棉纤维伸长促进基因*KCS*的生物信息研究

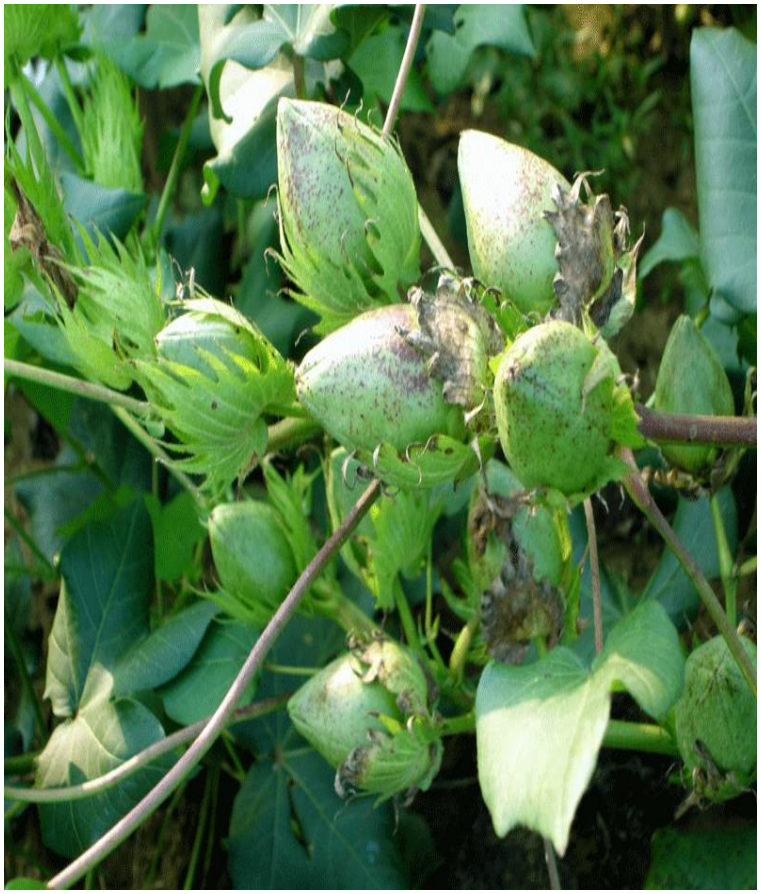
The bioinformatic research of cotton fibre elongation-promoting gene *KCS*

报告人：黄盖
组员：高晓雪、刘丹、屈颖
2014/1/10

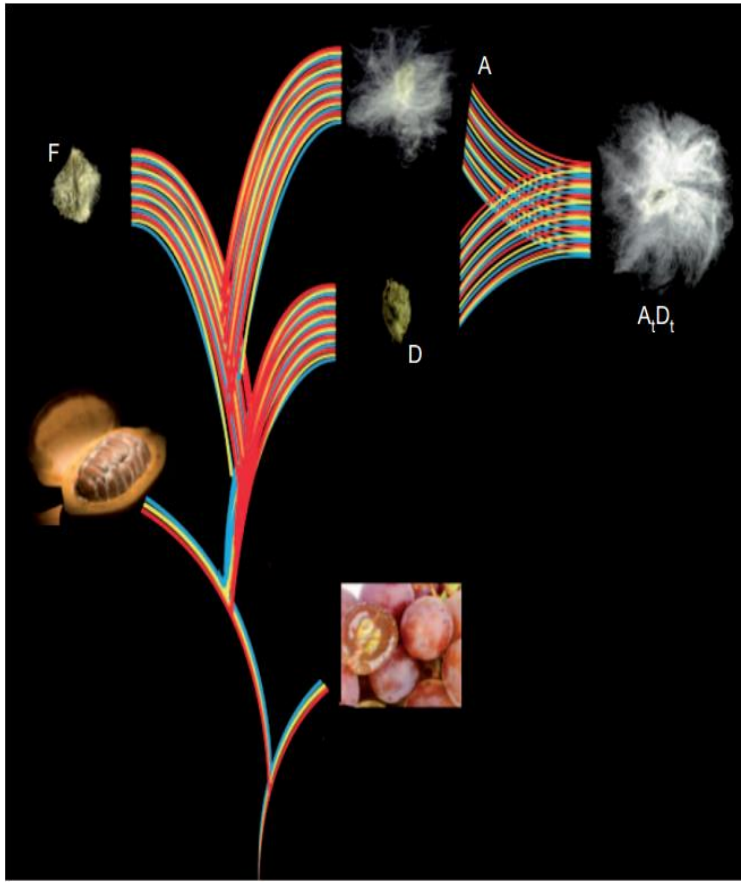
Content

Why we choose cotton gene KCS?

Cotton



About cotton evolution

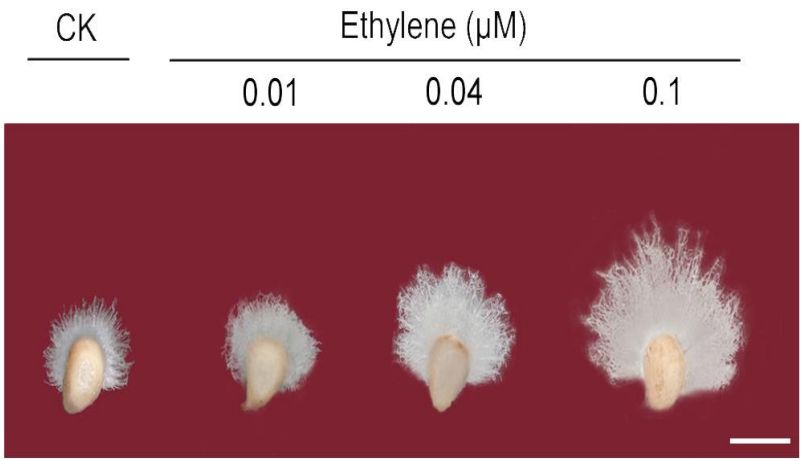


(Andrew H. et al.,2012)

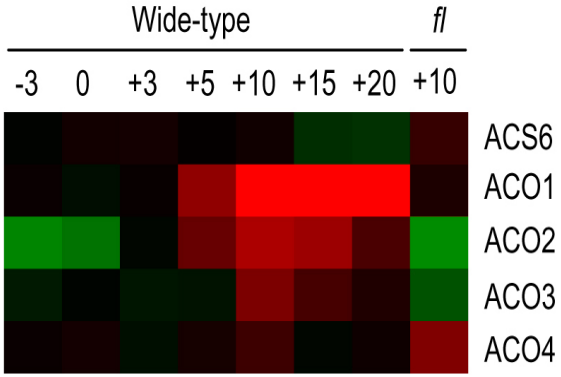
- Cotton is an important economic crop.
- Gossypium genus contains tetraploid and diploid species. The tetraploid cotton species AD-genome (such as G. hirsutum) are thougth to have formed by an allopolyploidzation of A and D-genome species.
- G. hirsutum (AD-genome) produce most widely used natural fiber, but G. raimondii (D-genome) is fibreless.
- The draft genome of G. raimondii is sequenced by our lab.(Kunbo Wang. et al.,2012)

