Protein Cell 2016, 7(5):373–382
DOI 10.1007/s13238-016-0260-y

Protein & Cell

RESEARCH ARTICLE

Transcriptome analyses of insect cells to facilitate baculovirus-insect expression

Kai Yu[1,2], Yang Yu[1,2], Xiaoyan Tang[1,2], Huimin Chen[1,2], Junyu Xiao[2,3]✉, Xiao-Dong Su[1,2]✉

[1] Biodynamic Optical Imaging Center, School of Life Science, Peking University, Beijing 100871, China
[2] State Key Laboratory of Protein and Plant Gene Research, Peking University, Beijing 100871, China
[3] Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China
✉ Correspondence: junyuxiao@pku.edu.cn (J. Xiao), xdsu@pku.edu.cn (X.-D. Su)

Received January 31, 2016 Accepted February 27, 2016

ABSTRACT

The High Five cell line (BTI-TN-5B1-4) isolated from the cabbage looper, *Trichoplusia ni* is an insect cell line widely used for baculovirus-mediated recombinant protein expression. Despite its widespread application in industry and academic laboratories, the genomic background of this cell line remains unclear. Here we sequenced the transcriptome of High Five cells and assembled 25,234 transcripts. Codon usage analysis showed that High Five cells have a robust codon usage capacity and therefore suit for expressing proteins of both eukaryotic- and prokaryotic-origin. Genes involved in glycosylation were profiled in our study, providing guidance for engineering glycosylated proteins in the insect cells. We also predicted signal peptides for transcripts with high expression abundance in both High Five and Sf21 cell lines, and these results have important implications for optimizing the expression level of some secretory and membrane proteins.

KEYWORDS High Five cell line, baculovirus-insect cell system, codon usage, glycosylation, signal peptide

INTRODUCTION

The baculovirus-insect cell expression system is one of the most popular platforms for recombinant protein expression. It is widely used for protein structure and function studies in academic laboratories, and facilitates massive protein production in industry (Kost et al. 2005). The two common cell lines in this binary system are Sf21 (IPLB-Sf21AE) from *Spodoptera frugiperda* (Vaughn et al. 1977), and High Five (BTI-TN-5B1-4) from ovarian tissues of *Trichoplusia ni* (cabbage looper) (Wickham et al.; Davis et al. 1992).

Protein expression in insect cells has several advantages such as high expression level and easy manipulation. In addition, difficult proteins especially eukaryotic proteins that need posttranslational processing usually fold better in insect cells than in the *E. coli* expression system (Brondyk 2009). However, compared to the mammalian cells, post-translational modifications are still limited in insect cells, with glycosylation as the most significant example (Jarvis 2003; Kost et al. 2005). Due to the defect in glycosylation, functions of some recombinant glycoproteins are impaired (Xu and Ng 2015). For example, the insect cells cannot produce sialylated N-linked glycans. In the past two decades, various efforts were made to import the mammalian glycosylation pathway related genes into the insect cells to engineer the required glycosylation modification (Castilho 2015). For example, Hollister first reported in 1998 that an engineered Sf9 cell line expressing the B4GALT1 gene could produce foreign glycoproteins with terminally galactosylated N-glycans (Hollister et al. 1998). It was also reported in 2001 that both Sf9 and High Five cells were engineered to produce sialylated proteins by adding the ST6GAL1 gene (Hollister and Jarvis 2001; Breitbach and Jarvis 2001). Hollister et al. (Hollister et al. 2002) later transformed a set of other genes to generate the SfSWT-1 cell line which produce biantennary, terminally α-2,6- and α-2,3- sialylated N-glycans. More work has been recently done to obtain more powerful insect cell lines.

Next generation sequencing technology is widely used recently in biological studies. Genomes and transcriptomes of different species are sequenced, which generate high-input information for genomic studies and molecular modifications. We believe similar information can also be explored to provide

Electronic supplementary material The online version of this article (doi:10.1007/s13238-016-0260-y) contains supplementary material, which is available to authorized users.

guidance to engineer new strains of insect cell line for expressing proteins with mammalian type of posttranslational modifications. Genome and transcriptome of Sf21 cell line have already been reported in 2014 and 2015, respectively (Kakumani et al. 2014; Kakumani et al. 2015); however, the global genomic information of the High Five cell line is still unknown.

We constructed and sequenced an mRNA library of the High Five cell line, assembled a reference transcriptome for function and expression studies. We analyzed some protein-expression-related problems by comparing our High Five transcriptome with the reported Sf21 transcriptome (Kakumani et al. 2015). In addition, we extracted codon usage information from their coding sequences and compared it with other expression systems and model species. We also annotated transcripts that may have glycosylation-related functions, and evaluated their expression abundance to generate the global view of glycogenes in High Five and Sf21 cell lines. High expression transcripts, which have predicted signal peptide sequences, were analyzed for predicting highly efficient signal peptide sequences for secretory protein expression.

RESULTS AND DISCUSSION

Reference transcriptome assembly

Considering the genome size of several reference-ready insects, a total 49.5 million 101 bp paired-end reads were sequenced and yield 4.95 Gb bases raw data, 48 million clean reads were kept after low quality reads were filtered. After reads trimming with Trimmomatic, we used Trinity pipeline to do the *de novo* transcriptome assembly and obtained 31,068 transcripts with an N50 value of 2,276 bp (Haas et al. 2013; Bolger et al. 2014).

In total, 39.4 Mb bases are assembled and the average transcript length is 1,269 bp. To reduce the redundancy of the assembly, cd-hit-est was used and transcript number was reduced to 25,234 under 90% sequence identity threshold. These transcripts data sets are the so-called ‘unigene’. 13,732 coding peptide sequences are predicted with Trans-Decoder. Detailed statistics numbers are shown in Table 1.

All clean reads have been submitted to NCBI SRA database under accession number SRP068276. Assembly version in this paper has been submitted to NCBI TSA database (GEEM01000000).

Assembly assessment

To evaluate the assembly quality, we employed several strategies for quality assessment. First, we used bwa (v0.7.10) (Li and Durbin 2010) to align all clean reads back to the assembly. 97.4% of the reads could be aligned and 95.3% are properly paired, indicating that the completeness of our assembly is high and very reliable. All transcripts were then aligned to SwissProt to check the proportion of transcripts that may be full-length or near full-length. As shown in Fig. 1A, about 6000 transcripts, with more than 10% of its contig length, could be aligned to a homolog in SwissProt. Among them, about 1/3 of the transcripts are fully aligned and more than 5000 have at least 30% sequence overlapping with known homolog. The result could be underestimated as the homologs of different species’ genes have different proportion in the SwissProt database, but we believe our data reached our expectation.

Taking expression values into consideration, we recalculated the N50 value after low expression contigs were eliminated (Fig. 1B) and plotted expression value distribution pattern in Fig. 1C. Ex in Fig. 1B means a subset of top x% highly expressed transcript, the ExN50 reached the max length at E95, showing that 8,147 transcripts are in the top 95% expression subset with the minimum TPM of 6.1. From this, we can conclude from these data that most of the extremely high expression transcripts in High Five cells are in the range from 600 to 1000 bp. Longer transcripts have more regular expression level. As shown in Fig. 1C, the transcript number was reduced to 10,047 after transcripts of low TPM values (below 5) were eliminated.

Function annotation

To annotate functions of transcripts and coding peptides, we searched homologous genes in SwissProt, TrEMBL90 and NCBI nr databases with blast. Among the 25,234 transcripts, 13,492 got blast hits in nr database, 9,639 and 13,767 have similar sequences in SwissProt and TrEMBL90 database, respectively. With 13,732 coding peptides, 9,427 and 12,198 got alignments in SwissProt and TrEMBL90 database, respectively.

