBank accession numbers of the papaya ESTs are EX227656–EX303501

Conclusion

Our study has shown that papaya possess a unique developmental process associated with sex chromosome compare with monoecious plant species, which could be regulated by complex sex association network. The cDNA collection is composed of a broad repertoire of genes that are involved in a variety of growth and development process in papaya including a rich sets of candidate targets related to sex chromosome. This set of genes is an important resource for understanding the genetic interactions underlying sex determination signaling and regulation and will eventually contribute to papaya breeding improvement and elucidate sex chromosome. These results and ongoing sequencing projects for additional particular species from the closely related genus such as monoecious *V. monoica*, trioecious *V. cundinamarcensis* and other dioecious species promise to reveal a greater level of functional diversity in sex-associated factors and will also facilitate to unfold the evolution of eukaryotic sex chromosome.

Acknowledgments

This project was supported by the University of Hawaii and US DoD W81XWH0520013 to M.A, Maui High Performance Computing Center to M.A., Hawaii Agriculture Research Center (HARC) to R.M and Q.Y., National Science Foundation Plant Genome Research Program grants to RM, QY, PHM (Award Nos. DBI0553417; DBI-0922545).

References

Alba R, Fei Z, Payton P, Liu Y, Moore SL, Debbie P, Cohn J, D'Ascenzo M, Gordon JS, Rose JKC, Martin G, Tanksley SD, Bouzayen M, Jahn MM, Giovannoni J (2004) ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. Plant J. 39:697-714.

Arpat A, Waugh M, Sullivan JP, Gonzales M, Frisch D, Main D, Wood T, Leslie A, Wing R, Wilkins T (2004) Functional genomics of cell elongation in developing cotton fibers. Plant Mol Biol. 54:911-929.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nat Genet. 25:25-29.

Barker G, Batley J, O' Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics.

19:421-422.

Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, Wing RA, Muehlbauer GJ, Kleinhofs A, Wise RP (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. Plant Physiol. 134:960-968.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674-3676.

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie J-M (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res. 9:950-959.

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell. 14:1457-1467.

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res. 11:1425-1433.

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.

Gschwend AR, Yu Q, Tong EJ, Zeng F, Han J, VanBuren R, Aryal R, Charlesworth D, Moore PH, Paterson AH, Ming R (2012) Rapid divergence and expansion of the X chromosome in papaya. Proc Natl Acad Sci U S A. 109:13716-13721.

Huang X (1996) An improved sequence assembly program. Genomics 33:21-31

Hughes A, Friedman R (2004) Expression patterns of duplicate genes in the developing root in *Arabidopsis thaliana*. J Mol Evol. 60: 247-256.

Kawasaki S, Borchert C, Deyholos M, Wang H, Brazille S, Kawai K, Galbraith D, Bohnert HJ (2001) Gene expression profiles during the initial phase of salt stress in rice. Plant Cell. 13:889-906.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms. Mol Ecol. 11:2453-2465.

Liu Z, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu Q, Pearl HM, Kim MS, Chartton JW, Stiles JI, Zee FT, Paterson AH, Ming R (2004) A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature. 427:348-352.

Mahalingam R, Gomez-Buitrago A, Eckardt N, Shah N, Guevara-Garcia A, Day P, Raina R, Fedoroff NV (2003) Characterizing the stress/defense transcriptome of *Arabidopsis*. Genome Biol. 4:R20.1-14.

Michalek W, Weschke W, Pleissner K-P, Graner A (2002) EST analysis in barley defines a unigene set comprising 4,000 genes. Theor Appl Genet. 104: 97-103.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ E et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* L.). Nature. 452:991-996.

Ming R, Moore PH, Zee F, Abbey CA, Ma H, Paterson AH (2001) Construction and characterization of a papaya BAC library as a foundation for molecular dissection of a tree-fruit genome. Theor Appl Genet. 102:892-899.

Moreno-Vazquez S, Ochoa OE, Faber N, Chao S, Jacobs JM, Maison-Neuve B, Kesseli RV, Michelimore RW (2003) SNP-based codominant markers for a recessive gene conferring resistance to corky root rot (*Rhizomonas suberifaciens*) in lettuce (*Lactuca sativa*). Genome. 46:1059-1069.

Na JK, Wang J, Murray JE, Gschwend AR, Zhang W, Yu Q, Navajas-Pérez R, Feltus FA, Chen C, Kubat Z, Moore PH, Jiang J, Paterson AH, Ming R (2012) Construction of physical maps for the sex-specific regions of papaya sex chromosomes. BMC Genomics. 13:176.

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol. 5:94-100.

Raju NL, Gnanesh BN, Lekha P, Jayashree B, Pande S, Hiremath PJ, Byregowda M, Singh NK, Varshney RK (2010) The first set of EST resource for gene discovery and marker development in pigeonpea (*Cajanus cajan* L.). BMC Plant Biol. 10:45

Ronning CM, Stegalkina SS, Ascenzi RA, Bougri O, Hart AL, Utterbach TR, Vanaken SE, Riedmuller SB, White JA, Cho J, Pertea GM, Lee Y, Karamycheva S, Sultana R, Tsai J, Quackenbush J, Griffiths HM, Restrepo S, Smart CD, Fry WE, Van Der Hoeven R, Tanksley S, Zhang P, Jin H, Yamamoto ML, Baker BJ, Buell CR (2003) Comparative analyses of potato expressed sequence tag libraries. Plant Physiol. 131:419-429.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle J, Shoemaker R (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome. 47:868-876.

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 106:411-422.

Varshney RK, Dubey A (2009a) Novel genomic tools and modern genetic and breeding approaches for crop improvement. J Plant Biochem Biotechnol. 18:127-138.

Varshney RK, Graner A, Sorrells E (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23:48-55.

Varshney RK, Hoisington DA, Tyagi AK (2006) Advances in cereal genomics applications in crop breeding. Trends Biotechnol. 11:490-499.

Varshney RK, Nayak SN, May GD, Jackson SA (2009b) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 27:522-530.

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett. 7:537-546.

Wang J, Na JK, Yu Q, Gschwend AR, Han J, Zeng F, Aryal R, VanBuren R, Murray JE, Zhang W, Navajas-Pérez R, Feltus FA, Lemke C, Tong EJ, Chen C, Wai CM, Singh R, Wang ML, Min XJ, Alam M, Charlesworth D, Moore PH, Jiang J, Paterson AH, Ming R (2012) Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci U S A. 109:13710-13715.

Wisman E, Ohlrogge J (2000) *Arabidopsis* microarray service facilities. Plant Physiol. 124:1468-1471.

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, Wang J (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 34:293-297.

Yu Q, Tong E, Skelton RL, Bowers JE, Jones MR, Murray JE, Hou S, Guan P, Acob RA, Luo MC, Moore PH, Alam M, Paterson AH, Ming R (2009) A physical map of the papaya genome with integrated genetic map and genome sequence. BMC Genomics. 10:371.

Zeng F, Zhang X, Cheng L, Hu L, Zhu L, Cao J, Guo X (2007) A draft gene regulatory network for cellular totipotency reprogramming during plant somatic embryogenesis. Genomics. 90:620-628.

Zeng F, Zhang X, Zhu L, Tu L, Guo X, Nie Y (2006) Isolation and characterization of genes associated to cotton somatic embryogenesis by suppression subtractive hybridization and macroarray. Plant Mol Biol. 60:167-183.