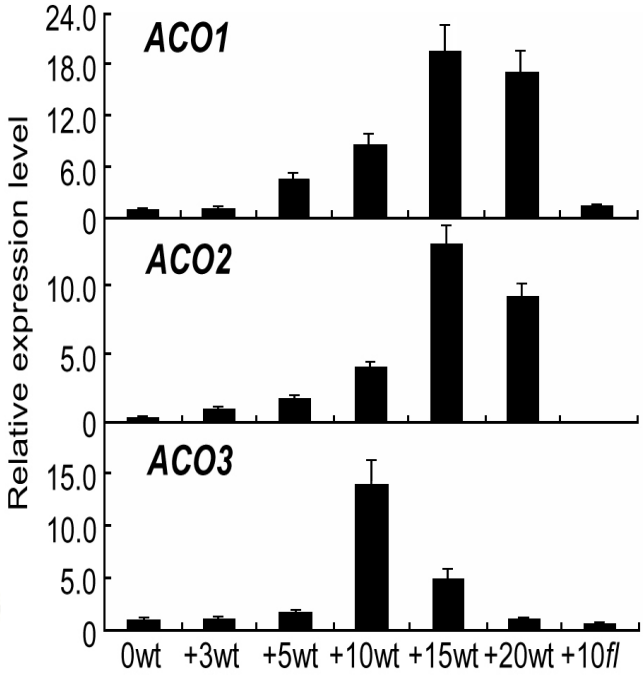
Previous Research of Our Lab

Ethylene



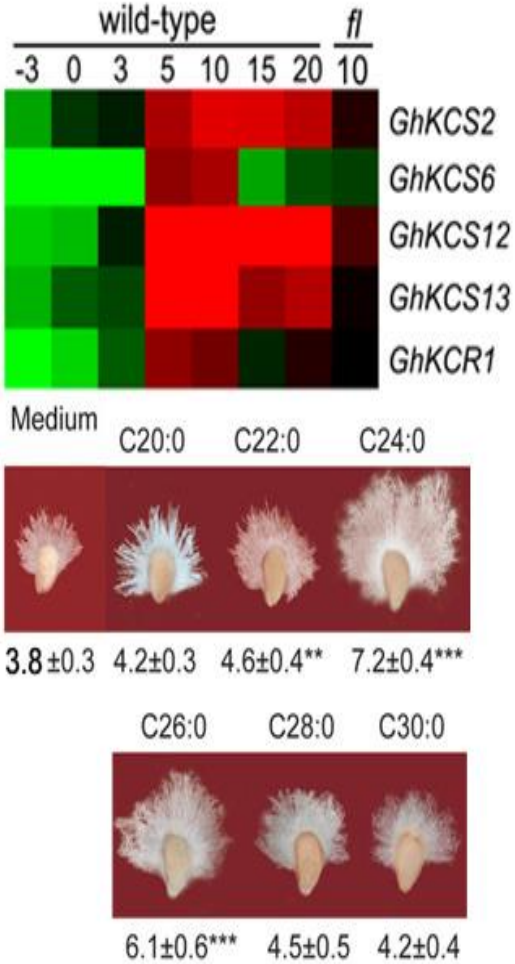
（Shi *et al*., 2006)

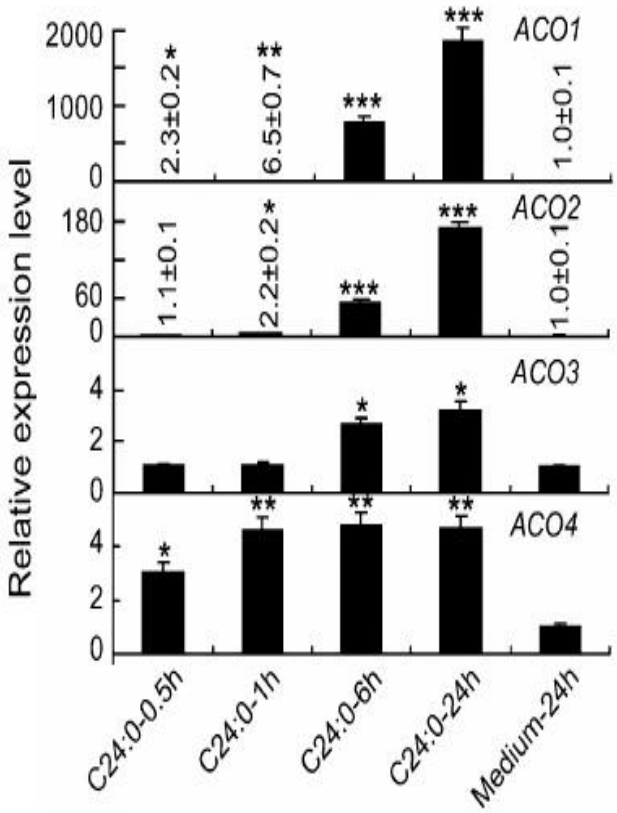




Previous Research of Our Lab

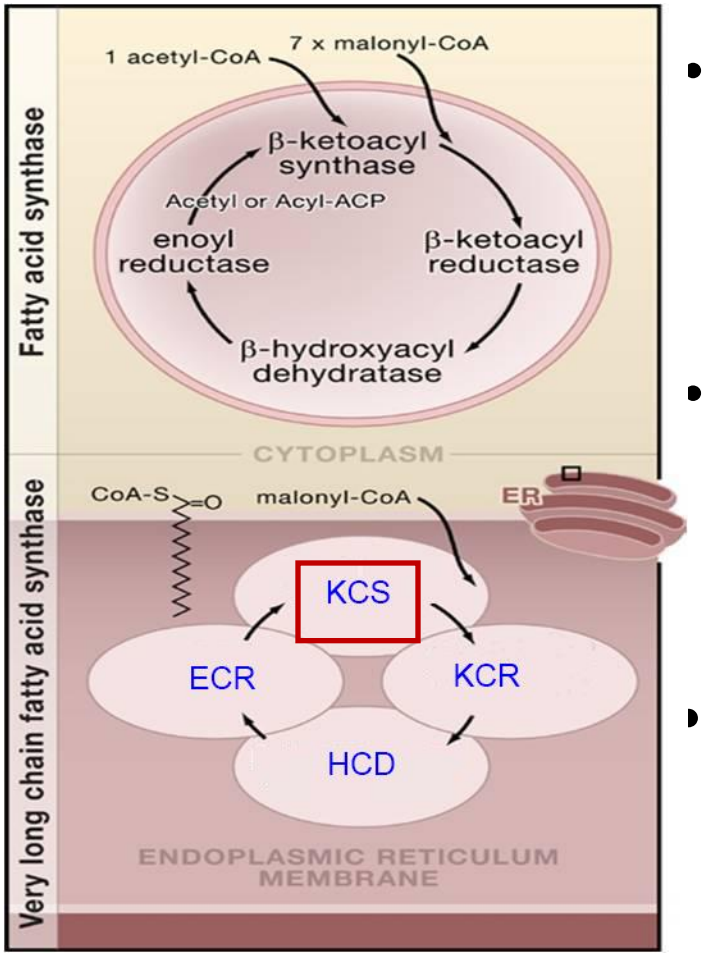
Very Long Chain Fatty Acids （VLCFAs）





(Qin et al.,2007)

About VLCFAs



(Riezman, 2007,revised)

- ◆VLCFAs (>18 carbons are essential components of plant lipids, suberins and cuticular waxes.
- ◆ VLCFAs is synthesized by four successive enzyme reactions, including **KCS**,KCR, HCD and ECR
- Previous results in our lab reveal that VCFAs may promote cotton fiber elongation by regulating ethylene biosynthesis.

UniProt database search

Gene names	Alias	Entry name	Length of amino acids
KCS1	EL1 At1g01120 T25K16.11	Q9MAM3_ARATH	528
KCS2	At4g34510 T4L20.90	O65677_ARATH	487
KCS3	At1g07720 F24B9.18	Q9LQP8_ARATH	478
KCS4	At1g19440 F18O14.21	Q9LN49_ARATH	516
KCS5	CER60 At1g25450 F2J7.9	Q9C6L5_ARATH	492
KCS6	CER6 EL6 At1g68530 T26J14.10	Q9XF43_ARATH	497
KCS7	At1g71160 F23N20.15	Q9C992_ARATH	460
KCS8	At2g15090 T15J14.13	Q4V3C9_ARATH	481
KCS9	At2g16280 F16F14.22	Q9SIX1_ARATH	512
KCS10	FDH EL4 At2g26250 T1D16.11	Q570B4_ARATH	550
KCS11	At2g26640 F18A8.1	O48780_ARATH	509
KCS12	At2g28630 T8O18.8	Q9SIB2_ARATH	476
KCS13	HIC At2g46720 T3A4.10	Q9ZUZ0_ARATH	466
KCS14	At3g10280 F14P13.12	Q9SS39_ARATH	459
KCS15	At3g52160 F4F15.270	Q9SUY9_ARATH	451
KCS16	EL2 At4g34250 F10M10.20	Q9SYZ0_ARATH	493
KCS17	At1g04220 F20D22.1	Q5XEP9_ARATH	528
KCS18	FAE1 At4g34520 T4L20.100	Q38860_ARATH	506
KCS19	At5g43760 MQD19.11	Q9FG87_ARATH	529
KCS20	At5g49070 K20J1.4	Q9FH27_ARATH	464
KCS21	At5g04530 T32M21.130	Q9LZ72_ARATH	464

The database contains 21 reviewed sequences of Arabidopsis thaliana and they belong to FAE protein family. Besides, none of their protein has 3D structure .