While executing GO annotation and EggNOG annotation, we also used protein sequence predicted from published

Table 1. Assembly statistics information

	Raw assembly	Duplicate removed assembly
Trinity ‘genes’	27,389	24,000
Trinity transcripts	31,068	25,234
GC content (%)	40.84	40.71
Median contig length (bp)	722	622
Average contig length (bp)	1269.6	1160.9
Total assembled bases (bp)	39,444,068	29,294,166

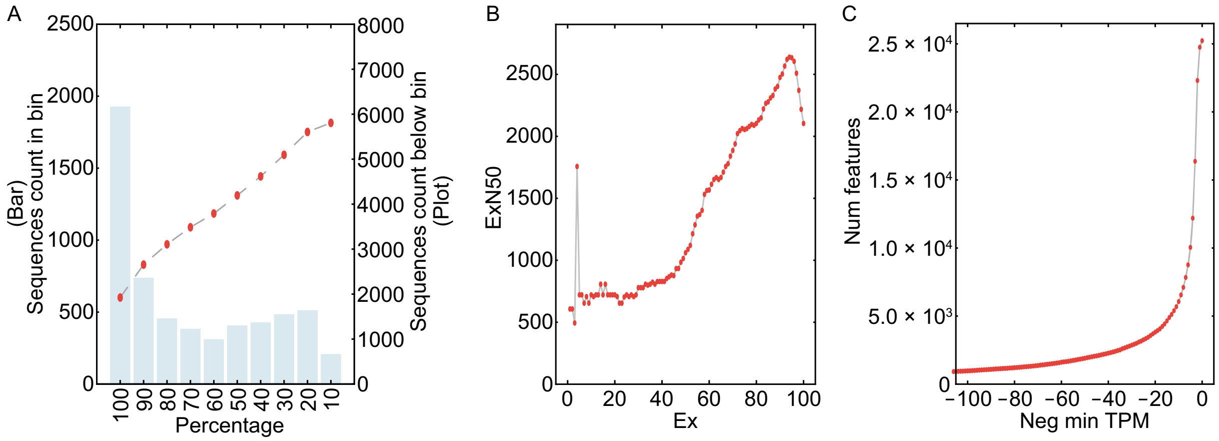


Figure 1. Assembly quality assessment. (A) Full-length transcript assessment. Bin on x-axis represent the percentage of the hit's length included in the alignment to the Trinity transcript. Left y-axis with bar plot is the transcript count in each bin and right y-axis with point plot is the accumulate count below that bin. (B) N50 of subset of transcript by decreasing the expression level. Ex is the top most expressed transcripts that represent x% of the data. ExN50 is the length of a transcript while the total length of transcripts shorter that it reached 50% of total length of all transcripts in this dataset. (C) Transcript count with a threshold of negative minimum TPM value.

Sf21 transcriptome data with the same analysis procedural. In High Five transcriptome, 13,447 transcripts have been annotated for GO terms and 13,459 have been annotated in EggNOG database. More specific function classification data are shown in Fig. 2 and Fig. 3, and detailed annotations for each transcript are described in supplemental file S1. From the global pattern of these figures, we can tell that the differences in GO terms and EggNOG categories between High Five and Sf21 cells are quite similar. Transcript numbers are higher for some function classes, including those related to intracellular trafficking, secretion, vesicular transport, posttranslational modification, protein turnover, transcription, translation, etc. Possessing plentiful genes with these functions make these insect cell lines an ideal host for protein production.

Codon usage among recombinant protein expression systems

Among different organisms, synonymous codons are usually utilized with different frequencies; a phenomenon generally referred to as the ‘codon usage bias’. Codon usage bias is a major factor that affects the level of recombinant protein expression (Holm 1986). We collected CDS sequences from the Sf21 and High Five cells and compared their codon usage preference to other commonly used expression hosts, including the prokaryotic system *E. coli* BL21, the eukaryotic expression system *S. cerevisiae* and mammalian system CHO. We also compared the codon preference of insect cells with five other species including human, mouse, drosophila, zebrafish and arabidopsis. RSCU (relative synonymous codon usage) values are used to compare the use of synonymous codons. RSCU values of all amino acids in 10 species’ CDS sequence were calculated separately for downstream studies as shown in Fig. 4.

With the characteristic of RSCU value, all x synonymous codon RSCU values of an amino acid always get a sum up value equals to x. For amino acids only encoded by one or two codons, there is no extreme distribution in it, but others are quite different. We calculated the RSCU range value of a set of codon related to one amino acid. By comparing this value, we can tell in which species this set of codon have greater bias. For example, the range value of arginine, isoleucine, leucine and proline in BL21 reached 2.29, 1.32, 2.81 and 1.63, while the minimum codon’s RSCU is only 0.13, 0.21, 0.21 and 0.49. This situation exactly indicates that optimization of codon usage is of great significance. For instance, if BL21 is used to produce recombinant protein, failure to avoid these minimum codons may dramatically reduce the expression level.

In comparison with the range and standard deviation values in all species, no matter which amino acid you are using, both High Five and Sf21’s range value are at a relatively low level. The homogeneity of codon usage in baculovirus-insect system could be an advantage for protein expression. Coding sequence cloned from most species could be normally expressed in insect cells without codon optimization. This robust property of codon usage in baculovirus-insect system made it a good platform for both eukaryotic and prokaryotic recombinant protein expression. But some previous publications claim that codon with lower RSCU value is intended to slow down the translation speed in order to produce well-folded proteins (Chaney and Clark 2015). More experimental evidence is required to show whether the usage of codon with relative higher RSCU values could become a disadvantage for expression of proteins of complex folding.