Names	Alias	Entry name	Family	Length of AA
KCS2		A9XUG6_GOSHI		529
KCS1	Gr10021139	M9Z5H0_GOSRA		533
KCS2	Gr10031660	M9Z380_GOSRA		529
KCS3	Gr10022901	M9ZC32_GOSRA		466
KCS4	Gr10018297	M9Z385_GOSRA		504
KCS5	Gr10014500	M9ZC12_GOSRA		496
KCS6	Gr10032226	M9ZC26_GOSRA		467
KCS7	Gr10016173	M9ZA63_GOSRA		510
KCS8	Gr10026783	M9Z6N0_GOSRA		535
KCS9	Gr10018147	M9Z5I3_GOSRA		504
KCS10	Gr10015926	M9Z5H7_GOSRA		531
KCS11	Gr10017624	M9ZC02_GOSRA		510
KCS	Gr10018148	M9Z5I6_GOSRA		418
KCS13	Gr10031991	M9Z390_GOSRA		492
KCS15	Gr10032475	M9ZC21_GOSRA		462
KCS16	Gr10000033	M9Z6L3_GOSRA		390
KCS17	Gr10028536	M9ZC07_GOSRA		537
KCS18	Gr10034062	M9ZA66_GOSRA	FAE1_CUT1_RppA	515
KCS19	Gr10010162	M9Z6K4_GOSRA		512
KCS20	Gr10019136	M9Z3A9_GOSRA		437
KCS21	Gr10017854	M9Z397_GOSRA		457
KCS	Gr10019659	M9ZA81_GOSRA		439
KCS	Gr10004055	M9Z6K7_GOSRA		533
KCS	Gr10036653	M9ZC17_GOSRA		515
KCS	Gr10040229	M9Z5K7_GOSRA		501
KCS	Gr10018150	M9ZA87_GOSRA		493
KCS	Gr10018149	M9ZA72_GOSRA		504
KCS	Gr10040228	M9Z3B3_GOSRA		501
KCS	Gr10019657	M9Z6M6_GOSRA		424
KCS	Gr10019616	M9Z5J9_GOSRA		449
KCS	Gr10033317	M9Z3A3_GOSRA		274
KCS	Gr10018195	M9Z6L8_GOSRA		285
KCS14	Gr10009484	M9Z5J2_GOSRA	ELO	282
KCS12	Gr10018194	M9ZA77_GOSRA		286

The database contains 33 KCS sequences of G. raimondii and 1 KCS sequence of G. hirsutum. Besides, none of them are reviewed and their sequences were submitted by our lab.

enome-wide analysis

Cotton database construction

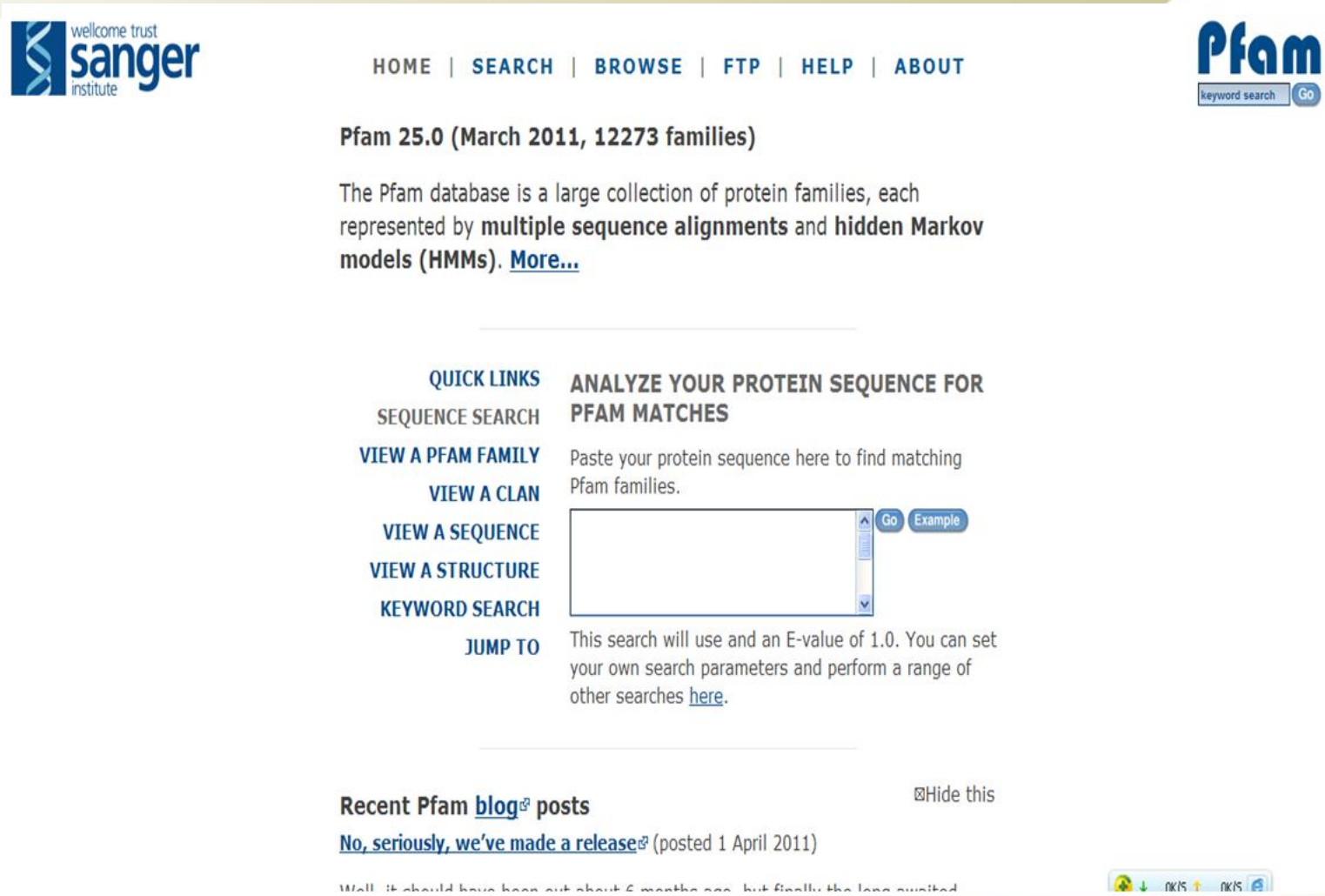
```
F:\blast\bin 的目录

2014/01/06  16:20    <DIR>          .
2014/01/06  16:20    <DIR>          ..
2013/03/13  06:34         5,488,640 blastdbcheck.exe
2013/03/13  06:34         6,592,512 blastdbcmd.exe
2013/03/13  06:34         4,174,336 blastdb_aliastool.exe
2013/03/13  06:34         9,608,704 blastn.exe
2013/03/13  06:34         9,604,608 blastp.exe
2014/01/06  19:12            88,207 blastpout
2013/03/13  06:34         9,597,440 blastx.exe
2013/03/13  06:34         9,463,808 blast_formatter.exe
2014/01/06  00:02             1,279 CDS.fas
2014/01/06  00:07            26,794 cdsout
2013/03/13  06:34         5,219,840 convert2blastmask.exe
2014/01/05  12:38       775,246,172 Cotton_D.all.anchored.fa
2014/01/05  19:03           337,026 Cotton_D.all.anchored.fa.nhr
2014/01/05  19:03            53,448 Cotton_D.all.anchored.fa.nin
2014/01/05  19:03       194,138,605 Cotton_D.all.anchored.fa.nsq
2013/03/13  06:34         9,780,224 deltablast.exe
2013/03/13  06:34         5,517,824 dustmasker.exe
2014/01/04  13:38        49,567,883 Gr-cds.fa.txt
2014/01/04  16:49         4,975,290 Gr-cds.fa.txt.nhr
2014/01/04  16:49           491,788 Gr-cds.fa.txt.nin
2014/01/04  16:49        11,348,776 Gr-cds.fa.txt.nsq
2014/01/04  17:07        18,193,161 Gr-PEP.fa.fasta
2014/01/04  17:08         5,016,266 Gr-PEP.fa.fasta.phr
2014/01/04  17:08           327,888 Gr-PEP.fa.fasta.pin
2014/01/04  17:08        15,083,278 Gr-PEP.fa.fasta.psq
2013/03/13  06:34            51,345 legacy_blast.pl
2013/03/13  06:34         6,111,232 makeblastdb.exe
2013/03/13  06:34         5,744,128 makembindex.exe
2013/03/13  06:34         5,085,184 makeprofiledb.exe
2014/01/06  19:12               554 PEP.fas
2013/03/13  06:34         9,755,136 psiblast.exe
2013/03/13  06:34         9,621,504 rpsblast.exe
2013/03/13  06:34         9,613,312 rpstblastn.exe
2013/03/13  06:34         5,727,744 segmasker.exe
2013/03/13  06:34         9,716,224 tblastn.exe
```

1. Download local blast software from NCBI.

2.Download G. raimondii genome sequence from CGP.

3.To execute blast in DOS system

To acquire conserved domain



pfam（http://pfam.sanger.ac.uk/）

Pfam search results

Significant Pfam-A Matches

Show or hide all alignments.

保守domain名称

Family	Description	Entry type	Clan
FAE1_CUT1_RppA	FAE1/Type III polyketide synthase-like p ...	Family	CL0046

```
#HMM     ylarrrrkvyLvdyacykpedelkvstetfleivkrvkkldeesleflrkilersGlgeetyvPrslleipeektlaeareEaeevlfgavdellaktkvkpkdigilvvnc
#MATCH   y+++r+r +yLvdyacy p+++lk++++ f+e +k++  +de ++eflrki+ersGlg+et +P+s++  p+++++a+ar+Eae v+fga+d l+a t+vkp+digilvvnc
#PP      689**********************************************************************************************************
#SEQ      VMTRKRFTYLVDYACYLPPQHLKADHRYFMEYAKEAADFDEPTMEFLRKIMERSGLGDETGAPPSMHCFPPRPSMAAARQEAELVMFGALDTLFASTNVKPRDIGILVVN
```

复制保守domain序列

FAE1_CUT1-RppA HMM:
ylarrrrkvyLvdyacykpedelkvstetfleivkrvkkldeesleflrkilersGlgeetyvPrslleipee
ktlaeareEaeevlfgavdellaktkvkpkdigilvvncslfsptPslsamvvnryklredvksynLsgm
GCsaglisidlakdllqvhkntlalvvstEnitlnwYvGnersmllsnclFRvGgaavllsnksadrrr
akykLvhvvRthkgaddkayrcvlqeeDeegkvGvslskdlvkvagealkknlttlgplvLPlsEkl
rflaslvarkl

To query cotton genome database

Length=290

Sequences producing significant alignments:		Score (Bits)	E Value
Cotton_D_gene_10017624	locus=scaffold163:642071:643603:+	427	4e-147
Cotton_D_gene_10018297	locus=scaffold71:210359:211873:+	425	2e-146
Cotton_D_gene_10016173	locus=scaffold172:1629641:1631173:-	422	3e-145
Cotton_D_gene_10036653	locus=scaffold109:1190172:1191719:+	418	1e-143
Cotton_D_gene_10034062	locus=scaffold4:3957287:3958834:-	417	2e-143
Cotton_D_gene_10010162	locus=scaffold130:399751:401289:-	417	4e-143
Cotton_D_gene_10015926	locus=scaffold166:1016377:1017972:+	414	6e-142
Cotton_D_gene_10018147	locus=scaffold152:1567657:1569171:+	413	8e-142
Cotton_D_gene_10028536	locus=scaffold43:1655946:1657559:+	409	1e-139
Cotton_D_gene_10018150	locus=scaffold152:1584096:1585466:+	405	1e-139
Cotton_D_gene_10031660	locus=scaffold84:3835910:3838652:-	408	2e-139
Cotton_D_gene_10021139	locus=scaffold31:112479:114389:-	408	2e-139
Cotton_D_gene_10018149	locus=scaffold152:1577823:1579337:+	404	4e-138
Cotton_D_gene_10014500	locus=scaffold72:1357197:1358687:+	402	2e-137
Cotton_D_gene_10031281	locus=scaffold8:314495:316978:-	390	2e-136
Cotton_D_gene_10004[illegible]	locus=scaffold[illegible]:[illegible]:[illegible]:-	[illegible]	5e-132
Cotton_D_gene_10000033	locus=scaffold1232:972:2144:+	380	3e-130
Cotton_D_gene_10026783	locus=scaffold67:784594:786550:+	376	6e-127
Cotton_D_gene_10007[illegible]	locus=scaffold[illegible]:[illegible]:[illegible]:5876:-	318	1e-107
Cotton_D_gene_10018148	locus=scaffold152:1574866:1576236:+	320	1e-106
Cotton_D_gene_10017854	locus=scaffold167:1150904:1152277:-	297	2e-097
Cotton_D_gene_10040228	locus=scaffold1:7996116:7997621:+	291	2e-094
Cotton_D_gene_10040229	locus=scaffold1:8009851:8011356:+	290	7e-094
Cotton_D_gene_10019136	locus=scaffold36:1496467:1497780:-	284	3e-092
Cotton_D_gene_10032226	locus=scaffold8:2307409:2308812:+	273	1e-087
Cotton_D_gene_10022901	locus=scaffold25:472748:474148:+	270	1e-086
Cotton_D_gene_10019616	locus=scaffold207:1262504:1263853:-	266	2e-085
Cotton_D_gene_10032475	locus=scaffold75:3586793:3588181:-	256	3e-081
Cotton_D_gene_10031[illegible]	locus=scaffold[illegible]:[illegible]:[illegible]298:691087:-	248	4e-081
Cotton_D_gene_10019659	locus=scaffold207:1692130:1693449:+	254	1e-080
Cotton_D_gene_10019657	locus=scaffold207:1657270:1658544:+	251	1e-079
Cotton_D_gene_10019658	locus=scaffold207:1664233:1664877:+	187	8e-058
Cotton_D_gene_10038985	locus=scaffold30:686293:686952:-	72.4	4e-015
Cotton_D_gene_10000813	locus=scaffold471:148816:149609:-	38.9	0.002
Cotton_D_gene_10026953	locus=scaffold260:135384:136656:-	39.3	0.003
Cotton_D_gene_10023921	locus=scaffold97:1006873:1009133:+	38.5	0.004
Cotton_D_gene_10002256	locus=scaffold414:316833:319857:-	37.4	0.009
Cotton_D_gene_10023920	locus=scaffold97:934662:936994:+	37.0	0.013
Cotton_D_gene_10029217	locus=scaffold96:1906837:1908121:-	37.0	0.014
Cotton_D_gene_10011223	locus=scaffold119:765522:768233:+	36.2	0.024
Cotton_D_gene_10007646	locus=scaffold236:85034:86759:-	35.0	0.059