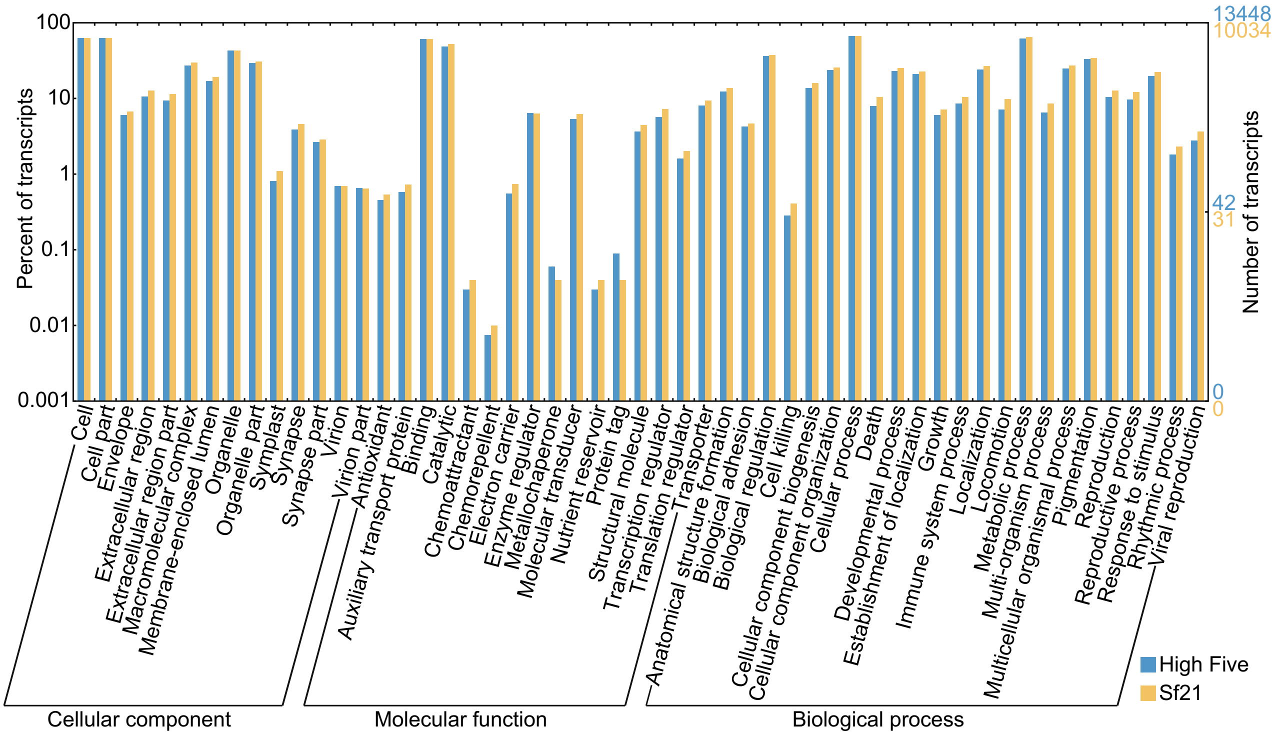


Figure 2. Gene Ontology of High Five and Sf21 transcriptome. Summarized in three main GO categories: Cellular component, Molecular function and Biological process. Right y-axis is the transcript count in that function item, left y-axis is the corresponding percentage of transcripts number.

Glycogene profile in protein expression insect cells

Post-translational modification is one of the most important characteristics of baculovirus-insect expression system. But the truncated N-glycosylation pathways in insect cells limit its application on some glycoprotein expression (Jarvis 2003). Several glyco-engineering modifications have been reported in the past two decades. Some modifications require importation of glycogenes into baculovirus-insect system. Glycosylation is mediated with complicated pathways and a number of genes are involved. Without a global gene map of the insect cells, we cannot thoroughly understand glycosylation related problems. Since GGDB and CAZy databases included genes associated with glycan synthesis procedural, we used them as references to find homologs of glycogenes in our High Five transcriptome and previous Sf21 transcriptome.

Here we identified 69 glycogenes in the High Five transcriptome and 72 in Sf21, with an overlap of 66 genes. Those genes are marked in blue with their expression value in Fig. 5A, and detailed informations are described in supplemental file S2. Glycogenes can be classified into several types according to their functions. Total gene counts of each type in High Five and Sf21 are shown in Fig. 5B. Insect cells have more or less homologs among most types. But for sialyltransferases and N-acetylgalactosaminyltransferases, no similar transcript was found in these two cell lines. That is the main reason why baculovirus-insect cannot produce complete mammalian N-glycosylation proteins. O-glycan modification is more complicated and not well studied. From the gene matrix, we found that some O-glycosylation required enzymes are detectable, such as OGT, POFUT1/2, XYLT1, POMT2, etc. Previous study suggests that O-mannosylation in insect species may occur more frequently than what is currently believed (Vandenborre et al. 2011). Understanding the profile of glycogenes in insect cells would be helpful for more detailed research on glycosylation.

We also compared the glycogenes' functioning mechanism and structure status in High Five and Sf21 (Fig. 5C). More than 50% of the glycosyltransferases function as inverting mechanism by catalyzing group transfer with inversion at anomeric reaction center of substrate, and less than 30% are for retentions. About 30% glycosyltransferase consist of two closely abutting β/α/β Rossmann domains, 20% consist of two β/α/β Rossmann domains that face each other and is flexibly linked. Remaining part have not yet been experimentally determined or studied (Lairson et al. 2008).

Here we complemented the glycogene database with data from baculovirus-insect expression system related cell lines. Our data would be valuable for introducing supplemental mammalian glycogesnes into insect cell lines and more efficiently modifying their glycosylation properties.

Highly expressed signal peptide containing genes

Baculovirus-insect system is a good platform for secretory protein expression. When there is a signal peptide fused to

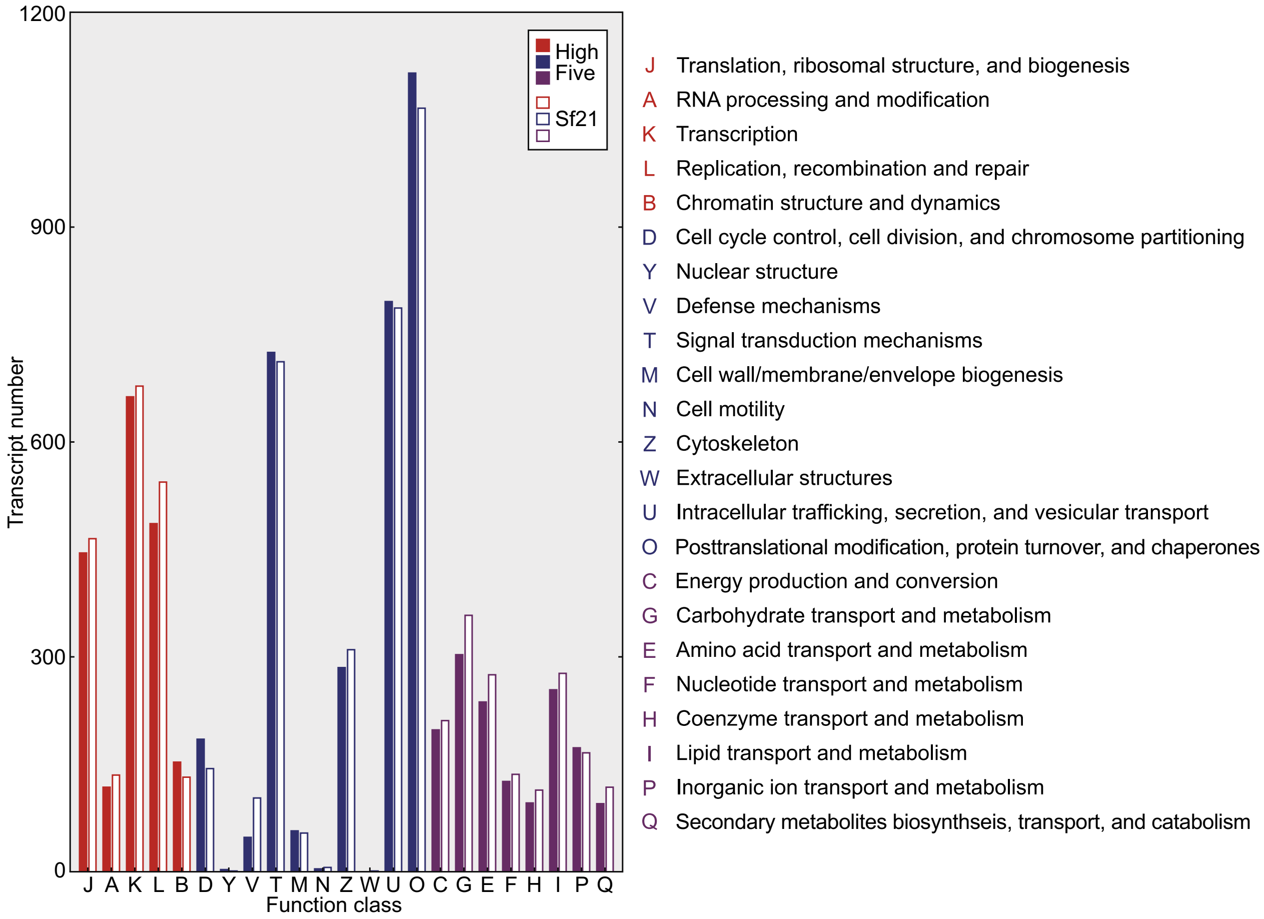


Figure 3. Transcript number in each EggNOG function classes. Divided into 3 parts by colors. (Red) Information storage and processing; (Blue) Cellular processes and signaling; (Purple) Metabolism.

the recombinant protein, product usually secreted to the outside environment through the secretory pathway. Optimization of signal peptide would be helpful to get a better yield (Olczak and Olczak 2006). Here we used SignalP software (Petersen et al. 2011) to predict all possible signal peptide in all protein sequences, and then sorted them with their transcript expression value. Fig. 6 showed all predicted transcripts that may have signal peptide sequence in High Five and Sf21 transcriptome. We identified signal peptide sequences from top 100 expressed transcripts and believe they are good candidates for higher protein production. Related peptide sequence, CDS sequence and functional annotation are described in supplementary file S3. Because higher expression value usually means higher protein amount, secretion efficiency could be closely linked with the amount of the signal peptide containing protein. Moreover, signal peptide sequence from insect itself is of the best choice because endogenous secretory signal peptide is more efficient than exogenous signal peptides (von Heijne and Abrahmsén 1989; Soejima et al. 2013). Combined with expression value and corresponding function of these proteins, we believe this approach would be useful and simplify the complexity of finding good signal peptides for protein expression.