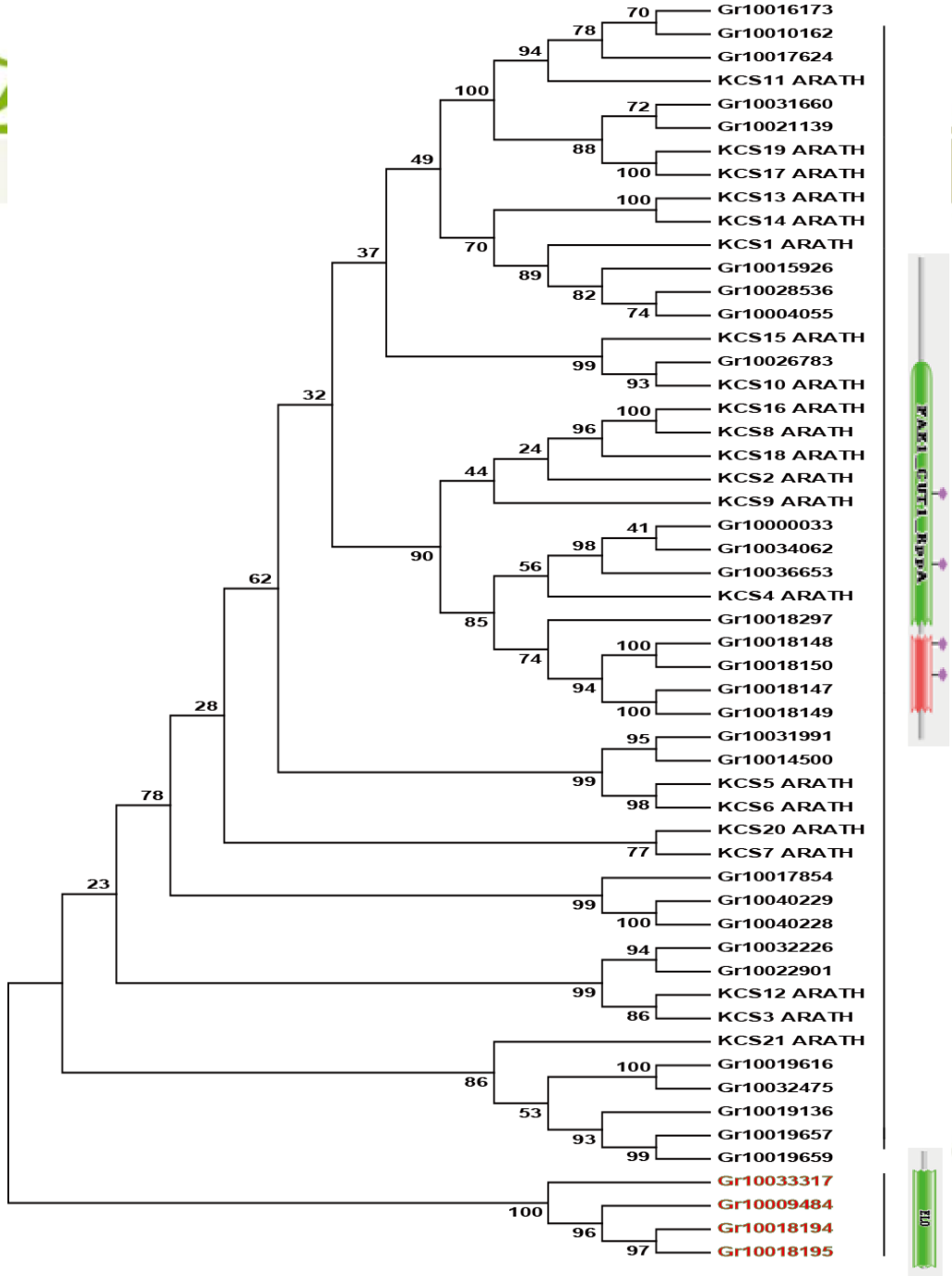
FAE CUT1-RppA+ACP Terminal

Lack ACP Terminal

Lack ACP Terminal

1. To query cotton genome database based on HMM model sequence
2. To retrieve the corresponding gene sequences ,then confirm the conserved domain in Pfam database and smart database . (http://smart.emblheidelberg.de/）
3. To further search the cotton database in order to find all KCS sequences.

Phylogeny construction and chromosome location

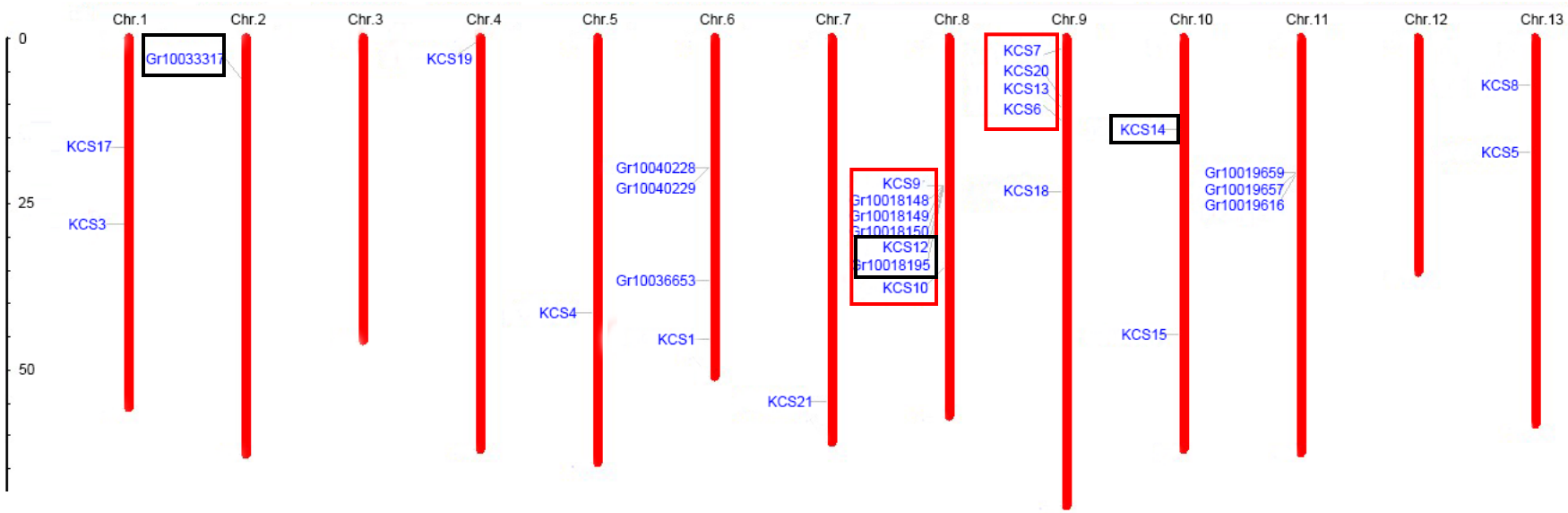


1.The phylogeny analysis suggests that all G. ranondii KCS genes are divided in **two major branches**.

2.In combination with Pfam prediction results, we know that the two major branches belong to **FAEFAE1_CUT1_RppA (PF08392 , PF08541) and Elo(PF01151, GNS1/SUR4 family)**

3.**FAE1_CUT1_RppA**: described as 3-ketoacyl-CoA synthases ; ACP_syn_III_C, ACP synthase III C terminal ; ELO1 Members of this family are involved in LCFAs elongation systems that produce the 26-carbon precursors for ceramide and sphingolipid synthesis

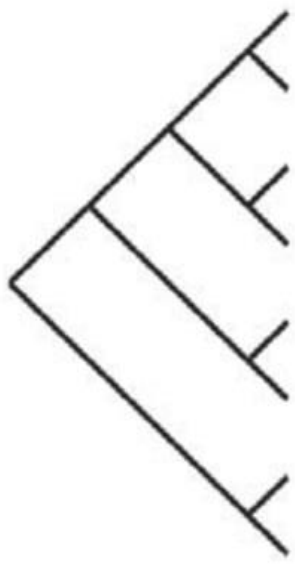
Candidate gene in chromosome location



29 G.raimondii KCS genes except GrKCS11 ， GrKCS16, GrKCS2, Gr10004055 were located on 11 chromosomes ,but chromosome 3 and 12 contain no KCS genes.