MATERIALS AND METHODS

RNA extraction and library construction

High Five cell line in this study was purchased from Life technologies, USA. Cells were cultured in serum free medium containing 0.5% Penicillin-Streptomycin (Gibco 15140-122) for 24 h at 27°C in suspension at a shaking speed of 110 rpm. When cell density reached 2.4×10^6/mL, 2×10^7 cells were collected by centrifugation. Total RNA was extracted with QIAGEN RNeasy Mini Kit (QIAGEN, 74104) immediately after cell collection. RNA quality was examined using Agilent 2100 RNA chip (standard setting). Poly-A tailed RNA was enriched and used to construct sequencing library with Illumina TruSeq RNA Sample Prep Kit (Illumina, RS-122-2001), following standard instruction. RNA-seq library was sequenced on Illumina HiSeq 2000 platform.

Assembly and statistics

After we got the raw reads, quality control was performed to remove poor-quality reads with in-house QC script. Then we used Trimmomatic (v0.32) (Bolger et al. 2014) software to trim low quality ends of all reads. Reference transcriptome was assembled with trimmed reads using Trinity (v2.0.6) (Grabherr et al. 2011; Haas et al. 2013).

To reduce the assembly redundancy, cd-hit-est (v4.6.1) (Fu et al. 2012) was used to cluster reads with 90% identity. At this time, the remaining contigs are the so called ‘unigenes’. Coding peptide sequences were predicted with TransDecoder.

We used the SwissProt database to check the integrity of transcripts to evaluate the quality of our assembly result. All transcript sequences were aligned to SwissProt using blastx (Camacho et al. 2009), and only the most similar target was kept with the e-value cutoff of 1 × 10^{-20}. Length coverage of aligned transcript was examined (Fig. 1A).

Function annotation

Function annotation was performed at both transcript and protein level. All transcripts were aligned to SwissProt, TrEMBL90 and NCBI nr database with blastx (Camacho et al. 2009; UniProt Consortium 2015). Predicted protein sequences were aligned to SwissProt, TrEMBL90 and EggNOG 4.1 with blastp (Powell et al. 2014). We also aligned protein sequences to Pfam28 database with hmmsearch (Finn et al. 2014). Subsequently, we annotated Gene Ontology (GO) and related pathways with blast2go and KOBAS (Conesa and Götz 2008; Xie et al. 2011). The GO annotation is presented in figure with WEGO (Ye et al. 2006). We also used SignalP (v4.1) (Petersen et al. 2011) for signal peptide prediction.

Figure 5. Glycogene profile of High Five and Sf21 cell line. (A) Heatmap represents the gene constituent of each species. Blue mark of Sf21 and High Five showed the expression value of each gene. Red mark only represent they have this gene. (B) Bar plot of glycogene categories. (C) Ring plot representing the properties of glycogenes in High Five and Sf21 cell lines.