KCS may be required in fiber initiation and elongation

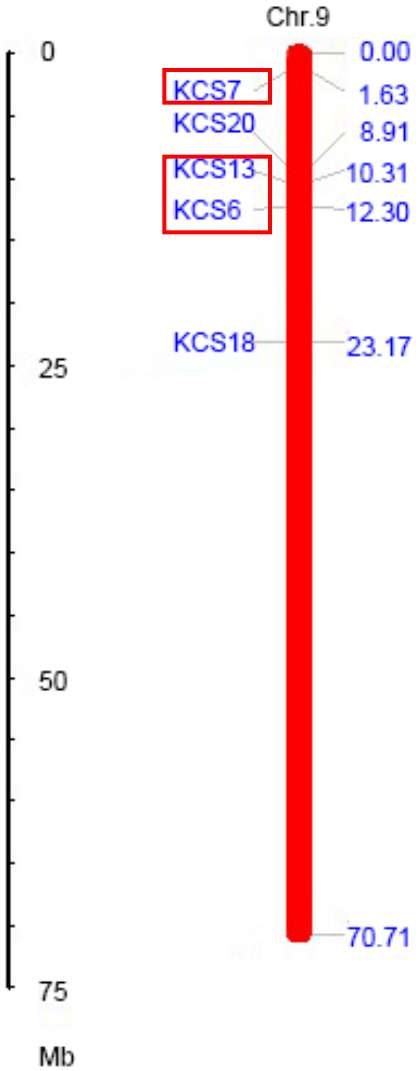
A



	RPKM	
	G. raimondii	*G. hirsutum*
KCS7	167	22
KCS2	4	73
KCS13	48	984
KCS6	3	288

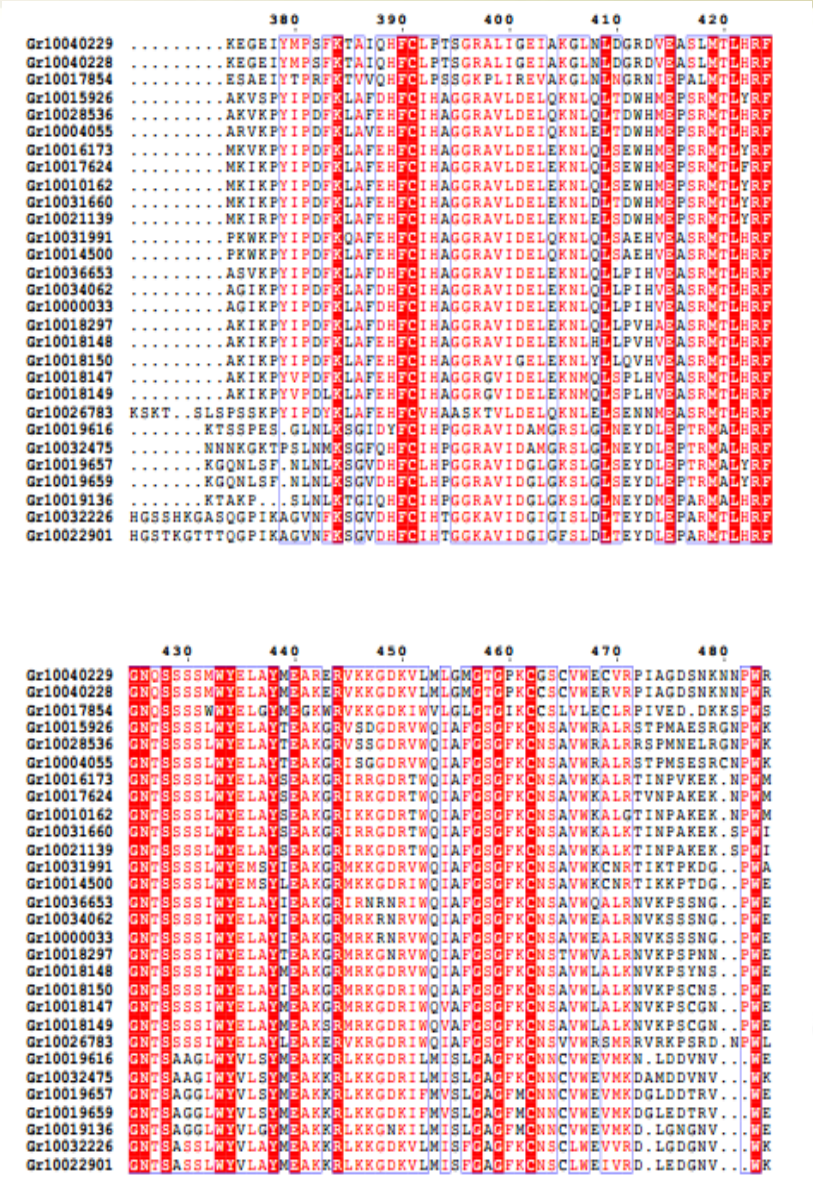
Chromosome location failed

B



Several 3-ketoacyl-CoA synthase (KCS) genes, including **KCS2, KCS13 and KCS6**, were **only expressed in G. hirsutum**, whereas intermediate levels of KCS7 transcripts were observed in both G. hirsutum and G. raimondii, indicating that high-level expression of Sus and KCS family genes may indeed be required for fiber cell initiation and elongation.(Kunbo Wang, et al. *2012.Nature Genetics.)*

Multiple sequence alignment



1. Align all KCS genes using ClustalX.

2.Resulting from poor alignment quality, we excluded all short sequences including Elo type protein and alignment again using ClustalX.

3.The alignment results suggest that FAE type KCS protein share some strict conserved amiono acids.

4.Based on Netural Evolution of Kimura, most amino acids mutants are neutral, few mutants are harmful and these sites are functional.

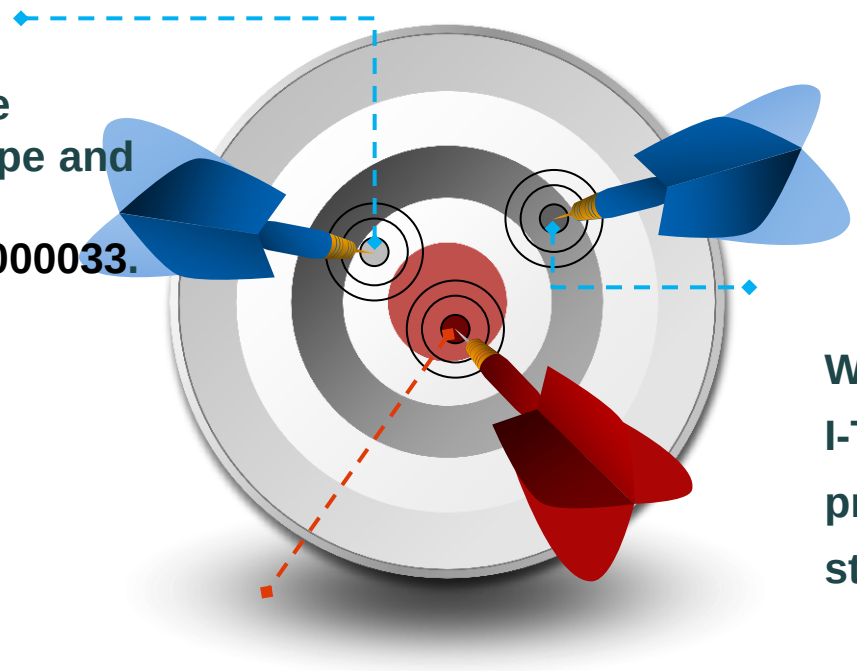
3D structure prediction

In order to find KCS proteins' active sites and study functional mechanism, we used popular methods to predict our target proteins' structures.

3D structure prediction

Obtain Squence

We extracted each one sequence from FAE type and Elo type , such as Gr10018148 and Gr10000033.