▶

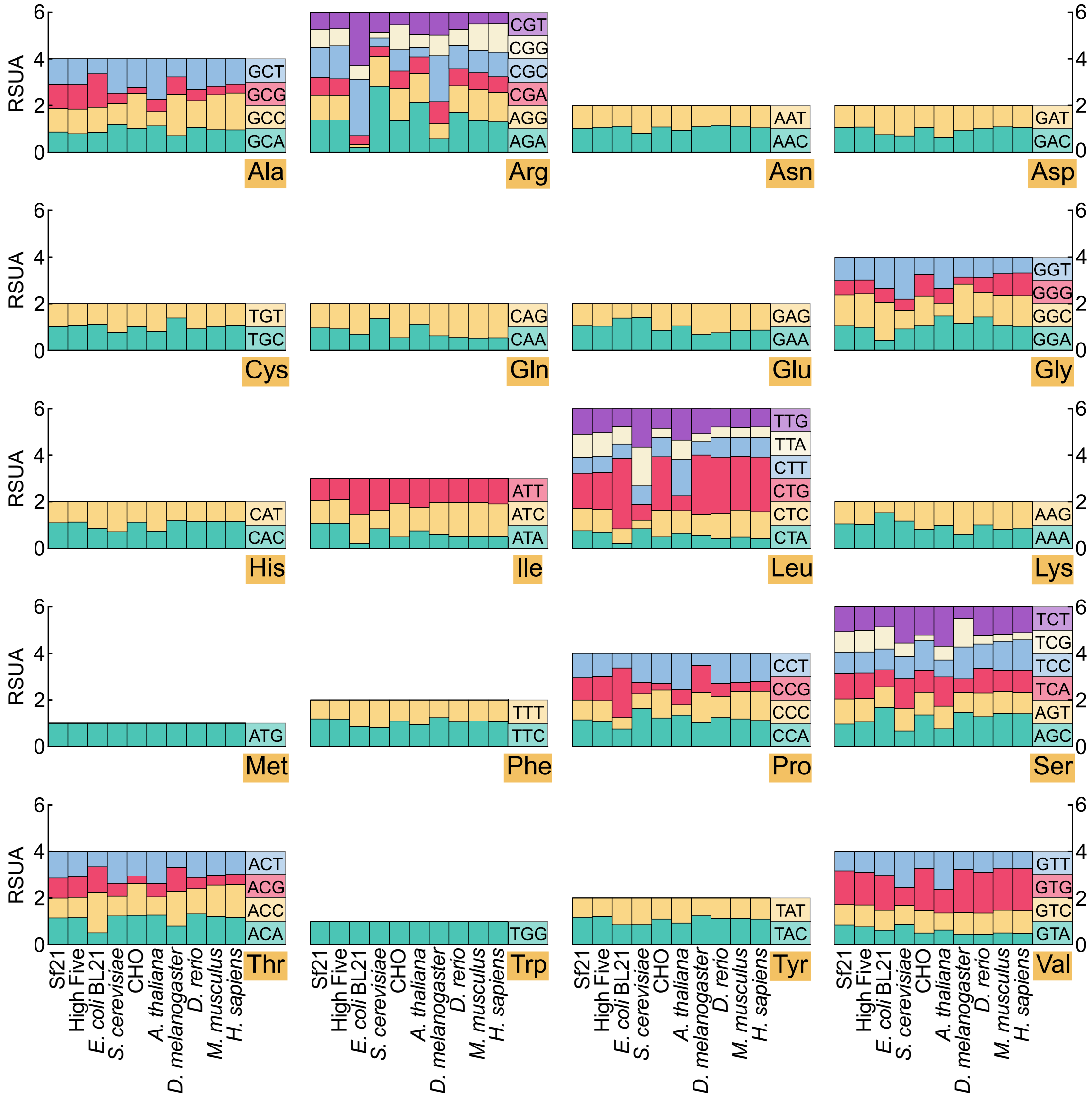
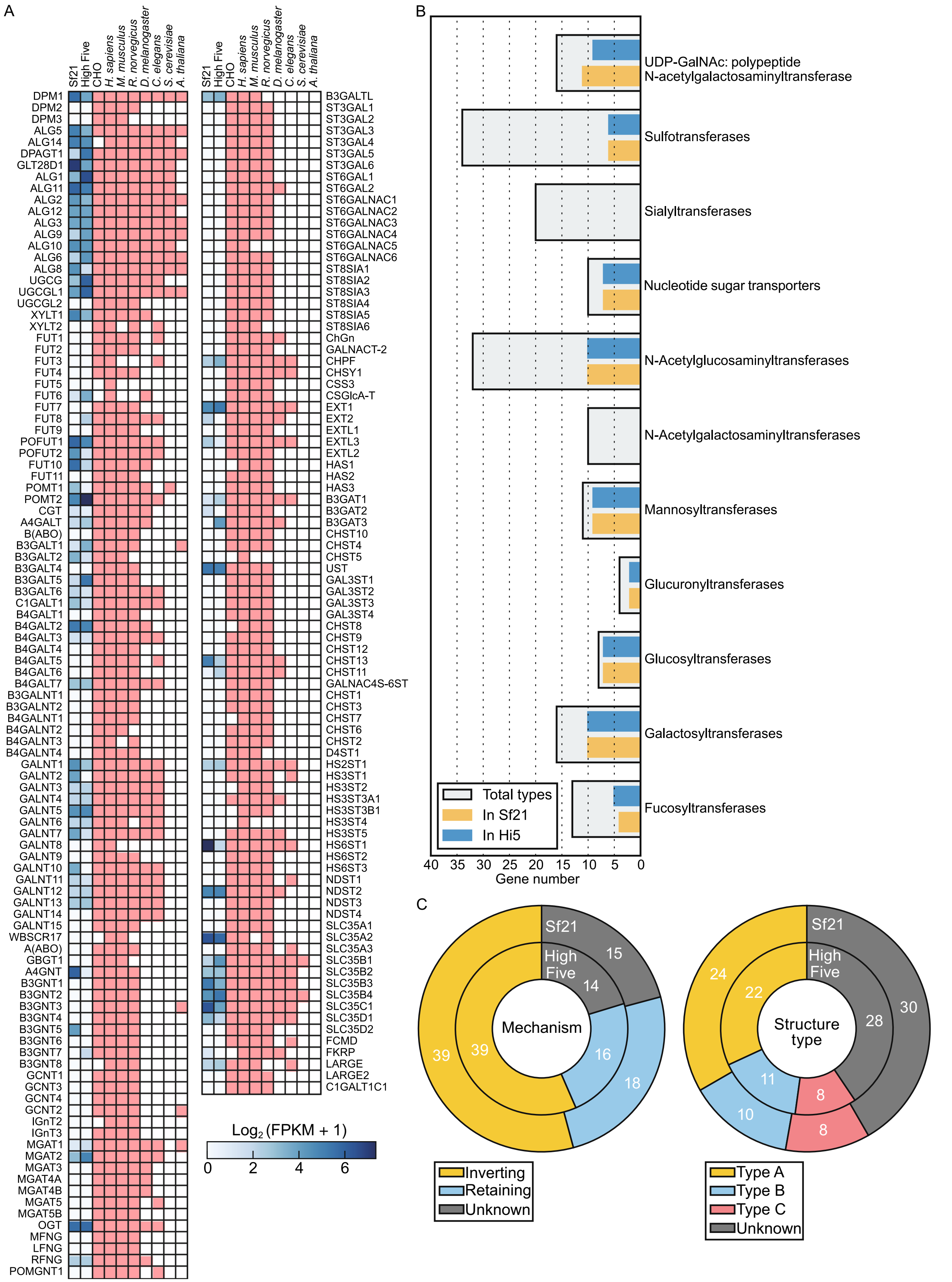
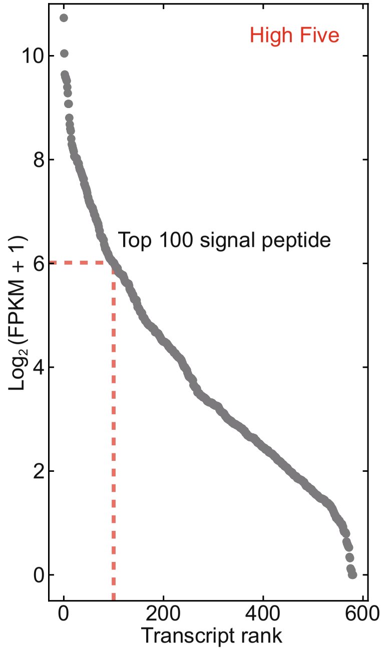
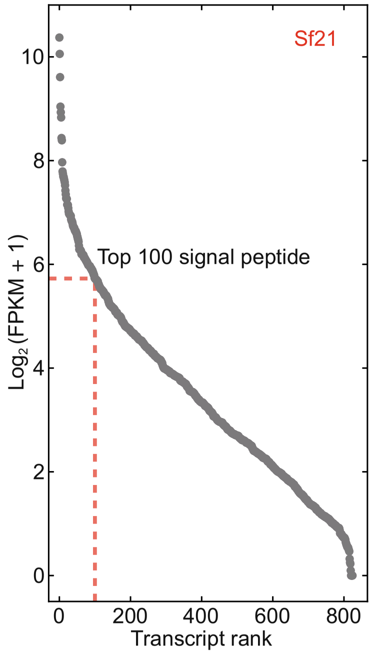


Figure 4. Codon usage of 20 amino acids across 10 different species. Each subplot is an amino acid, x-axis is different species and y-axis is the RSCU value of codons coding that amino acid.


A
Sf21
High Five
CHO
H. sapiens
M. musculus
R. norvegicus
D. melanogaster
C. elegans
S. cerevisiae
A. thaliana
DPM1
DPM2
DPM3
ALG5
ALG14
DPAGT1
GLT28D1
ALG1
ALG11
ALG2
ALG12
ALG3
ALG9
ALG10
ALG6
ALG8
UGCG
UGCGL1
UGCGL2
XYLT1
XYLT2
FUT1
FUT2
FUT3
FUT4
FUT5
FUT6
FUT7
FUT8
FUT9
POFUT1
POFUT2
FUT10
FUT11
POMT1
POMT2
CGT
A4GALT
B(ABO)
B3GALT1
B3GALT2
B3GALT4
B3GALT5
B3GALT6
C1GALT1
B4GALT1
B4GALT2
B4GALT3
B4GALT4
B4GALT5
B4GALT6
B4GALT7
B3GALNT1
B3GALNT2
B4GALNT1
B4GALNT2
B4GALNT3
B4GALNT4
GALNT1
GALNT2
GALNT3
GALNT4
GALNT5
GALNT6
GALNT7
GALNT8
GALNT9
GALNT10
GALNT11
GALNT12
GALNT13
GALNT14
GALNT15
WBSCR17
A(ABO)
GBGT1
A4GNT
B3GNT1
B3GNT2
B3GNT3
B3GNT4
B3GNT5
B3GNT6
B3GNT7
B3GNT8
GCNT1
GCNT3
GCNT4
GCNT2
IGnT2
IGnT3
MGAT1
MGAT2
MGAT3
MGAT4A
MGAT4B
MGAT5
MGAT5B
OGT
MFNG
LFNG
RFNG
POMGNT1
B3GALTL
ST3GAL1
ST3GAL2
ST3GAL3
ST3GAL4
ST3GAL5
ST3GAL6
ST6GAL1
ST6GAL2
ST6GALNAC1
ST6GALNAC2
ST6GALNAC3
ST6GALNAC4
ST6GALNAC5
ST6GALNAC6
ST8SIA1
ST8SIA2
ST8SIA3
ST8SIA4
ST8SIA5
ST8SIA6
ChGn
GALNACT-2
CHPF
CHSY1
CSS3
CSGlcA-T
EXT1
EXT2
EXTL1
EXTL3
EXTL2
HAS1
HAS2
HAS3
B3GAT1
B3GAT2
B3GAT3
CHST10
CHST4
CHST5
UST
GAL3ST1
GAL3ST2
GAL3ST3
GAL3ST4
CHST8
CHST9
CHST12
CHST13
CHST11
GALNAC4S-6ST
CHST1
CHST3
CHST7
CHST6
CHST2
D4ST1
HS2ST1
HS3ST1
HS3ST2
HS3ST3A1
HS3ST3B1
HS3ST4
HS3ST5
HS6ST1
HS6ST2
HS6ST3
NDST1
NDST2
NDST3
NDST4
SLC35A1
SLC35A2
SLC35A3
SLC35B1
SLC35B2
SLC35B3
SLC35B4
SLC35C1
SLC35D1
SLC35D2
FCMD
FKRP
LARGE
LARGE2
C1GALT1C1
Log2 (FPKM + 1)
0
2
4
6
B
UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase
Sulfotransferases
Sialyltransferases
Nucleotide sugar transporters
N-Acetylglucosaminyltransferases
N-Acetylgalactosaminyltransferases
Mannosyltransferases
Glucuronyltransferases
Glucosyltransferases
Galactosyltransferases
Fucosyltransferases
Total types
In Sf21
In Hi5
40
35
30
25
20
15
10
5
0
Gene number
C
Sf21
High Five
Mechanism
15
14
16
18
39
39
Inverting
Retaining
Unknown
Sf21
High Five
Structure type
24
22
30
28
11
10
8
8
Type A
Type B
Type C
Unknown



01. (0.53) MIATLLYLLIHLRRA
02. (0.84) MRLFIFALLTVCAHA
03. (0.83) MLKMRWSILALACVVVAVYA
04. (0.59) MIFLLYNLHLALN
05. (0.86) MKSVAVFLCLVAAASA
06. (0.81) MWSKFVVMLAVSCVANA
07. (0.88) MFEFRILLCSLLIISCTIA
08. (0.81) MKSLVLAVVSLLAISSINC
09. (0.89) MKVILFTAIALLGAVLA
10. (0.69) MKGVIFVCLLSLSVSFSLP
11. (0.88) MSRLRFCLLLAVFCSCTQYGLG
12. (0.86) MMTSVSLLVALLVSLQCVFQTSA
13. (0.82) MKITLILGIIASFSLALA
14. (0.77) MGAFAAALGFILLVAAARS
15. (0.67) MYLVILFAATALLGSSRS
16. (0.60) MKVWGAHGEFCARHQWEVIVATLALLACAAS
17. (0.84) MVMKYVICLSALAAVAFA
18. (0.85) MYLRALLLLSLSTLCLS
19. (0.68) MEGFKALILLQLVYLCGC
20. (0.58) MKQIIFCVIVALSSVIA
21. (0.89) MAPLFLIFAVFISATLG
22. (0.67) MKLIYLTLLVLPAALLCFQSANT
23. (0.79) MCRTECVLALVALSFGVAHS
24. (0.86) MKRGLLVLLAMCSLACA
25. (0.89) MELFRISSLFALFLILDVTSA
26. (0.92) MKMLILAVSCLLAITGSLA
27. (0.77) MSGPARARFLALHLVPATLLLLLRNPTPTSA
28. (0.87) MILNIKMFAVLSVILLPLFLSQQTEA
29. (0.85) MGRSLAVLGCLCALAALAHS
30. (0.74) MLLAFGRSIGLIFFVCLLSCVVS
31. (0.90) MKYFILLAVGVLLLSGWSHA
32. (0.87) MAVTMALFWVSLVLSVIGANA
33. (0.87) MQKLRVFLLALAALCISCEA
34. (0.71) MLNGKFLLCVVLVAAWNEVQP
35. (0.68) MEVHYLVFVLCAIWQGTNA
36. (0.77) MAGAGGFSRRFLHVFLDCLLLLQLAEYVPTASA
37. (0.81) MKSWGLLLVAALAIGSSFG
38. (0.81) MRLHMIAVLTLLGVVNYAAG
39. (0.87) MFRMRLPFVLVIILMIHVSVS
40. (0.83) MKSSWRTFSFAILFAASLLNTAHS
41. (0.55) MAPVPSLFITLILSIVLSVNA
42. (0.82) MIRSRGDTMSRTLARLLLALSIVAQCRA
43. (0.84) MSVQRVRTFRVLLVVLAFVNTSRA
44. (0.92) MGLLNIVLFLVLVAACCA
45. (0.50) MARVHTVPQFLHFFLYFSFVLCCFTSFVSA
46. (0.83) MSSIKLNLVLCSVILSILVCVNT
47. (0.84) MIYSTWISQFYSVLLLICSVITYVQS
48. (0.76) MFAYLNALIVLCFCFKNVAS
49. (0.87) MSAKLLSLILIIATAVTA
50. (0.59) MALAHGGQLVLVALALALLTLG
51. (0.90) MSLYVVCALLLLTPIRA
52. (0.81) MDSQVSVVIIAFLCILGTNA
53. (0.84) MELKTVNLFLLIALSICFVHNAIA
54. (0.92) MYKLNVFHLLFLATIVQVALA
55. (0.89) MNSILLVFAGILAVCLPASA
56. (0.76) MLRSLFVALAVSAVFA
57. (0.68) MKQLCIVVALVSVFGLSFQ
58. (0.75) MESLSRIVLGILILQIYVHA
59. (0.78) MCYYIIFVVFSIFVTDCTG
60. (0.79) MKFKLLLFLTIILTYYEGEA
61. (0.83) MAKIVFLLVLLCISASFA
62. (0.83) MTLRCVFAVLALAGATFA
63. (0.78) MFNSYFLGILICVAGSATA
64. (0.68) MSLKMSSVALLAYFTLNLIYCNA
65. (0.59) MMSRIYCSIFLLGLLLPLTTA
66. (0.83) MHIYSQTLVFVLLIVSCVLC
67. (0.77) MGARSRELVLLLAVCALALSEPVSKA
68. (0.76) MKQLLSIFFSVSTFSYVLS
69. (0.69) MMMSFSTNFCISLSFLTIVITCIVLSNDNYAQA
70. (0.61) MARVPGSVYVLLLCSFVTETVLS
71. (0.63) MFKWAVFAVLALGSVLGPMREVEA
72. (0.80) MRSLILIFLVLYLGYEVHG
73. (0.93) MILRMMSFCVFLFFLFSFNQVLC
74. (0.62) MLRLLVLTAITAA
75. (0.86) MFSPRRTILLGVILACTIIVVLG
76. (0.85) MKTILLVVTQILILASVES
77. (0.59) MAVLADINRIFSVLLSIIVITRCNA
78. (0.52) MHVAVVLVLMALVCVEG
79. (0.79) MEVKAWVILTLTALSIHGSCG
80. (0.78) MLFPKTALFYGFLIALLLEPISG
81. (0.53) MVDFVASVVLVALVLELAG
82. (0.70) MRGSWVICLCVVKLALC
83. (0.87) MLLKTLVFSLLAVAVLA
84. (0.84) MMLRILFILIVTIVKVTHS
85. (0.86) MYKIIVIALFITLVHG
86. (0.80) MSILIKMFFRVICLYLLFTTQISC
87. (0.56) MGKLLIVFLAFAATCVA
88. (0.79) MYLNNVISSVLLLIATSFSITSC
89. (0.60) MFRKVVLGSVGVAALIPAIGAA
90. (0.84) MFFSPFIITILLSLVNGILS
91. (0.87) MAGMLWALLILQLFFNAEG
92. (0.51) MNSLLRMFCISCLRSRCLS
93. (0.75) MGQARYIACALLMLCPLALT
94. (0.71) MEHALSFLLSLLLFSLLLQLLCT
95. (0.75) MKYHLLSIILVITLSISCES
96. (0.60) MNKYIYCFVLSAVLASVRA
97. (0.85) MYCFIVLCLVLILLQMNINSSSC
98. (0.59) MAFCRVVFPLLVLSIVSCSA
99. (0.84) MRTMASVWFSVLVLVSVQQVFG
100. (0.66) MLEVINGFLLVYFVVLCSLSLLVPQLVKPISA