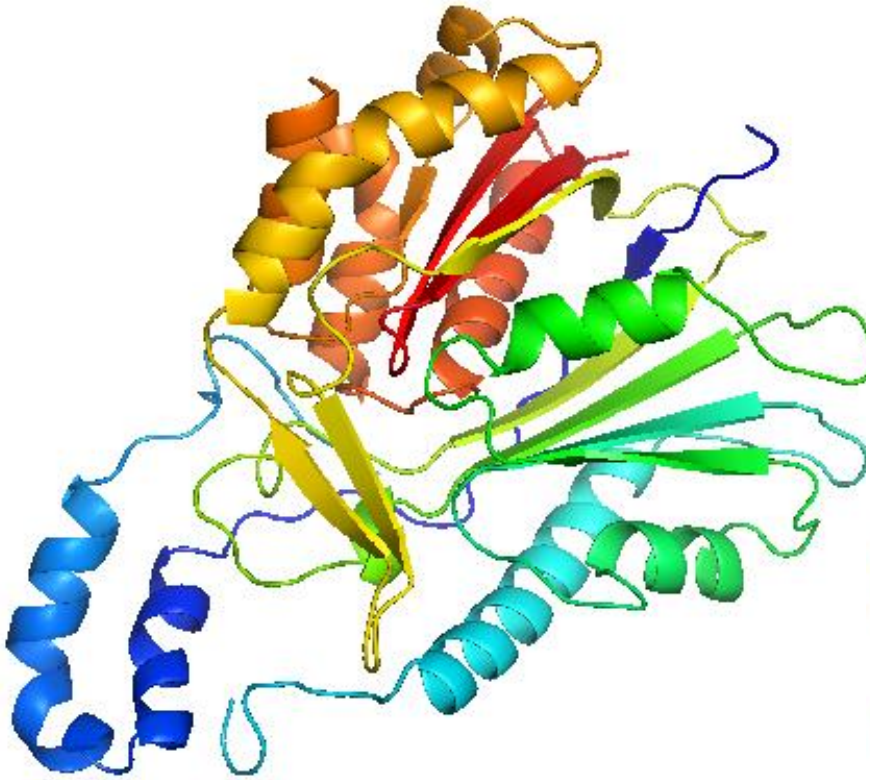
Prediction

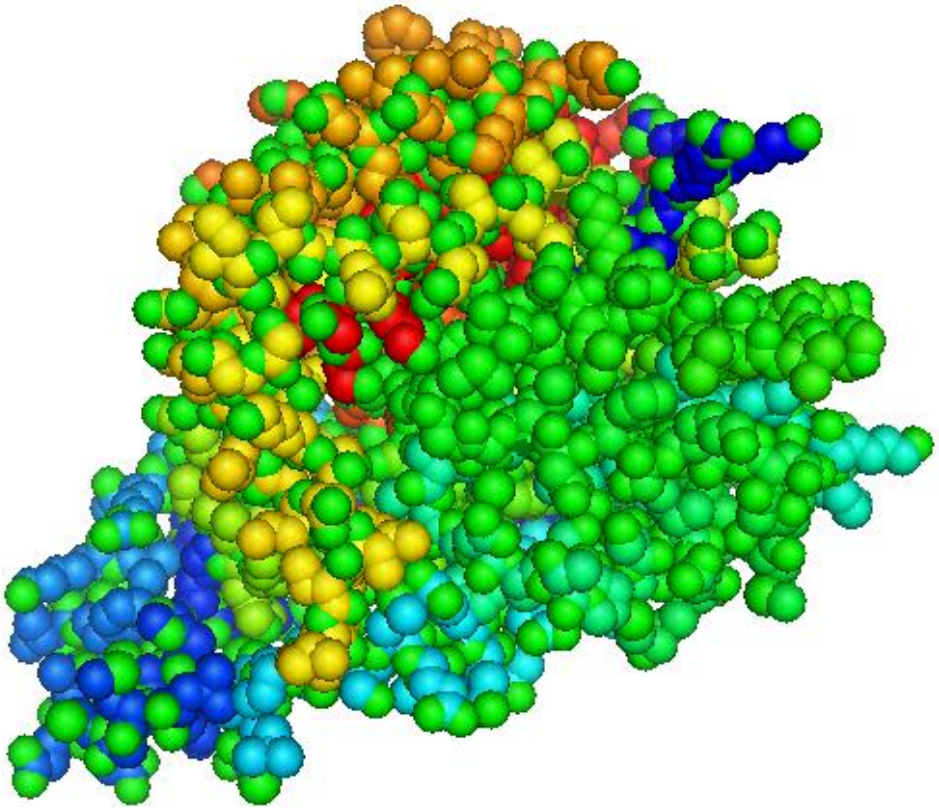
We used Phyre and I-TASSER methods to predict KCS protein 3D sturcuture.

Analysis

To analysis some details about protein structure

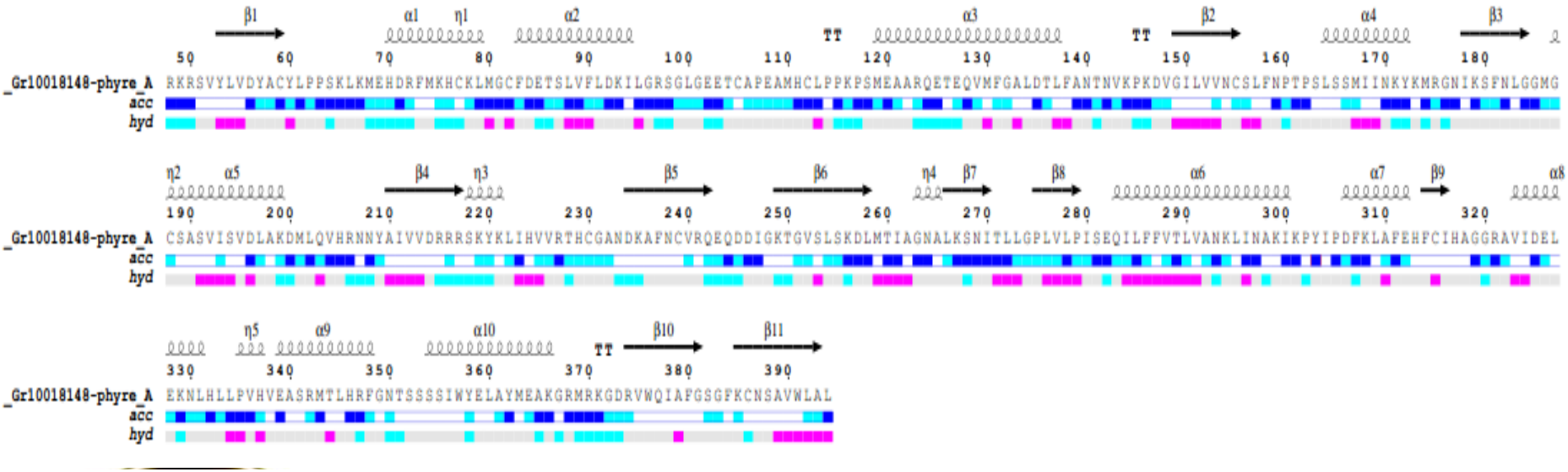
Analysis of Gr10018148 Protein structure





Gr10018148-phyre

The secondary structure of Gr10018148 protein

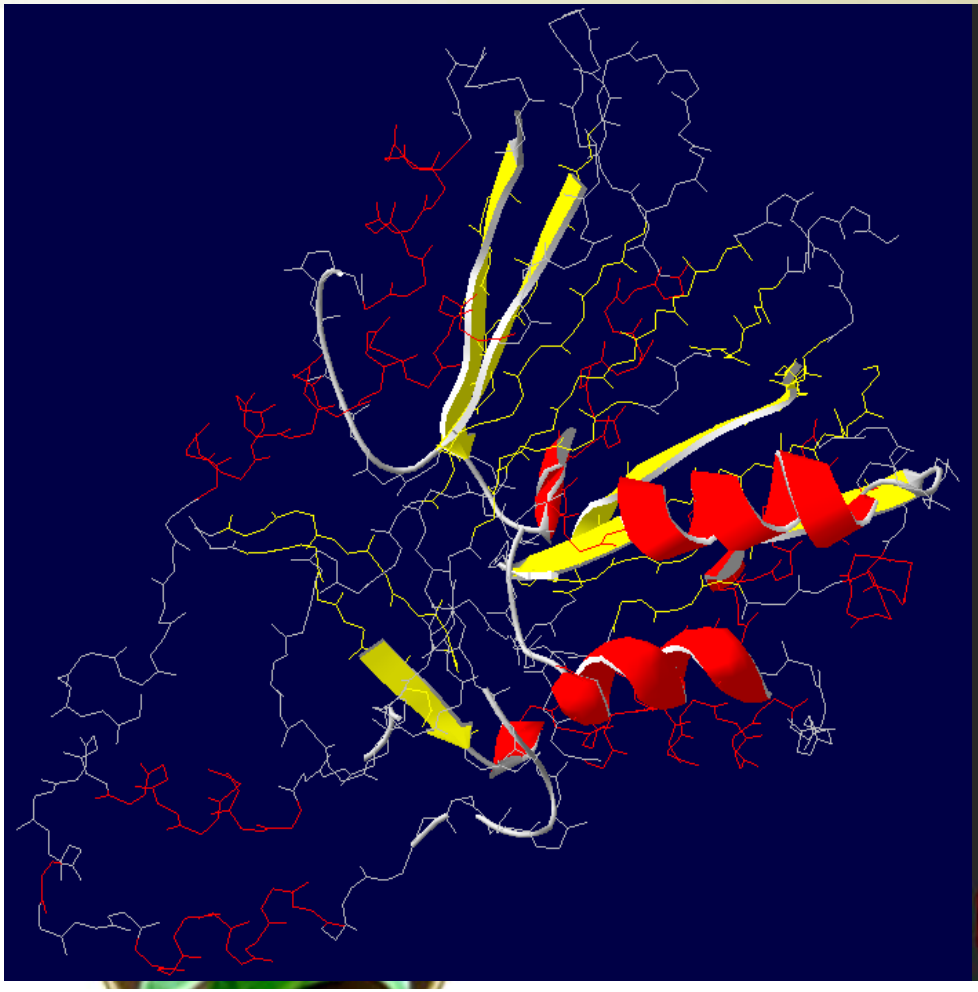


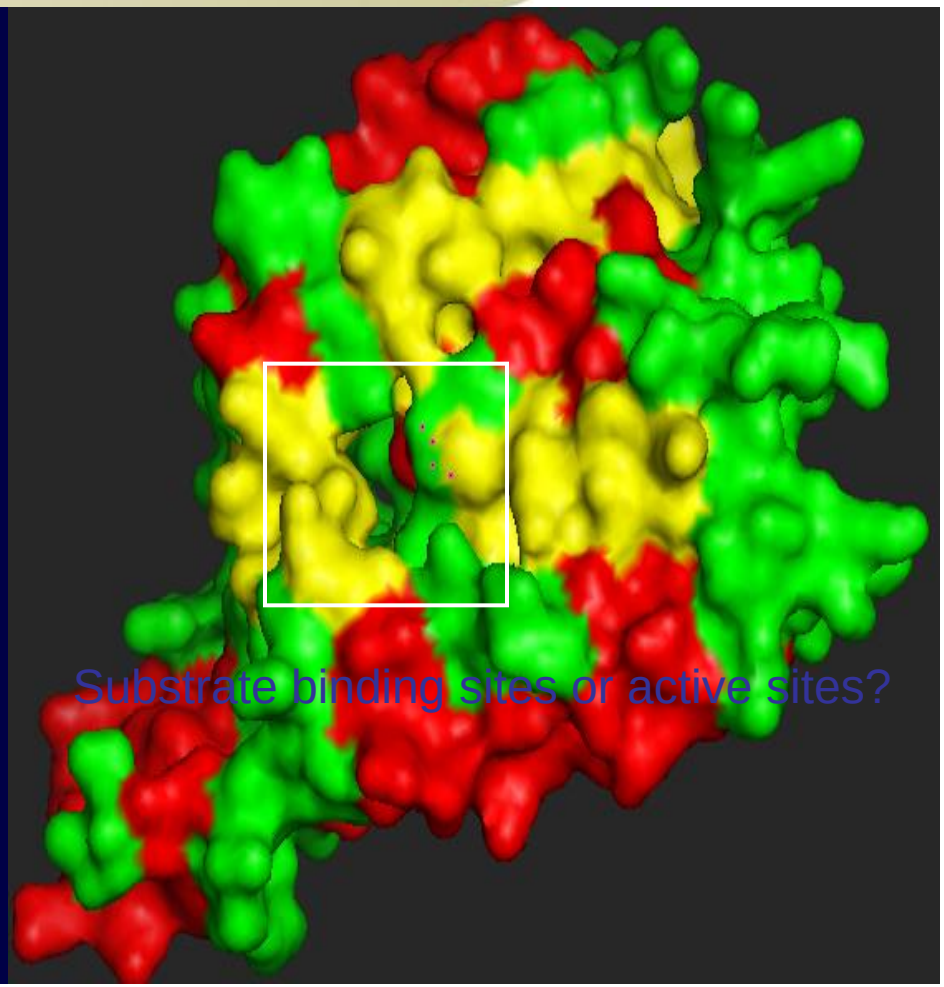
Acc: accessibility, white is buried; cyan is intermediate; blue is accessible; blue with red borders is highly exposed.
Hyd: hydropath, pink is hydrophobic; grey is intermediate; cyan is hydrophilic

Conserved domain in all KCS proteins

```
                                α4                  β3          η2        α5
Gr10018148                                                                          ....
                      160       170       180       190       200       210
Gr10018148            FNPTPSLSSMIINKYKMRGNIKSFNLGGMGCSASVISVDLAKDMLQVHRNNYAIVV....
Gr10040229            FCPSPSLSSIIVNKYSMRSDVKSFNLSGMGCSAGAIGIDLAQNLLKTNNNCYAIVISTEI
Gr10040228            FCPSPSLSSIIVNKYSMRSDVKSFNLSGMGCSSGAIGIDLAQNLLKTNNNCYAIVISTEI
Gr10017854            FCPSPSLSSIIINKYSMKSDIKSYNLSGMGCSAGTIGVDLAQNLLKTHENKTAIVLSTEI
Gr10015926            FNPTPSLSAMIVNHYKLRTNINSYNLGGMGCSAGLISIDLAKNLLQSNPNTYALVVSTEN
Gr10028536            FNPTPSLSAMIVKHYKLRTDIKSYNLGGMGCSAGLISVELAKNLLQANPNTYAVVVSTEN
Gr10004055            FNPTPSLSAMVINHYKLRTDIKSYNLGGMGCSAGLISVELAKNLLRANPNTYAVVVSTEN
Gr10016173            FNPTPSLSAMVINHYKLRGNIQSYNLGGMGCSAGLISIDLAKHLLQVHPNSYALVISMEN
Gr10017624            FNPTPSLSAMVINHYKLRGNIQSYNLGGMGCSAGLLSIDLAKNLLQVHPNSYALVISMEN
Gr10010162            FNPTPSLSAMVINHYKLRGNIQSYNLGGMGCSAGLISIDLAKNLLQVHPNTYALVISMEN
Gr10031660            FNPTPSLSAVVVNRYKFRGNILSYNLGGMGCSAGLISIDLAKQLLRVHPNSYALVVSMEN
Gr10021139            FNPTPSLSAMIVNRYKLRGNILSYNLGGMGCSAGLISIDLAKQMLQVHPNSYALVVSMEN
Gr10031991            FSPTPSLSAMVINKYKLRSNIKSFNLSGMGCSAGLISIDLARDLLQVHPNSNAVVVSTEI
Gr10014500            FSPTPSLSAMVINKYKLRSNIKSFNLSGMGCSAGLISIDLARDLLQVHPNSNAVVVSTEI
Gr10036653            FNPTPSLSAMIVNKYKLRGNIRSFNLGGMGCSAGVIAVDLAKDLLQVHRNTYAVVVSTEN
Gr10034062            FNPTPSLSAMIVNKYKLRGNIRSFNLGGMGCSAGVIAVDLAKDMLQVHRNTYAVVVSTEN
Gr10000033            FNPTPSLSAMIVNKYKLRGNIRSFNLGGMGCSAGVIAVDLAKDMLQVHRNTYAVVVSTEN
Gr10018297            FNPTPSLSAMIINKYKLRGNIRSFNLGGMGCSAGVIAIDLAKDMLQVHRNSYAVVVSTEN
Gr10018150            FNPTPSLSSMIINKYKMRGNIKSFNLGGMGCSASVISVDLAKDMLQVHRNNYAIVVSTEN
Gr10018147            FNPTPSLTAMIINKYKMRGNIKSFNLSGMGCSAGVIAIDLAKDMLQVHRNNYAVVFSTEN
Gr10018149            FNPTPSLTAMIINKYKMRGNNKSFNLSGMGCSASVIAIDLAKDMLQVYRNNYAVVFSTEN
Gr10026783            FNPTPSLSAMIINHYKMRGNILSYNLGGMGCSAGIIAVDLARDMLQANPNNYAVVVSSEM
Gr10019616            ISSVPSIPARVINRYKMREDVKVFNLSGMGCSASLIAVDLVNHLFQTYKNQFAIVVSSES
Gr10032475            ITSVPSLPARVINRYKMRDDVKVFNLSGMGCSASVIAVDLVHHLFKTYMNSFAVIVSSES
Gr10019657            FSPAPSLTSRIINRYKMRHNIKSFSLSGMGCSASMLAIDLVQQLFKTYKNQFAIVVSTES
Gr10019659            FSPAPSLTSRIINRYKMRDNIKSFSLSGMGCSASMVAIDMVQQLFKTYKNQFAIVVSTES
Gr10019136            FSPSPSLTARIVNRYKMRDNIKSFSLSGMGCSASMVAIDLVQNLFKSYKNAFAVVVSSET
Gr10032226            LSTVPSLCSRIINHYKMRPDIKSFNLTGMGCSASLISLDIVRNVFKSYKNKFALLVTSES
Gr10022901            ITAPPCLSSRIINHYKMRQDIKCFNLTGMGCSASLISLDIVRNVFKSYKNKYALLVTSES
consensus>70          f.p.Psls..!!n.Ykmr.#i.s%nL.GMGCSa..i.i#la.dll....N.yAvvvs.e.
```

Why these domains are conserved?





Red: helix Yellow: sheet Loop: green

3D structure prediction suggest thatStrict domains locate on two helix ,five sheets and some loops.

Summary

★ Local cotton database construction is helpful to query target sequence and conserved domain prediction contributes to research loss of function in protein family.

★ Phylogeny construction and chromosome location suggest that there are some KCS genes that are required for fiber initiation and elongation.

★ Protein structure prediction can provide useful information to design point mutants.

★ Bioinformatics analysis is useful, but we should be careful.

Acknowdgement

- Prof. Luo
- Prof. Zhu
- All members in ABC class, especially G01 group.
- Our lab members