01. (0.50) MRYSLTYCSALLKISFLFLMA
02. (0.89) MKGVAVLLCLVAGACA
03. (0.48) MMIVSKRAAWLLLLTTRGYNRCCA
04. (0.48) MLLLLSQLLGQVVKG
05. (0.78) MNMFEPRAGRVVLLFIITLQLFEFCAS
06. (0.82) MSTVLLAVIAVILCVLFRILNVNS
07. (0.84) MSGPARARFIALQLVPATLLLLSWTPTPSSA
08. (0.92) MKVIILAVSCLLAVSGALA
09. (0.79) MFSAAIFMCFATFFILSSNA
10. (0.65) MSKSVLCFVIAVLCFASVNT
11. (0.81) MKSLVLAVVSLLAISSINC
12. (0.82) MIKMRWSILALACVVVAVYA
13. (0.73) MRSVSVVLILVVAAMASS
14. (0.55) MHVIIMDGWSHLLLVCNVGTQLAYI
15. (0.71) MGERSRELILLLAVCALALS
16. (0.82) MVTKYAICLFAIATVAFA
17. (0.88) MKAILFALAGLLAVCLPASA
18. (0.73) MSDWYWRVIALKATFALTFIFCNVRA
19. (0.80) MDFRILLLWMFAVVTRNYA
20. (0.89) MLGSLKIVLLLGVIYLCRA
21. (0.88) MDLKFIFCLFLAFISSASS
22. (0.69) MLISKQSASVSRALNFGKMKCSALLLVLMLCVLSDA
23. (0.90) MAWIYLILSVFISTAVA
24. (0.83) MAGAGGILRRFLLVFLDCLLLLQLAQY
25. (0.90) MRGRAAQIAALVILCSLLGTTHA
26. (0.51) MKHWTAWMVTAFIFVVLVGVEELTA
27. (0.86) MELFKVSSLFALFFFLDVTSA
28. (0.89) MCKLSVFHLLFLATIVQVTIA
29. (0.85) MTLPCVFAVLALVGATFA
30. (0.73) MEVHYLVFVLSVLWHGTDA
31. (0.49) MFQTILYTVVSLVIVQA
32. (0.78) MVFLSLSTIIFILSLRAVTC
33. (0.88) MSRLRFCFLLAVLCSCLQNGYG
34. (0.47) MSKLCKVIFLYIRLAFDLMVDGIFS
35. (0.84) MMLVSVVVTALLVLRSIEG
36. (0.76) MWRVLLALGAACATARA
37. (0.77) MKILICITLLLSIFANYDEA
38. (0.67) MYKNMWCVFVAILGLAAG
39. (0.90) MLRIAVFVFVALLCTVMC
40. (0.90) MTNTLVFCLIALFCCTDLSFA
41. (0.88) MYLFVIHLSMLLSVLA
42. (0.85) MSLTMKSVYLLVLVSVCQA
43. (0.48) MQFARSSCCRLYYNIVIFQLVLHICIERVLP
44. (0.78) MRIAALFALFAIGLA
45. (0.83) MGAFAAALGFLLLIAAARS
46. (0.61) MRLDIIALLILYLLNININIC
47. (0.84) MHFQSIAAILFLTLINYAVC
48. (0.87) MRVILFTAIALLGAALA
49. (0.73) MYSLTVVGVLLAVCGGAFS
50. (0.83) MALFSRKIYLLGAFLVLVNAISVVHS
51. (0.57) MVGLCTSFFLYVSVVYS
52. (0.78) MVASRATFVALVCALALASA
53. (0.84) MAKLVSLVLCVVLVYMISA
54. (0.85) MSVRRLRSFRVLLVVLVFANTSRA
55. (0.95) MGLVMYRLWMWCVALALAALWFAPAQA
56. (0.77) MLRSLFVALAVSAVFA
57. (0.83) MKPAFISICKIALFVLVFMSVISHVVIG
58. (0.84) MNLGHAVVLFNAIFIALVNS
59. (0.74) MKHFLILFTIVQSTLS
60. (0.70) MHLLIVLSAIVYASVDIVQG
61. (0.83) MRPAFVVILVVLLNQCDA
62. (0.63) MKLLYLTLLVLPAALLCFQSANTILARA
63. (0.66) MPIRKEVFWIVIICIISQCYG
64. (0.79) MKVLVVLAGLLAISAA
65. (0.78) MTAYNFRSLLLAVIVLSNCITWTAS
66. (0.78) MAVARAVLVLLASTAQA
67. (0.81) MTIAYIHISVILLFASAAHA
68. (0.89) MKLLSSLLFVLMILHCSYA
69. (0.50) MISCWMSSISSSASS
70. (0.79) MHGSARALGLALAVLALAYARA
71. (0.74) MFKLLVLILCLGFNYCLIES
72. (0.76) MDLTASLLTFTMFWKLSWS
73. (0.80) MELRRFTLCMALFVLAFAHCSES
74. (0.72) MKYILVSIILVLTFSNSCKS
75. (0.46) MLTMVPIFVVSWAGRLMA
76. (0.57) MFRMAQQAFFCTLALS
77. (0.59) MHKIFSFIGSIFFSFSSFSG
78. (0.76) MMSGLVRSMSVSTVTLVFCCLALMCTG
79. (0.47) MEIPKLMILLCCSYYGFA
80. (0.80) MVLNIKMFGVLSVFLLPLLLSQETKA
81. (0.56) MESRSSSLHRLSLDFFLSLFLSNSLLA
82. (0.78) MAFCRVVFPLLVLSVISCYA
83. (0.78) MYLRISRNFYFVLFFAFALVSVTT
84. (0.90) MRPIIFFKFLLLFSLFALTFS
85. (0.82) MWKIILFCTFLSVSQA
86. (0.49) MKMLISLNKKSLITLYFSFLRPCLVLS
87. (0.85) MCKTQCVLALLALSVGVAHS
88. (0.77) MHVVIPLLLVWALCGVRG
89. (0.92) MKITLLLVAVCLAVSLVSA
90. (0.46) MATMLNGAAVMDAALLLIAGNES
91. (0.67) MQRYLLLYTFVLIVPFAYQ
92. (0.64) MCQLMFYIVVSIFSVLKTATG
93. (0.81) MVPNLNLLHVICFTLLFGLCVRTHA
94. (0.60) MDAILAALVALMALAAAMA
95. (0.82) MLKKLLKVCAVFMIFFKCAAS
96. (0.78) MFRIIFSSLLITTVLC
97. (0.54) MAYESSIVYTLIICFFITDTALS
98. (0.92) MQRLLKILVLAIVCTAVRA
99. (0.74) MFTTIFMLLYFLNFESVSS
100. (0.85) MYFYALFLLLLSTCTISHG

◀ **Figure 6. Highly expressed predicted signal peptides.** Plot on the left are the expression values of those transcripts which have a predicted signal peptide. Sequences on the right are the top 100 signal peptide sequence with the signalP score in the brackets.

Since some protein sequences are 5′ truncated or predicted starting from an upstream point, we manually checked the prediction of top expression of transcripts containing signal peptides.

To compare the constituent of glycogenes, we collected glycogene sequences from GGDB as reference (Narimatsu 2004). Published Sf21 transcriptome was downloaded from TSA database with accession No. GCTM00000000 (Kakumani et al. 2015). Protein sequences were aligned to this reference with blastp. Blast results of glycogenes were manually checked by aligning sequences to the TrEMBL90 and nr databases, to eliminate false positive hits. Classification by function, mechanism and structure were based on information from GGDB and CAZy database (Narimatsu 2004; Lombard et al. 2014).

Codon usage evaluation

We used our predicted coding sequences in codon usage evaluation. CDS sequences from other species were downloaded from Ensembl and NCBI genome databases. Only those sequences longer than 100 amino acids (300 bp CDS sequence) were used for calculation. All codons were counted with our script by shifting a simulated reading frame from 5′ end to 3′ end. With the count data of all codons, we calculated RSCU value according to the formula as described (Cannarozzi and Schneider 2012).

Expression analysis

To evaluate the transcript abundance, Bowtie (v1.0.0) and RSEM (v1.2.15) were used for sequence alignment and calculation of TPM (Transcripts Per Million) and FPKM (Fragments Per Kilobase of transcript per Million mapped reads) (Langmead et al. 2009; Li and Dewey 2011). This value could measure the transcripts abundance with transcript length, thus we can compare it among different genes or samples.

ACKNOWLEDGMENTS

We thank Dr. Yun Zhang and the High-Throughput Sequencing Center of Peking University for assistance with Illumina sequencing. This work was supported by the National Basic Research Program (973 Program) (No. 2014CB910102 to X.S.).

ABBREVIATIONS

GO, Gene Ontology; TPM, Transcripts Per Million; FPKM, Fragments Per Kilobase of transcript per Million mapped reads; RSCU, relative synonymous codon usage.

COMPLIANCE WITH ETHICS GUIDELINES

Kai Yu, Yang Yu, Xiaoyan Tang, Huimin Chen, Junyu Xiao and Xiao-Dong Su declare that they have no conflict of interest. This article does not contain any studies with human or animal subjects performed by the any of the authors.

OPEN ACCESS

REFERENCES

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi:10.1093/bioinformatics/btu170

Breitbach K, Jarvis DL (2001) Improved glycosylation of a foreign protein by Tn-5B1-4 cells engineered to express mammalian glycosyltransferases. Biotechnol Bioeng 74:230–239. doi:10.1002/bit.1112

Brondyk WH (2009) Selecting an appropriate method for expressing a recombinant protein. Methods Enzymol 463:131–147. doi:10.1016/S0076-6879(09)63011-1

Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi:10.1186/1471-2105-10-421

Cannarozzi GM, Schneider A (eds) (2012) Codon Evolution. Oxford University Press

Castilho A (ed) (2015) Glyco-Engineering. Springer New York, New York, NY

Chaney JL, Clark PL (2015) Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophys 44:143–166. doi:10.1146/annurev-biophys-060414-034333

Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:619832. doi:10.1155/2008/619832

Davis TR, Trotter KM, Granados RR, Wood HA (1992) Baculovirus Expression of Alkaline Phosphatase as a Reporter Gene for Evaluation of Production, Glycosylation and Secretion. Bio/Technology 10:1148–1150. doi:10.1038/nbt1092-1148

Finn RD, Bateman A, Clements J et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. doi:10.1093/nar/gkt1223

Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. doi:10.1093/bioinformatics/bts565

Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. doi:10.1038/nbt.1883

Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512. doi:10.1038/nprot.2013.084

Hollister JR, Jarvis DL (2001) Engineering lepidopteran insect cells for sialoglycoprotein production by genetic transformation with mammalian 1,4-galactosyltransferase and 2,6-sialyltransferase genes. Glycobiology 11:1–9. doi:10.1093/glycob/11.1.1

Hollister J, Grabenhorst E, Nimtz M et al (2002) Engineering the Protein N-Glycosylation Pathway in Insect Cells for Production of Biantennary, Complex N-Glycans †. Biochemistry 41:15093–15104. doi:10.1021/bi026455d

Hollister JR, Shaper JH, Jarvis DL (1998) Stable expression of mammalian beta 1,4-galactosyltransferase extends the N-glycosylation pathway in insect cells. Glycobiology 8:473–480

Holm L (1986) Codon usage and gene expression. Nucleic Acids Res 14:3075–3087

Jarvis DL (2003) Developing baculovirus-insect cell expression systems for humanized recombinant glycoprotein production. Virology 310:1–7. doi:10.1016/S0042-6822(03)00120-X

Kakumani PK, Malhotra P, Mukherjee SK, Bhatnagar RK (2014) A draft genome assembly of the army worm, Spodoptera frugiperda. Genomics 104:134–143. doi:10.1016/j.ygeno.2014.06.005

Kakumani PK, Shukla R, Todur VN et al (2015) De novo transcriptome assembly and analysis of Sf21 cells using illumina paired end sequencing. Biol Direct 10:44. doi:10.1186/s13062-015-0072-7

Kost TA, Condreay JP, Jarvis DL (2005) Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nat Biotechnol 23:567–575. doi:10.1038/nbt1095

Lairson LL, Henrissat B, Davies GJ, Withers SG (2008) Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem 77:521–555. doi:10.1146/annurev.biochem.76.061005.092322

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi:10.1186/gb-2009-10-3-r25

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323. doi:10.1186/1471-2105-12-323

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. doi:10.1093/bioinformatics/btp698

Lombard V, Golaconda Ramulu H, Drula E et al (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495. doi:10.1093/nar/gkt1178

Narimatsu H (2004) Construction of a human glycogene library and comprehensive functional analysis. Glycoconj J 21:17–24. doi:10.1023/B:GLYC.0000043742.99482.01

Olczak M, Olczak T (2006) Comparison of different signal peptides for protein secretion in nonlytic insect cell system. Anal Biochem 359:45–53. doi:10.1016/j.ab.2006.09.003

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786. doi:10.1038/nmeth.1701

Powell S, Forslund K, Szklarczyk D et al (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res 42:D231–D239. doi:10.1093/nar/gkt1253

Soejima Y, Lee J, Nagata Y et al (2013) Comparison of signal peptides for efficient protein secretion in the baculovirus-silkworm system. Open Life Sci 8:1–7. doi:10.2478/s11535-012-0112-6

UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212. doi:10.1093/nar/gku989

Vandenborre G, Smagghe G, Ghesquière B et al (2011) Diversity in Protein Glycosylation among Insect Species. PLoS One 6:e16682. doi:10.1371/journal.pone.0016682

Vaughn JL, Goodwin RH, Tompkins GJ, McCawley P (1977) The Establishment of Two Cell Lines from the Insect Spodoptera frugiperda (Lepidoptera; Noctuidae). In Vitro 13:213–217

von Heijne G, Abrahmsén L (1989) Species-specific variation in signal peptide design. Implications for protein secretion in foreign hosts. FEBS Lett 244:439–446

Wickham TJ, Davis T, Granados RR, et al Screening of insect cell lines for the production of recombinant proteins and infectious virus in the baculovirus expression system. Biotechnol Prog 8:391–6. doi: 10.1021/bp00017a003

Xie C, Mao X, Huang J et al (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39:W316–W322. doi:10.1093/nar/gkr483

Xu C, Ng DTW (2015) Glycosylation-directed quality control of protein folding. Nat Rev Mol Cell Biol 16:742–752. doi:10.1038/nrm4073

Ye J, Fang L, Zheng H et al (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34:W293–W297. doi:10.1093/nar/gkl031